



**UNIVERSIDAD
NACIONAL DE
LOJA**



Facultad de la Energía, las Industrias y los Recursos Naturales No Renovables

CARRERA DE INGENIERÍA EN SISTEMAS

“Aplicación de algoritmos de clasificación para determinar cómo influye la definición del tema de un proyecto de titulación en su éxito o fracaso”

“Trabajo de titulación previo a la obtención del título de Ingeniería en Sistemas”

Autor:

- Pablo Leonardo Valdivieso Orellana

Director:

- Ing. Oscar Miguel Cumbicus Pineda, Mg. Sc.

LOJA-ECUADOR

2019

Certificación del director

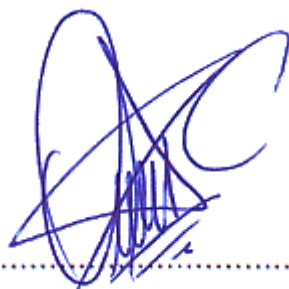
Ing. Oscar Miguel Cumbicus Pineda

DOCENTE DE LA CARRERA DE INGENIERÍA EN SISTEMAS

CERTIFICA:

Que el egresado Pablo Leonardo Valdivieso Orellana autor del presente trabajo de titulación, cuyo tema versa sobre “APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN PARA DETERMINAR CÓMO INFLUYE LA DEFINICIÓN DEL TEMA DE UN PROYECTO DE TITULACIÓN EN SU ÉXITO O FRACASO”, ha sido dirigido, orientado y discutido bajo mi asesoramiento y reúne a satisfacción los requisitos exigidos en una investigación de este nivel por lo cual autorizo su presentación y sustentación.

Loja, 14 de marzo del 2019



.....
Ing. Oscar Miguel Cumbicus Pineda, Mg. Sc.
DIRECTOR DEL TRABAJO DE TITULACIÓN

Autoría

Yo **PABLO LEONARDO VALDIVIESO ORELLANA** declaro ser autor del presente trabajo de titulación y eximo expresamente a la Universidad Nacional de Loja y a sus representantes jurídicos de posibles reclamos o acciones legales por el contenido del mismo.

Adicionalmente acepto y autorizo a la Universidad Nacional de Loja, la publicación de mi trabajo de titulación en el Repositorio Institucional - Biblioteca Virtual.

Firma: 

Cedula: 1150025748.

Fecha: 05/07/2019.

CARTA DE AUTORIZACIÓN DE TESIS POR PARTE DEL AUTOR, PARA LA CONSULTA, REPRODUCCIÓN PARCIAL O TOTAL Y PUBLICACIÓN ELECTRÓNICA DEL TEXTO COMPLETO.

Yo PABLO LEONARDO VALDIVIESO ORELLANA, declaro ser el autor del trabajo de tesis: “**APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN PARA DETERMINAR CÓMO INFLUYE LA DEFINICIÓN DEL TEMA DE UN PROYECTO DE TITULACIÓN EN SU ÉXITO O FRACASO**”, como requisito para optar al grado de: **INGENIERO EN SISTEMAS**; autorizo al Sistema Bibliotecario de la Universidad Nacional de Loja para con fines académicos, muestre al mundo la producción intelectual de la Universidad, a través de la visibilidad de su contenido de la siguiente manera en el Repositorio Digital Institucional:

Los usuarios pueden consultar el contenido de este trabajo en el RDI, en las redes de información del país y del exterior, con las cuales tenga convenio la Universidad. La Universidad Nacional de Loja, no se responsabiliza por plagio o copia del trabajo de titulación que realice un tercero.

Para constancia de esta autorización, en la ciudad de Loja, el día cinco del mes de julio del dos mil diecinueve.

Firma:



Autor: Pablo Leonardo Valdivieso Orellana

Cédula: 1150025748

Dirección: Loja (Ciudad Alegría, Condamine y Calle 8)

Correo Electrónico: plvaldiviesoo@unl.edu.ec

Celular: 0992248115

DATOS COMPLEMENTARIOS

Director de Trabajo de Titulación: Ing. Oscar Miguel Cumbicus Pineda, Mg. Sc.

Tribunal de Grado: Ing. Hernán Leonardo Torres Carrión, Mg. Sc.

Ing. Edwin René Guamán Quinche, Mg. Sc.

Ing. Valeria del Rosario Herrera Salazar, Mg. Sc.

Agradecimiento

A mi madre por su apoyo al brindarme la posibilidad de concretar una carrera profesional. Al Ing. Oscar Cumbicus quien con su experiencia y conocimiento fue mi guía en la elaboración y desarrollo del trabajo de titulación. A los docentes de la Carrera por los conocimientos impartidos en mi formación académica.

Dedicatoria

A mi madre, Elisa.

A mi abuelita, Imelda.

A mis hermanos, Jonathan y Jossué.

Índice de Contenidos

Índice General

Índice General	VI
Índice de Figuras	IX
Índice Tablas	XIII
a. Título	1
b. Resumen.....	2
c. Introducción	4
d. Revisión de Literatura.....	6
1. Conceptos Preliminares	6
1.1. Inteligencia Artificial.	6
1.2. Aprendizaje Automático.	6
1.3. Minería de datos.....	7
1.4. Minería de texto.....	7
1.5. Tokenización.	8
1.6. Bolsa de palabras.	8
1.7. Nube de Palabras.	8
1.8. Algoritmos de clasificación.....	9
1.8.1. Algoritmos de Clasificación Supervisada.....	9
1.8.2. Algoritmos de Clasificación No supervisada.....	10
1.9. Herramientas de software para la aplicación de algoritmos de clasificación.....	12
1.9.1. Openrefine.	12
1.9.2. R Studio.	13
1.9.3. IBM SPSS Statistics.	13
1.9.4. RapidMiner Studio.	13
2. Trabajos Relacionados.....	13
e. Materiales y Métodos	16
1. Materiales.	16
2. Métodos.....	17
2.1. Método Deductivo.	17
2.2. Método Inductivo.....	17
2.3. Método Científico.	18
2.4. Método Empírico.	18
3. Técnicas.....	18

3.1.	Observación.....	18
3.2.	Entrevista.....	18
3.3.	Muestreo aleatorio simple.....	18
4.	Metodologías.....	18
f.	Resultados.....	24
1.	Depuración del conjunto de datos obtenido de los proyectos de titulación.	24
1.1.	Evaluación del conjunto de datos.....	24
1.2.	Categorización de los proyectos de titulación.....	27
1.3.	Eliminación de información innecesaria.....	30
2.	Aplicación de algoritmos de clasificación al conjunto de datos de los proyectos de titulación.....	36
2.1.	Búsqueda de algoritmos de clasificación.....	36
2.2.	Selección de los algoritmos de clasificación.....	38
2.3.	Aplicación de los algoritmos de clasificación seleccionados.....	40
2.3.1.	Pre procesamiento de los datos.....	40
2.3.2.	Nube de Palabras y Agrupación Jerárquica.....	51
2.3.3.	Reglas de Asociación Apriori.....	54
2.3.4.	K-means y K-medoids.....	57
2.3.5.	CHAID Exhaustivo.....	61
2.3.6.	Random Forest.....	66
3.	Evaluación de los algoritmos aplicados.....	68
3.1.	Evaluación de los resultados.....	68
3.1.1.	Dendograma.....	72
3.1.2.	Reglas de Asociación Apriori.....	75
3.1.3.	K-means.....	80
3.1.4.	K-medoids.....	83
3.1.5.	CHAID Exhaustivo.....	84
3.1.6.	Random Forest.....	91
3.2.	Evaluación de los algoritmos aplicados.....	92
3.3.	Introducción de nuevos datos.....	96
3.3.1.	Aplicación de los algoritmos de Clasificación.....	100
3.3.2.	Evaluación de los algoritmos.....	118
g.	Discusión.....	122
1.	Desarrollo de la Propuesta Alternativa.....	122
2.	Valoración técnica económica ambiental.....	127
h.	Conclusiones.....	129

i.	Recomendaciones.....	130
j.	Bibliografía.....	131
k.	Anexos.....	137
	Anexo 1.....	137
	Anexo 2.....	138
	Anexo 3.....	141
	Anexo 4.....	142
	Anexo 5.....	144
	Anexo 6.....	145
	Anexo 7.....	147
	Anexo 8.....	148
	Anexo 9.....	149
	Anexo 10.....	150
	Anexo 11.....	154

Índice de Figuras

Figura 1. Proceso de la metodología para la aplicación de los algoritmos.....	23
Figura 2. Conjunto de datos de los proyectos aprobados para su desarrollo.....	24
Figura 3. Muestreo aleatorio simple.....	26
Figura 4. Números aleatorios obtenidos.....	27
Figura 5. Conjunto de datos categorizado.....	30
Figura 6. Cargar datos en Openrefine.....	32
Figura 7. Reemplazo de palabras mal escritas.....	32
Figura 8. Conversión de texto a minúsculas.....	33
Figura 9. Estandarización de las tildes.....	33
Figura 10. Reemplazo de caracteres especiales.....	34
Figura 11. Eliminación de texto alfanumérico.....	34
Figura 12. Separación de variable tema en columnas.....	35
Figura 13. Estandarización de palabras semejantes.....	35
Figura 14. Términos más frecuentes.....	42
Figura 15. Inspección de la bolsa de palabras.....	43
Figura 16. Nueva inspección de la bolsa de palabras.....	45
Figura 17. Términos más frecuentes sin conectores.....	46
Figura 18. Dendograma con nuevas relaciones.....	47
Figura 19. Dendograma con relaciones obvias.....	48
Figura 20. Reemplazo de UNL.....	49
Figura 21. Dendograma sin relaciones válidas.....	50
Figura 22. Términos más frecuentes con siglas.....	50
Figura 23. Reemplazo de ciudad Loja.....	51
Figura 24. Dendograma sin relaciones.....	52
Figura 25. <i>Resultados de k-means en todo el conjunto de datos.</i>	58
Figura 26. Resultados de k-medoids en todo el conjunto de datos.....	59
Figura 27. Resultados de k-means en los proyectos de éxito.....	59
Figura 28. Resultados de k-medoids en los proyectos de éxito.....	60
Figura 29. Resultados de k-means en los proyectos de fracaso.....	60
Figura 30. Resultados de k-medoids en los proyectos de fracaso.....	60
Figura 31. Separar columnas del archivo con tema sin conectores.....	63
Figura 32. Establecer el espacio como separador en columna del tema.....	63
Figura 33. Conjunto de datos con las columnas separadas sin conectores.....	64
Figura 34. Conjunto de datos con las columnas separadas en SPSS.....	64

Figura 35. Variables independientes y dependiente en CHAID Exhaustivo.	65
Figura 36. Criterios del algoritmo CHAID Exhaustivo.	65
Figura 37. Árbol obtenido con el algoritmo CHAID Exhaustivo.	66
Figura 38. Carga del conjunto de datos en RapidMiner.	67
Figura 39. Establecer variable dependiente en el conjunto de datos.	67
Figura 40. Aplicación del algoritmo Random Forest.	68
Figura 41. Nube de Palabras de todo el conjunto de datos.	69
Figura 42. Nube de Palabras de los proyectos de éxito.	69
Figura 43. Nube de Palabras de los proyectos de fracaso	70
Figura 44. Representación gráfica de la palabra web.	71
Figura 45. Dendograma de todo el conjunto de datos.	72
Figura 46. Dendograma de los proyectos de éxito.	72
Figura 47. Dendograma de los proyectos de fracaso.	73
Figura 48. Representación gráfica de la relación móvil aplicación.	74
Figura 49. Reglas de Asociación de todo el conjunto de datos.	75
Figura 50. Reglas relacionadas de todo el conjunto de datos.	76
Figura 51. Reglas de Asociación de los proyectos de éxito.	76
Figura 52. Reglas relacionadas de los proyectos de éxito.	77
Figura 53. Reglas de Asociación de los proyectos de fracaso.	77
Figura 54. Reglas relacionadas de los proyectos de fracaso.	78
Figura 55. Representación gráfica de la relación gestión-sistema.	79
Figura 56. Representación gráfica de la relación diseño-construcción-prototipo.	81
Figura 57. Representación gráfica de la relación territorio-Loja-estudio.	82
Figura 58. Representación gráfica de la relación aplicación-móvil-desarrollo.	82
Figura 59. Representación gráfica de la palabra Loja.	83
Figura 60. Primer nivel del árbol de CHAID Exhaustivo.	84
Figura 61. Nodo 4 del árbol de CHAID Exhaustivo.	85
Figura 62. Segundo nivel del árbol de CHAID Exhaustivo.	85
Figura 63. Nodo 1 del árbol de CHAID Exhaustivo.	86
Figura 64. Nodo 6 del árbol de CHAID Exhaustivo.	86
Figura 65. Nodo 2 del árbol de CHAID Exhaustivo.	86
Figura 66. Nodo 8 del árbol de CHAID Exhaustivo.	87
Figura 67. Tercer nivel del árbol de CHAID Exhaustivo.	87
Figura 68. Nodo 5 del árbol de CHAID Exhaustivo.	88
Figura 69. Nodo 12 del árbol de CHAID Exhaustivo.	88

Figura 70. Nodo 11 del árbol de CHAID Exhaustivo.	89
Figura 71. Nodo 7 del árbol de CHAID Exhaustivo.	89
Figura 72. Nodo 14 del árbol de CHAID Exhaustivo.	89
Figura 73. Nodo 9 del árbol de CHAID Exhaustivo.	90
Figura 74. Nodo 15 del árbol de CHAID Exhaustivo.	90
Figura 75. Nodo 16 del árbol de CHAID Exhaustivo.	91
Figura 76. Árbol obtenido con el algoritmo Random Forest.	92
Figura 77. Nueva Nube de Palabras de todo el conjunto de datos.	97
Figura 78. Nueva Nube de Palabras de los proyectos de éxito.	97
Figura 79. Nueva Nube de Palabras de los proyectos de fracaso.	98
Figura 80. Nueva representación gráfica de la palabra Loja.	99
Figura 81. Nueva representación gráfica de la palabra aplicación.	99
Figura 82. Dendograma del nuevo conjunto de datos.	100
Figura 83. Dendograma de nuevos proyectos de éxito.	101
Figura 84. Dendograma de nuevos proyectos de fracaso.	101
Figura 85. Nueva representación gráfica de la relación AEIRNNR-utilizando.	102
Figura 86. Reglas de Asociación del nuevo conjunto de datos.	103
Figura 87. Reglas relacionadas del nuevo conjunto de datos.	103
Figura 88. Reglas de Asociación de los nuevos proyectos de éxito.	104
Figura 89. Reglas relacionadas de los nuevos proyectos de éxito.	104
Figura 90. Reglas de Asociación de los nuevos proyectos de fracaso.	105
Figura 91. Reglas relacionadas de los nuevos proyectos de fracaso.	105
Figura 92. Nueva representación gráfica de la relación gestión-sistema.	106
Figura 93. Nueva representación gráfica de la relación web-sistema.	107
Figura 94. Nueva representación gráfica de la relación móvil-aplicación.	107
Figura 95. Resultados de k-means en el nuevo conjunto de datos.	108
Figura 96. Resultados de k-means en los nuevos proyectos de éxito.	108
Figura 97. Resultados de k-means en los nuevos proyectos de fracaso.	108
Figura 98. Nueva representación gráfica de la relación territorio-Loja-sistema.	110
Figura 99. Nueva representación gráfica de la relación aprendizaje-virtual-entorno.	111
Figura 100. Nueva representación gráfica de la relación implementación-desarrollo-sistema.	111
Figura 101. Nueva representación gráfica de la relación sistema-gestión-web.	112
Figura 102. Nueva representación gráfica de la relación software-desarrollo-UNL. ...	112
Figura 103. Resultados de k-medoids en el nuevo conjunto de datos.	113

Figura 104. Resultados de k-medoids en los nuevos proyectos de éxito.....	113
Figura 105. Resultados de k-medoids en los nuevos proyectos de fracaso.....	113
Figura 106. Primer nivel del nuevo árbol de CHAID Exhaustivo.....	114
Figura 107. Segundo nivel del nuevo árbol de CHAID Exhaustivo.....	115
Figura 108. Tercer nivel del nuevo árbol de CHAID Exhaustivo.....	116
Figura 109. Nuevo árbol obtenido al aplicar Random Forest.....	117
Figura 110. Captura del conjunto de datos de los proyectos.....	151
Figura 111. Temas de los proyectos en celdas combinadas.....	152
Figura 112. Proyectos enumerados para contabilizarlos.....	152

Índice Tablas

TABLA I MATERIALES USADOS.....	16
TABLA II CARACTERES REEMPLAZADOS O ELIMINADOS.....	31
TABLA III ESTUDIOS SELECCIONADOS	36
TABLA IV OBTENCIÓN DE ALGORITMOS DE CLASIFICACIÓN	37
TABLA V COMPARATIVA DE LOS ALGORITMOS DE CLASIFICACIÓN	39
TABLA VI LIBRERÍAS UTILIZADAS EN RSTUDIO	40
TABLA VII CARGA DE PROYECTOS EN RSTUDIO	40
TABLA VIII GENERACIÓN DE LA BOLSA DE PALABRAS	41
TABLA IX OBTENCIÓN DE TÉRMINOS MÁS FRECUENTES	41
TABLA X CONVERSIÓN A DATAFRAME DE LOS TÉRMINOS FRECUENTES.....	42
TABLA XI CARGA DE TEMAS EN MAYÚSCULAS	44
TABLA XII CONTROL DE CONECTORES EN EL CORPUS.....	44
TABLA XIII APLICACIÓN DEL DENDOGRAMA	46
TABLA XIV NUEVA APLICACIÓN DEL DENDOGRAMA	48
TABLA XV CARGA DE TEMAS CON SIGLAS	49
TABLA XVI CARGA DEL CONJUNTO DE DATOS PRE PROCESADO	51
TABLA XVII APLICACIÓN NUBE DE PALABRAS EN TODO EL CONJUNTO DE DATOS.....	52
TABLA XVIII NUEVA APLICACIÓN DEL DENDOGRAMA AL CONJUNTO DE DATOS	53
TABLA XIX OBTENCIÓN DE PROYECTOS CON ESTADO DE ÉXITO	53
TABLA XX OBTENCIÓN DE PROYECTOS CON ESTADO DE FRACASO.....	54
TABLA XXI NUEVO ARCHIVO DE TODOS LOS PROYECTOS SIN CONECTORES	54
TABLA XXII APLICACIÓN DE APRIORI EN TODO EL CONJUNTO DE DATOS	55
TABLA XXIII NUEVO ARCHIVO DE PROYECTOS DE ÉXITO SIN CONECTORES..	55
TABLA XXIV APLICACIÓN DE APRIORI EN LOS PROYECTOS DE ÉXITO	56
TABLA XXV NUEVO ARCHIVO DE PROYECTOS DE FRACASO SIN CONECTORES	56
TABLA XXVI APLICACIÓN DE APRIORI EN LOS PROYECTOS DE FRACASO	57
TABLA XXVII OBTENCIÓN DE TODOS LOS PROYECTOS SIN CONECTORES.....	57
TABLA XXVIII APLICACIÓN DE K-MEANS EN TODO EL CONJUNTO DE DATOS..	58
TABLA XXIX APLICACIÓN DE K-MEDOIDS EN TODO EL CONJUNTO DE DATOS	59
TABLA XXX APLICACIÓN DE LA REPRESENTACIÓN GRÁFICA DE LAS RELACIONES.....	61

TABLA XXXI OBTENCIÓN DE VARIABLE TEMA SIN CONECTORES	62
TABLA XXXII NUEVO ARCHIVO DE TEMAS SIN CONECTORES Y CON ESTADO	62
TABLA XXXIII COMPARATIVA DE RESULTADOS DE LAS NUBES DE PALABRAS	70
TABLA XXXIV COMPARATIVA DE RESULTADOS DEL DENDOGRAMA	74
TABLA XXXV COMPARATIVA DE RESULTADOS DE APRIORI.....	78
TABLA XXXVI COMPARATIVA DE RESULTADOS DE K-MEANS	80
TABLA XXXVII COMPARATIVA DE RESULTADOS DE K-MEDOIDS	83
TABLA XXXVIII COMPARATIVA DE NUEVOS RESULTADOS DE NUBES DE PALABRAS.....	98
TABLA XXXIX COMPARATIVA DE NUEVOS RESULTADOS DEL DENDOGRAMA	102
TABLA XL COMPARATIVA DE NUEVOS RESULTADOS DE APRIORI.....	106
TABLA XLI COMPARATIVA DE NUEVOS RESULTADOS DE K-MEANS	109
TABLA XLII COMPARATIVA DE NUEVOS RESULTADOS DE K-MEDOIDS.....	113
TABLA XLIII VALORACIÓN ECONÓMICA DEL PROYECTO.....	127

a. Título

Aplicación de algoritmos de clasificación para determinar cómo influye la definición del tema de un proyecto de titulación en su éxito o fracaso.

b. Resumen

Una mala definición del tema de un proyecto de titulación en muchos casos acarrea problemas en el desarrollo del proyecto. En la Facultad de Energía, las Industrias y los Recursos Naturales No Renovables (FEIRNNR) de todos los proyectos de titulación aprobados para su desarrollo aproximadamente el 20% de los proyectos han fracasado, generando inseguridad y diferentes puntos de vista por parte de los estudiantes al momento de elegir el tema de su proyecto. En el presente trabajo se determinó cómo influye la definición del tema de un trabajo de titulación en su posterior éxito o fracaso utilizando algoritmos de clasificación, que son técnicas de Inteligencia Artificial. Se inició realizando la limpieza del conjunto de datos obtenido en la Secretaria General de la FEIRNNR, luego se evaluó el conjunto de datos y se determinó las variables útiles en la categorización de los proyectos, categorización que tuvo dos clases: éxito o fracaso.

La herramienta utilizada para la limpieza de la información fue Openrefine y las herramientas de software empleadas para la aplicación de los algoritmos de clasificación fueron RStudio, SPSS Statistics y RapidMiner.

Los algoritmos y técnicas de clasificación aplicadas al conjunto de datos fueron: Dendograma, K-means, K-medoids, Apriori, CHAID Exhaustivo y Random Forest. La evaluación de estos algoritmos de clasificación se realizó comparando los resultados obtenidos por cada uno, de tal forma que los resultados de todo el conjunto de datos que coinciden con los casos de éxito o fracaso fueron considerados los más relevantes, pero los resultados de los casos de éxito a su vez no debían coincidir con los resultados de fracaso debido a que no permiten determinar el éxito o fracaso de los proyectos. El algoritmo de clasificación supervisada que proporcionó un porcentaje de eficiencia y resultados relevantes fue CHIAD Exhaustivo, donde se obtuvo que el 32% de los proyectos aprobados para su desarrollo en la FEIRNNR fracasaron y que si las palabras “**gestión**” y “**sistema**” están presentes en el tema de un trabajo de titulación tiende a culminar con éxito, pero si en el tema están presentes las palabras “**móvil**” y “**aplicación**” el trabajo de titulación tiende a fracasar, por lo que se concluyó que la definición del tema del trabajo de titulación si influye en el éxito o fracaso del mismo.

Abstract

A poor theme definition for a degree project leads to problems in the project development in most of the cases. In the Faculty of Energy, Industries and Non-Renewable Natural Resources (FEIRNNR) approximately the 20% of the approved degree projects have failed causing unsureness in students and different points of view at the moment they have to choose their degree theme project. In the present work, how the definition of the theme for the degree project has influence in the future success or failure of the project was determined by using classification algorithms which are artificial intelligence techniques. The work started by cleaning the obtained dataset from the General Secretariat in the Faculty of Energy, Industries and Non-Renewable Natural Resources (FEIRNNR), then, the dataset was evaluated and the variables useful in the categorization of the projects were determined the one that had two categories: Success and Failure.

Openrefine was used as the tool for cleaning the information and RStudio, SPSS Statistics and RapidMiner were the tools used for the application of classification algorithms.

Dendogram, K-means, K-medoids, Apriori, Exhaustive CHAID and Random Forest were the algorithms and classification techniques applied to the dataset. The evaluation of these algorithms of classification was carried out by comparing the obtained results in each classification, so the results of the complete dataset which matched with success or failure cases were considered as the most relevant but the results of the success cases should not coincide with failure results due to they do not allow to determine the success or failure of the projects.

CHAID Exhaustive was the supervised algorithm of classification which provided a percentage of efficiency and relevant results in which it was found that the 32% of the approved projects to be develop in General Secretariat in the Faculty of Energy, Industries and Non-Renewable Natural Resources (FEIRNNR) failed and that if the words "management" and "system" are present in the theme of a degree work the project tends to culminate successfully, but if the works "mobile" and "application" are present in the theme, the degree work project tends to fail, so it was concluded that the definition of the theme of the degree work project have influence to the success or failure of it.

c. Introducción

Al momento de elaborar un proyecto de titulación, es muy importante establecer aspectos como: la definición del tema, la identificación del problema y el planteamiento de la hipótesis, por lo que es recomendable que se tenga al menos algún conocimiento del tema planteado, para determinar acertadamente las técnicas y metodologías a emplear en el desarrollo del proyecto. Prepararse antes de realizar un proyecto de titulación permitirá al investigador centrarse en el problema a resolver y conocer cómo desarrollar el estudio, por ello se debe investigar e indagar sobre el tema que se pretende desarrollar[1].

La definición del tema de un proyecto de titulación es el primer obstáculo para la elaboración del mismo, en muchos casos la definición del tema implica demasiado tiempo en su elaboración, y una mala definición del tema acarrea problemas en el desarrollo del proyecto, que a su vez pueden ocasionar cambios en objetivos, prorrogas o bien el fracaso del proyecto de titulación.

En un conjunto de datos como el existente en la Secretaria General de la FEIRNNR de todos los proyectos de titulación aprobados para su desarrollo de las Carreras de la Facultad, los algoritmos de clasificación permitieron encontrar información oculta a simple vista recuperando información útil, mediante el uso de herramientas capaces de extraer información de grandes volúmenes de datos[2].

De todo lo mencionado anteriormente surge una preocupación que impulsó la investigación de determinar cómo influye la definición del tema de un proyecto de titulación en el posterior éxito o fracaso del proyecto. Para validar esta hipótesis se utilizaron algoritmos de clasificación que son técnicas de inteligencia artificial que nos permiten tratar estos problemas.

El presente trabajo de titulación consta de varias secciones, entre ellas la de Revisión de Literatura que abarca la temática de los algoritmos de clasificación, se detalla su clasificación, su uso en la minería de datos, aprendizaje automático, minería de texto, su importancia en la inteligencia artificial y las herramientas empleadas para su aplicación. La sección materiales y métodos que abarca los materiales, métodos, técnicas y metodologías empleadas para la realización del trabajo. Otra sección es la de Resultados que tiene la siguiente estructura: Depurar el conjunto de datos, aplicar los algoritmos de clasificación en el conjunto de datos y evaluar los algoritmos aplicados.

La depuración del conjunto de datos consistió en evaluar la validez de dichos datos, realizar la categorización de los proyectos de titulación y limpiar la información irrelevante del conjunto de datos. Al aplicar los algoritmos de clasificación se realizó el pre procesamiento de los datos para eliminar las relaciones obvias del conjunto de datos, se hizo una búsqueda de los algoritmos de clasificación aplicables a este trabajo, se seleccionó los algoritmos de clasificación más óptimos para ser empleados en el conjunto de datos y se aplicó los algoritmos de clasificación seleccionados. Para la evaluación de los algoritmos aplicados, se realizó una evaluación de los resultados y posterior a ello se evaluó los algoritmos de clasificación en base a los resultados previamente evaluados, por último, se realizó la introducción de nuevos datos con los que se realizó el mismo procedimiento aplicado al primer conjunto de datos.

También están las secciones de Discusión, Conclusiones y Recomendaciones, en la sección de Discusión se presenta una explicación de cómo todo el proceso realizado conlleva al cumplimiento de los objetivos planteados en este trabajo, en la sección Conclusiones se expone las deducciones de las experiencias obtenidas durante el cumplimiento de los objetivos y en la sección Recomendaciones se detallan sugerencias para un mejor desarrollo de temas similares al presente trabajo.

d. Revisión de Literatura

En la presente sección se describe el concepto de minería de datos, minería de texto, bolsa de palabras y tokenización. También el concepto de algoritmos de clasificación y de los algoritmos de clasificación supervisada y no supervisada aplicados en este trabajo de titulación. Otros conceptos descritos son los de los algoritmos de clasificación y las herramientas usadas para llevar a cabo la aplicación de los algoritmos.

1. Conceptos Preliminares

1.1. Inteligencia Artificial.

La inteligencia artificial surge en la década de 1940, tiene como principal objetivo imitar el comportamiento humano con el fin de obtener el mejor resultado esperado, es un campo de estudio muy amplio y en constante cambio que busca una comprensión profunda de la inteligencia y de la capacidad de esta a través de la comprensión de sus límites y alcances[3]. También permite dar solución a problemas referentes al análisis de la información con el fin de optimizar el aprendizaje y la toma de decisiones[4],

Entre las principales aplicaciones tenemos: el aprendizaje automático, los agentes de conversacionales y el razonamiento casual. Y algunas de las principales áreas son: aprendizaje automático, procesamiento del lenguaje natural, visión por computadoras, robótica y reconocimiento automático del habla[3]. La inteligencia artificial es la rama de investigación usada en este trabajo, permitió identificar de manera más específica las áreas a desarrollar en el trabajo.

1.2. Aprendizaje Automático.

El aprendizaje automático es una rama de la inteligencia artificial que desarrolla técnicas capaces de extraer de forma automática conocimiento subyacente en la información[5] y busca lograr que las computadoras aprendan[3][6]. Este método científico permite usar computadoras y otros dispositivos con capacidad computacional para que aprendan a extraer patrones y relaciones entre los datos[7], también incluye la reorganización de conocimientos previos y la adquisición de nuevos conocimientos mediante la experimentación y la observación[8]. Los principios del aprendizaje automático y la minería de datos son generar un modelo a partir de ejemplos y usarlo para resolver el problema[5]. Existen 5 paradigmas fundamentales del aprendizaje automático que son: algoritmos evolutivos, redes neuronales, simbolismo, redes bayesianas y razonamiento por analogía[3]. Los algoritmos de aprendizaje automático se agrupan en función a su

salida y son el aprendizaje supervisado y no supervisado[6]. En la investigación el aprendizaje automático fue utilizado para comprender la clasificación de los algoritmos de clasificación.

1.3. Minería de datos.

La minería de datos es una solución para el análisis de fenómenos no explícitos en bases de datos[9] y la búsqueda de patrones ocultos entre los datos, para posteriormente ser usados en la predicción de comportamientos a futuro, es decir mediante la aplicación de técnicas de Inteligencia Artificial y de Aprendizaje Automático identifica información relevante que de otra manera permanecería oculta[8][10]. Esta herramienta también es conocida como Descubrimiento de Conocimiento de Bases de datos, debido a que permite analizar grandes bases de datos y obtener una descripción de las tendencias y correlaciones entre los datos, facilitando así la toma de decisiones[9][11]. Además, haciendo uso de diferentes algoritmos a partir de datos pre procesados resuelve problemas de agrupamiento automático, clasificación, asociación y detección de patrones secuenciales, ya que proporcionan nuevos conocimientos[12][13]. Las técnicas de la minería de datos pueden ser descriptivas o predictivas, y generalmente se dividen en cinco categorías: métodos estadísticos, análisis de clúster, arboles de decisión y reglas de decisión, reglas de asociación y detección de fraudes[12][13]. El proceso de la minería de datos consta de cinco partes: selección de la información, transformación de los datos, aplicación de las técnicas, interpretación de los resultados y la incorporación del nuevo conocimiento[14]. Las fases fundamentales de la minería de datos proporcionaron una guía para la óptima aplicación de los algoritmos de clasificación.

1.4. Minería de texto.

Es el descubrimiento de la información desconocida por medio de la búsqueda de patrones en el texto y la extracción de datos[15], procesa información no estructurada y extrae índices numéricos del texto[16]. Permite que la información sea accesible para varios algoritmos de minería de datos[16] y realizar búsquedas eficientes de un tema, siendo así de importancia para el proceso de investigación y en las actividades que requieran el uso de información[15]. Las tecnologías de la minería de texto pueden ser utilizadas de forma individual o conjunta y son: recuperación de información, extracción de información, clasificación, clustering y generación automática de resúmenes[16]; estas tecnologías están comprendidas dentro de tres actividades fundamentales que son: la recuperación de la información, extracción de la información mediante el

procesamiento del lenguaje natural y minería de datos en busca de asociaciones entre los datos[17]. La información que se extrae es clara e indicada en el texto para optimizar el consumo por equipos directamente, sin necesitar de un humano[15]. En la investigación sus fases fueron una guía para el pre procesamiento de los datos y la búsqueda de relaciones entre las palabras de los temas.

1.5. Tokenización.

La tokenización conocida como limpieza de datos o depuración de los datos[18], es el proceso de separar una cadena de texto en oraciones o palabras[17][19], frases, símbolos y otros elementos, todos estos elementos son la entrada para el procesamiento posterior en la minería de texto[17] y son denominados como “tokens”[20][19]. También se encarga del pre procesamiento de los datos para reducir el ruido mediante técnicas o remover información no relevante[18], es decir palabras repetidas frecuentemente sin sentido, palabras usadas para unir otras palabras en una oración y palabras vacías que no contribuyen al contexto[17] ni a una apropiada separación de las palabras[20]. Permite comprender y establecer que caracteres deben ser eliminados y reemplazados en la limpieza del conjunto de datos.

1.6. Bolsa de palabras.

Es un tipo de representación de documentos, consiste en eliminar las relaciones semánticas y sintácticas entre las palabras[21] y mostrar la información en forma de vectores[22], estos vectores que se construyen representan a los documentos[23] como el conjunto de todas las palabras que aparecen en él[24], sin tomar en cuenta el orden, estructura o significado de las palabras[22]. El tamaño de los vectores es igual al número de palabras en el vocabulario de la colección de documentos y cada elemento del vector indica la presencia o ausencia de cada término en el documento[23], de esta manera se determina la frecuencia con la que aparece cada palabra en el documento y se asigna un peso a cada término en función de su importancia[22] para futuros tratamientos de esa información[21]. Fue útil en la investigación para la representación gráfica de las palabras y la aplicación de los algoritmos de clasificación no supervisados.

1.7. Nube de Palabras.

Es una combinación de varios tamaños de fuentes en una única visualización[25] donde las palabras más frecuentes de un texto resaltan por ocupar más lugar en la representación[26]. Tiene la función específica de presentar una descripción visual de una colección de datos tipo texto[25] basado en el conteo de las frecuencias de palabras

o frases[27]. Las nubes de palabras pueden ser usadas para fines analíticos y comunicar patrones de texto, por tal motivo existe una demanda creciente del uso de nubes de palabras que indica la necesidad del usuario de visualizar una importante clase de información tipo texto[25] de forma rápida y visualmente rica[26]. Dio a conocer las palabras más usadas en todo el conjunto de datos, como también en los proyectos de éxito y fracaso.

1.8. Algoritmos de clasificación

Uno de los conceptos más importantes relacionados con las técnicas de minería de datos es el concepto de aprendizaje automático, cuyo fin es inducir conocimiento a partir de datos[28]. Las técnicas de aprendizaje automático se utilizan regularmente para resolver problemas de clasificación en diversos campos de manera rápida y precisa, proporcionando una significativa ventaja para la investigación[29]. Estas técnicas o algoritmos de clasificación permiten generar conocimiento a partir de datos[30] y se clasifican en supervisados y no supervisados[31], donde el primero es más preciso debido a que los clasificadores trabajan con datos ya entrenados[32], es decir tienen ya una clase o etiqueta asignada de forma correcta[33]. Los algoritmos no supervisados presentan un menor desempeño al no presentar sus datos una etiqueta, pero minimizan el trabajo de clasificación trabajando por medio de palabras semilla y calculando la orientación semántica de las frases[32]. Brindó una mejor comprensión de la aplicación de los algoritmos de clasificación, ayudando a identificar como usar los algoritmos de clasificación supervisados y los no supervisados.

1.8.1. Algoritmos de Clasificación Supervisada.

Estos algoritmos trabajan con un conjunto de datos o de entrenamiento etiquetado que permite construir modelos confiables[24] para clasificar nuevas observaciones[34] o interpretar la información para transformarla a conocimiento[30]. En un contexto con datos muy diversos y desconocidos, el contar con datos etiquetados se vuelve una práctica muy costosa e inimaginable[24] debido a que los algoritmos usan la clase que facilita un experto en la base de datos como atributo[30]. Estos algoritmos tienen como objetivo determinar cuál es la clase a la que pertenece una nueva muestra sin clases, en base a las clases de las que ya se tiene conocimiento y patrones de entrada y salida[35][36]. Estos algoritmos fueron usados en el trabajo debido a que proporcionan resultados más óptimos haciendo uso del conjunto de datos ya categorizado.

- **CHAID Exhaustivo**

CHAID es un acrónimo de Chi-squared Automatic Interaction Detector, que en español significa detector automático de interacciones mediante Chi cuadrado[37]. CHAID Exhaustivo es una variación de CHAID, esta variación permite realizar una clasificación óptima de la variable predictora debido a que sigue fundiendo categorías de la clase hasta que queden dos super categorías[37][38], por lo que tarda más tiempo en realizar los cálculos[38]. Luego de ello examina la serie de funciones del predictor y busca el conjunto de categorías que brinda la mayor asociación con la variable respuesta y calcula un nivel crítico p corregido para esa asociación, de esta forma encuentra la mejor división para cada predictor comparando los niveles p corregidos[38]. Este algoritmo dio a conocer el porcentaje de éxito o fracaso en base a la posición en la que deben estar las palabras.

- **Random Forest.**

Permite la representación de un árbol diseñado al azar[39], consiste en un conjunto de árboles de decisión[40] que realizan la clasificación tomando como entrada a un vector de características y genera la etiqueta de la clase que recibió la mayoría de los votos en la clasificación[39][35]. Los árboles se construyen de la siguiente manera: al dividir un nodo se escoge el corte realizado de forma óptima en un subconjunto aleatorio de características[40]. Otro dato es que el clasificador es más preciso cuando aumenta el sesgo y disminuye la varianza[40]. En este trabajo permitió identificar la posición más influyente del tema de un proyecto y las palabras que influyen en esa posición.

1.8.2. Algoritmos de Clasificación No supervisada.

Estos algoritmos no requieren un conjunto de datos entrenado, ni de la intervención de humanos para elaborar un conjunto de datos categorizado[36][31]. La meta de los algoritmos de clasificación no supervisada es encontrar patrones considerando la distribución y composición de los datos[36] y organizar por si solo los diferentes valores de entrada dependiendo de si su entrada es binaria o continua[4]. Los algoritmos se basan en la agrupación de los datos y para ello generalmente se usa alguna medida métrica euclidiana[31], pero usualmente antes de aplicar los algoritmos se realiza el pre procesamiento de la información para realizar una agrupación más óptima[24]. Al no requerir un conjunto de datos categorizado, permitió una clasificación de las palabras relacionadas sin que influya la etapa de categorización de los proyectos.

- **Agrupación Jerárquica o Dendograma.**

Muestra la secuencia de los clústeres que se están conformando de acuerdo a la medida de distancia empleada[41], este método de clasificación es de tipo jerárquico por lo cual se obtiene un número creciente de clases anidadas[42]. Consiste en representar gráficamente el número de clústeres que se observan en los distintos niveles del dendograma frente a los niveles de fusión a los que los clústeres se unen en cada nivel[41]. También este método se basa en la existencia de pequeños saltos en los niveles de fusión[41] para organizar los datos en subcategorías que se van dividiendo en otras hasta llegar a un nivel de detalle deseado[42]. Fue útil para realizar el pre procesamiento de los datos y obtener relaciones obvias entre las palabras.

- **Reglas de Asociación Apriori.**

El algoritmo Apriori es el algoritmo de reglas de asociación más utilizado[43], busca posibles relaciones entre elementos para descubrir hechos que ocurren dentro de un conjunto de datos[44][8] y para ello identifica conjuntos de elementos frecuentes de donde luego se derivan las reglas de asociación[45]. Este algoritmo considera cada posible combinación de pares atributo-valor y se denomina cada par como ítem y a su vez el conjunto de ítems se denomina ítem-sets[44]. Para obtener resultados óptimos se debe establecer la condición de confianza para eliminar ítem-sets con un nivel por debajo del valor establecido[44][36]. Otro aspecto a tener en cuenta, es que un ítem-set puede dar más de una regla de asociación, así como también ninguna[44]. Las reglas de asociación ligan cualquier atributo y no solamente las clases de un conjunto de datos[36] por lo que presentan las siguientes medidas de fiabilidad: soporte, soporte absoluto, confianza, elevación, soporte multiplicado por confianza y grupo de reglas[8]. Este algoritmo proporciona relaciones entre las palabras basándose en reglas, por lo cual también se obtuvo un porcentaje de confianza por cada regla obtenida.

- **K-means.**

Este algoritmo es de los más importantes y más utilizados para realizar agrupamiento de datos[24][36] debido a su simplicidad, escalabilidad y velocidad de convergencia[46]. Tiene como objetivo dividir un conjunto de N elementos en un número k de grupos[24] y también minimizar el error cuadrático de la distancia euclidiana con respecto al centroide del agrupamiento al cual pertenece el elemento[36]. Es eficiente porque permite procesar patrones de forma secuencial, consiste en determinar k centros para cada grupo y luego la distancia entre cada centro determina el resultado del agrupamiento[24] debido a que cada elemento es asignado al centroide más

cercano[40]. El proceso del algoritmo se resume en seleccionar los centroides iniciales, asignar los elementos del conjunto de datos a su centroide más cercano, calcular nuevos centros y volver a asignar los objetos a su centroide más cercano hasta que el algoritmo converge[43][47]. En la investigación nos permitió obtener las tres palabras más cercanas a cada centro establecido, obteniendo de esta manera las palabras más relacionadas de cada clúster.

- **K-medoids (PAM).**

Las siglas del algoritmo k-medoids PAM significan Partitioning Around Medoids[41], es un algoritmo que se cambia ligeramente del algoritmo k-means[47]. Este algoritmo se basa en particiones que divide el conjunto de datos en grupos buscando minimizar la distancia entre los objetos que se van añadiendo a un grupo con respecto a un centroide[41]. A diferencia del k-means que usa la media del grupo, el k-medoids usa el punto medio del grupo y es más robusto ante datos atípicos y el ruido[45]. El algoritmo abarca dos fases para construir y mejorar grupos, en la primera se construye grupos usando una distancia y en la segunda se intercambia pares de elementos (i,h), donde i es un centro y h no lo es[46]. Un medoide es el objeto de un grupo cuya disimilitud media a todos los objetos en el grupo es mínima, es decir es el punto ubicado más hacia el centro en todo el grupo[41]. Al ser este algoritmo una variación del algoritmo k-means se buscó obtener igualmente las relaciones más influyentes en cada medoide.

1.9. Herramientas de software para la aplicación de algoritmos de clasificación.

Actualmente existen diversas herramientas tecnológicas que permiten identificar comportamientos o patrones entre los datos[48]. Algunas de esas herramientas son: Openrefine, R Studio, IBM SPSS Statistics y RapidMiner Studio.

1.9.1. Openrefine.

Es una herramienta desarrollada por Google para el tratamiento de datos desordenados, se encarga de limpiarlos y transfórmalos de un formato a otro. También es de ayuda para explorar grandes conjuntos de datos con facilidad y permite vincular y ampliar el conjunto de datos con varios servicios web[49]. Esta herramienta permitió hacer el reemplazo y eliminación de los caracteres establecidos en la limpieza, además de realizar el reemplazo de las relaciones obvias obtenidas en el pre procesamiento de los datos.

1.9.2. R Studio.

Es un entorno de desarrollo para R e incluye una consola y un editor de resaltado de sintaxis que admite la ejecución de código, así como herramientas para el historial, depuración y administración de datos[50]. R proporciona una amplia variedad de técnicas estadísticas tales como clasificación y técnicas gráficas altamente extensibles para la manipulación de datos. Una de las ventajas de R es la excelente calidad de los gráficos que produce[51]. Se realizó la aplicación de los algoritmos de clasificación no supervisados y la obtención de los resultados con sus respectivas gráficas.

1.9.3. IBM SPSS Statistics.

Esta herramienta puede ejecutar estadísticas descriptivas, regresión, estadística, clasificación y mucho más. Para ello permite crear gráficos, tablas y árboles de decisión. Posee una interfaz sencilla que facilita la utilización de la amplia gama de variedades que ofrece[52]. Permite aplicar el algoritmo CHAID Exhaustivo y además ampliar la visualización del gráfico obtenido debido al gran tamaño del gráfico.

1.9.4. RapidMiner Studio.

Esta herramienta posee una interfaz visual atractiva que permite la creación de modelos predictivos, posee una amplia gama de más de 1500 algoritmos y funciones de aprendizaje automático. Permite conectarse a los datos ya sea en almacenamiento local o en la nube, e inclusive documento, redes sociales y aplicaciones empresariales. Además, es posible explorar y visualizar el contenido de un conjunto de datos para identificar problemas como valores perdidos y valores atípicos[53]. En el trabajo esta herramienta sirvió para aplicar el algoritmo Random Forest y para obtener una mejor visualización del resultado del algoritmo.

2. Trabajos Relacionados.

Se describe a continuación estudios relacionados al presente trabajo que sirven de ayuda para comprender la forma en la que se desarrollan temas similares y determinar cómo influye la definición del tema de un proyecto de titulación en su éxito o fracaso mediante la aplicación de algoritmos de clasificación. Según [54] las técnicas de clasificación se utilizan frecuentemente para la solución de diferentes problemas, estas técnicas de aprendizaje automático son las más usadas cuando existe una gran cantidad de datos, en los cuales se encuentran patrones ocultos y presentan ruido entre los patrones.

En el estudio [55] se determina el grado de éxito y fracaso de estudiantes mexicanos para el examen nacional de egreso en ingeniería, para ello hace uso de información que permanece en las instituciones educativas acerca de la deserción, fracaso y rendimiento académico, también presenta que la minería de datos es una técnica que permite deducir fenómenos en el ámbito educativo mediante el análisis de variables que tienden a ser factores influyentes en el rendimiento académico.

Según [56] la minería de texto utiliza técnicas de minería de datos para descubrir conocimientos de datos textuales no estructurados, en este estudio se utiliza varias técnicas de clasificación para identificar habilidades de docentes.

El estudio [57] utiliza la categorización automática de tweets mediante la aplicación de algoritmos de clasificación supervisada, el proceso realizado es el pre procesamiento de los datos, la aplicación de los algoritmos y la presentación, análisis y comparación de los resultados obtenidos.

En el estudio [43] se recolecta información mediante cuestionarios para luego ser procesada usando Weka, en esta herramienta aplica técnicas de minería de datos siguiendo la metodología Proceso de Extracción del Conocimiento para facilitar la interpretación de resultados y encontrar conocimiento útil que sirva para una apropiada toma de decisiones en el tratamiento de la deserción escolar.

El estudio [44] usa el lenguaje R para la extracción, almacenamiento, representación, aplicación de los algoritmos de minería de texto y visualización de resultados. El propósito de este estudio es demostrar la eficacia de las técnicas de minería de texto para encontrar patrones ocultos en datos no estructurados, tal como textos expresados en lenguaje natural.

En el estudio [45] se presentan diversas secciones entre las que tenemos: clustering, importación y exportación de datos, regresión, análisis de sentimientos, reglas de asociación, minería de texto, entre otras. Además, en cada sección presenta en detalle la forma en la se aplica diversas técnicas de minería de datos, por ejemplo, en la sección de minería de texto muestra la aplicación de algoritmos como k-means y k-medoids para encontrar relaciones entre las palabras.

Según el estudio [58] actualmente se presenta como problema para muchas empresas el recopilar, explorar y aprovechar la información que generan de manera no estructurada, lo que hace necesario el uso de la minería de textos. Las técnicas de la minería de textos permiten la explotación analítica de grandes volúmenes de

información, en este estudio se explican sobre estas técnicas y las etapas que se deben seguir para su aplicación, también presenta dos casos prácticos en el que se usa las técnicas explicadas introduciendo datos reales.

e. Materiales y Métodos

Este trabajo de titulación fue desarrollado en la FEIRNNR de la Universidad Nacional de Loja (UNL), quien supervisó el desarrollo del trabajo fue el director del presente trabajo de titulación, además para el cumplimiento del primer objetivo intervinieron en el trabajo las secretarías de la Biblioteca y de la Secretaría General de la FEIRNNR de la UNL aportando información útil mediante entrevistas. También intervino en el presente trabajo la secretaria de la Carrera de Ingeniería en Sistemas aportando información que ayudo al cumplimiento del tercer objetivo en el ingreso de nuevos datos. Para el desarrollo del trabajo se utilizó datos de los proyectos de titulación aprobados para su desarrollo de la FEIRNNR de la UNL y de los trabajos de titulación presentes en la biblioteca de la misma facultad. Además, en el desarrollo del presente trabajo se usó las herramientas Openrefine, RStudio, SPSS Statistics y RapidMiner. Los diferentes materiales, métodos y técnicas científicas descritas en esta sección sirvieron para cumplir con los objetivos planteados y responder a la pregunta de investigación: “¿Cómo ayuda la aplicación de algoritmos de inteligencia artificial a determinar el éxito o fracaso de un proyecto de titulación tomando como base su tema?”.

1. Materiales.

Para el desarrollo del trabajo de titulación se emplearon materiales y herramientas que permitieron llevar a cabo el desarrollo de los objetivos, en la TABLA I son presentados los materiales y herramientas utilizados.

TABLA I
MATERIALES USADOS

Hardware	
Computadora	Fue empleada para la aplicación de los algoritmos de clasificación y la redacción del trabajo.
Impresora	Se usó para imprimir el trabajo en las ocasiones que fue necesario hacerlo.
Dispositivos de almacenamiento	Se utilizó para almacenar los resultados y la redacción del trabajo durante su desarrollo.
Software	
Openrefine	Se utilizó para realizar la depuración del conjunto de datos.

R Studio	Se empleo para aplicar los algoritmos de clasificación no supervisados.
IBM SPSS Statistics	Permitió aplicar el algoritmo CHAID Exhaustivo y obtener una óptima visualización de su resultado.
RapidMiner Studio	Fue usado para la aplicación del algoritmo Random Forest.
Varios	
Papel	Se usó para la presentación del trabajo de titulación.
Internet	Permitió acceder a la información usada en la elaboración del presente trabajo.
Transporte	Necesario para el traslado a los diferentes lugares que permitieron culminar el trabajo.

2. Métodos.

2.1. Método Deductivo.

Este método va de lo general a lo particular, parte de conocimientos generales e información recopilada para determinar un caso de estudio particular. Los pasos empleados en este trabajo son los siguientes:

- Establecer el tema a desarrollar.
- Tutorías guiadas por el director del trabajo.
- Definición de los objetivos.
- Realización de la revisión de literatura.
- Seleccionar y aplicar los algoritmos de clasificación.
- Evaluar los resultados y los algoritmos.
- Determinar si influye o no la definición de un proyecto de titulación en su éxito o fracaso.

2.2. Método Inductivo.

El método inductivo va de los casos particulares a conocimientos generales, se utilizó para estructurar la revisión de literatura con información relacionada al tema y solucionar problemas que se pudieran ocasionar en el método deductivo.

2.3. Método Científico.

Con este método se buscó, analizó y sintetizó los conceptos presentes en la revisión de literatura que dan fundamento teórico al proceso investigativo.

2.4. Método Empírico.

Este método hace uso de la experiencia y la complementación sensorial para realizar una óptima evaluación, categorización y pre procesamiento del conjunto de datos en la introducción de los nuevos datos.

3. Técnicas.

3.1. Observación.

Se empleó este método para realizar la evaluación de los resultados y de los algoritmos de clasificación, fue de ayuda para seleccionar los resultados más influyentes en el éxito o fracaso de los proyectos y asignar los porcentajes de cada relación con respecto al algoritmo CHAID Exhaustivo.

3.2. Entrevista.

Esta técnica fue muy beneficiosa para la realización del proyecto, debido a que permitió obtener información que ayudó a comprender como estaban estructurados los conjuntos de datos, para así realizar una correcta evaluación y categorización de los datos. Las entrevistas se realizaron a las secretarias de la Biblioteca, de la Carrera de Ingeniería en Sistemas y de la Secretaría General de la FEIRNNR de la UNL.

3.3. Muestreo aleatorio simple.

Con el uso de esta técnica se evaluó la validez del conjunto de datos, consistió en establecer una muestra de todos los proyectos que se encontraban previamente clasificados y escoger al azar los proyectos a ser evaluados, ayudando a establecer así cuantos y cuales proyectos evaluar.

4. Metodologías.

Se emplearon dos metodologías para la realización del presente trabajo, la primera se utilizó para la elaboración de la revisión de literatura y la búsqueda de los algoritmos de clasificación. Y la segunda metodología se utilizó para obtener los resultados de la aplicación de los algoritmos de clasificación. Las metodologías empleadas se describen a continuación:

Metodología para la obtención de información.

Para la realización del presente trabajo fue necesario conocer los conceptos útiles, estos conceptos son presentados en la sección de revisión de literatura y fueron desarrollados empleando una metodología que a su vez tuvo como base los estudios de Villavicencio[59], Banegas[60] y García[61]. Los dos primeros estudios presentan la aplicación del método de revisión sistemática de Barbara Kitchenham, en donde algunos de los pasos fundamentales que se usaron son: planificación de la revisión, desarrollo de un protocolo de revisión, desarrollo de la revisión, selección de los estudios primarios, síntesis de datos y publicación de resultados. El tercer estudio presenta un protocolo para realizar una revisión sistemática, cuyos pasos más importantes fueron: preguntas de investigación, criterios de inclusión y exclusión, fuentes de búsqueda, cadenas de búsqueda, consultas, extracción de datos y escribir resultados. Para la realización de la revisión de literatura y selección de algoritmos de clasificación se estableció una metodología basada en algunas etapas fundamentales de la revisión sistemática presentadas anteriormente, las actividades seleccionadas fueron: criterios de inclusión y exclusión, fuentes de búsqueda, cadenas de búsqueda, ejecución de consultas, selección de los estudios primarios y presentación de resultados; revisar el Anexo 11 para obtener mayor información de esta metodología. Las fases para la obtención de información son presentadas a continuación:

- El criterio de inclusión más importante fue que los estudios sean posteriores al año 2014 para sustentar el presente trabajo.
- Para la obtención de los estudios se seleccionó como fuente de búsqueda a Google Académico, debido a que es una fuente conocida y permite obtener los estudios en orden por relevancia, sin limitaciones de idioma e inclusive establecer un intervalo de tiempo. Además, permite realizar una búsqueda avanzada donde se indica funciones como: que el artículo contenga la cadena de búsqueda, que contenga el artículo al menos una palabra de la cadena de búsqueda, entre otras funciones.
- Antes de realizar la búsqueda de los documentos fue necesario definir palabras y cadenas de búsqueda de manera clara y precisa, esto nos permitió obtener resultados óptimos al encontrar estudios adecuados para realizar el fundamento teórico del presente trabajo, algunas de las cadenas empleadas para la revisión de literatura fueron: inteligencia artificial, algoritmos de clasificación

supervisados y no supervisados, minería de datos y textos, aprendizaje automático y minería de textos.

- Las consultas en la fuente de búsqueda previamente seleccionada se las realizó haciendo uso de las palabras clave y las cadenas de búsqueda definidas.
- Se eliminó los estudios duplicados que se obtuvieron al consultar en la fuente de búsqueda, luego se seleccionó los estudios revisando el texto completo de cada estudio aplicando los criterios de inclusión y exclusión.
- Dos casos se tomaron en cuenta para la presentación de resultados. En caso de la revisión de literatura se analizó, sintetizó y parafraseo los conceptos buscados, y en caso de la búsqueda de algoritmos se presentó en una tabla los estudios seleccionados que cumplan con los criterios de inclusión y exclusión.

Metodología para la minería de datos.

Para el cumplimiento de los objetivos del presente trabajo, se utilizó como base la metodología CRISP-DM y la metodología KDD. La primera es una de las metodologías más usadas para analizar grandes conjuntos de datos y descubrir información valiosa, está compuesta por seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e interpretación. La metodología KDD es un proceso enfocado a descubrir patrones útiles y comprensibles a partir de un conjunto de datos, está compuesta por seis fases: sistema de información, preparación de los datos, minería de datos, patrones, evaluación o interpretación, y presentación del conocimiento. Partiendo de lo explicado anteriormente, las fases de la metodología empleada son las siguientes: comprensión de la información, pre procesamiento de los datos, búsqueda y selección de los algoritmos, aplicación de los algoritmos y evaluación de los algoritmos. Se describe cada una de las fases de la metodología empleada a continuación:

- **Comprensión de la información**, esta fase consistió en identificar las variables útiles para el desarrollo del trabajo, se evaluó la validez del conjunto de datos con el que se trabajó y se realizó la categorización de los proyectos. Para identificar estas variables se realizó una entrevista a la secretaria de la Secretaría General de la FEIRNNR. Una vez identificadas las variables útiles y las que no son útiles, se procedió a descartar las variables no útiles ya que no fueron utilizadas para realizar la categorización de los proyectos, posteriormente se evaluó la validez del conjunto de datos, usando el muestreo aleatorio simple para establecer una muestra de los proyectos que ya se encontraron

categorizados, comparándolos con los proyectos del conjunto de datos de la biblioteca.

- En el **pre procesamiento del conjunto de datos** se realizó la depuración de los datos, estandarizando el contenido de la variable “TEMA”. Para ello se utilizó la herramienta Openrefine que fue desarrollada por Google para realizar el refinamiento de datos, se procedió a eliminar y reemplazar tildes y otros caracteres(tokenización), y por último se realizó una exploración de los datos para localizar y reemplazar oraciones como “Universidad Nacional de Loja” por “UNL”, ya que las palabras que forman “UNL” obviamente se encuentran relacionadas y producen ruido al alterar las relaciones en todo el conjunto de datos.
- Para la **búsqueda de los algoritmos** de clasificación se buscó estudios similares al presente trabajo, que cumplieran con algunos criterios establecidos y de los cuales se escogió los algoritmos más usados. A estos algoritmos se los comparó en base a características establecidas a partir de los resultados y conclusiones presentes en los estudios analizados, estas características fueron: porcentaje de eficiencia, representación gráfica, efectividad en grandes conjuntos de datos, especificar cantidad de agrupamientos y disponibilidad en las herramientas elegidas para la aplicación de los algoritmos. Los algoritmos escogidos fueron los que cumplieron con más características, es decir los que presentan más ventajas en relación a otros algoritmos.
- La **aplicación de los algoritmos de clasificación** se la realizó en el conjunto de datos pre procesado, para aplicar los algoritmos no supervisados se utilizó la herramienta RStudio, para el algoritmo CHAID Exhaustivo se utilizó la herramienta SPSS Statistics y para el algoritmo Random Forest se usó la herramienta RapidMiner. Para aplicar los algoritmos de clasificación supervisados seleccionamos las variables “TEMA” y “ESTADO” del conjunto de datos, la primera variable que tiene los temas de los proyectos de titulación se la separó en diferentes columnas, tantas como palabras tenía el tema sin tomar en cuenta los conectores (por, para, con, etc.) de cada tema. Los algoritmos de clasificación no supervisados se aplicaron en todo el conjunto de datos y posterior a ello se procedió a separar el conjunto de datos en dos subconjuntos uno por categoría, para aplicar los algoritmos en cada uno de los subconjuntos, es decir aplicar cada algoritmo solamente en los proyectos que culminaron con éxito y luego en los proyectos que fracasaron.

- La fase final consistió en **evaluar los resultados** proporcionados por los algoritmos y la eficiencia de los mismos. La evaluación de resultados se realizó comparando los resultados de todo el conjunto de datos con los obtenidos en cada clase. Cuando un resultado está presente en la clase éxito y en la clase fracaso, es descartado debido a que no indica si influye en el éxito o el fracaso del proyecto de titulación. Por otra parte, los algoritmos de clasificación supervisados generalmente son evaluados en base a su porcentaje de eficiencia, pero en el caso del algoritmo Random Forest este no presenta ese porcentaje, así que se lo evaluó igual que los algoritmos de clasificación no supervisada. Usualmente la **evaluación de los algoritmos de clasificación** no supervisados se realiza usando las etiquetas de cada instancia o se evalúan de acuerdo a la interpretación humana. En este caso los proyectos poseen etiquetas, es decir se encuentran categorizados, pero no es posible realizar una evaluación en base a las etiquetas debido a que los algoritmos fueron aplicados en dos subconjuntos uno por categoría, por lo que no tendría sentido evaluar de esta forma. Tomando en cuenta lo anterior, los algoritmos de clasificación no supervisada se evaluaron de acuerdo a la interpretación humana usando como referencia al algoritmo CHAID Exhaustivo, ya que es el único algoritmo que presenta un porcentaje de eficiencia.

En la Figura 1 se observa todo el proceso de la metodología usada para la minería de datos.

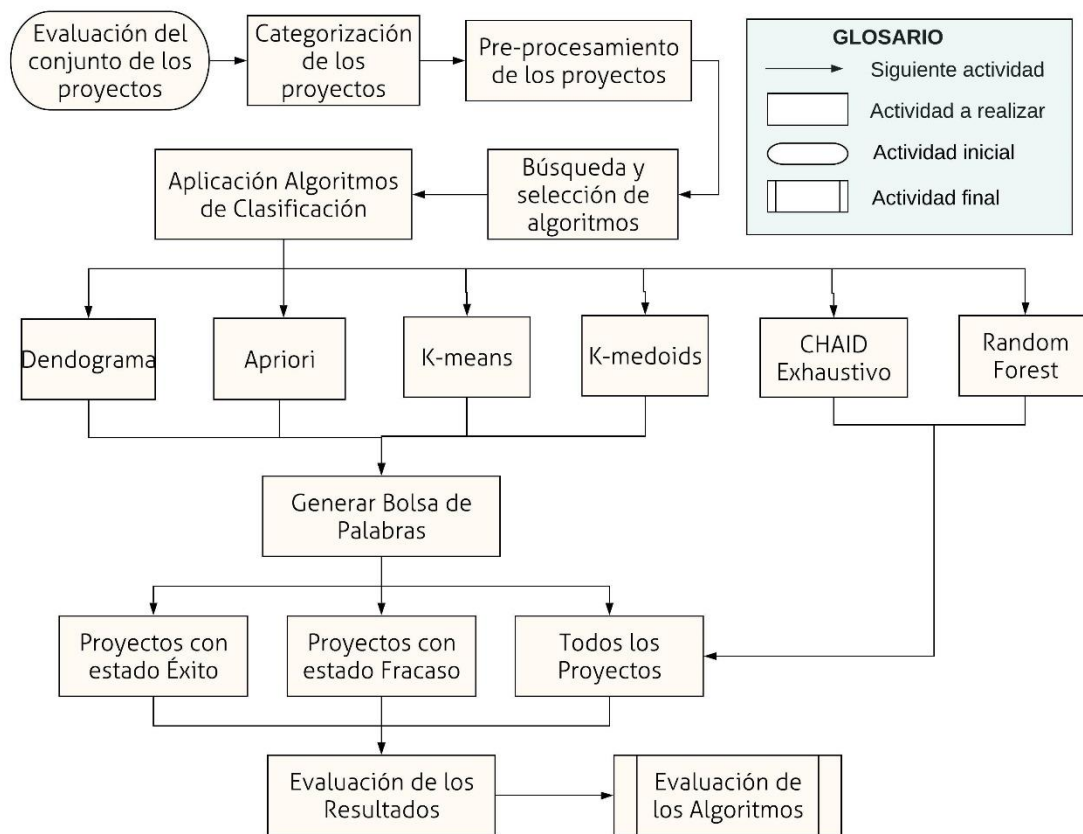


Figura 1. Proceso de la metodología para la aplicación de los algoritmos.

f. Resultados

En esta sección se presentan los resultados obtenidos del proceso mostrado en la Figura 1, donde se detallan las actividades de la metodología utilizada para el cumplimiento de los objetivos del presente trabajo de titulación.

1. Depuración del conjunto de datos obtenido de los proyectos de titulación.

1.1. Evaluación del conjunto de datos.

El conjunto de datos evaluado fue el obtenido de la Secretaría General de la FEIRNRR, en el **Anexo 1 se observa la solicitud para obtener el conjunto de datos** y en el **Anexo 7 se visualiza la certificación de haber obtenido el conjunto de datos**. La categorización de los proyectos de titulación se realizó tomando en cuenta la información obtenida del **Anexo 2 que presenta la entrevista realizada acerca del conjunto de datos**.

REGISTRO DE APROBACIÓN DE PROYECTOS DE TESIS									
INGRESO	PROYECTO	NOMBRE DE LA TESIS	POSTULANTES	ESPECIALIDAD	FECHA DE APROBACIÓN	DIRECTOR	TIEMPO DE EJECUCIÓN	OBSERVACIONES	FECHA DE GRADO
	133	DISEÑO Y CONSTRUCCIÓN DE UN AEROGENERADOR EXPERIMENTAL MODULAR PARA APLICACIÓN RURAL	Oscar Iván Cabrera González Darvin Rigoberto Cuenca Guinde	Ingeniería Electromecánica	11/05/2006	Ing. Jorge Luis Maldonado Correa	11 meses		15/09/2006
	134	METODOLOGÍA PARA LA APLICACIÓN DE ENYESOS NO DESTRUCTIVOS MEDIANTE ULTRASONIDO	Vicente Manuel Álvarez Vega, Carlos Manuel Jimbo Muñoz, Galo Fernando Medina Piviera, Alex Patricio Ordóñez Páico	Ingeniería Electromecánica	19/05/2006	Ing. Flaminio Gonzalo Rieffro Cruz	10 meses		grado: 19/05/06
anulado	135	APLICACIÓN PARA EL CONTROL DE GARANTÍAS DE EQUIPOS Y COMPONENTES DE COMPUTACIÓN APOYADOS EN LOS PROCESOS DE FACTURACIÓN E INVENTARIO, UTILIZANDO COMO METODOLOGÍA INTEGRADORA DE APLICACIONES PARA UN RÁPIDO DESARROLLO Y UML COMO METODOLOGÍA DE MODELADO	Mónica María Díaz Samaniego, Pablo Osvaldo Córdova García	Ingeniería en Sistemas	30/06/2006	Franco Salcedo López	6 meses	anulado HCA: 5 agosto 2009	
ANULADO	136	AUTOMATIZACIÓN DE FERTIRRIGACIÓN PARA LOS TERRENOS DEL PROYECTO RECURSOS FITOGENÉTICOS DEL ÁREA AGROPECUARIA Y	Hernán Diego León Abarca, Juan Carlos Rodríguez Daquintero, Juan Carlos Villamagua Armiños	Ingeniería Electromecánica	29/06/2006	Ing. Darwin Geovanny Tapia Peralta	11 meses	ANULADO Of. 180-CA 08/07/2014	
	137	ELABORACIÓN DE UN PORTAL WEB PARA LA CÁMARA DE LA CONSTRUCCIÓN DE LOJA	Yadira Jhoana Banegas Michay y César Leonardo Delgado Correa	Ingeniería en Sistemas	18/07/2006	Ing. Hernán Leonardo Torres Carrión	9 meses		
	138	DISEÑO DE UN PLAN PILOTO DE TELEMETRÍA Y CONTROL DEL CONSUMO DE ENERGÍA ELÉCTRICA	Eggar Geovanny Alvarez chavez Jaiko Patricio Cabrera Medina	Ingeniería Electromecánica	18/07/2006	Ing. Juan Carlos Solano Jiménez	10 meses	Reglamento Estructuración de las Áreas	13/11/2006
	139	DISEÑO Y CONSTRUCCIÓN DE UN TABLEO DIDÁCTICO DE TRANSFERENCIA DE ENERGÍA ELÉCTRICA PARA EL LABORATORIO DE ELECTRICIDAD	Weslio Antonio Ortega Guamán y Milton Leodán León Aguilera	TECNOLOGÍA EN ELECTRICIDAD	19/07/2006	Ing. Juan Carlos Dicho Alarzo HCA: 10/06/2011 cambia director a 2. Ing. José Espinoza 3. Ing. Norman Augusto Jiménez León	5 meses	Reglamento Estructuración de las Áreas	29/04/2011
	140	SISTEMA INFORMÁTICO PARA LA ELABORACIÓN Y RESOLUCIÓN DE EMERGENCIAS EN LÍNEA, UTILIZANDO LA TECNOLOGÍA JAVA Y EL PROTOCOLO SSL (CAPA SEGURA DE SOCKETS) PARA EL ÁREA DE ENERGÍA LAS INDUSTRIAS Y LOS RECURSOS NATURALES NO RENOVABLES DE LA UNIVERSIDAD NACIONAL DE LOJA	María Magdalena Armiños JumboSilvia Consuelo Ordóñez Escobar	Ingeniería en Sistemas	20/07/2006	Ing. Vilman Patricio Chamba ZaragozínZaragozín		Reglamento Estructuración de las Áreas	30/03/2006
	141	DISEÑO E IMPLEMENTACIÓN DE LA BIBLIOTECA VIRTUAL PARA EL ÁREA DE ENERGÍA LAS INDUSTRIAS	MARÍA FERNANDA CASTILLO TRIZAPAY	Sistemas	10/09/2006	Ing. Yvett Palacios	10 meses	Reglamento Estructuración de las Áreas	

Figura 2. Conjunto de datos de los proyectos aprobados para su desarrollo.

El conjunto de datos de los proyectos de titulación aprobados para su desarrollo se muestra en la Figura 2, donde se observa que contó con los siguientes títulos de las columnas:

- Ingreso.
- Proyecto.
- Tema.
- Autores.
- Especialidad.

- Director.
- Fecha de aprobación
- Tiempo de ejecución.
- Observaciones.
- Fecha de Grado.

En el campo “Tema” se encontró un gran número de errores de ortografía y caligrafía en las palabras, el campo “Especialidad” presentó de igual forma errores de esa naturaleza. Otro aspecto importante es que un gran número de proyectos de titulación se encontraban pintados de color rojo y con la palabra anulado o abandonado en alguno de sus campos indicando que estos proyectos habían fracasado, pero para realizar una correcta asignación de los estados al resto de los proyectos, se procedió a eliminar los campos innecesarios, tales como:

- Autores.
- Director.

Al campo “Especialidad” se lo mantiene en el conjunto de datos para que sea posible identificar a que carrera pertenece cada proyecto de titulación. Luego para descartar los proyectos de titulación que aún siguen en ejecución se analizó los campos “Fecha de aprobación” y “Tiempo de ejecución”, en donde se evaluó que el tiempo entre la fecha de aprobación y la fecha de ejecución estén fuera de la fecha actual para descartar los proyectos que aún cuentan con tiempo para su desarrollo. Posterior a ello también se procedió a eliminar los siguientes campos por ser innecesarios:

- Fecha de aprobación.
- Tiempo de ejecución.

Sin tomar en cuenta el campo “Especialidad”, los campos restantes que nos ayudaron a determinar el estado final del proyecto son:

- Ingreso.
- Proyecto.
- Tema.
- Observaciones.
- Fecha de Grado.

Una vez establecidos los campos útiles para el desarrollo del proyecto se procedió a evaluar la validez del conjunto de datos, para ello se tomó en cuenta la información obtenida de las entrevistas y se realizó lo siguiente:

- Identificar los proyectos que se encuentran categorizados, ya que en su mayoría se encuentran pintados de un color que indica su estado de éxito o fracaso, como se muestra en la Figura 2.
- Establecer una muestra de los proyectos anteriormente identificados para compararlos con los proyectos del conjunto de datos de la biblioteca. Los proyectos categorizados como exitosos deben estar presentes en el conjunto de datos de la biblioteca para una validez exitosa del conjunto de datos y los proyectos categorizados como fracasados no deben estar.
- Determinar la validez del conjunto de datos con respecto a los resultados obtenidos.

Muestra

Se identificó que, de los 1502 proyectos presentes en el conjunto de datos 1023 proyectos se encuentran ya categorizados, por lo que esta cantidad es el tamaño de la población y se les asignó un identificador del 1 al 1023, de ahí para establecer una muestra se usó el muestreo aleatorio simple.

$$n = \frac{N * Z_a^2 * p * q}{d^2 * (N - 1) + Z_a^2 * p * q}$$

Figura 3. Muestreo aleatorio simple.

Usando la fórmula de la Figura 3 se estableció como parámetros: N=1023 por ser el tamaño de la población, $Z_a = 1.96$ que indica el 95% de nivel de confianza, $d = 5$ es el error muestral deseado, $p = 0.5$ y $q = 0.5$ que indican que cada individuo tiene 50% de probabilidad de pertenecer o no a la muestra. Con los parámetros anteriores se obtuvo que el tamaño de la muestra es de 279 proyectos, para seleccionar al azar los proyectos a ser evaluados se usó el número asignado a cada proyecto y la función **sample(1:1023,279,replace=F)** en la **herramienta RStudio**, donde 1:1023 es el número de proyectos clasificados o rango de números a ser elegidos, 279 es la muestra y **replace=F** es para indicar que no se repitan los números aleatorios. En la Figura 4 se observan los números aleatorios obtenidos.

```

> sample(1:1023,279,replace=F)
[1] 534 689 635 929 210 655 843 671 67 539 346 917 793 1020 518 224 276 154
[19] 32 887 205 706 443 982 1011 634 657 743 170 831 662 661 741 371 246 735
[37] 239 492 958 602 792 370 187 631 947 628 59 919 824 70 1012 191 758 778
[55] 852 220 610 601 292 458 332 802 699 155 400 313 767 619 730 896 691 71
[73] 258 425 133 362 110 766 497 750 777 251 197 558 412 174 855 149 175 680
[91] 895 837 543 989 844 959 639 459 822 925 725 867 714 322 472 185 293 343
[109] 646 493 115 702 957 65 300 789 218 330 49 203 173 405 728 369 287 283
[127] 936 875 594 315 554 392 963 637 851 349 973 966 850 424 891 826 99 96
[145] 722 583 450 427 703 484 48 83 674 167 506 675 652 479 530 950 409 827
[163] 603 968 682 498 830 241 288 782 560 74 615 139 804 347 62 146 333 206
[181] 525 127 20 665 10 513 63 505 876 672 402 471 723 903 55 825 81 466
[199] 357 742 140 629 41 19 314 854 893 166 660 727 984 481 243 58 381 745
[217] 131 946 640 562 248 627 898 15 507 998 245 86 654 902 395 575 85 410
[235] 931 298 134 884 690 383 467 796 650 751 669 449 312 696 101 738 685 160
[253] 335 406 451 999 620 550 954 378 870 523 169 874 773 695 310 828 267 97
[271] 204 142 678 119 444 261 990 666 368

```

Figura 4. Números aleatorios obtenidos.

Los proyectos con los identificadores 619, 83, 381 y 86 fueron proyectos culminados con éxito, pero no se encontraban presente en el conjunto de datos de la biblioteca, por lo que se determinó que estaban mal categorizados. De los 279 proyectos evaluados solo 4 se encontraron mal categorizados, lo que representa que el 98.56% de los proyectos se encuentran bien categorizados. Con el resultado anterior se determina que el conjunto de datos es válido y por lo tanto es factible usarlo en el presente trabajo.

1.2. Categorización de los proyectos de titulación.

Luego de haber validado el conjunto de datos se procedió a la categorización de los proyectos, por ello en el conjunto de datos se añadió el campo “Estado” para asignarle a cada proyecto una clase que puede ser de éxito o fracaso. Antes de asignar las clases, se usó la columna “Especialidad” para descartar los proyectos de titulación de maestrías y tecnologías, ya que tal como nos lo indican los **Anexos 4, 5 y 9 que presentan la certificación e información obtenida del conjunto de datos de la biblioteca**, estas especialidades ya no se ofertan en la facultad y no representan información útil.

Por último, se elimina el campo “Especialidad” debido a que no es útil para determinar el estado final de los proyectos. El resto de campos del conjunto de datos fueron indispensables para determinar el estado de los proyectos, por lo que a continuación se habla detalladamente de cada uno de ellos.

En los campos “Ingreso” y “Proyecto” nos encontramos con dos casos:

- Cuando se encuentran vacíos, no fueron de utilidad para establecer el estado del proyecto.
- Cuando contiene la palabra anulado o abandonado, nos permitió establecer el fracaso del proyecto.

En el campo “Fecha de grado” nos encontramos con tres casos diferentes:

- Se encuentra vacío: No se establece el estado del proyecto.
- Tiene la fecha, pero con la palabra complejo: Se establece que el proyecto fracasó, ya que se graduó por la modalidad de examen complejo.
- Tiene simplemente la fecha: Se determina que son proyectos que concluyeron con éxito.

En el campo “Observaciones” hay dos casos, vacío o con información, en el segundo caso se encuentra información importante del proyecto, información tal como:

- El proyecto fue anulado.
- El proyecto fue abandonado.
- Prorroga.
- Cambio de objetivos.
- Renuncia al proyecto.

De esto lo que nos ayuda a establecer el fracaso del proyecto son:

- El proyecto fue anulado.
- El proyecto fue abandonado.
- Renuncia al proyecto.

Estos casos, tanto de “Observaciones” como de “Fecha de grado” están relacionados con el estado final del proyecto de la siguiente manera:

- Sin información de fracaso en observaciones y fecha de grado vacío.
- Con información de fracaso en observaciones y fecha de grado vacío.
- Con información de fracaso en observaciones y fecha de grado con complejo.
- Sin información de fracaso en observaciones y fecha de grado con complejo.
- Sin información de fracaso en observaciones y con fecha de grado.
- Información incompleta del tema y sin datos en el resto de columnas.

De los casos anteriores los que establecen que fracasó el proyecto son los siguientes:

- Con información de fracaso en observaciones y fecha de grado vacío.
- Con información de fracaso en observaciones y fecha de grado con complejo.
- Sin información de fracaso en observaciones y fecha de grado con complejo.

El caso que se establece los proyectos que culminaron con éxito es el siguiente:

- Sin información de fracaso en observaciones y con fecha de grado.

Hay que tomar en cuenta que existen dos casos particulares que fueron tratados de diferente manera al resto, y son las siguientes:

- Información incompleta del tema y sin datos en el resto de columnas, en este caso al no contar con información completa y no tener semántica en el campo del tema se descartó a los proyectos pertenecientes a este caso.
- Sin información de fracaso en observaciones y fecha de grado vacío, este caso al no contar con información útil para determinar el éxito o fracaso del proyecto se lo evaluó en base al conjunto de datos obtenido de la biblioteca, tal como indica el **Anexo 3 referente a la solicitud para obtener ese conjunto de datos**. Tomando como base el **Anexo 2** y el **Anexo 4 en los cuales se describen las entrevistas realizadas**, a todos los proyectos de titulación pertenecientes a este caso, se los buscó si existen en el conjunto de datos de la biblioteca para asignarles el estado de que culminaron con éxito, por otra parte, a los proyectos de este caso que no se encuentren en ese conjunto de datos se les asignara que fracasaron.

Una vez realizada la categorización de los proyectos, se elimina los campos innecesarios, dejando así los campos “Tema” y “Estado” para obtener el archivo “0_Conjunto_de_Datos_m.csv” y proceder a la limpieza del conjunto de datos, el conjunto de datos obtenido se lo visualiza en la Figura 5, donde se obtuvo 1226 proyectos de los cuales 834 son proyectos con estado de éxito y 392 proyectos con estado de fracaso.

Diseno y construccion de un tablero didactico con transformadores de medida	Fracaso
Implementacion de una guia de practicas de redes lan con un router linksys	Fracaso
Control domotico del sistema de luces del aula magna del Area de la Energia las Industria y los Recursos Naturales No Renovables de la Universidad Nacional de Loja	Fracaso
Desarrollo de una aplicacion web que permita administrar programaciones de eventos basados en reglas para un hogar inteligente haciendo uso de un computador de bajo costo	Fracaso
Diseno de red electrico subterranea para la parroquia Malacatos canton Loja	Fracaso
Implementacion de practica de laboratorio de electronico basadas en arduino	Fracaso
Estimacion de reserva y eleccion de la mejor alternativo en la explotacion de caolines enla concesion mineria los crueceros barrio Cola parroquia Guachanama canton Paltas	Exito
Sistema de gestion medica para el departamento de bienestar estudiantil y policlinico de Motupe	Exito
Generador de contenido academico hipermedia basados en sml	Exito
Sistema de gestion administrativo para el Area de la Energia las Industria y los Recursos Naturales no Renovables de la Universidad Nacional de Loja	Exito

Figura 5. Conjunto de datos categorizado.

1.3. Eliminación de información innecesaria.

La limpieza de información innecesaria del conjunto datos consiste en limpiar información no útil que pueda afectar las relaciones entre las palabras, se la realizó en el campo tema del conjunto de datos, por lo cual, se observó que caracteres pueden incidir para una mala aplicación de los algoritmos de clasificación, y se llegó a determinar los siguiente:

- El conjunto de datos cuenta en el campo tema con letras mayúsculas y minúsculas, esto afecta considerablemente al no ser reconocidos los caracteres como iguales cuando se compara una cadena. Por ello para estandarizar, a todas las letras se las convirtió a minúsculas.
- Las tildes en las palabras afectan de igual forma al reconocer un carácter con tilde y otro sin tilde como diferentes. En el conjunto de datos existen palabras con errores ortográficos, por lo que se estandariza reemplazando todas las vocales con tilde por vocales sin tilde.
- El carácter punto y coma no solo está como separador de los campos, también se encuentra dentro de los temas, por ello se lo eliminó.
- Se eliminó caracteres como los signos de puntuación e interrogación que no deben ser reconocidos como palabras. También fueron eliminadas

combinaciones de letras y símbolos que no son palabras del lenguaje español, lenguaje en el que están escritos los temas de los proyectos.

- La letra “ñ” forma parte de algunos temas, pero representa en R un problema al no presentarse correctamente en las gráficas y palabras como “Diseño” se encuentra en algunos temas como “Diseno”, por tal motivo fue reemplazada por la letra “n”.

A continuación, en la TABLA II se presenta todos los caracteres que se reemplazaron o eliminaron en el conjunto de datos:

TABLA II
CARACTERES REEMPLAZADOS O ELIMINADOS

	Carácter Original	Carácter Actual
Caracteres Reemplazados		
Tildes	á	a
	é	e
	í	i
	ó	o
	ú	u
Letras	ñ	n
Caracteres o Cadenas Eliminadas		
Cadenas	(YY)	(?Y)
	(??)	(Y?)
Paréntesis	()
Comillas	"	'
Puntos	:	.
Slash y backslash	/	\
Guiones	-	—
Comas y tildes	,	´
Mas, y punto y coma	+	;
Símbolos y Números	\$	1 2 3 4 5 6 7 8 9 0

Para limpiar el conjunto de datos se usó la herramienta de software Openrefine, este software fue creado por Google para realizar específicamente el refinamiento de un conjunto de datos, usa lenguajes como Python o GREL (Lenguaje de Expresión de Refinamiento General), este segundo lenguaje fue el empleado en la depuración del conjunto de datos. Se cargó el conjunto de datos en Openrefine, seleccionando la codificación UTF-8, tal como se muestra en la Figura 6.

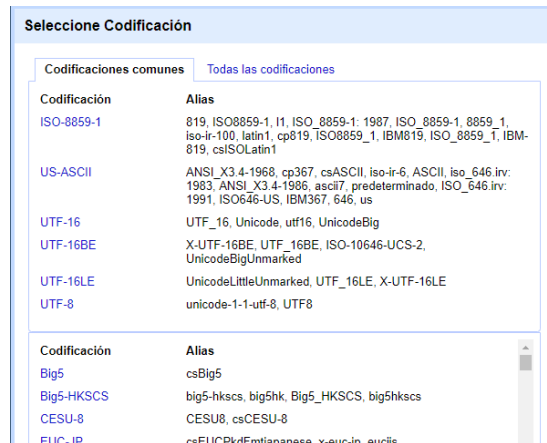


Figura 6. Cargar datos en Openrefine.

Una vez que se cargó el conjunto de datos, se corrigió palabras mal escritas tales como: “pilteas” a “piletas”, “codificacion” a “codificación”, “pirmarios” a “primarios”, entre otras. En la Figura 7 se observa cómo se corrigió a las palabras en cuanto a la semántica más no en ortografía, ya que se reemplazaron las tildes.

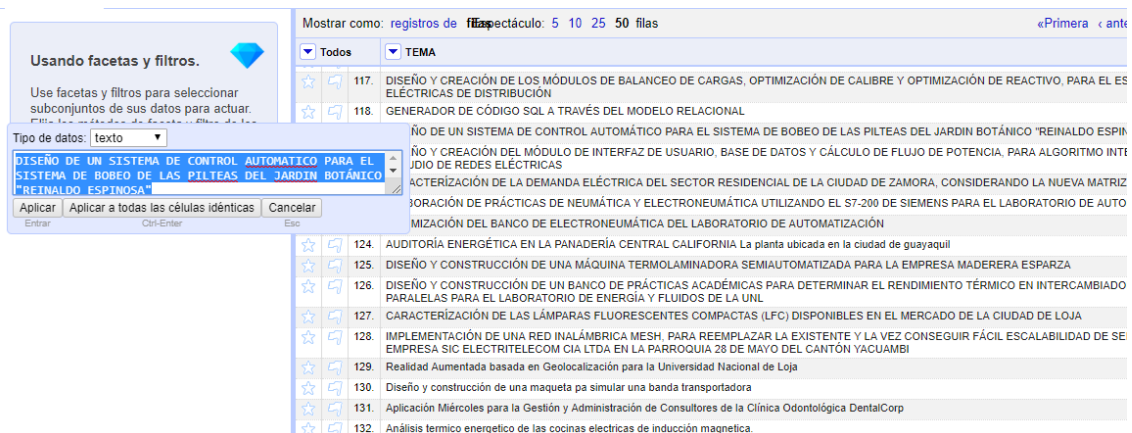


Figura 7. Reemplazo de palabras mal escritas.

Antes de reemplazar las tildes es necesario estandarizar primero la nomenclatura de los temas convirtiendo todos los caracteres a minúsculas, solo se debe elegir la opción “a minúsculas”, como se indica en la Figura 8.

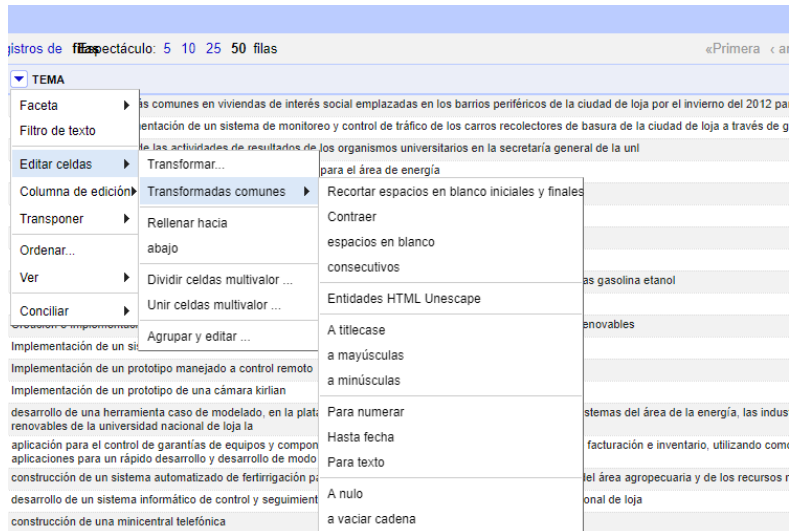


Figura 8. Conversión de texto a minúsculas.

La Figura 9 muestra que para reemplazar las tildes se empleó el comando `value.replace("valor actual", "valor nuevo")`, por ejemplo: `value.replace("á", "a")`. Esto servirá para estandarizar palabras como: casos en los que la palabra "informática" lleva tilde y casos en los que no.

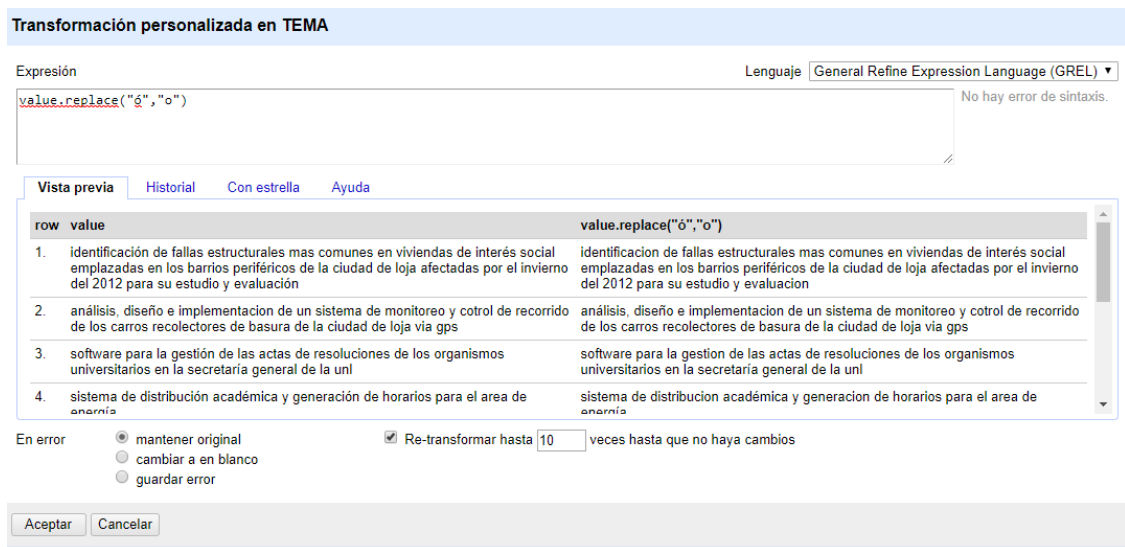


Figura 9. Estandarización de las tildes.

Se debe eliminar caracteres especiales, tales como "(" o ")" que en el caso de "(eerssa)" se encuentra en otros temas sin paréntesis, para estandarizar estos casos se eliminaron los caracteres que se indican en la TABLA //. Para eliminar los caracteres especiales se los reemplazó por un espacio, para al final no unir palabras de forma indebida, ya que por ejemplo en el caso de eliminar una coma dos palabras pueden unirse afectando así la calidad del conjunto de datos, por ello se utilizará espacios en la eliminación de

caracteres. Nuevamente se empleará el comando: **value.replace(“valor actual”, “valor nuevo”)**, tal como se observa en la Figura 10.

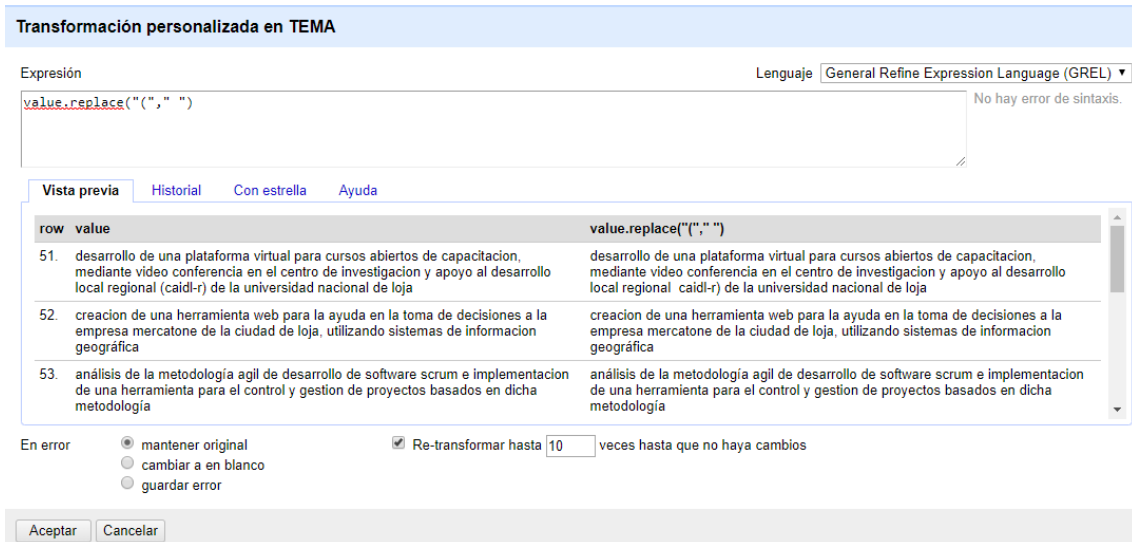


Figura 10. Reemplazo de caracteres especiales.

También se encontraron combinaciones alfanuméricas como: “200.OO” que fueron eliminadas del conjunto de datos. Antes de reemplazar palabras redundantes por una palabra estándar, se usaron las opciones de “Quitar espacios al inicio y final” y “Contraer espacios consecutivos”, como se indica en la Figura 11.

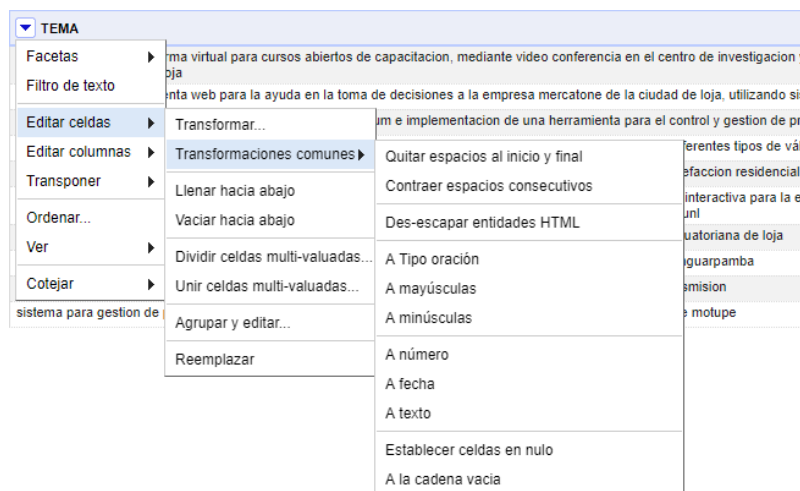


Figura 11. Eliminación de texto alfanumérico.

Por último, se procedió a eliminar palabras redundantes y nombres propios que producen ruido en el conjunto de datos, como ejemplo tenemos: la palabra análisis y la palabra analizar que implican lo mismo. Para los nombres propios se eliminó los espacios entre dos o más nombres, es decir si está presente la frase “bernardo

valdivieso” quedó como “bernardovaldivieso” para indicar que todo el nombre se refiere a la misma institución. Para realizar el reemplazo de las palabras redundantes de manera más eficiente se debió separar en columnas las palabras de cada tema, tal como muestra la Figura 12.

TEMA 1	TEMA 2	TEMA 3	TEMA 4	TEMA 5	TEMA 6	TEMA 7	TEMA 8	TEMA 9	TEMA 10	TEMA 11	TEMA 12	TEMA 1
identificación	de	fallas	estructurales	mas	comunes	en	viviendas	de	interés	social	emplazadas	en
análisis,	diseño	e	implementacion	de	un	sistema	de	monitoreo	y	control	de	recorrido
software	para	la	gestion	de	las	actas	de	resoluciones	de	los	organismos	universitario
sistema	de	distribucion	académica	y	generacion	de	horarios	para	el	area	de	energía
sistema	informático	de	gestion	deportiva	para	la	liga	deportiva	universitaria	de	loja	
desarrollo	del	mapa	virtual	de	la	ciudad	de	loja	y	sus	cantones	
estudios	de	costos	marginales	en	distribucion	para	la	empresa	eléctrica	regional	del	sur
elaboracion	de	kits	de	electronica	de	potencia	y	guía	de	prácticas	para	estudiantes
análisis	de	los	parámetros	de	eficiencia	de	los	motores	de	combustion	interna	para
diseño	y	construccion	de	un	tablero	didáctico	de	transferencia	eléctrica	automatica	desde	un

Figura 12. Separación de variable tema en columnas.

En la Figura 13 se observa que luego de separar las palabras de los temas de los proyectos, se aplicó algoritmos presentes en Openrefine que encuentran palabras semejantes, facilitando así el reconocimiento de palabras redundantes y el reemplazo de estas palabras por una sola que represente a estas.

Agrupar y editar valores en la columna "TEMA 1"

Esta función le permite encontrar agrupaciones de diferentes valores que pueden ser representaciones alternativas de la misma cosa. Por ejemplo, "New York" y "new york" probablemente se refieren al mismo concepto, solo se presenta diferencia en la capitalización. De la misma manera "Gödel" y "Godel" probablemente se refieren a la misma persona. [Más información ...](#)

Método: Función: 13 clusters Encontrado

Número de valores	Número de filas	Valores en la agrupación	¿Unir?	Nuevo valor de las celdas
4	67	<ul style="list-style-type: none"> análisis (38 rows) análisis (20 rows) análisis, (8 rows) análisis, (1 rows) 	<input type="checkbox"/>	análisis
4	13	<ul style="list-style-type: none"> calculo (5 rows) cálculo, (4 rows) cálculo (3 rows) calculo, (1 rows) 	<input type="checkbox"/>	calculo
2	6	<ul style="list-style-type: none"> levantamiento (5 rows) levantamiento, (1 rows) 	<input type="checkbox"/>	levantamiento
2	15	<ul style="list-style-type: none"> auditoria (9 rows) auditoria (6 rows) 	<input type="checkbox"/>	auditoria
2	65	<ul style="list-style-type: none"> implementacion (63 rows) implementacion, (2 rows) 	<input type="checkbox"/>	implementacion
2	403	<ul style="list-style-type: none"> diseño (378 rows) diseño, (25 rows) 	<input type="checkbox"/>	diseño
2	17	<ul style="list-style-type: none"> metodología (13 rows) 	<input type="checkbox"/>	metodología

Valores en la agrupación: 2 — 4

Filas en la agrupación: 0 — 410

Longitud promedio de los valores: 6.5 — 15

Varianza de los valores: 0 — 0.5

Figura 13. Estandarización de palabras semejantes.

Ya que se terminó la limpieza del conjunto de datos, lo exportamos en un archivo con formato CSV, este archivo reemplazó al archivo anterior puesto que se encuentra estandarizado y conservó el mismo nombre “0_Conjunto_de_Datos_m.csv”.

2. Aplicación de algoritmos de clasificación al conjunto de datos de los proyectos de titulación.

2.1. Búsqueda de algoritmos de clasificación.

El presente trabajo se realizó aplicando diversos algoritmos de clasificación, para encontrar estos algoritmos se procedió a la búsqueda de estudios relacionados aplicando la metodología planteada en la sección anterior.

Los criterios de inclusión establecidos son:

- Estudios a partir del año 2014.
- Estudios que estén relacionados con al menos 3 de los 4 temas generales de la revisión de literatura, estos temas son: aprendizaje automático, minería de texto, minería de datos y algoritmos de clasificación.

El criterio de exclusión es todos los estudios que no cumplan con los criterios de inclusión. Algunas de las palabras clave definidas para realizar la búsqueda en la fuente Google Académico son: algoritmos de clasificación, clasificación no supervisada, aprendizaje automático y minería de texto, minería de datos y algoritmos de clasificación, entre otras. Luego de realizar la búsqueda y selección de estudios se obtuvo los siguientes:

TABLA III
ESTUDIOS SELECCIONADOS

Id	Estudios analizados	Año de publicación
E1	Aplicación de árboles de clasificación a la detección precoz de abandono en los estudios universitarios de administración y dirección de empresas[62].	2017
E2	Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos[55].	2015
E3	Aplicación de Algoritmos de Clasificación de Minería de Textos para el Reconocimiento de Habilidades de E-tutores Colaborativos[56].	2014
E4	Minería de Datos para segmentación de clientes en la empresa tecnológica Master PC[47].	2015

E5	Técnicas de aprendizaje de máquina utilizadas para la minería de texto[36].	2015
E6	Algoritmo de clustering basado en el concepto de densidad atómica[63].	2016
E7	Modelo de Sentiment Analysis para la clasificación de noticias en tiempo real en el Mercado de Valores de Buenos Aires[64].	2014
E8	Algoritmos de clustering y aprendizaje automático aplicados a Twitter[65].	2016
E9	R and Data Mining Examples and Case Studies[45].	2015
E10	Técnicas estadísticas en Minería de textos[58].	2017
E11	Evaluación de la decisión de obtener el título profesional con la elaboración de la tesis mediante técnicas multivariantes Caso Universidad Nacional Agraria La Molina[37].	2017
E12	Minería de datos aplicada para la identificación de factores de riesgo en alumnos[43].	2017

Una vez obtenidos los estudios de la TABLA III se buscó los algoritmos usados en cada estudio, se contabilizó la cantidad de veces en las que fue usado cada algoritmo y se los presentó en una tabla comparativa. Algunos estudios presentan gran cantidad de algoritmos de clasificación que no aparecen en otros estudios, por lo que estos algoritmos al final son descartados y no son presentados en la tabla comparativa, puesto que representan información no útil. La tabla comparativa obtenida es la siguiente:

TABLA IV
OBTENCIÓN DE ALGORITMOS DE CLASIFICACIÓN

ALGORITMOS	ESTUDIOS												Total	
	1	2	3	4	5	6	7	8	9	10	11	12		
AID	x													1
CHAID Exhaustivo	x							x				x		3
CART	x													1
QUEST	x													1
J48 (C4,5)		x	x											2
NaiveBayes		x	x		x		x	x						5

Reglas de Asociación			x		x	x									x	4
Apriori																
KNN				x		x				x						3
SVM				x		x		x	x							4
PAUM				x												1
K-means					x	x	x		x	x					x	6
K-medoids					x		x			x						3
SOM					x											1
Random Forest						x		x		x						3
Dendograma										x	x					2
Regresión logística															x	1

En la TABLA IV se observa que los algoritmos de clasificación más utilizados en los estudios son:

- Árboles de Clasificación J48 (C4.5).
- NaiveBayes.
- Clustering K-Means.
- Dendograma.
- Árboles de Clasificación CHAID Exhaustivo.
- Clustering K-Medoids.
- Clustering K-Vecinos Mas Cercanos (KNN).
- Reglas de Asociación Apriori.
- Árboles de Clasificación Random Forest.
- Máquinas de Soporte Vectorial (SVM).

Con los algoritmos encontrados se realizó otra comparativa que permitió identificar los mejores algoritmos para la realización del proyecto y posterior a ello aplicarlos en el conjunto de datos.

2.2. Selección de los algoritmos de clasificación.

Para seleccionar los algoritmos de clasificación, se comparó los algoritmos hallados en el punto anterior de acuerdo a características establecidas en base a los resultados y conclusiones de los estudios de la TABLA III. Las características planteadas son:

- PE= Porcentaje de eficiencia.
- GR= Representación gráfica.

- EC= Efectivo en conjuntos grandes de datos.
- CA= Especificar cantidad de agrupamientos.
- DH= Disponibilidad en herramientas elegidas.

Cabe destacar que para la asignación de las características a cada algoritmo se tomó en cuenta que deban cumplirse en las herramientas de software elegidas para la aplicación de los algoritmos. Tanto las características de los algoritmos como las herramientas elegidas, son tomadas de la presentación de resultados y conclusiones del **Anexo 11**. Tomando en cuenta lo anterior la tabla comparativa es la siguiente:

TABLA V
COMPARATIVA DE LOS ALGORITMOS DE CLASIFICACIÓN

Algoritmos	PE	GR	EC	CA	DH
CHAID Exhaustivo	x	x	x		x
J48 (C4,5)	x	x			
NaiveBayes	x	x			
Reglas de Asociación Apriori		x	x		x
KNN			x		x
SVM	x				x
K-means			x	x	x
K-medoids			x	x	x
Random Forest		x	x		x
Dendograma		x	x		x

Los algoritmos que abarcan más características fueron los seleccionados para ser aplicados en este trabajo de titulación. La tabla comparativa se realizó de la siguiente manera: para el algoritmo CHAID Exhaustivo que está presente en los estudios E1, E8 y E11, el estudio E1 establece que este algoritmo posee porcentaje de eficiencia, es efectivo en grandes conjuntos de datos y tiene representación gráfica, el estudio E8 no presenta características, mientras que el algoritmo E11 presenta que el algoritmo posee porcentaje de eficiencia; por lo tanto en la tabla comparativa se seleccionó las características con las que cumple el algoritmo. Para el resto de algoritmos se realizó la asignación de características de igual manera que el algoritmo CHAID Exhaustivo.

Como se observa en la TABLA V, los algoritmos que coinciden con más de dos características son: Dendograma, Reglas de Asociación Apriori, K-means, K-medoids,

CHAID Exhaustivo y Random Forest; por lo tanto, estos algoritmos son los aplicados en el conjunto de datos.

2.3. Aplicación de los algoritmos de clasificación seleccionados.

Estando ya limpio el conjunto de datos, está listo para aplicar los algoritmos de clasificación. Primero se utilizó la herramienta RStudio para el pre procesamiento de los datos, en la TABLA VI se presenta las librerías de R empleadas para la elaboración del presente proyecto.

TABLA VI
LIBRERÍAS UTILIZADAS EN RSTUDIO

1	library(NLP)
2	library(tm)
3	library (twitterR)
4	library(SnowballC)
5	library(ggplot2)
6	library(RColorBrewer)
7	library(wordcloud)
8	library(fpc)
9	library(igraph)
10	library(rgl)
11	library(arules)
12	library(grid)
13	library(arulesViz)

2.3.1. Pre procesamiento de los datos.

Se exploró los datos para comprender mejor la información contenida y poder aplicar correctamente los algoritmos de clasificación, para ello se trabajó con todos los temas de los proyectos de titulación sin tomar en cuenta la columna “Estado” que se encuentra presente en el conjunto de datos. Primeramente, en la TABLA VII se muestra la carga de datos de una variable y la posterior selección de la clase “Tema”.

TABLA VII
CARGA DE PROYECTOS EN RSTUDIO

```
1 procesamiento_datos<-read.table("0_Conjunto_de_Datos_m.csv", head(TRUE),  
2 sep=";")  
3 mi_vector<-procesamiento_datos[,1]
```

Para el tratamiento de los datos es necesario realizar conversiones en el siguiente orden:

- El vector de los temas “mi_vector” convertir a data frame “df”.
- Crear una variable de tipo Corpus “miCorpus” con la información de la data frame “df”. La creación del Corpus nos permite manipular la información de los temas, por ejemplo: eliminar palabras no deseadas en los temas de los proyectos de titulación.
- Al Corpus convertirlo en tipo documento y guardarlo en la misma variable “miCorpus”.
- Por último, se convierte la variable “miCorpus” a matriz tipo documento “tdm” para obtener todas las palabras empleadas en el conjunto de datos, estas palabras se encuentran como etiquetas de las filas y las columnas son las instancias en las que aparece la palabra.

El código usado para estas conversiones se muestra en la TABLA VIII.

TABLA VIII

GENERACIÓN DE LA BOLSA DE PALABRAS

```

1 df <- do.call("rbind", lapply(mi_vector, as.data.frame))
2 miCorpus <- Corpus (VectorSource(df$`X[[i]]`))
3 miCorpus<-tm_map(miCorpus,stemDocument)
4 tdm <- TermDocumentMatrix(miCorpus,
5 control=list(wordLengths=c(1,Inf)))

```

La variable “tdm” es la bolsa de palabras del conjunto de datos ya que contiene todas las palabras presentes en los temas de los proyectos de titulación. La bolsa de palabras no muestra las palabras que aparecen con mayor frecuencia en los temas, por ello para una primera exploración de los datos se creó una matriz en donde cada fila tiene como nombre la palabra y frente a ella en una columna con su respectiva frecuencia o número de veces que aparece la palabra en el conjunto de datos. Partiendo de lo anterior, se obtuvo las palabras que aparecen más de 149 veces en el conjunto de datos y se guardaron en la variable “Frecuencia_Terminos”, el código empleado para la obtención de las palabras se muestra en la TABLA IX.

TABLA IX

OBTENCIÓN DE TÉRMINOS MÁS FRECUENTES

```

1 Frecuencia_Terminos<- rowSums(as.matrix(tdm))
2 Frecuencia_Terminos<- subset(Frecuencia_Terminos,
3 Frecuencia_Terminos>=150)

```


Para realizar una interpretación de los datos contenidos en “Frecuencia_Terminos”, se graficó el contenido de la variable, pero antes fue necesario convertir la variable a tipo dataframe, en la TABLA X se indica el código utilizado para realizar este proceso.

TABLA X

CONVERSIÓN A DATAFRAME DE LOS TÉRMINOS FRECUENTES

```

1 df_frecuencias <- data.frame(term=names(Frecuencia_Terminos),
2                               freq=Frecuencia_Terminos)
3 ggplot(df_frecuencias, aes(x=term, y=freq))+geom_bar(stat="identity")
4 +xlab("Terminos")+ylab("Cantidad de Palabras")+coord_flip()

```

El gráfico obtenido de la variable “Frecuencia_Terminos” es:

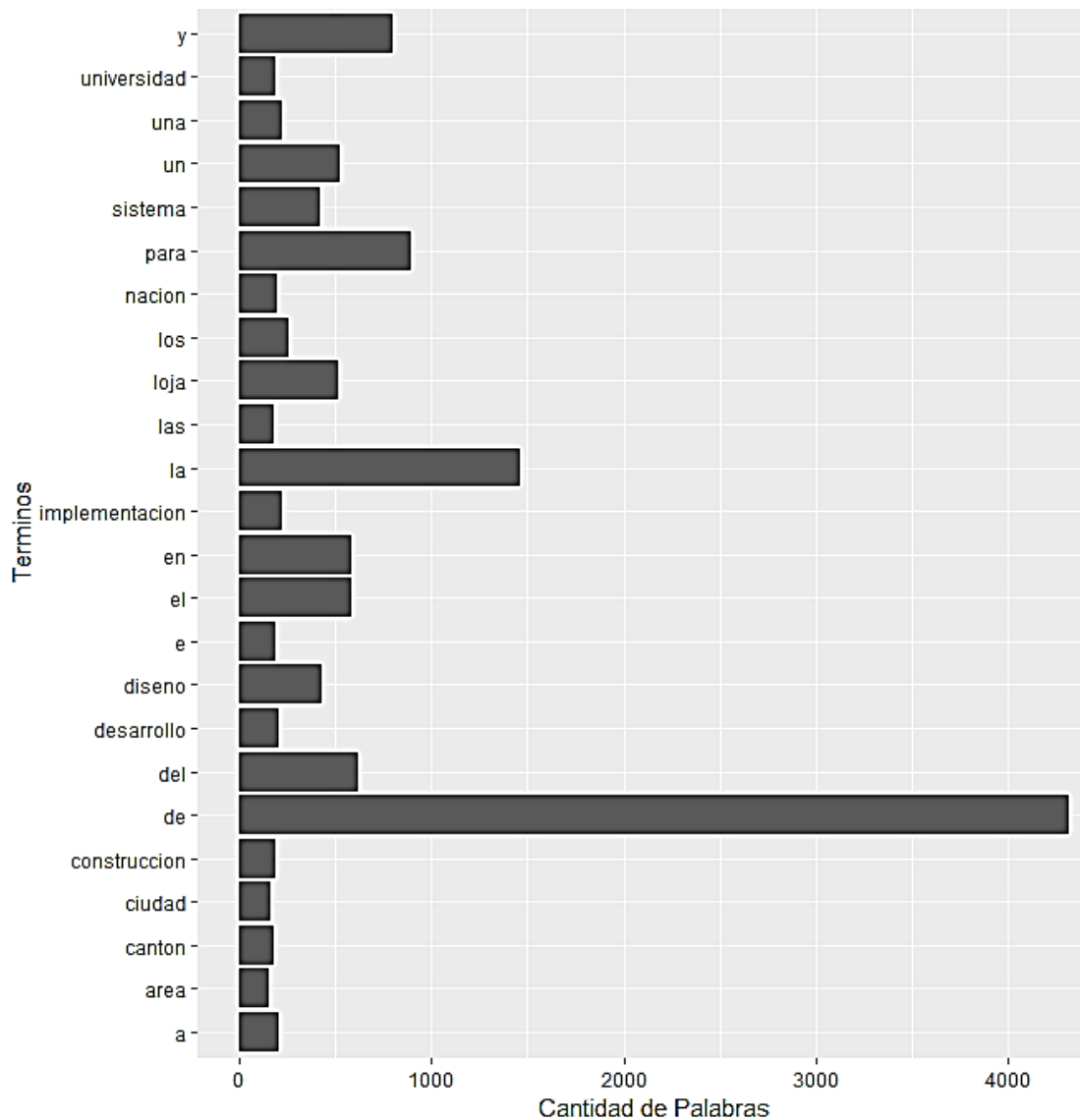


Figura 14. Términos más frecuentes.

En la Figura 14 se puede observar como las palabras que más veces se repiten son los conectores entre las oraciones, tales como: y, de, de, para, entre otros. Estos conectores no permiten identificar con facilidad cuales son las palabras más usadas en los temas de los proyectos. También se observa que la palabra "nacion" se encuentra presente en el gráfico a pesar de no estar presente en el conjunto de datos, por ello se inspeccionó la bolsa de palabras. En la Figura 15 se presenta el resultado de haber realizado la inspección.

```
> rownames(tdm)
[1] "core"
[4] "despliegue"
[7] "ipv"
[10] "los"
[13] "protocolo"
[16] "version"
[19] "barrio"
[22] "cola"
[25] "eleccion"
[28] "explotacion"
[31] "mineria"
[34] "reserva"
[37] "diseno"
[40] "ferroso"
[43] "material"
[46] "area"
[49] "jabonillo"
[52] "sistema"
[55] "distribucion"
[58] "estudio"
[61] "sur"
[64] "con"
[67] "funciona"
[70] "las"
[73] "parametro"
[76] "desd"
[79] "tablero"
[82] "agropecuario"
[85] "fitogenetico"
[88] "recurso"
[91] "almacenamiento"
[94] "diesel"
[97] "iess"
[100] "bomba"
[103] "efecto"
[106] "industria"
"de"
"dispositivo"
"la"
"nacion"
"swit"
"y"
"canton"
"conces"
"en"
"guachanama"
"palta"
"a"
"el"
>manual"
"traccion"
"existente"
"malacato"
"yeso"
"electrico"
"marginal"
"analisi"
"eficiencia"
"gasolina"
"mezcla"
"que"
"didactico"
"transferencia"
"automatizado"
"natur"
"renov"
"automatizacion"
"energetico"
"accionado"
"caudal"
"energia"
"laboratorio"
"del"
"internet"
"loja"
"para"
"universidad"
"alternativo"
"caolin"
"crucero"
"estimacion"
"mejor"
"parroquia"
"construccion"
"ensayo"
"maquina"
"una"
"exploracion"
"minera"
"costo"
"empresa"
"region"
"combust"
"etanol"
"interna"
"motor"
"automatica"
"generador"
"un"
"fertirrigacion"
"proyecto"
"terreno"
"ciudad"
"hospital"
"alternativa"
"dobl"
"hidraulicament"
"mecanicament"
```

Figura 15. Inspección de la bolsa de palabras.

Luego de la inspección de la bolsa de palabras se constató que las palabras se estaban cargando de forma indebida, para solucionar esto se transformó en Openrefine todas las palabras a mayúsculas y se cargó nuevamente el nuevo conjunto de datos que contiene solo mayúsculas, como se indica en la TABLA XI.

TABLA XI

CARGA DE TEMAS EN MAYÚSCULAS

```
1 procesamiento_datos<-read.table("1_Conjunto_de_Datos_M.csv",
2                                 head(TRUE),sep=";")
3 mi_vector<-procesamiento_datos[,1]
```

En base a las observaciones obtenidas de la Figura 14, luego de cargar los temas de los proyectos de titulación se elimina los conectores, como se indica en la TABLA XII para que no interfieran entre las relaciones de las palabras de los temas.

TABLA XII

CONTROL DE CONECTORES EN EL CORPUS

```
1 df <- do.call("rbind", lapply(mi_vector, as.data.frame))
2 miCorpus <- Corpus(VectorSource(df$`X[[i]]`))
3 miCorpus<- tm_map(miCorpus,removeWords,c("PARA","Y","UN"
4      , "LA","DE","DEL","UNA","UNO","UNOS","EL","LOS"
5      , "LAS","CON","UNAS","EN","A","E","QUE"
6      , "POR","SU","ES","O","U","S","AL","COMO"))
7 miCorpus<-tm_map(miCorpus,stemDocument)
8 tdm <- TermDocumentMatrix(miCorpus, control=list(wordLengths= c(1,Inf)))
```

Nuevamente se inspeccionó el conjunto de datos para verificar que las palabras se cargaron correctamente en la bolsa de palabras.

```

> rownames(tdm)
 [1] "core"           "despliegue"      "dispositivos"
 [4] "internet"      "ipv"             "loja"
 [7] "nacional"      "protocolo"       "swit"
[10] "universidad"  "version"         "alternativo"
[13] "barrio"       "canton"          "caolines"
[16] "cola"         "concesion"       "cruceiros"
[19] "eleccion"     "estimacion"      "explotacion"
[22] "guachanama"  "mejor"           "mineria"
[25] "paltas"      "parroquia"       "reserva"
[28] "construccion" "diseno"          "ensayo"
[31] "ferrosos"    "manual"          "maquina"
[34] "materiales"  "traccion"        "area"
[37] "existentes"  "exploracion"     "jabonillo"
[40] "malacatos"   "minera"          "sistema"
[43] "yesos"       "costo"           "distribucion"
[46] "electrico"   "empresa"         "estudio"
[49] "marginales"  "regional"        "sur"
[52] "analisis"    "combustion"      "eficiencia"
[55] "etanol"      "funciona"        "gasolina"
[58] "interna"     "mezclas"         "motores"
[61] "parametros"  "automatica"      "desde"
[64] "didactico"   "generador"       "tablero"
[67] "transferencia" "agropecuario"   "automatizado"
[70] "fertirrigacion" "fitogeneticos"  "natural"
[73] "proyecto"    "recursos"        "renovables"
[76] "terreno"     "almacenamiento" "automatizacion"
[79] "ciudad"      "diesel"          "energetico"
[82] "hospital"    "iess"            "recurso"
[85] "accionado"   "alternativas"   "bombas"
[88] "caudal"     "doble"           "efecto"
[91] "energia"     "hidraulicamente" "industrias"
[94] "laboratorio" "mecanicamente"  "naturales"

```

Figura 16. Nueva inspección de la bolsa de palabras.

Se observa en la Figura 16 que las palabras contenidas en la bolsa de palabras se cargaron correctamente, luego de haber corregido que las palabras se carguen de forma indebida y de remover los conectores en las oraciones, se graficó nuevamente las palabras más frecuentes en los temas de los proyectos de titulación con el código empleado en la TABLA IX y la TABLA X.

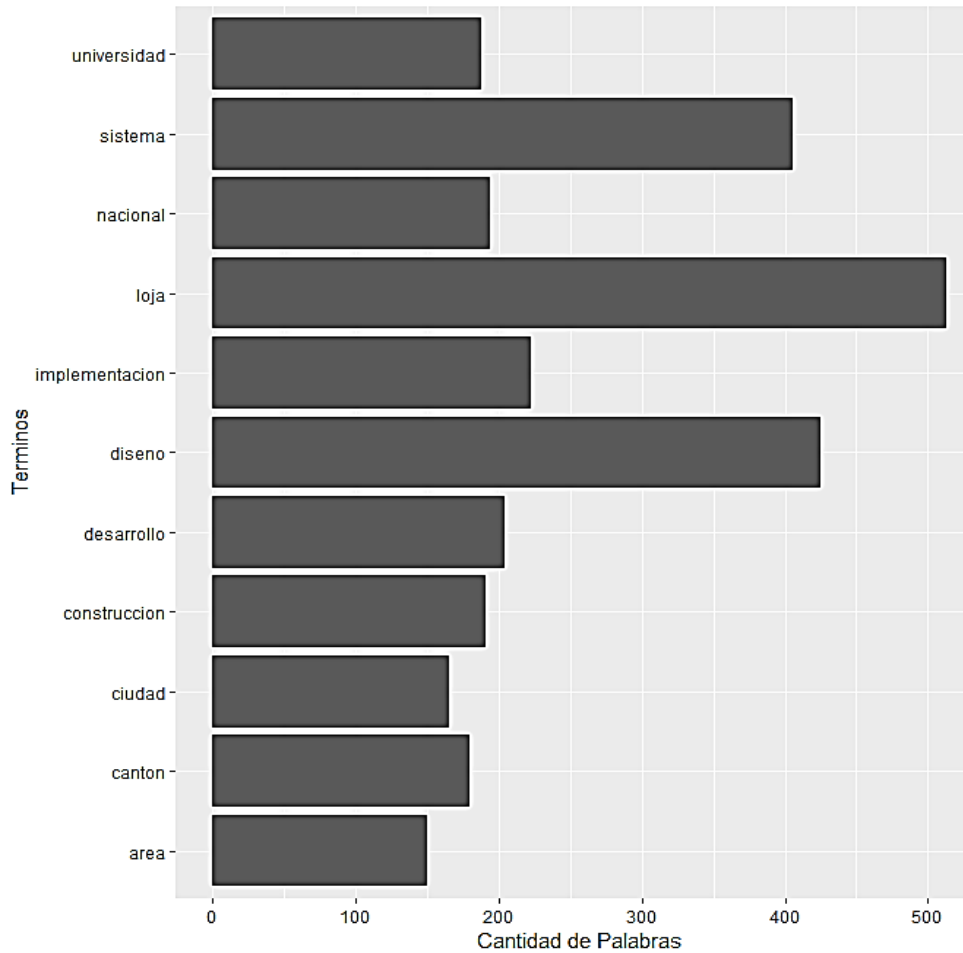


Figura 17. Términos más frecuentes sin conectores.

En la Figura 17 se observa como las tres palabras más usadas en los temas de los proyectos son: Loja, diseño y sistema. El código usado para la aplicación del dendograma se observa en la TABLA XIII, donde se usó la función “removeSparseTerms” para remover los términos menos frecuentes y se estableció un nivel de 0.95 debido a que entre mayor es el porcentaje más son los términos pocos frecuentes que se eliminan, luego se cargó estos términos en una matriz para calcular la distancia entre los términos y se aplicó la función hclust de R para finalmente guardar los resultados en la variable “fit”.

TABLA XIII

APLICACIÓN DEL DENDOGRAMA

```

1 clustering_tdm<- removeSparseTerms(tdm, sparse=0.95)
2 clustering_m<- as.matrix(clustering_tdm)
3 terminos_del_cluster<- dist(scale(clustering_m))
4 fit<- hclust(terminos_del_cluster, method="ward.D")

```

```

5 plot(fit)
6 rect.hclust(fit, k=10)
7 (groups <- cutree(fit, k=10))

```

Luego se graficó la variable “fit” de la TABLA XIII, en donde el gráfico obtenido encierra las palabras que se encuentran relacionadas.

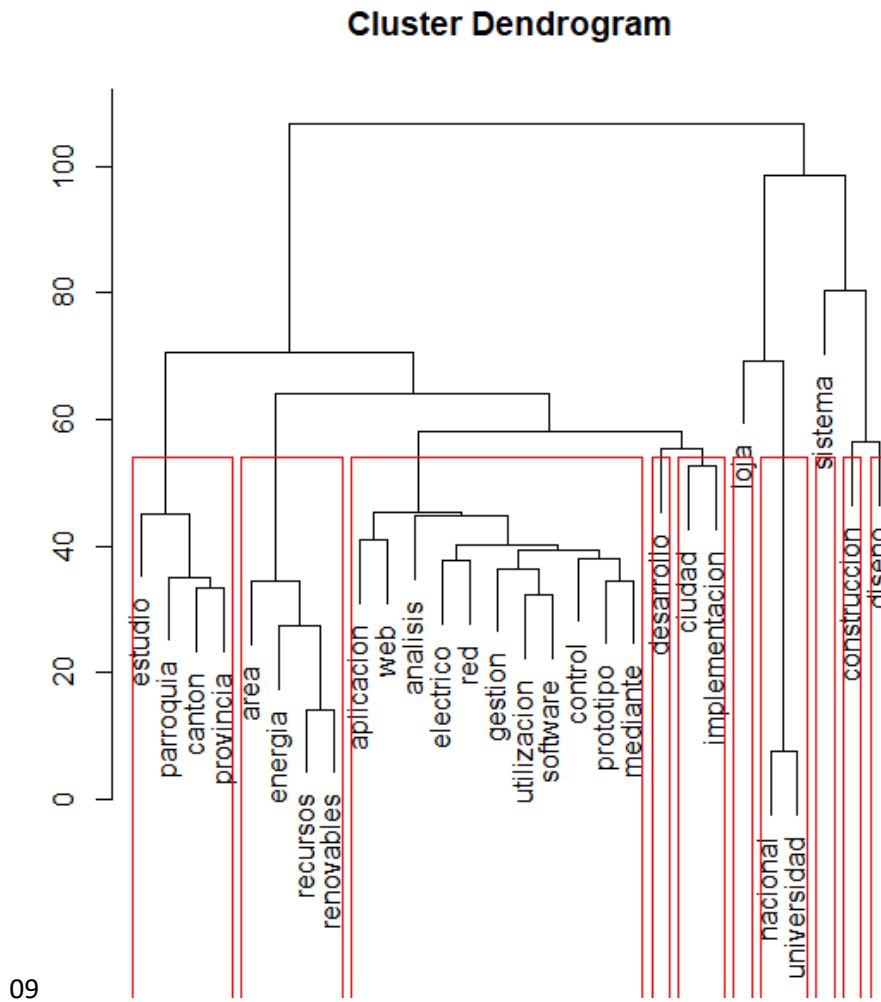


Figura 18. Dendrograma con nuevas relaciones.

Las relaciones obtenidas en la Figura 18 no muestran mucha claridad, debido a que algunas palabras que ni siquiera son semejantes se encuentran agrupadas dentro de una misma relación. Para mejorar la obtención de las relaciones entre las palabras se aumentó el nivel de remoción de términos a 0.97, como se indica en la TABLA XIV.

TABLA XIV

NUEVA APLICACIÓN DEL DENDOGRAMA

```

1 clustering_tdm<- removeSparseTerms(tdm, sparse=0.97)
2 clustering_m<- as.matrix(clustering_tdm)
3 terminos_del_cluster<- dist(scale(clustering_m))
4 fit<- hclust(terminos_del_cluster, method="ward.D")
5 plot(fit)
6 rect.hclust(fit, k=10)
7 (groups <- cutree(fit, k=10))

```

En la Figura 19 se observa que se confirman algunas relaciones anteriores, e inclusive se mejoran.

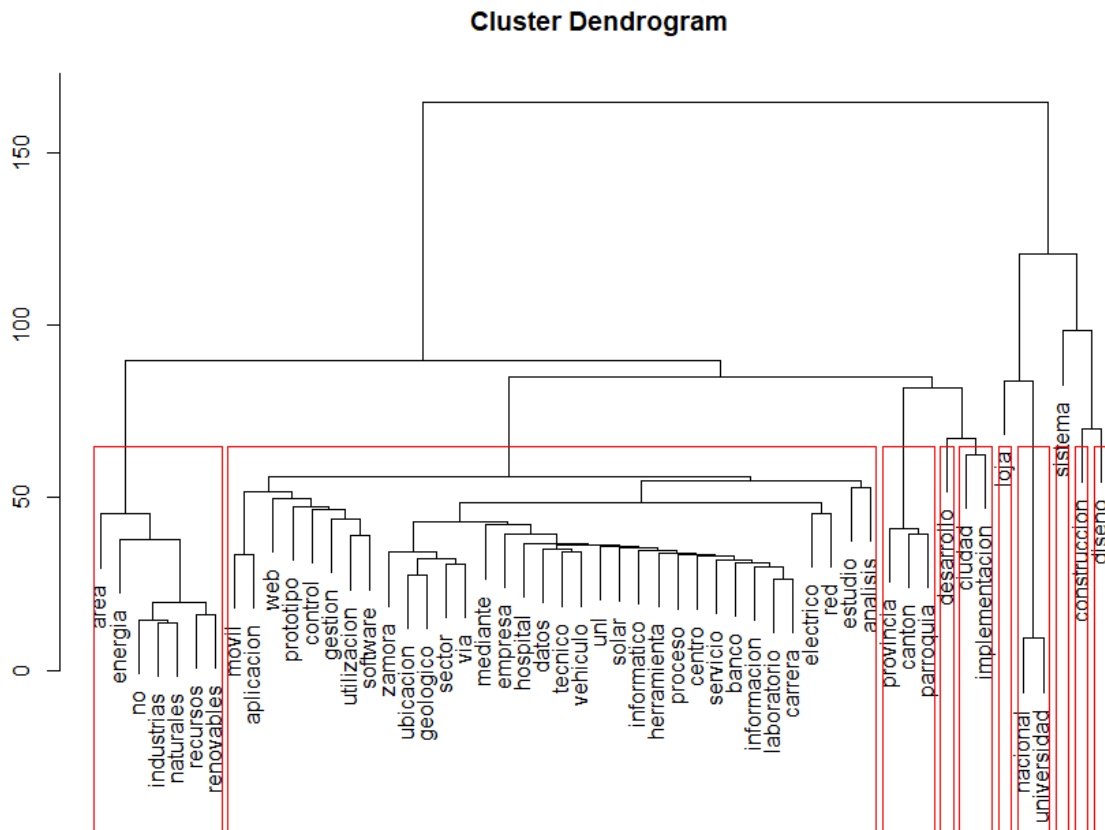


Figura 19. Dendrograma con relaciones obvias.

También se evidencia que oraciones como “universidad nacional loja” influyen de manera errónea en el conjunto de datos, ya que toda la oración se refiere a una sola cosa, a estos casos se los denomina relaciones obvias. En el conjunto de datos ya se encuentran presentes las siglas “uni”, por ello haciendo uso nuevamente de la herramienta Openrefine se reemplazó todas las oraciones que eran representadas por siglas en el conjunto de datos, como en este caso las más evidentes son: “area energia

industrias recursos naturales no renovables” por “aeirnr”, y “universidad nacional loja” por “unl”, como se indica en la Figura 20. Otras relaciones evidentes son “canton”, “parroquia” y “provincia”, así que se decidió reemplazar estas palabras semejantes por la palabra “territorio”.

Transformación personalizada en TEMA

Expresión Lenguaje Lenguaje de expresión de refinamiento general (GREL)

`value.replace("UNIVERSIDAD NACIONAL DE LOJA", "UNL")` No hay error de sintaxis.

Vista previa Historial Con estrella Ayuda

fila	valor	value.replace ("UNIVERSIDAD NAC ...
1.	DESPLIEGUE DEL PROTOCOLO DE INTERNET VERSIÓN IPV PARA LOS DISPOSITIVOS CORE Y SWIT DE LA UNIVERSIDAD NACIONAL DE LOJA	DESPLIEGUE DEL PROTOCOLO DE INTERNET VERSIÓN IPV PARA LOS DISPOSITIVOS CORE Y SWIT DE LA UNL
2.	ESTIMACIÓN DE RESERVA Y ELECCIÓN DE LA MEJOR ALTERNATIVO EN LA EXPLOTACION DE CAOLINES EN LA CONCESION MINERIA LOS CRUCEROS BARRIO COLA PARROQUIA GUACHANAMA CANTON PALTAS	ESTIMACION DE RESERVA Y ELECCION DE LA MEJOR ALTERNATIVO EN LA EXPLOTACION DE CAOLINES EN LA CONCESION MINERIA LOS CRUCEROS BARRIO COLA PARROQUIA GUACHANAMA CANTON PALTAS
3.	DISEÑO Y CONSTRUCCIÓN DE UNA MAQUINA MANUAL PARA EL ENSAYO A LA TRACCIÓN DE MATERIALES FERROSOS	DISEÑO Y CONSTRUCCIÓN DE UNA MAQUINA MANUAL PARA EL ENSAYO A LA TRACCIÓN DE MATERIALES FERROSOS
4.	ELECCIÓN DEL SISTEMA DE EXPLORACIÓN PARA LOS YESOS EXISTENTES EN EL DEPÓSITO MINERALÍFICO PARROQUIA MAI ACATOS CANTÓN LOJA	ELECCIÓN DEL SISTEMA DE EXPLORACIÓN PARA LOS YESOS EXISTENTES EN EL DEPÓSITO MINERALÍFICO PARROQUIA MAI ACATOS CANTÓN LOJA

En error mantener original Re-transformar hasta veces hasta que no haya cambios cambiar a en blanco guardar error

Aceptar Cancelar

Figura 20. Reemplazo de UNL.

Luego de haber hecho los reemplazos se carga el nuevo conjunto de datos tal como se indica en la TABLA XV y se controla los conectores con el código usado en la TABLA XII.

TABLA XV
CARGA DE TEMAS CON SIGLAS

```

1 procesamiento_datos<-read.table("2_Conjunto_de_Datos_UNDL.csv",
2                               head(TRUE),sep=";")
3 mi_vector<-procesamiento_datos[,1]

```

Nuevamente realizamos el dendograma con el código de la TABLA XIV, al brindarnos este un mejor resultado que el de la TABLA XIII.

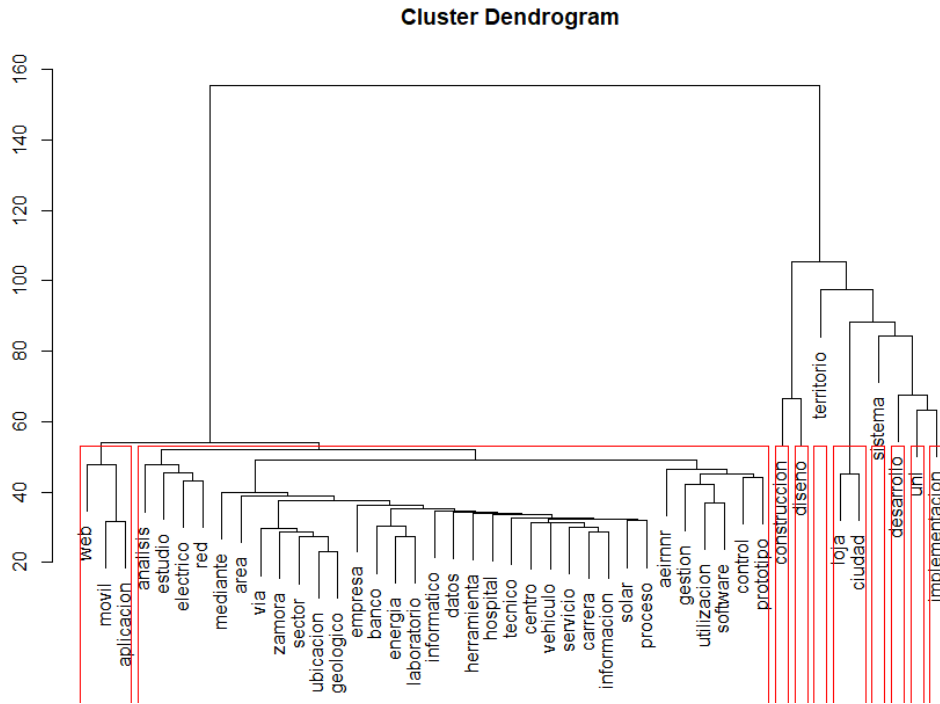


Figura 21. Dendrograma sin relaciones válidas.

La Figura 21 muestra que las palabras “web”, “movil” y “aplicacion” están relacionadas, también se observa la relación entre las palabras “ciudad” y “loja”. Ya que estas dos últimas palabras hacen referencia a un mismo lugar, se procedió a realizar una exploración de los datos usando una nube de palabras, obteniendo así el gráfico con el código de la TABLA IX y la TABLA X.

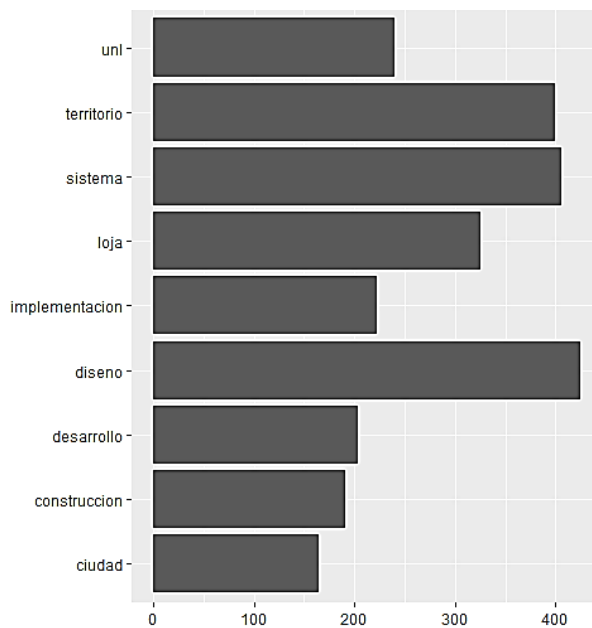


Figura 22. Términos más frecuentes con siglas.

Se observa en la Figura 22 que la palabra “ciudad” se encuentra en el conjunto de datos casi la mitad de veces que la palabra “loja”, dando así a entender que cuando está la palabra “loja” no siempre está la palabra “ciudad”, por tal motivo en la Figura 23 se muestra cómo se usó nuevamente Openrefine para reemplazar la oración “ciudad loja” por la palabra “ciudadloja”.

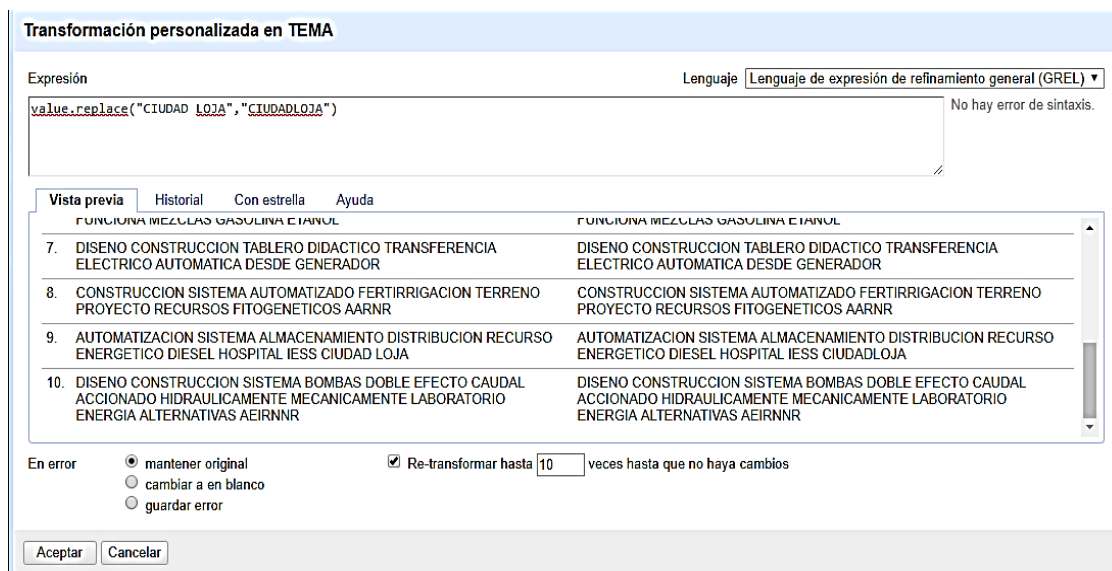


Figura 23. Reemplazo de ciudad Loja.

En la TABLA XVI se indica la carga del nuevo conjunto de datos estandarizado, donde también se controló los conectores con el código usado en la TABLA XII.

TABLA XVI

CARGA DEL CONJUNTO DE DATOS PRE PROCESADO

```

1 procesamiento_datos<-read.table("2_1_Conjunto_de_Datos_UNDL.csv",
2                               head(TRUE),sep=";")
3 mi_vector<-procesamiento_datos[,1]

```

2.3.2. Nube de Palabras y Agrupación Jerárquica.

Al estar ya corregidas todas las asociaciones entre las palabras semejantes se procedió a la obtención de la nube de palabras para conocer las palabras más usada en los temas y del dendograma que es la representación de la agrupación jerárquica. Se aplicaron tres veces: en todo el conjunto de datos, en los proyectos con estado de éxito y en los proyectos con estado de fracaso. Primero para la obtención de la nube de palabras de todo el conjunto de datos el código empleado se muestra en la TABLA XVII.

TABLA XVII

APLICACIÓN NUBE DE PALABRAS EN TODO EL CONJUNTO DE DATOS

```

1 m<-rowSums(as.matrix(tdm))
2 Frecuencia_Terminos_descendente<- sort(m, decreasing=TRUE)
3 color<- brewer.pal(9, "BuGn")
4 color<- color[-(1:4)]
5 set.seed(375)
6 Niveles_de_gris<- gray( (Frecuencia_Terminos_descendente+10) /
7 (max(Frecuencia_Terminos_descendente)+10) )
8 wordcloud(words=names(Frecuencia_Terminos_descendente),
9 freq=Frecuencia_Terminos_descendente,
10 min.freq=3, random.order=F,colors=color)

```

La nube de palabras nos presenta las palabras más frecuentes en orden descendente, conforme la frecuencia de las palabras va disminuyendo estas también disminuyen su tamaño y color. Por lo tanto, las palabras más usadas en el conjunto de datos son las que poseen un mayor tamaño y un color más intenso.

Luego para la obtención del dendrograma el código empleado nuevamente es el de la TABLA XIV.

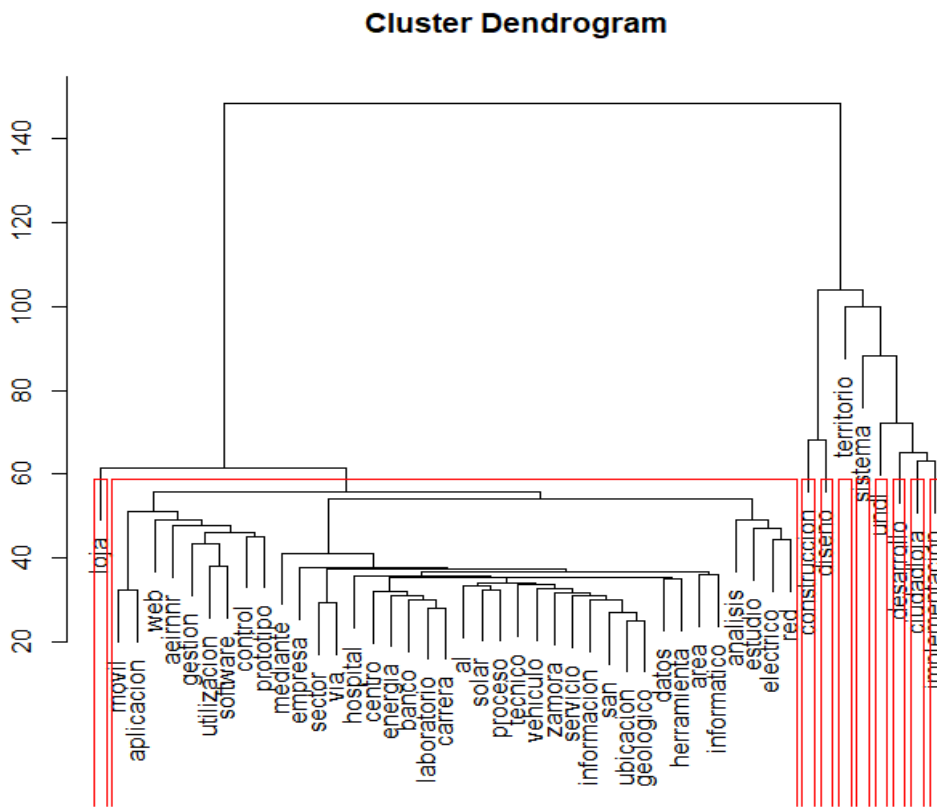


Figura 24. Dendrograma sin relaciones.

En la Figura 24 se observa que, al corregir todas las asociaciones entre palabras semejantes, no se obtienen relaciones nuevas entre las palabras. Por ello se aumenta el número de grupos a 20, como se indica en la TABLA XVIII.

TABLA XVIII

NUEVA APLICACIÓN DEL DENDOGRAMA AL CONJUNTO DE DATOS

```

1 clustering_tdm<- removeSparseTerms(tdm, sparse=0.97)
2 clustering_m<- as.matrix(clustering_tdm)
3 terminos_del_cluster<- dist(scale(clustering_m))
4 fit<- hclust(terminos_del_cluster, method="ward.D")
5 plot(fit)
6 rect.hclust(fit, k=20)
7 (groups <- cutree(fit, k=20))

```

Luego para aplicar la nube de palabras y el dendograma de los proyectos con estado de éxito, fue necesario obtener los proyectos de titulación con estados de éxito, tal como se muestra en la TABLA XIX.

TABLA XIX

OBTENCIÓN DE PROYECTOS CON ESTADO DE ÉXITO

```

1 matriz_exito<- subset(procesamiento_datos, procesamiento_datos[,2]=='EXITO')
2 mi_vector<-matriz_exito[,1]
3 df <- do.call("rbind", lapply(mi_vector, as.data.frame))
4 miCorpus <- Corpus(VectorSource(df$`X[[i]]`))
5 miCorpus<- tm_map(miCorpus,removeWords,c("PARA","Y","UN"
6           ,"LA","DE","DEL","UNA","UNO"
7           ,"UNOS","EL","LOS","LAS","CON","UNAS","EN"
8           ,"A","E","QUE","POR","SU","ES","O","U","S"
9           ,"AL","COMO"))
10 miCorpus<-tm_map(miCorpus,stemDocument)
11 tdm <- TermDocumentMatrix(miCorpus, control=list(wordLengths=c(1,Inf)))

```

La obtención de la nube de palabras de todos los temas pertenecientes a los proyectos de titulación con estado de éxito, se realizó con el código de la TABLA XVII y luego para la obtención del dendograma se usó el código de la TABLA XVIII, ya que con ese código se obtuvo nuevas relaciones entre las palabras.

Por último, para aplicar la nube de palabras y el dendograma en los proyectos con estado de fracaso fue necesario obtener los proyectos de titulación con estados de fracaso, para ello se utilizó el código de la TABLA XX.

TABLA XX

OBTENCIÓN DE PROYECTOS CON ESTADO DE FRACASO

```
1 matriz_fracaso<- subset(procesamiento_datos, procesamiento_datos[,2]==
2     'FRACASO')
3 mi_vector<-matriz_fracaso[,1]
4 df <- do.call("rbind", lapply(mi_vector, as.data.frame))
5 miCorpus <- Corpus(VectorSource(df$`X[[i]]`))
6 miCorpus<- tm_map(miCorpus,removeWords,c("PARA","Y","UN"
7     ,"LA","DE","DEL","UNA","UNO"
8     ,"UNOS","EL","LOS","LAS","CON","UNAS","EN"
9     ,"A","E","QUE","POR","SU","ES","O","U","S"
10    ,"AL","COMO"))
11 miCorpus<-tm_map(miCorpus,stemDocument)
12 tdm <- TermDocumentMatrix(miCorpus, control=list(wordLengths=c(1,Inf)))
```

Para aplicar la nube de palabras se volvió a hacer uso del código mostrado en la TABLA XVII y para aplicar el dendograma se empleó nuevamente el código de la TABLA XVIII.

2.3.3. Reglas de Asociación Apriori.

En el conjunto de datos también se aplicaron reglas de asociación, pero primeramente fue necesario generar un nuevo conjunto de datos en el cual no se encuentren conectores para que no interfieran en las relaciones entre las palabras. Fue necesario crear un nuevo archivo debido a que para aplicar reglas de asociaciones usando R no se trabaja con una variable de tipo Corpus sino de tipo Transactions, para ello se cargó el ultimo conjunto de datos estandarizado "2_1_Conjunto_de_Datos_UNL.csv", se eliminó los conectores de ese conjunto de datos y luego se generó un nuevo archivo "3_Conjunto_de_Datos_RA_TODO.csv", este proceso se muestra en la TABLA XXI.

TABLA XXI

NUEVO ARCHIVO DE TODOS LOS PROYECTOS SIN CONECTORES

```
1 procesamiento_datos<-read.table("2_1_Conjunto_de_Datos_UNDL.csv",
2     head(TRUE),sep=";")
3 mi_vector<-procesamiento_datos[,1]
4 df <- do.call("rbind", lapply(mi_vector, as.data.frame))
5 miCorpus <- Corpus(VectorSource(df$`X[[i]]`))
6 miCorpus<- tm_map(miCorpus,removeWords,c("PARA","Y","UN"
7     ,"LA","DE","DEL","UNA","UNO"
8     ,"UNOS","EL","LOS","LAS","CON","UNAS","EN"
9     ,"A","E","QUE","POR","SU","ES","O","U","S"
10    ,"AL","COMO"))
```

```

11 miCorpus<-tm_map(miCorpus,stemDocument)
12 s<-"TEMA"
13 for (i in 1:length(miCorpus)) {
14     s<-rbind(s,(strwrap(miCorpus[[i]], width=500)))
15 }
16 write.csv (s, file="3_Conjunto_de_Datos_RA_TODO.csv")

```

Luego al archivo "3_Conjunto_de_Datos_RA_TODO.csv" se lo cargó como tipo Transactions para proceder a aplicar el algoritmo Apriori. En la TABLA XXII se muestran los parámetros empleados en este algoritmo, los cuales son de soporte y de confianza, se eligieron niveles bajo de soporte debido a que el conjunto de datos al ser tan grande aún posee sinónimos que afectan en la búsqueda de patrones entre las palabras y se eligió un alto nivel de confianza para tener fiabilidad en los resultados.

TABLA XXII

APLICACIÓN DE APRIORI EN TODO EL CONJUNTO DE DATOS

```

1 transacciones <- read.transactions(file = "3_Conjunto_de_Datos_RA_TODO.csv",
2                                 format="basket",sep=" ",rm.duplicates = TRUE)
3 reglas<-apriori(transacciones,parameter = list(support=0.04,confidence=0.5))
4 inspect(head(sort(reglas,by="lift")))
5 plot(reglas,method = "graph",control = list(type="items")
6      ,main="Palabras Asociadas")

```

Al igual que en la aplicación de las nubes de palabras y el agrupamiento jerárquico, las reglas de asociación serán aplicadas no solo en todo el conjunto de datos, sino también en los proyectos con estado de éxito y los proyectos con estado de fracaso.

Primero para los proyectos con estado de éxito fue necesario generar un archivo "3_Conjunto_de_Datos_RA_EXITO.csv" que no tiene conectores y en el cual se encuentran solo los temas de los proyectos de titulación con estado de éxito, como se indica en la TABLA XXIII.

TABLA XXIII

NUEVO ARCHIVO DE PROYECTOS DE ÉXITO SIN CONECTORES

```

1 procesamiento_datos<-read.table("2_1_Conjunto_de_Datos_UNDL.csv",
2                                 head(TRUE),sep=";")
3 matriz_exito<- subset(procesamiento_datos, procesamiento_datos[,2]=='EXITO')
4 mi_vector<-matriz_exito[,1]
5 df <- do.call("rbind", lapply(mi_vector, as.data.frame))
6 miCorpus <- Corpus(VectorSource(df$`X[[i]]`))
7 miCorpus<- tm_map(miCorpus,removeWords,c("PARA","Y","UN")

```

```

8           ,"LA","DE","DEL","UNA","UNO"
9           ,"UNOS","EL","LOS","LAS","CON","UNAS","EN"
10          ,"A","E","QUE","POR","SU","ES","O","U","S"
11          ,"AL","COMO"))
12 miCorpus<-tm_map(miCorpus,stemDocument)
13 s<-"TEMA"
14 for (i in 1:length(miCorpus)) {
15   s<-rbind(s,(strwrap(miCorpus[[i]], width=500)))
16 }
17 write.csv (s, file="3_Conjunto_de_Datos_RA_EXITO.csv")

```

Cargamos el archivo "3_Conjunto_de_Datos_RA_EXITO.csv" y aplicamos las reglas de asociación con el código de la TABLA XXIV.

TABLA XXIV

APLICACIÓN DE APRIORI EN LOS PROYECTOS DE ÉXITO

```

1 transacciones <- read.transactions(file = "3_Conjunto_de_Datos_RA_EXITO.csv"
2                                   ,format="basket",sep=" ",rm.duplicates = TRUE)
3 reglas<-apriori(transacciones,parameter = list(support=0.04,confidence=0.5))
4 inspect(head(sort(reglas,by="lift")))
5 plot(reglas,method = "graph",control = list(type="items")
6       ,main="Reglas de Exito")

```

Para obtener las reglas de asociación de los proyectos con estado de fracaso, fue necesario crear un nuevo archivo "3_Conjunto_de_Datos_RA_FRACASO.csv", el cual no contenga conectores entre las palabras. El proceso de obtención del archivo se observa en la TABLA XXV.

TABLA XXV

NUEVO ARCHIVO DE PROYECTOS DE FRACASO SIN CONECTORES

```

1 procesamiento_datos<-read.table("2_1_Conjunto_de_Datos_UNDL.csv"
2                                   , head(TRUE),sep=";")
3 matriz_exito<- subset(procesamiento_datos,
4                       procesamiento_datos[,2]=='FRACASO')
5 mi_vector<-matriz_exito[,1]
6 df <- do.call("rbind", lapply(mi_vector, as.data.frame))
7 miCorpus <- Corpus(VectorSource(df$ X[[i]]`))
8 miCorpus<- tm_map(miCorpus,removeWords,c("PARA","Y","UN"
9                                           ,"LA","DE","DEL","UNA","UNO"
10                                          ,"UNOS","EL","LOS","LAS","CON","UNAS","EN"
11                                          ,"A","E","QUE","POR","SU","ES","O","U","S"
12                                          ,"AL","COMO"))
13 miCorpus<-tm_map(miCorpus,stemDocument)

```

```

14 s<-"TEMA"
15 for (i in 1:length(miCorpus)) {
16     s<-rbind(s,(strwrap(miCorpus[[i]], width=500)))
17 }
18 write.csv (s, file="3_Conjunto_de_Datos_RA_FRACASO.csv")

```

Se carga el archivo "3_Conjunto_de_Datos_RA_FRACASO.csv" como tipo Transactions, y como se indica en la TABLA XXVI aplicamos el algoritmo.

TABLA XXVI

APLICACIÓN DE APRIORI EN LOS PROYECTOS DE FRACASO

```

1 transacciones<-read.transactions(file=3_Conjunto_de_Datos_RA_FRACASO.csv"
2     ,format="basket",sep=" ",rm.duplicates = TRUE)
3 reglas<-apriori(transacciones,parameter = list(support=0.04,confidence=0.5))
4 inspect(head(sort(reglas,by="lift")))
5 plot(reglas,method = "graph",control = list(type="items")
6     ,main="Reglas de Fracaso")

```

2.3.4. K-means y K-medoids.

Los algoritmos k-means y k-medoids se aplicaron en todo el conjunto de datos, también en solo los proyectos con estado de éxito y en solo los proyectos con estado de fracaso. Para la aplicación de los algoritmos se trabajó con el ultimo conjunto de datos estandarizado "2_1_Conjunto_de_Datos_UNL.csv", a este conjunto de datos se lo cargó nuevamente en R y se controló que los conectores entre las palabras no se carguen en la bolsa de palabras. El proceso detallado anteriormente se muestra en la TABLA XXVII.

TABLA XXVII

OBTENCIÓN DE TODOS LOS PROYECTOS SIN CONECTORES

```

1 procesamiento_datos<-read.table("2_1_Conjunto_de_Datos_UNDL.csv"
2     , head(TRUE),sep=";")
3 mi_vector<-procesamiento_datos[,1]
4 df <- do.call("rbind", lapply(mi_vector, as.data.frame))
5 miCorpus <- Corpus(VectorSource(df$`X[[i]]`))
6 miCorpus<- tm_map(miCorpus,removeWords,c("PARA","Y","UN"
7     ,"LA","DE","DEL","UNA","UNO"
8     ,"UNOS","EL","LOS","LAS","CON","UNAS","EN"
9     ,"A","E","QUE","POR","SU","ES","O","U","S"
10    ,"AL","COMO"))
11 miCorpus<-tm_map(miCorpus,stemDocument)
12 tdm <- TermDocumentMatrix(miCorpus, control=list(wordLengths=c(1,Inf)))

```


Los primeros pasos para la aplicación de los algoritmos son iguales a los primeros pasos realizados en los dendogramas, estos pasos consisten en remover las palabras que se encuentran más separadas y generar una matriz con los términos restantes, se estableció un nivel de 0.97 en la función “removeSparseTerms” al igual que en los dendogramas debido a que con este nivel ya se demostró obtener mejores resultados. Luego se transpone la matriz obtenida, se establece la semilla en 122 para sortear los centroides iniciales y el número de clústeres en 9 para que los resultados no sean redundantes, el haber establecido la semilla y el número de clústeres permite que los resultados obtenidos sean repetibles. Una vez realizado lo anterior se aplica el algoritmo k-means y se centran los clústeres. Por último, imprimimos las tres palabras más influyentes de cada clúster, es decir las más cercanas de cada centroide para evitar redundancia en los resultados. Todo este proceso es mostrado en la TABLA XXVIII.

TABLA XXVIII

APLICACIÓN DE K-MEANS EN TODO EL CONJUNTO DE DATOS

```

1 clustering_tdm<- removeSparseTerms(tdm, sparse=0.97)
2 clustering_m<- as.matrix(clustering_tdm)
3 kmeans_m<-t(clustering_m)
4 set.seed(122)
5 k <- 9
6 kmeans_resultados<- kmeans(kmeans_m, k)
7 round(kmeans_resultados$centers, digits=3)
8 for (i in 1:k) {
9   cat(paste("kmeans_todo ", i, ": ", sep=""))
10  s <- sort(kmeans_resultados$centers[i,], decreasing=T)
11   cat(names(s)[1:3], "\n")
12 }
```

Como se estableció 9 clústeres en la aplicación del algoritmo, se obtuvo las relaciones entre las palabras mostradas en la Figura 25.

```

kmeans_todo 1: desarrollo sistema implementacion
kmeans_todo 2: aplicacion desarrollo movil
kmeans_todo 3: diseno construccion prototipo
kmeans_todo 4: diseno construccion aeirnr
kmeans_todo 5: ciudadloja analisis implementacion
kmeans_todo 6: unl implementacion diseno
kmeans_todo 7: sistema diseno gestion
kmeans_todo 8: territorio loja estudio
kmeans_todo 9: electrico diseno sistema
```

Figura 25. Resultados de k-means en todo el conjunto de datos.

Por otra parte, para el algoritmo k-medoids luego de haber obtenido la matriz se aplica al algoritmo, el cual recibe dos parámetros: en el primer parámetro se ingresa la transpuesta de la matriz y en el segundo parámetro se ingresa la métrica “manhattan”. Una vez aplicado el algoritmo, obtenemos el número de clústeres generados y se cargan los medoides en una variable para luego presentar los medoides como resultados. Se muestra en la TABLA XXIX el proceso descrito anteriormente.

TABLA XXIX

APLICACIÓN DE K-MEDOIDS EN TODO EL CONJUNTO DE DATOS

```

1 clustering_tdm<- removeSparseTerms(tdm, sparse=0.97)
2 clustering_m<- as.matrix(clustering_tdm)
3 kmedias_resultados<-pamk(t(clustering_m),metric="manhattan")
4 k <-(kmedias_resultados$nc)
5 kmedias_resultados<- kmedias_resultados$pamobject
6 for (i in 1:k) {
7   cat(paste("kmedoids_todo ", i, ": "))
8   cat(colnames(kmedias_resultados$medoids)
9     [which(kmedias_resultados$medoids[i,]==1)], "\n")
10 }

```

Las palabras centros de cada medoide obtenido mediante el algoritmo se observan en la Figura 26.

```

kmedoids_todo 1 :
kmedoids_todo 2 : loja
kmedoids_todo 3 : diseno

```

Figura 26. Resultados de k-medoids en todo el conjunto de datos.

Primero para aplicar los algoritmos en todos los temas de los proyectos de titulación con estado de éxito es necesario obtener solamente esos proyectos, para ello se usó el código de la TABLA XIX. Ya que se obtuvo los datos se procedió a la aplicación del algoritmo k-means haciendo uso del código mostrado en la TABLA XXVIII. Obteniendo así los resultados mostrados en la Figura 27.

```

kmeans_exito 1: desarrollo sistema un1
kmeans_exito 2: un1 sistema diseno
kmeans_exito 3: ciudadloja sistema implementacion
kmeans_exito 4: software loja implementacion
kmeans_exito 5: red diseno un1
kmeans_exito 6: territorio loja estudio
kmeans_exito 7: aplicacion desarrollo web
kmeans_exito 8: diseno construccion prototipo
kmeans_exito 9: sistema diseno loja

```

Figura 27. Resultados de k-means en los proyectos de éxito.

La aplicación del algoritmo k-medoids en el conjunto de datos de los proyectos con estado de éxito se realizó con el código de la TABLA XXIX, y los resultados obtenidos se presentan en la Figura 28.

```
kmedoids_exito 1 :  
kmedoids_exito 2 : loja  
kmedoids_exito 3 : diseno
```

Figura 28. Resultados de k-medoids en los proyectos de éxito.

En la aplicación del algoritmo k-means en los proyectos que han fracasado, es necesario obtener específicamente esos proyectos, así que para obtenerlos se empleó el código usado anteriormente en la TABLA XX. Luego de ello se utilizó el código de la TABLA XXVIII para aplicar el algoritmo k-means, y los resultados obtenidos se presentan en la Figura 29.

```
kmeans_fracaso 1: aplicacion movil desarrollo  
kmeans_fracaso 2: territorio loja sistema  
kmeans_fracaso 3: ciudadloja estudio electrico  
kmeans_fracaso 4: diseno sistema implementacion  
kmeans_fracaso 5: un1 sistema implementacion  
kmeans_fracaso 6: desarrollo un1 plataforma  
kmeans_fracaso 7: sistema desarrollo web  
kmeans_fracaso 8: analisis diseno implementacion  
kmeans_fracaso 9: construccion diseno sistema
```

Figura 29. Resultados de k-means en los proyectos de fracaso.

El código utilizado para aplicar el algoritmo k-medoids es el mismo de la TABLA XXIX, en la Figura 30 se muestran los resultados obtenidos.

```
kmedoids_fracaso 1 :  
kmedoids_fracaso 2 : diseno  
kmedoids_fracaso 3 : sistema  
kmedoids_fracaso 4 :
```

Figura 30. Resultados de k-medoids en los proyectos de fracaso.

Por último, se pudo obtener una representación gráfica de la cantidad de proyectos que contienen los patrones o relaciones obtenidas con los algoritmos: agrupamiento jerárquico, k-means, k-medoids y Apriori. Como ejemplo en la TABLA XXX se presenta el código empleado para realizar lo detallado anteriormente con la relación “kmeans_todo 3”.

TABLA XXX

APLICACIÓN DE LA REPRESENTACIÓN GRÁFICA DE LAS RELACIONES

```
1 procesamiento_datos<-read.table("2_1_Conjunto_de_Datos_UNDL.csv"
2                               , head(TRUE),sep=";")
3 matriz_exito<-subset(procesamiento_datos, procesamiento_datos[,2]=='EXITO')
4 todo_exito<-length(matriz_exito[,1])
5 Temas_Sistema<-(subset(matriz_exito,grep("DISENO",matriz_exito[,1])>0))
6 palabra_1<-length(Temas_Sistema[,1])
7 (length(Temas_Sistema[,1])*100)/length(matriz_exito[,1])
8 Temas_Sistema<-(subset(Temas_Sistema,grep("CONSTRUCCION"
9                          ,Temas_Sistema[,1])>0))
10 palabra_2<-length(Temas_Sistema[,1])
11 (length(Temas_Sistema[,1])*100)/length(matriz_exito[,1])
12 Temas_Sistema<-(subset(Temas_Sistema,grep("PROTOTIPO"
13                          ,Temas_Sistema[,1])>0))
14 palabra_3<-length(Temas_Sistema[,1])
15 (length(Temas_Sistema[,1])*100)/length(matriz_exito[,1])
16 dfinal<- data.frame(term=c(3,2,1,0), freq=c(todo_exito,palabra_1
17                                             ,palabra_2,palabra_3))
18 ggplot(dfinal, aes(x=c("TOTAL_EXITO","DISENO","DISENO-CONSTRUCCION"
19                       ,"DISENO-CONSTRUCCION-PROTOTIPO"), y=c(todo_exito
20                       ,palabra_1,palabra_2,palabra_3)))+geom_bar(stat="identity")+
21 xlab("Palabras de la Relacion") +ylab("Cantidad de Proyectos")
```

2.3.5. CHAID Exhaustivo.

Para la aplicación del algoritmo CHAID Exhaustivo fue necesario modificar el último conjunto de estandarizado "2_1_Conjunto_de_Datos_UNL.csv", se necesita que los temas se encuentren sin conectores y con los estados de cada proyecto, debido a que este algoritmo al ser de clasificación supervisada necesita que los datos tengan etiqueta o clase para poder clasificar. Se realizó lo siguiente para obtener el nuevo conjunto de datos: se cargaron los últimos datos estandarizados usando el código de la TABLA XVI, luego se obtuvo la columna de los temas en la variable "mis_temas" y la columna de los estados en la variable "mi_estado", para eliminar los conectores de los temas a la variable "mis_temas" se la convirtió a tipo Corpus y se removió los conectores, como se indica en la TABLA XXXI.

TABLA XXXI

OBTENCIÓN DE VARIABLE TEMA SIN CONECTORES

```
1 mis_temas<-procesamiento_carrera[,1]
2 mi_estado<-procesamiento_carrera[,2]
3 df <- do.call("rbind", lapply(mis_temas, as.data.frame))
4 miCorpus <- Corpus(VectorSource(df$ X[[i]]`))
5 miCorpus<- tm_map(miCorpus,removeWords,c("PARA","Y","UN"
6           ,"LA","DE","DEL","UNA","UNO"
7           ,"UNOS","EL","LOS","LAS","CON","UNAS","EN"
8           ,"A","E","QUE","POR","SU","ES","O","U","S"
9           ,"AL","COMO"))
10 miCorpus<-tm_map(miCorpus,stemDocument)
```

De este Corpus se obtuvo los temas de los proyectos sin conectores y se los unió a los estados de cada proyecto, pero en la columna “mis_estados” los estados se convirtieron en valores numéricos por lo que fue necesario convertirlos nuevamente los de valor 1 a “EXITO” y los de valor 2 a “FRACASO”. Por último, guardamos este nuevo conjunto de datos como “4_Conjunto_de_Datos_Sin_Conectores.csv”. El proceso anterior se muestra en la TABLA XXXII

TABLA XXXII

NUEVO ARCHIVO DE TEMAS SIN CONECTORES Y CON ESTADO

```
1 for (i in 1:length(miCorpus)) {
2   if(i==1){
3     s<-(strwrap(miCorpus[[i]], width=500))
4   }else{
5     s<-rbind(s,(strwrap(miCorpus[[i]], width=500)))
6   }
7 }
8 carrera<-cbind(s,mi_estado)
9 carrera[,2] = ifelse(carrera[,2] == "1", "EXITO", carrera[,2])
10 carrera[,2] = ifelse(carrera[,2] == "2", "FRACASO", carrera[,2])
11 write.csv (carrera, file="4_Conjunto_de_Datos_Sin_Conectores.csv")
```

Al nuevo conjunto de datos se lo limpió en Openrefine como se muestra en la Figura 31, eliminando la columna que crea R por defecto al crear un archivo CSV.

js_Sin_Conectores.csv [Enlace permanente](#) Abrir... Exportar Ayuda

1226 registros

Mostrar como: filas registrosMostrar: 5 10 25 50 registros

Exportar proyecto

Delimitado por tabulaciones

Delimitado por comas

Tabla HTML

Excel (.xls)

Excel en XML (.xlsx)

Hoja de cálculo ODF

Configurar exportación ...

Plantilla ...

QuickStatements

Columna	Columna 1	Columna 2	Columna 3	Columna 4	Columna 5	Columna 6	Columna 7	Columna 8	Columna 9	Columna 10	Columna 11	Columna 12	Columna 13	Columna 14
1	DESPLIEGUE	PROTOCOLO	INTERNET	VERSION	IPV	DISPOSITIVOS	CORE	NACION	CRUCER	MALACA	ETANOL	AGROPE	CIUDAD	MECANI
2	ESTIMACION	RESERVA	ELECCION	MEJOR	ALTERNATIVO	EXPLOTACION	CAOLI							
3	DISENO	CONSTRUCCION	MAQUINA	MANUAL	ENSAYO	TRACCION	MATEF							
4	ELECCION	SISTEMA	EXPLORACION	YESOS	EXISTENTES	AREA	MINER							
5	ESTUDIO	COSTO	MARGINALES	DISTRIBUCION	EMPRESA	ELECTRICO	REGIO							
6	ANALISIS	PARAMETROS	EFICIENCIA	MOTORES	COMBUSTION	INTERNA	FUNCI							
7	DISENO	CONSTRUCCION	TABLERO	DIDACTICO	TRANSFERENCIA	ELECTRICO	AUTOI							
8	CONSTRUCCION	SISTEMA	AUTOMATIZADO	FERTIRRIGACION	TERRENO	PROYECTO	RECUJ							
9	AUTOMATIZACION	SISTEMA	ALMACENAMIENTO	DISTRIBUCION	RECURSO	ENERGETICO	DIESE							
10	DISENO	CONSTRUCCION	SISTEMA	BOMBAS	DOBLE	EFECTO	CAUDAL	ACCIONADO	HIDRAULICAMENTE	MECANI				

Figura 33. Conjunto de datos con las columnas separadas sin conectores.

La Figura 34 muestra como una vez listo el conjunto de datos apto para la aplicación del algoritmo CHAID, fue cargado en la herramienta SPSS.

	Columna21	Columna22	Columna23	Columna24	Columna25	Columna26
1	DESPLIEGUE	PROTOCOLO	INTERNET	VERSION	IPV	DISPOSITIVOS
2	ESTIMACION	RESERVA	ELECCION	MEJOR	ALTERNATIVO	EXPLOTACION
3	DISENO	CONSTRUCCION	MAQUINA	MANUAL	ENSAYO	TRACCION
4	ELECCION	SISTEMA	EXPLORACION	YESOS	EXISTENTES	AREA
5	ESTUDIO	COSTO	MARGINALES	DISTRIBUCION	EMPRESA	ELECTRICO
6	ANALISIS	PARAMETROS	EFICIENCIA	MOTORES	COMBUSTION	INTERNA
7	DISENO	CONSTRUCCION	TABLERO	DIDACTICO	TRANSFERENCIA	ELECTRICO
8	CONSTRUCCION	SISTEMA	AUTOMATIZADO	FERTIRRIGACION	TERRENO	PROYECTO
9	AUTOMATIZACION	SISTEMA	ALMACENAMIENTO	DISTRIBUCION	RECURSO	ENERGETICO
10	DISENO	CONSTRUCCION	SISTEMA	BOMBAS	DOBLE	EFECTO
11	DETERMINACION	DEMANDA	POTENCIA	CLIENTES	RESIDENCIALES	ZONAS
12	DISENO	MECANICO	SISTEMA	TRANSPORTE	VACIADO	MATERIAL
13	DISENO	CONSTRUCCION	BANCO	MEDIR	FUERZA	IMPACTO
14	METODOLOGIA	CONTRASTACION	INSTRUMENTOS	MEDICION	ELECTRICIDAD	
15	DISENO	CONSTRUCCION	SISTEMA	HIDRAULICO	VIVIENDA	RURAL
16	PLANTA	TRATAMIENTO	RESIDUOS	SOLIDOS	AGUAS	GRISES
17	DISENO	CALCULO	CONSTRUCCION	DOBLADORA	SISTEMA	OLEOHIDRAULICO
18	DISENO	CONSTRUCCION	CASA	BIOCLIMATICA	DESMONTABLE	LABORATORIO
19	METODOLOGIA	DISENO	CONSTRUCCION	MALLAS	PUESTA	TIERRA
20	DISENO	IMPLEMENTACION	BANCO	PRUEBAS	DETERMINAR	CURVAS
21	DISENO	EVALUACION	TERMICA	PROTIPO	COLECTOR	SOLAR
22	DISENO	CONSTRUCCION	PROTOTIPO	ROBOT	CONTROL	RARIO
23	ESTIMACION	POTENCIAL	EOLIOENERGETICO	AEIRNIR	IMPLEMENTACION	AEROGENERADOR
24	DISENO	CONSTRUCCION	PROTOTIPO	MOTOR	STIRLING	TIPO

Figura 34. Conjunto de datos con las columnas separadas en SPSS.

Luego de cargar los datos se clasificó usando el árbol CHAID Exhaustivo, pero primero se especificó que la variable "mi_estado" es la variable dependiente y el resto de columnas las variables independientes, en la Figura 35 se presenta lo explicado.

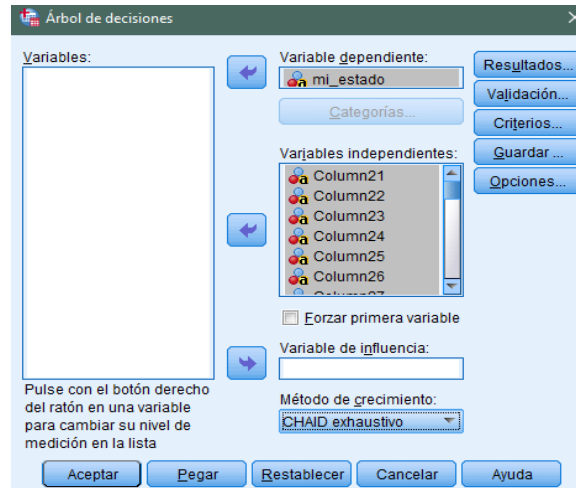


Figura 35. Variables independientes y dependiente en CHAID Exhaustivo.

Se configuró los criterios del algoritmo como se muestra en la Figura 36, para la división de nodos y el cambio mínimo de frecuencias se usó 0.05 y 0.001 respectivamente, estos valores son bajos para generar un árbol con pocos nodos ya que el conjunto de datos es extenso. También se estableció el número máximo de iteraciones en 100 para controlar el crecimiento del árbol hasta haber alcanzado ese número de iteraciones y se seleccionó la opción de Razón de Verosimilitud porque brinda mejores resultados al ser más robusto que el Pearson. Un último criterio establecido es el de ajuste de valores que permite corregir los valores de los criterios de división de nodos y fusión de categorías en caso de ser necesario.

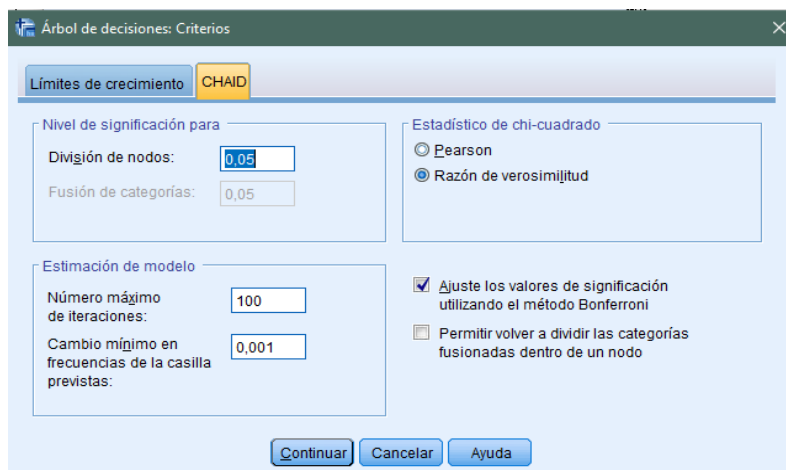


Figura 36. Criterios del algoritmo CHAID Exhaustivo.

El árbol obtenido cuenta con un porcentaje de eficiencia de 80.8% y una profundidad de 3, con 17 nodos de los cuales 11 son nodos terminales. Se visualiza el árbol completo en la Figura 37.

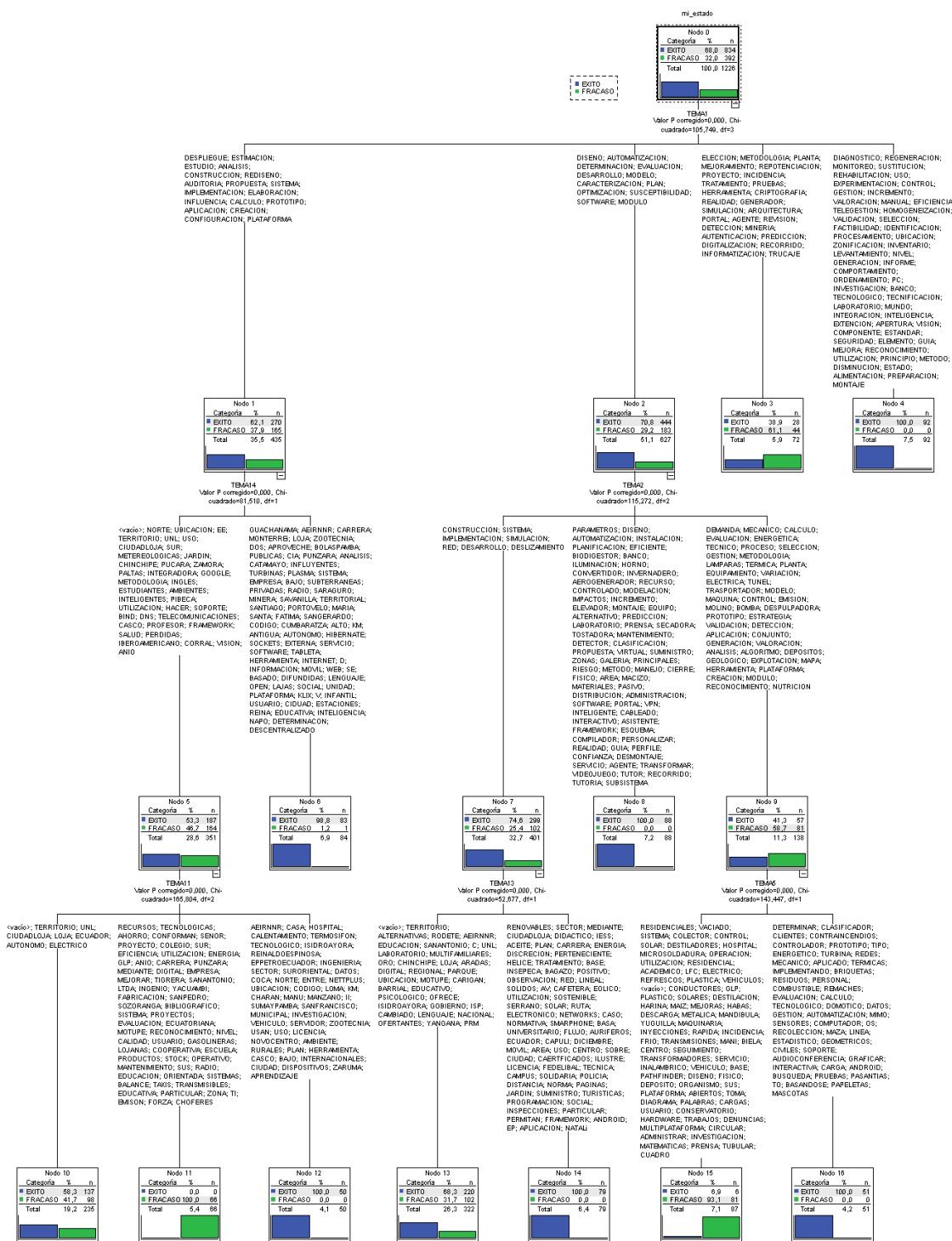


Figura 37. Árbol obtenido con el algoritmo CHAID Exhaustivo.

2.3.6. Random Forest.

El algoritmo fue aplicado en el conjunto de datos "5_Conjunto_de_Datos_SCo_SP.csv" usado en el algoritmo CHAID Exhaustivo, debido a que este conjunto de datos es óptimo para la aplicación de algoritmos de clasificación supervisada. Para la aplicación del

algoritmo se usó la herramienta RapidMiner, por lo que primeramente fue necesario cargar los datos en la herramienta, como se indica en la Figura 38.

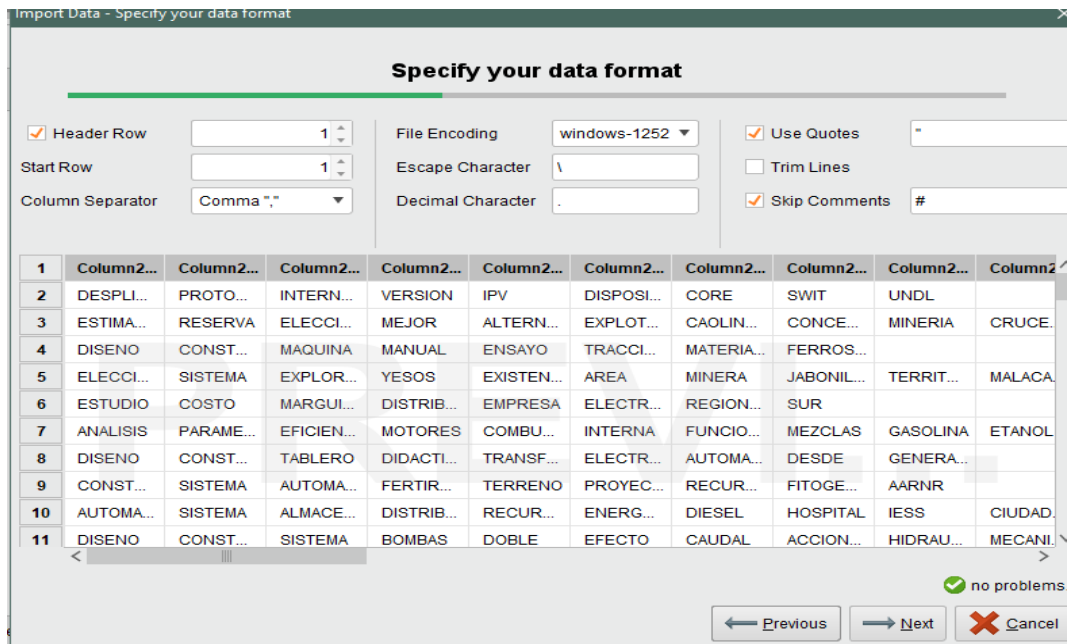


Figura 38. Carga del conjunto de datos en RapidMiner.

Es necesario establecer como tipo binomial y label a la columna dependiente para que la herramienta reconozca que solo hay dos clases en esa columna y que esa columna es la variable dependiente, de esta manera la herramienta no da problemas al momento de aplicar el algoritmo, lo descrito anteriormente se muestra en la Figura 39.

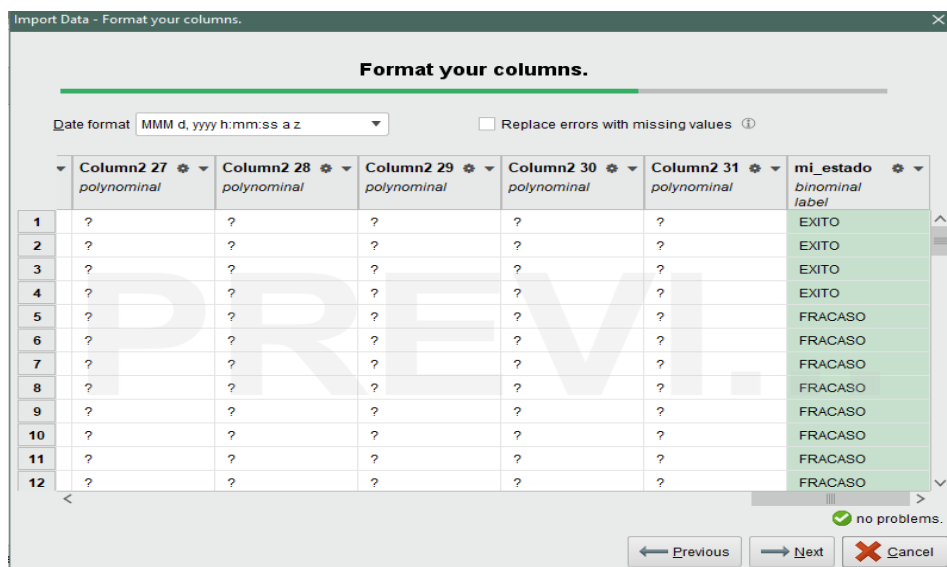


Figura 39. Establecer variable dependiente en el conjunto de datos.

En la Figura 40 se observa cómo en el área de Process se encuentra el conjunto de datos importado, adicional a ello se debe añadir el algoritmo Random Forest que se encuentra dentro de la categoría Tree y esta a su vez se encuentra en la categoría Modeling.

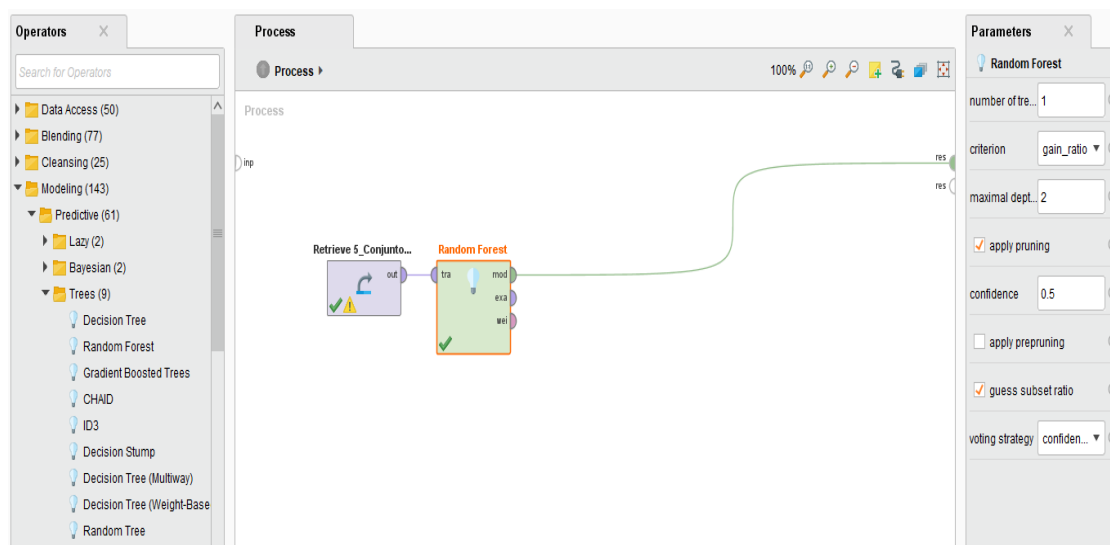


Figura 40. Aplicación del algoritmo Random Forest.

Los criterios empleados más influyentes en el algoritmo son el número de árboles y el máximo de divisiones, en el primero se puso uno para obtener solamente el árbol de la columna más influyente y en el segundo criterio se colocó dos en base a que solo existen dos clases, de éxito o fracaso. También se empleó un 0.5 de nivel de confianza debido a que es el nivel más alto que permite la herramienta. El árbol obtenido es presentado en la evaluación de resultados de este algoritmo.

3. Evaluación de los algoritmos aplicados.

3.1. Evaluación de los resultados.

Los resultados obtenidos en cada algoritmo se evaluaron colocando en una tabla a los más relevantes y posterior a ello dar mayor importancia a las palabras o relaciones que están presentes en los resultados de uno de los estados y en los resultados de todo el conjunto de datos. Estas palabras o relaciones son representadas gráficamente usando el código de la TABLA XXX.

Antes de proceder a evaluar los resultados de los algoritmos, se evaluó las representaciones gráficas de los datos obtenidas con las nubes de palabras para

conocer las palabras más influyentes en todo el conjunto de datos y en el éxito o fracaso de los proyectos. Con la nube de palabras se obtuvo lo siguiente:

- **Nube de Palabras.**

Las nubes de palabras muestran las palabras más usadas en el conjunto de datos, para ello usan la frecuencia con la que aparecen las palabras.



Figura 41. Nube de Palabras de todo el conjunto de datos.

En la Figura 41 se observa que las palabras más usadas en todo el conjunto de datos son: “territorio”, “diseno”, “sistema”, “unl”, “implementacion”, “construccion”, “desarrollo”, “loja”, “ciudadloja” y “web”.



Figura 42. Nube de Palabras de los proyectos de éxito.

La Figura 42 presenta que las palabras usadas con mayor frecuencia en los proyectos de titulación que culminaron con éxito son: “territorio”, “diseno”,

“sistema”, “implementacion”, “construccion”, “desarrollo”, “unl” y “loja”, “ciudadloja” y “web”.



Figura 43. Nube de Palabras de los proyectos de fracaso

Se observa en la Figura 43 que las palabras más frecuentes en los proyectos de titulación con estado de fracaso, y son: “diseño”, “sistema”, “territorio”, “unl”, “implementación”, “desarrollo”, “ciudadloja”, “construcción”, “loja” y “estudio”. Las palabras obtenidas tanto en el fracaso son las mismas que se obtuvieron en el éxito, por lo que pierden relevancia al no ayudar a determinar si influyen en el fracaso. Las palabras que son más usadas en todo el conjunto de datos y en los proyectos con estado de éxito y fracaso se presentan en la TABLA XXXIII.

TABLA XXXIII

COMPARATIVA DE RESULTADOS DE LAS NUBES DE PALABRAS

Todo el conjunto de datos	Proyectos con éxito	Proyectos con fracaso
sistema	territorio	diseño
diseño	diseño	sistema
territorio	sistema	territorio
unl	implementación	unl
desarrollo	construcción	implementación
implementación	desarrollo	desarrollo
construcción	unl	ciudadloja
loja	loja	construcción

ciudadloja	ciudadloja	loja
web	web	estudio

La palabra “web” es la única palabra que se encuentra entre las más usadas en todo el conjunto de datos y en los proyectos culminados con éxito, por lo tanto, se determina que los proyectos de titulación que contienen la palabra “web” tienden a culminar con éxito; en el caso de los proyectos culminados con fracaso se encontró algunas coincidencias con todo el conjunto de datos pero al coincidir también con las de éxito se descartaron al no ayudar a determinar la influencia de las palabras en el estado final de los proyectos. La representación gráfica de la palabra “web” se observa en la Figura 44.

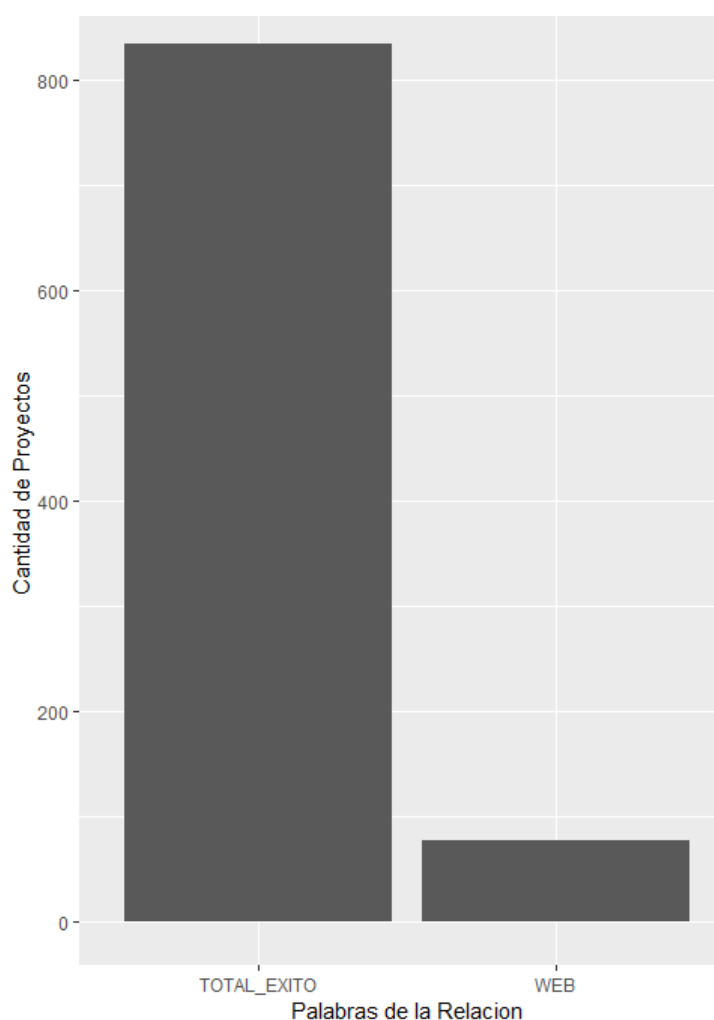


Figura 44. Representación gráfica de la palabra web.

Una vez evaluados los resultados obtenidos con la nube de palabras se procedió a evaluar los resultados de los algoritmos de clasificación que fueron aplicados.

3.1.1. Dendrograma.

Los dendogramas realizan clústers para agrupar las palabras y presentarlas en forma jerárquica.

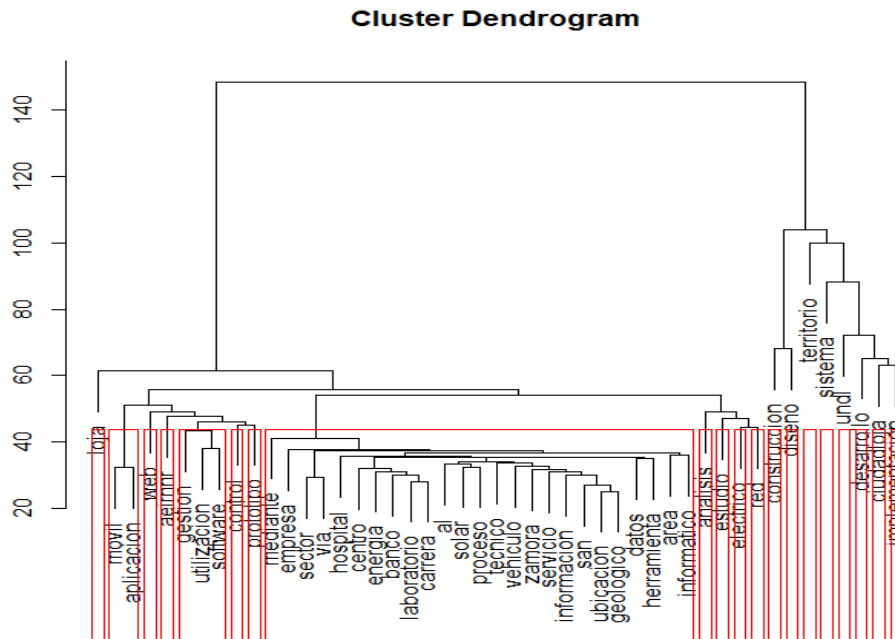


Figura 45. Dendrograma de todo el conjunto de datos.

La Figura 45 presenta el dendrograma de todo el conjunto de datos, las relaciones obtenidas son: “movil-aplicacion” y “gestión-utilizacion-software”.

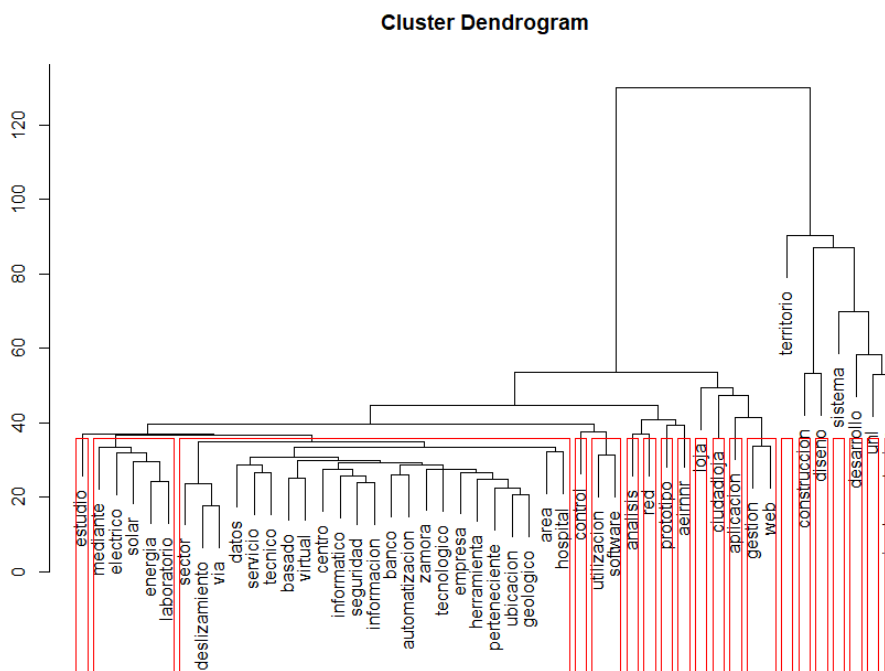


Figura 46. Dendrograma de los proyectos de éxito.

En la Figura 46 del dendrograma de los proyectos con estado de éxito se obtuvo las siguientes relaciones: “electrico-solar-energia-laboratorio-mediante”, “gestion-web” y “utilizacion-software”.

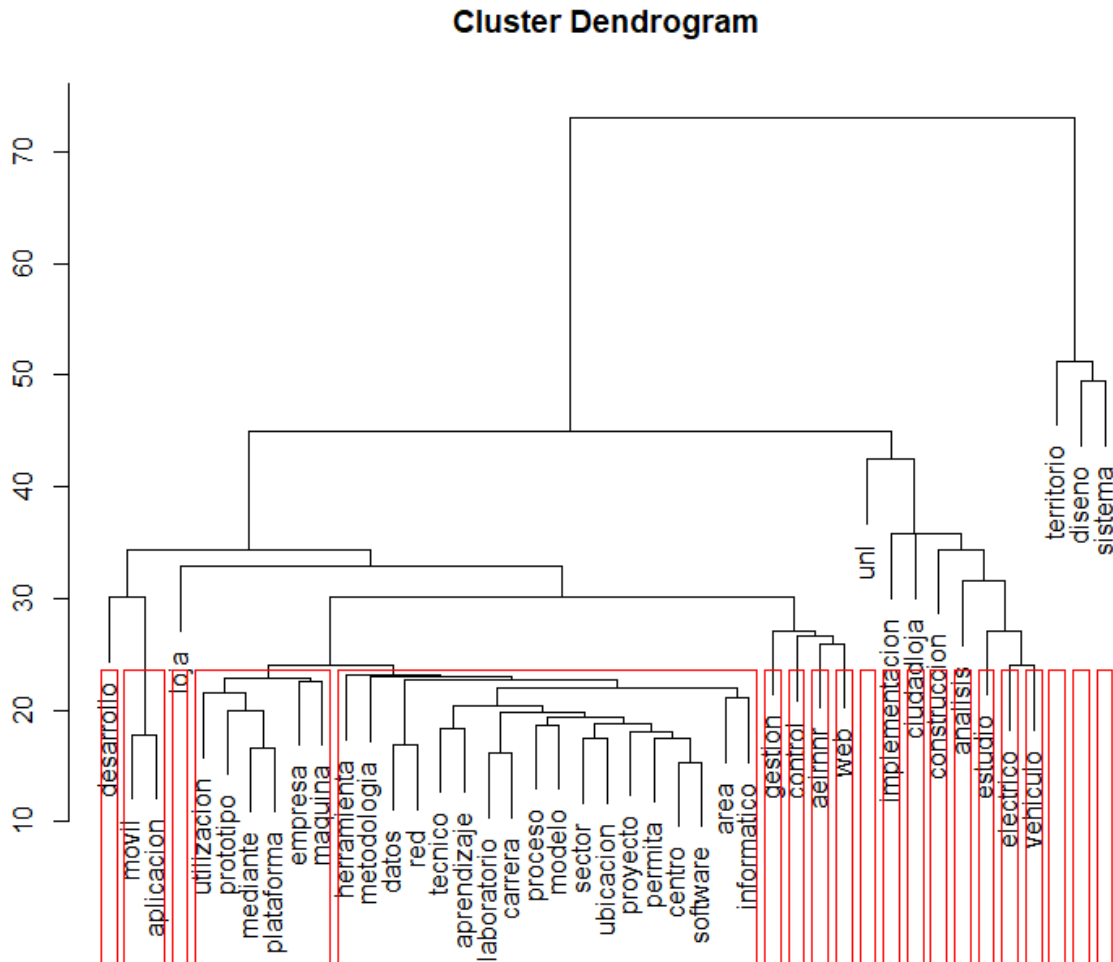


Figura 47. Dendrograma de los proyectos de fracaso.

Se observa que en la Figura 47 del dendrograma de los proyectos que culminaron con estado de fracaso que las relaciones obtenidas son: “movil-aplicacion” y “empresa-utilizacion-maquina-prototipo-mediante-plataforma”. Las relaciones obtenidas de los dendrogramas son evaluadas en la TABLA XXXIV que es mostrada a continuación.

TABLA XXXIV

COMPARATIVA DE RESULTADOS DEL DENDOGRAMA

Todo el conjunto de datos	Proyectos con éxito	Proyectos con fracaso
movil-aplicacion	electrico-solar-energia-laboratorio-mediante	movil-aplicacion
gestion-utilizacion-software	gestion-web	empresa-utilizacion-maquina-prototipo-mediante-plataforma
	utilizacion-software	

La relación “móvil-aplicación” es la única coincidencia encontrada, esta se encuentra en el dendograma de todo y en el dendograma de fracaso, por lo cual se demuestra que esta relación tiene influencia en el fracaso de los proyectos de titulación.

En la Figura 48 se observa la representación gráfica de la relación “móvil-aplicación”.

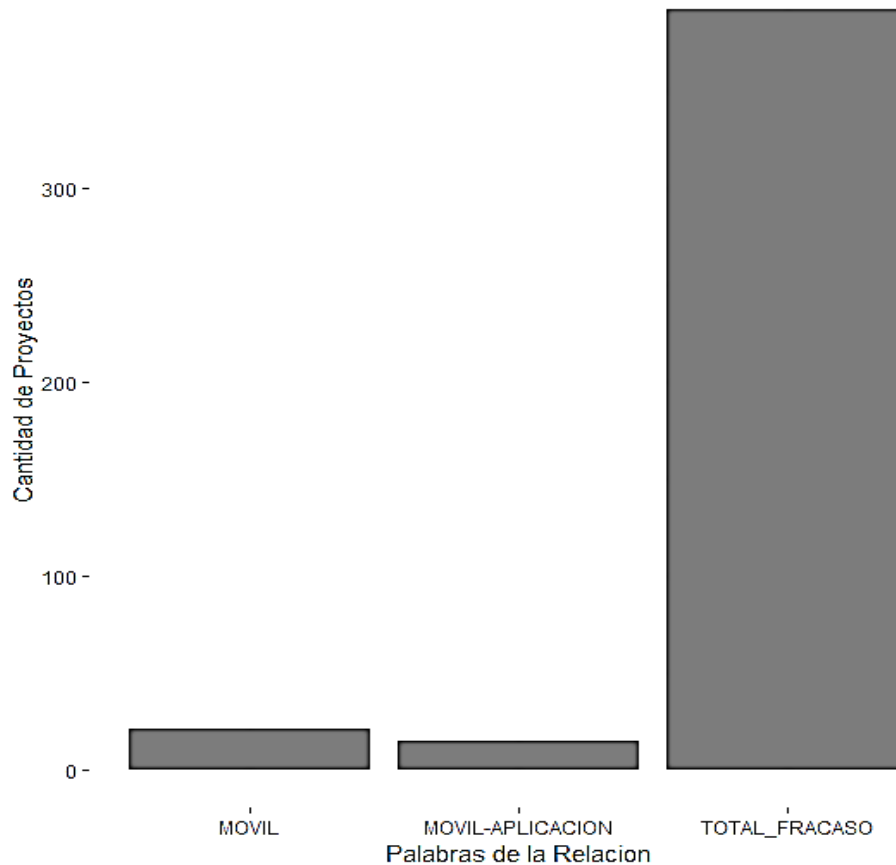


Figura 48. Representación gráfica de la relación móvil aplicación.

3.1.2. Reglas de Asociación Apriori.

Las reglas de asociación obtenidas de todo el conjunto de datos se pueden observar en la Figura 49, donde también se observan diversos valores que se detallan a continuación:

- **Lhs**, es la palabra independiente de la relación, es decir cuando se encuentra en el tema también estará presente en el tema la palabra perteneciente a la columna rhs.
- **Rhs**, es la palabra dependiente de la relación, ya que depende de lhs para estar presente en el tema de un proyecto.
- **Support**, es el número de veces en las que aparece una relación en el tema de un proyecto dividido para el número total de proyectos.
- **Confidence**, es la probabilidad de que rhs esté cuando esta lhs.
- **Lift**, compara la frecuencia observada de una regla con la frecuencia esperada simplemente por azar, cuanto más se aleje el valor de lift de 1 mayor es la evidencia de que la regla representa un patrón real.
- **Count**, es la cantidad de veces en las que esa relación se encuentra en el conjunto de datos.

	lhs	rhs	support	confidence	lift	count
[1]	{SOLAR}	=> {DISENO}	0.03017945	0.8222222	2.388731	37
[2]	{GEOLOGICO}	=> {TERRITORIO}	0.03099511	0.9047619	5.159247	38
[3]	{GESTION}	=> {SISTEMA}	0.05383361	0.7096774	2.202695	66
[4]	{CONSTRUCCION}	=> {DISENO}	0.13376835	0.8631579	2.507658	164
[5]	{CONSTRUCCION, SISTEMA}	=> {DISENO}	0.03262643	0.9090909	2.641103	40

Figura 49. Reglas de Asociación de todo el conjunto de datos.

Se representan las reglas de asociación obtenidas en la Figura 50.

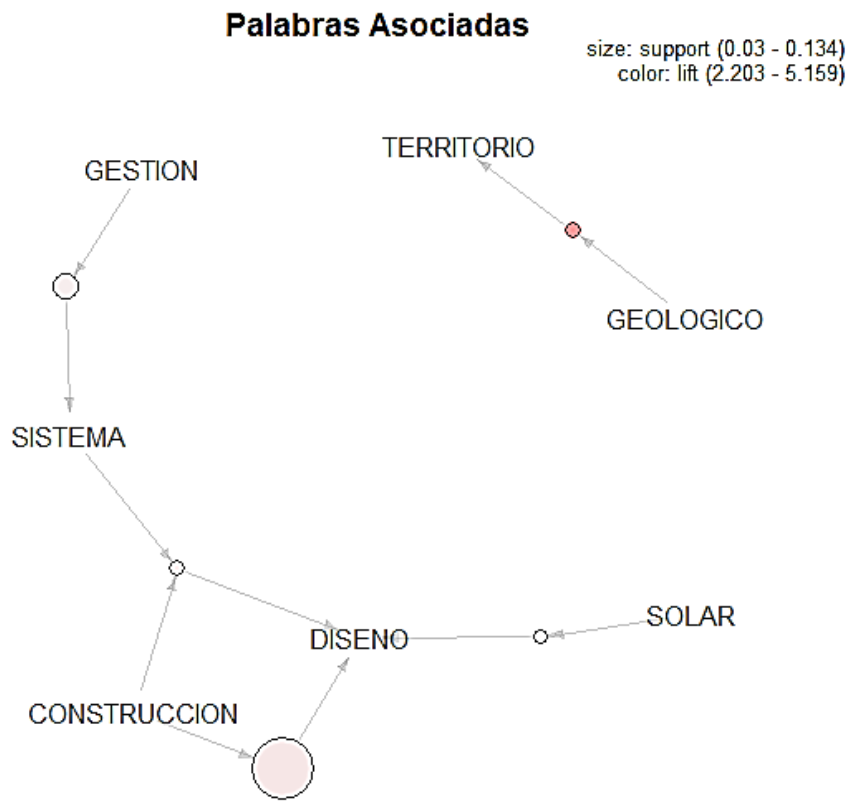


Figura 50. Reglas relacionadas de todo el conjunto de datos.

Entre las relaciones que se observan tenemos que cuando en el tema de un proyecto de titulación esta la palabra “GEOLOGICO”, esta va acompañada de la palabra “TERRITORIO”. También se observa que la palabra “SISTEMA” se encuentra acompañada de la palabra “GESTION”. Otra relación obtenida es que cada vez que se encuentra la palabra “DISENO” también están presente las palabras “SOLAR” o “CONSTRUCCION” o las palabras “SISTEMA” y “CONSTRUCCION”.

Las reglas de asociación de los proyectos de titulación con estado de éxito se observan en la Figura 51.

	lhs	rhs	support	confidence	lift	count
[1]	{GESTION}	=> {SISTEMA}	0.05515588	0.7796610	2.435346	46
[2]	{LOJA}	=> {TERRITORIO}	0.10431655	0.7131148	3.694023	87
[3]	{CONSTRUCCION}	=> {DISENO}	0.15107914	0.8750000	2.499144	126

Figura 51. Reglas de Asociación de los proyectos de éxito.

Estas reglas se representan en la Figura 52.

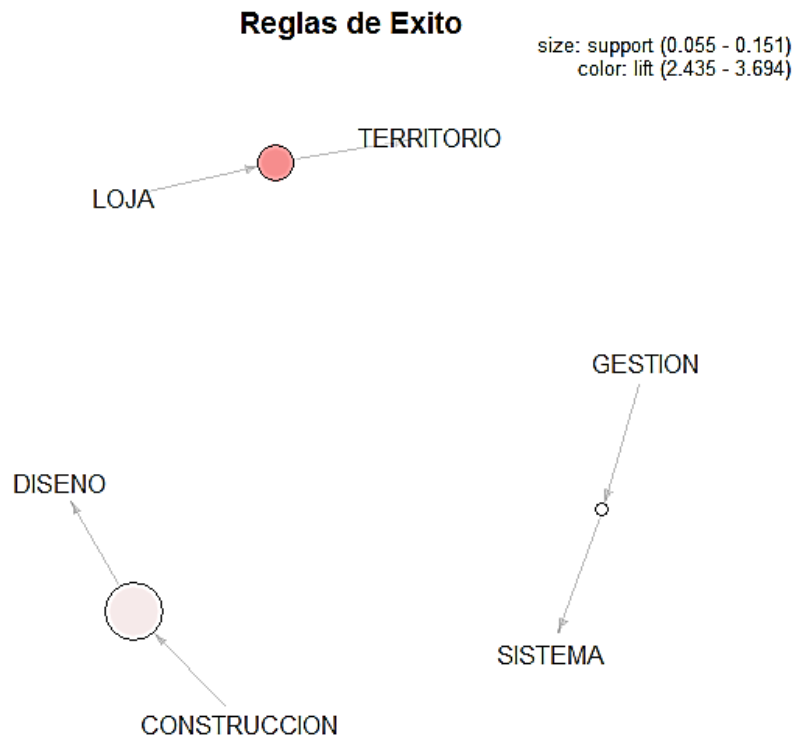


Figura 52. Reglas relacionadas de los proyectos de éxito.

Se observa que las palabras “LOJA” y “TERRITORIO” están relacionadas, es decir siempre que se encuentra en el tema la palabra “LOJA” también se encuentra la palabra “TERRITORIO”. También es visible que la palabra “SISTEMA” está presente en un tema cuando también se encuentra la palabra “GESTION”, y que al estar presente en el tema la palabra “CONSTRUCCION” también se encuentra la palabra “DISENO”.

Las reglas de asociación de todos los proyectos con estado de fracaso se observan en la Figura 53.

	lhs	rhs	support	confidence	lift	count
[1]	{MAQUINA}	=> {DISENO}	0.03826531	1.0000000	3.015385	15
[2]	{MOVIL}	=> {APLICACION}	0.03316327	0.7222222	9.762452	13
[3]	{CARRERA}	=> {UNL}	0.03571429	0.7777778	4.065185	14
[4]	{CONSTRUCCION}	=> {DISENO}	0.09693878	0.8260870	2.490970	38

Figura 53. Reglas de Asociación de los proyectos de fracaso.

Se representan las reglas obtenidas en la Figura 54.

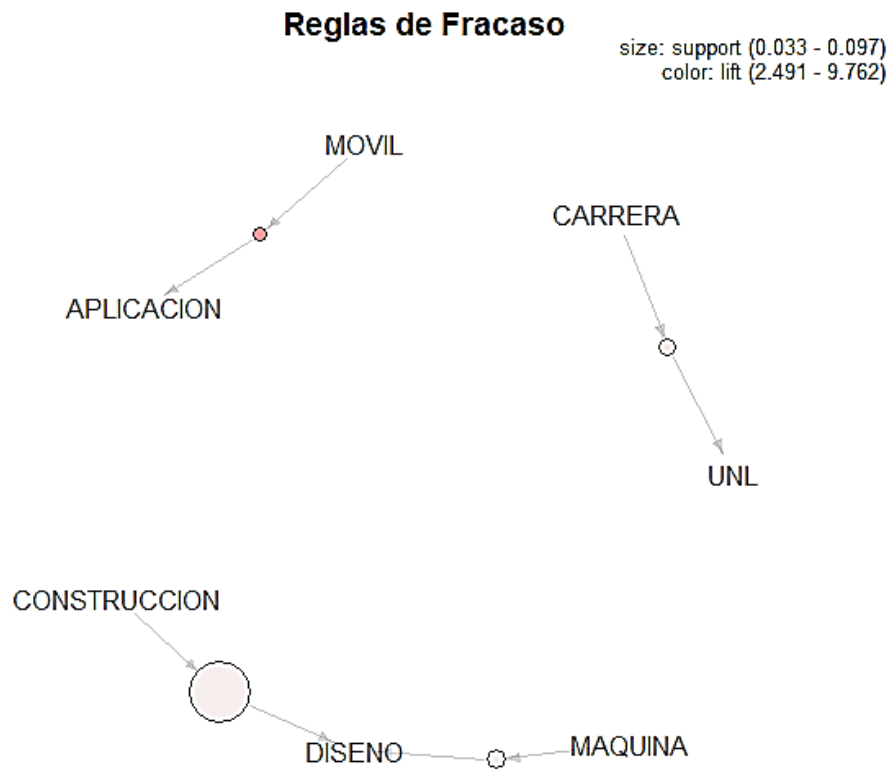


Figura 54. Reglas relacionadas de los proyectos de fracaso.

Entre las relaciones que se observan en la Figura 54 tenemos que la palabra “APLICACION” está presente en un tema cuando la palabra “MOVIL” también lo está y cuando se encuentra la palabra “CARRERA” está también la palabra “UNL”. Otras relaciones observables son que cuando está la palabra “MAQUINA” se encuentra también la palabra “DISENO” y que la palabra “DISENO” forma parte del tema de un proyecto cuando en el tema también se encuentra la palabra “CONSTRUCCION”. En la TABLA XXXV se presentan todas las relaciones obtenidas con el algoritmo.

TABLA XXXV

COMPARATIVA DE RESULTADOS DE APRIORI

Todo el conjunto de datos	Proyectos con éxito	Proyectos con fracaso
SOLAR	CONSTRUCCION	MAQUINA
DISENO	DISENO	DISENO
GESTION	GESTION	MOVIL
SISTEMA	SISTEMA	APLICACION

GEOLOGICO TERROTORIO	LOJA TERRITORIO	CARRERA UNL
CONSTRUCCION DISENO		CONSTRUCCION DISENO
SISTEMA DISENO		

La relación "GESTION-SISTEMA" es la única de las relaciones de éxito que coincide con una de las relaciones de todo el conjunto de datos, demostrando así la influencia que tiene esta relación en los temas de los proyectos de titulación. De las relaciones de fracaso no se encontró ninguna coincidencia con las relaciones de todo el conjunto de datos, pero si con una relación de éxito por lo que fue descartado al no ayudar a determinar cómo influye en el estado final de un proyecto de titulación.

En la Figura 55 se observa la representación gráfica de la relación "GESTION-SISTEMA".

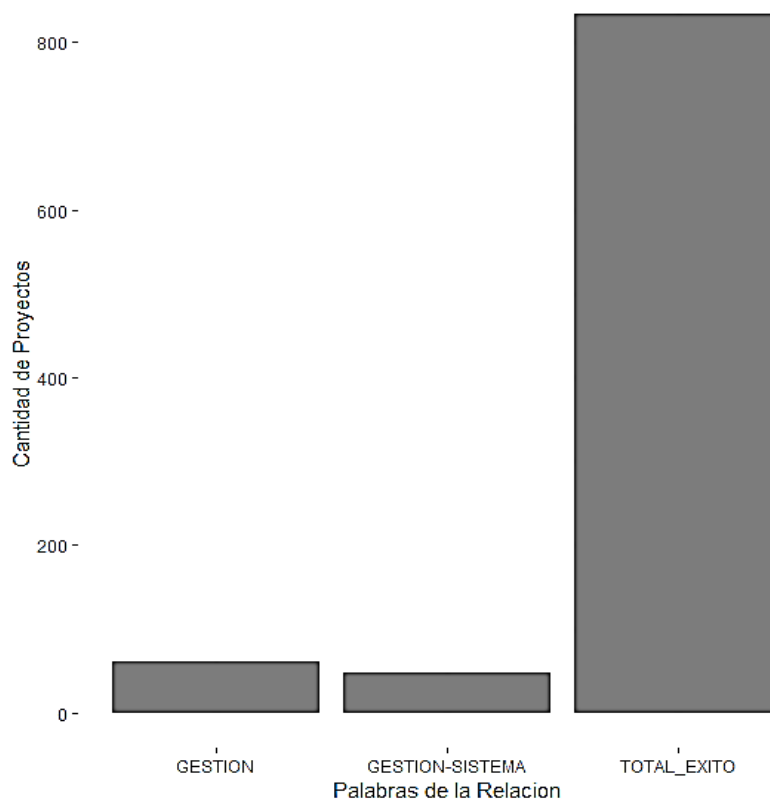


Figura 55. Representación gráfica de la relación gestión-sistema.

3.1.3. K-means.

Las relaciones obtenidas al aplicar k-means en todo el conjunto de datos se pueden observar en la Figura 25, las obtenidas de los proyectos de titulación con estado de éxito en la Figura 27 y las obtenidas de los proyectos con estado de fracaso en la Figura 29. Las relaciones de la aplicación de k-means a todo el conjunto de datos tienen el nombre de kmeans_todo, las obtenidas de los proyectos con estado de éxito se llaman kmeans_exito y las de los proyectos con estado de fracaso se denominan kmeans_fracaso. En la TABLA XXXVI se presentan las relaciones obtenidas al aplicar el algoritmo k-means.

TABLA XXXVI
COMPARATIVA DE RESULTADOS DE K-MEANS

No.	Kmeans_todo	Kmeans_exito	Kmeans_fracaso
1	desarrollo sistema implementacion	desarrollo sistema unl	aplicacion movil desarrollo
2	aplicacion desarrollo movil	unl sistema diseno	territorio loja sistema
3	diseno construccion prototipo	ciudadloja sistema implementacion	ciudadloja estudio electrico
4	diseno construccion aeirnr	software loja implementacion	diseno sistema implementacion
5	ciudadloja implementacion analisis	red diseno unl	unl sistema implementacion
6	implementacion unl diseno	territorio loja estudio	desarrollo unl plataforma
7	sistema diseno gestion	aplicacion desarrollo web	sistema desarrollo web

8	territorio loja estudio	diseño construcción prototipo	análisis diseño implementación
9	eléctrico diseño sistema	sistema diseño loja	construcción diseño sistema

Las relaciones de éxito kmeans_exito 6 y kmeans_exito 8 coinciden con dos relaciones de todo el conjunto de datos, kmeans_todo 8 y kmeans_todo 3 respectivamente, estas relaciones de éxito demuestran tener relevancia al influir también en todo el conjunto de datos; la relación que presenta kmeans_exito 6 es: “territorio-loja-estudio” y la que presenta kmeans_exito 8 es: “diseño-construcción-prototipo”. También una relación de fracaso coincide con una relación de todo el conjunto de datos, estas relaciones son kmeans_fracaso 1 y kmeans_todo 2; la relación que presenta kmeans_fracaso 1 es: “aplicación-movil-desarrollo”. El resto de relaciones de éxito y fracaso no coinciden con alguna de las relaciones de todo el conjunto de datos por lo que no serán tomadas en cuenta al no presentar gran influencia.

La representación gráfica de la relación de éxito kmeans_exito 8 y kmeans_todo 3 (“DISEÑO CONSTRUCCION PROTOTIPO”) se observa en la Figura 56.

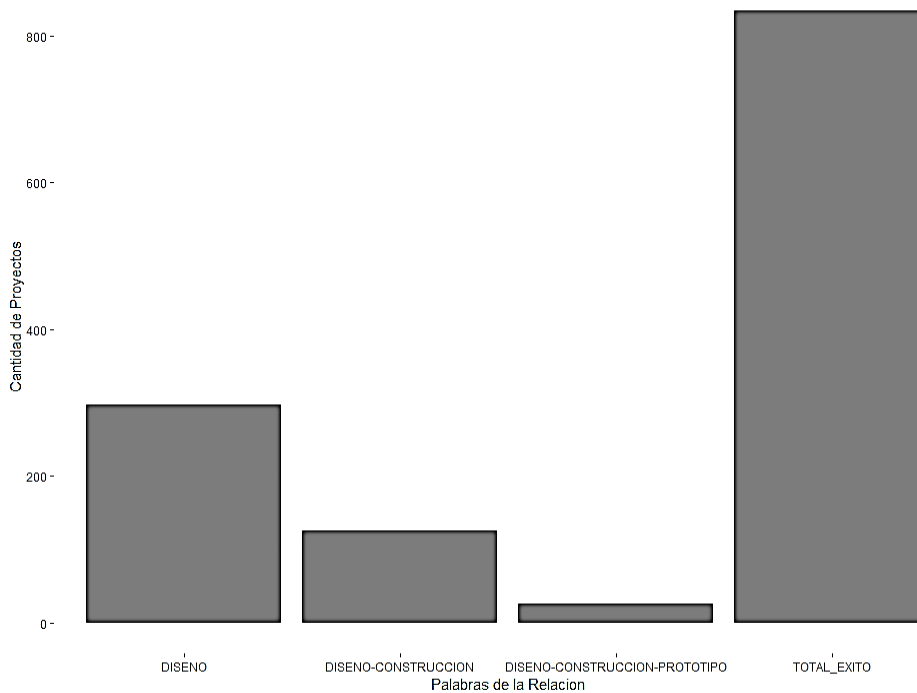


Figura 56. Representación gráfica de la relación diseño-construcción-prototipo.

La representación gráfica de la relación de éxito kmeans_exito 6 y kmeans_todo 8 ("TERRITORIO LOJA ESTUDIO") se visualiza en la Figura 57.

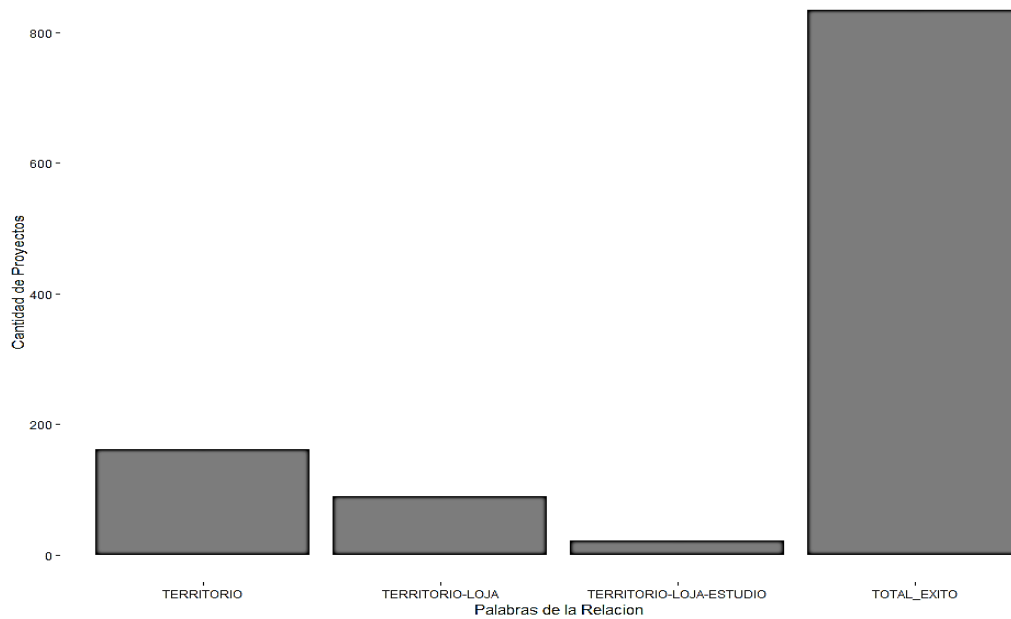


Figura 57. Representación gráfica de la relación territorio-Loja-estudio.

En la Figura 58 se observa la representación gráfica de la relación de fracaso kmeans_fracaso 1 y kmeans_todo 2 ("APLICACION MOVIL DESARROLLO").

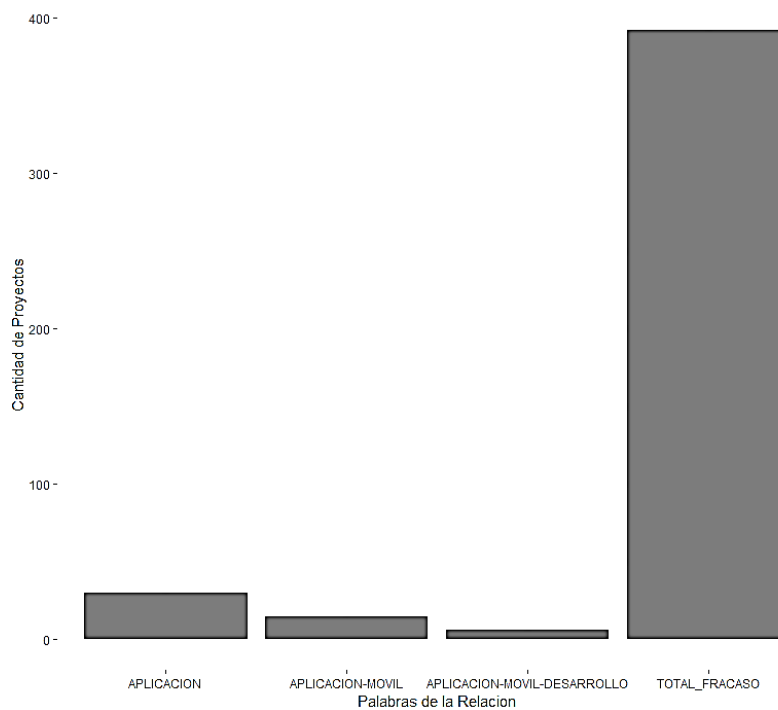


Figura 58. Representación gráfica de la relación aplicación-móvil-desarrollo.

3.1.4. K-medoids.

Las relaciones obtenidas de la aplicación de k-medoids a todo el conjunto de datos se denomina kmedoids_todo, a los proyectos con estado de éxito kmedoids_exito y a los proyectos con estado de fracaso kmedoids_fracaso. Las relaciones de kmedoids_todo se pueden observar en la Figura 26, de kmedoids_exito en la Figura 28 y de kmedoids_fracaso en la Figura 30. Las relaciones obtenidas con el algoritmo k-medoids se observan en la TABLA XXXVII.

TABLA XXXVII
COMPARATIVA DE RESULTADOS DE K-MEDOIDS

No.	Kmedoids_todo	Kmedoids_exito	Kmedoids_fracaso
2	loja	loja	diseno
3	diseno	diseno	sistema

Las relaciones kmedoids_exito 2 y kmedoids_exito 3 coinciden con las relaciones de todo el conjunto de datos kmedoids_todo 2 y kmedoids_todo 3 pero la relación kmedoids_exito 3 también coincide con la relación de fracaso kmedoids_fracaso 2, perdiendo así esta relevancia y determinando que solamente la relación kmedoids_exito 2 influye en el éxito de los proyectos de titulación. También se observa que la relación de fracaso restante no coincide con alguna de las relaciones de kmedoids_todo, por lo que al no influir en todo el conjunto de datos pierde relevancia. La palabra que presenta la relación kmedoids_exito 2 es: “loja”, la representación gráfica de esta palabra se presenta en la Figura 59.

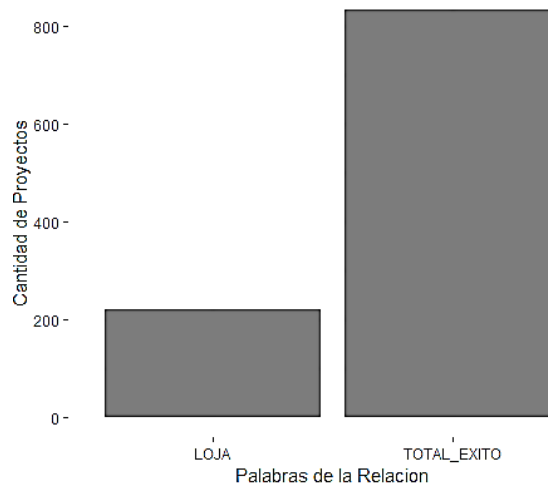


Figura 59. Representación gráfica de la palabra Loja.

3.1.5. CHAID Exhaustivo.

La visualización del árbol que se observa en la Figura 37 no es la más adecuada debido a su gran tamaño por ello se realizará la descripción de cada uno de los caminos para llegar a los nodos terminales que posean una gran diferencia en los porcentajes de éxito y fracaso. Se observa en la Figura 60 que la columna 1 es el nodo raíz, representando así que las palabras presentes en la posición 1 del tema de un proyecto son las más influyentes para determinar el estado de los proyectos de titulación. También se observa que el primer nivel del árbol posee 4 nodos de los cuales 2 son nodos terminales.

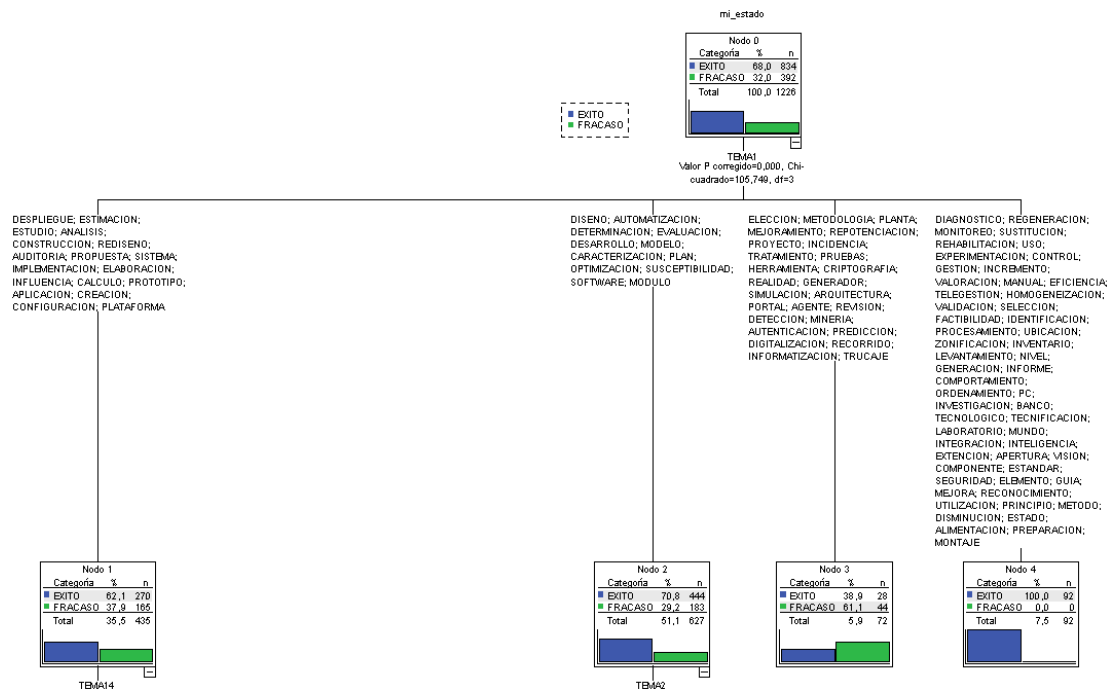


Figura 60. Primer nivel del árbol de CHAID Exhaustivo.

En el primer nivel del árbol el nodo terminal que presenta una gran diferencia en los porcentajes de los estados es el nodo 4, así que se empezó detallando este nodo.

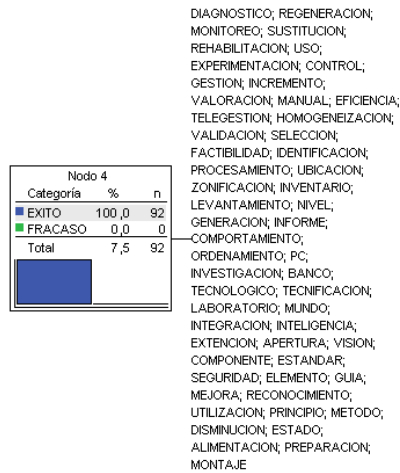


Figura 61. Nodo 4 del árbol de CHAID Exhaustivo.

El nodo 4 se muestra en la Figura 61 e indica que, si una de las palabras pertenecientes a este nodo se encuentra en la posición 1 del tema de un proyecto de titulación sin tomar en cuenta los conectores, el proyecto de titulación tiene un 100% de posibilidades de terminar con éxito. Ya que en el nivel 1 del árbol no se encontraron más nodos terminales con gran diferencia en los porcentajes se evaluó el siguiente nivel del árbol.

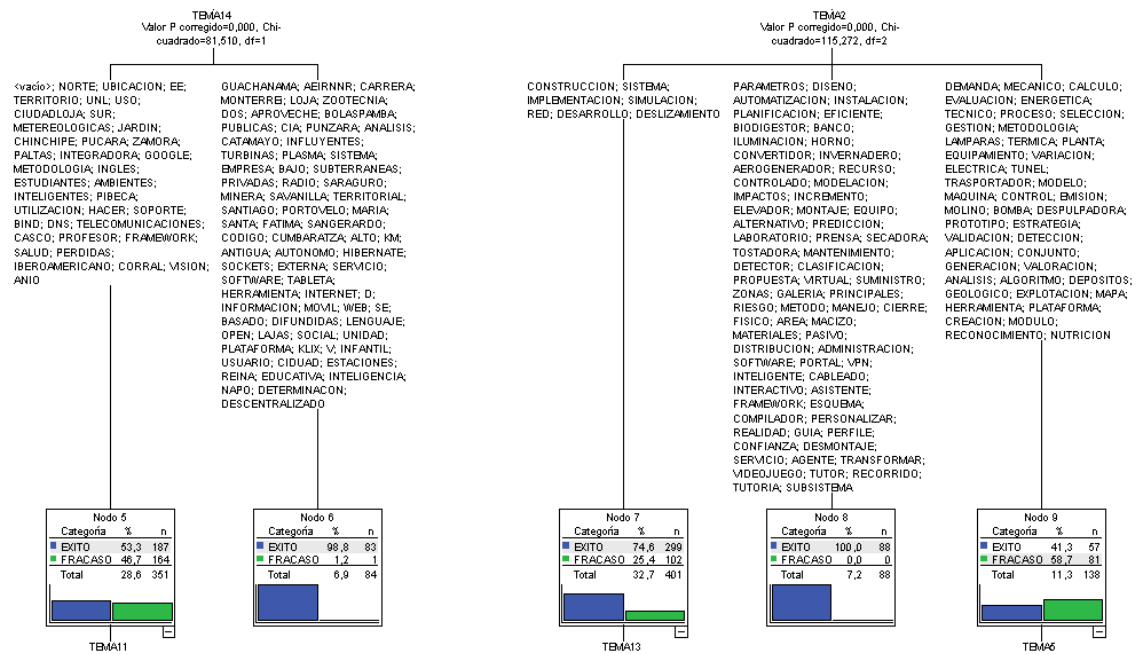


Figura 62. Segundo nivel del árbol de CHAID Exhaustivo.

En la Figura 62 se muestra el nivel 2 del árbol, donde se observa que los nodos 6 y 8 poseen una mayor diferencia respecto al estado final, por ello se evaluaron ambos por

separado debido a que son precedidos por diferentes nodos. El nodo 6 es precedido por el nodo 1 del primer nivel del árbol, así que se evalúan ambos.

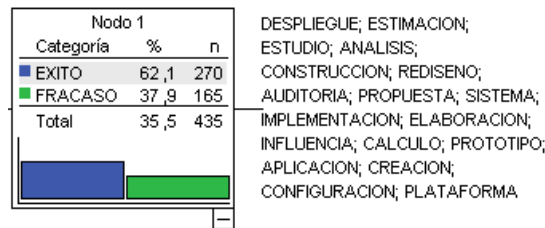


Figura 63. Nodo 1 del árbol de CHAID Exhaustivo.

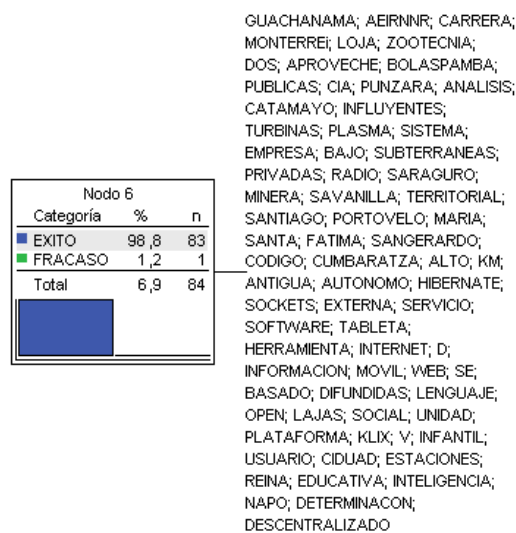


Figura 64. Nodo 6 del árbol de CHAID Exhaustivo.

Se observa en la Figura 63 y la Figura 64 que el nodo 6 es perteneciente a la columna 14, lo que indica que, si una de las palabras contenidas en el nodo 1 se encuentra en la posición 1 y una de las palabras contenidas en el nodo 6 se encuentra en la posición 14 del tema de un proyecto de titulación, este tiene un 98.8% de posibilidades de culminar con éxito. El nodo 8 en cambio es precedido por el nodo 2, por lo que se evaluaron ambos.

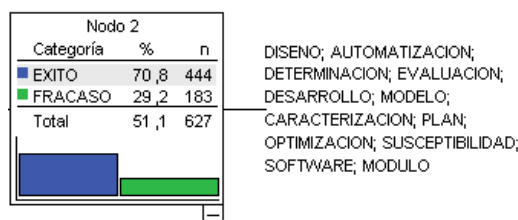


Figura 65. Nodo 2 del árbol de CHAID Exhaustivo.

PARAMETROS; DISEÑO; AUTOMATIZACION; INSTALACION; PLANIFICACION; EFICIENTE; BIODIGESTOR; BANCO; ILUMINACION; HORNO; CONVERTIDOR; INVERNADERO; AEROGENERADOR; RECURSO; CONTROLADO; MODELACION; IMPACTOS; INCREMENTO; ELEVADOR; MONTAJE; EQUIPO; ALTERNATIVO; PREDICCIÓN; LABORATORIO; PRENSA; SECADORA; TOSTADORA; MANTENIMIENTO; DETECTOR; CLASIFICACION; PROPUESTA; VIRTUAL; SUMINISTRO; ZONAS; GALERIA; PRINCIPALES; RIESGO; METODO; MANEJO; CIERRE; FISICO; AREA; MACIZO; MATERIALES; PASIVO; DISTRIBUCION; ADMINISTRACION; SOFTWARE; PORTAL; VPN; INTELIGENTE; CABLEADO; INTERACTIVO; ASISTENTE; FRAMEWORK; ESQUEMA; COMPILADOR; PERSONALIZAR; REALIDAD; GUIA; PERFIL; CONFIANZA; DESMONTAJE; SERVICIO; AGENTE; TRANSFORMAR; VIDEOJUEGO; TUTOR; RECORRIDO; TUTORIA; SUBSISTEMA

Nodo 8		
Categoría	%	n
EXITO	100,0	88
FRACASO	0,0	0
Total	7,2	88

Figura 66. Nodo 8 del árbol de CHAID Exhaustivo.

La Figura 65 y la Figura 66 presentan que el nodo 8 es perteneciente a la columna 2 y tomando en cuenta que el nodo 2 pertenece a la columna 1, se determina que siempre que una de las palabras del nodo 8 se encuentre en la posición 2 del tema de un proyecto y también una de las palabras contenidas en el nodo 2 se encuentre en la posición 1 del tema de un proyecto, este proyecto tendrá un 100% de probabilidad de culminar con éxito. Al no encontrarse más nodos terminales con gran diferencia entre las probabilidades de culminar con éxito o fracaso se evaluó el último nivel del árbol.

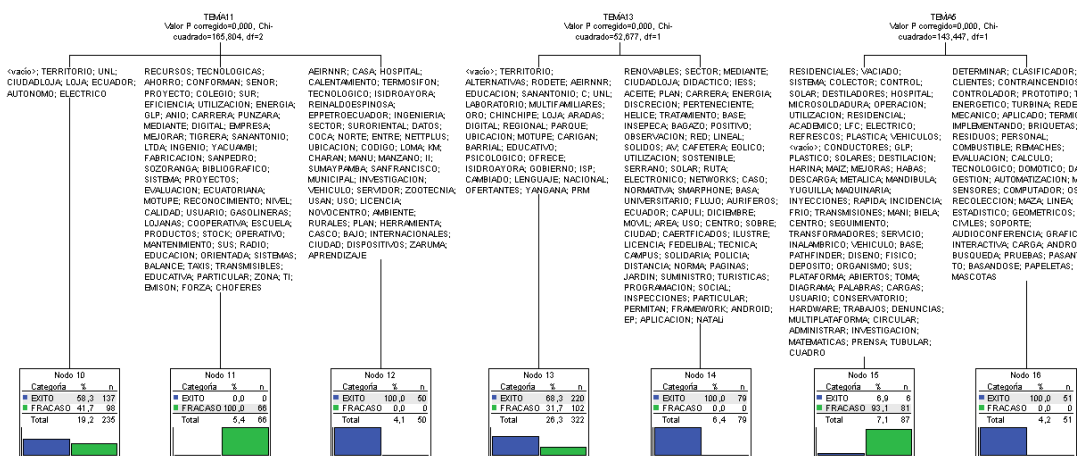


Figura 67. Tercer nivel del árbol de CHAID Exhaustivo.

El nivel 3 del árbol es presentado en la Figura 67, donde se identifica a los nodos 11, 12, 14, 15 y 16 como los nodos en los que existe una mayor diferencia entre las probabilidades de culminar con éxito o fracaso, debido a ello se evaluó cada uno de los

nodos. El primer nodo analizado es el nodo 12 que se muestra en la Figura 69, este nodo es precedido por el nodo 5 que es presentado en la Figura 68 y este nodo a su vez es precedido por el nodo 1, por lo tanto, se evaluó los tres nodos.

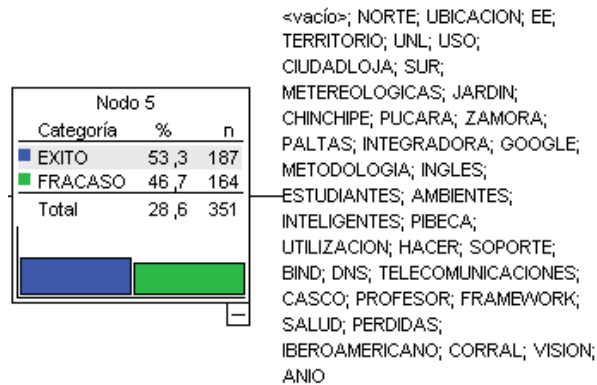


Figura 68. Nodo 5 del árbol de CHAID Exhaustivo.

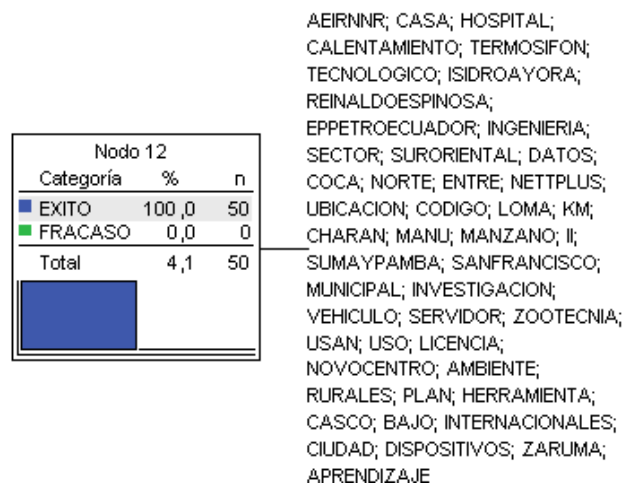


Figura 69. Nodo 12 del árbol de CHAID Exhaustivo.

Al nodo 1 se lo puede visualizar en la Figura 63, en donde se observa que pertenece a la columna 1, el nodo que le sigue es el nodo 5 que pertenece a la columna 14 y el nodo siguiente es el nodo 12 que es perteneciente a la columna 11. También es visible que si una de las palabras contenidas en el nodo 1 se encuentran en la posición 1 y una de las palabras contenidas en el nodo 5 se encuentran en la posición 14 y una de las palabras contenidas en el nodo 12 se encuentran en la posición 11 del tema de un proyecto de titulación sin tomar en cuenta los conectores, el proyecto tiene un 100% de probabilidad de culminar con éxito. El siguiente nodo perteneciente al tercer nivel que se evaluó es el nodo 11, el cual es presentado en la Figura 70.

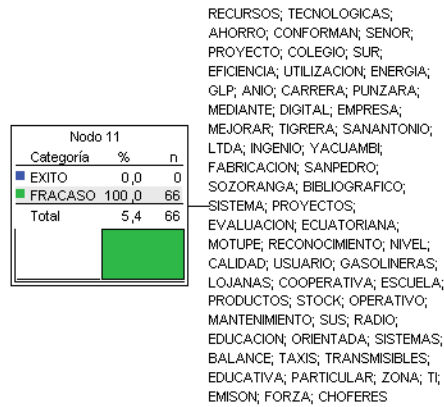


Figura 70. Nodo 11 del árbol de CHAID Exhaustivo.

Este nodo es precedido por los mismos nodos que es precedido el nodo 12, por lo que se determinó que, si una de las palabras contenidas en el nodo 1 y en el nodo 5 se encuentran en la posición 21 y 14 respectivamente, y una de las palabras contenidas en el nodo 11 se encuentra en la posición 11 del tema de un proyecto de titulación, el proyecto tiene un 100% de probabilidad de fracasar. El próximo nodo del nivel 3 que es evaluado es el nodo 14.

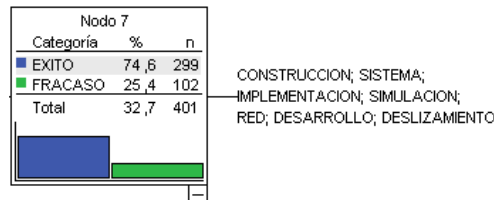


Figura 71. Nodo 7 del árbol de CHAID Exhaustivo.

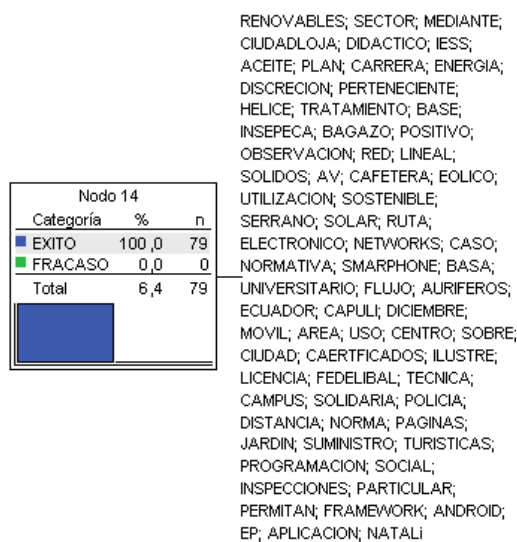


Figura 72. Nodo 14 del árbol de CHAID Exhaustivo.

La Figura 65 presenta el nodo 2 que es el primer nodo para llegar al nodo 14 y pertenece a la columna 1, el nodo 7 es mostrado en la Figura 71, es el segundo nodo para llegar al nodo 14 y pertenece a la columna 2, y por último el nodo 14 pertenece a la columna 13 y es mostrado en la Figura 72. Al observar los nodos se determinó que si una de las palabras contenidas en el nodo 2 se encuentra en la posición 1 del tema de un proyecto y una de las palabras del nodo 7 se encuentra en la posición 2 y una de las palabras del nodo 14 está en la posición 13 de un proyecto de titulación, entonces el proyecto tiene 100% de probabilidad de culminar con éxito. El siguiente nodo del nivel 3 a ser evaluado es el nodo 15.

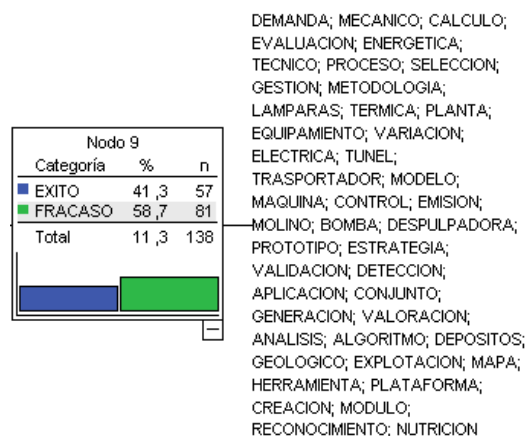


Figura 73. Nodo 9 del árbol de CHAID Exhaustivo.

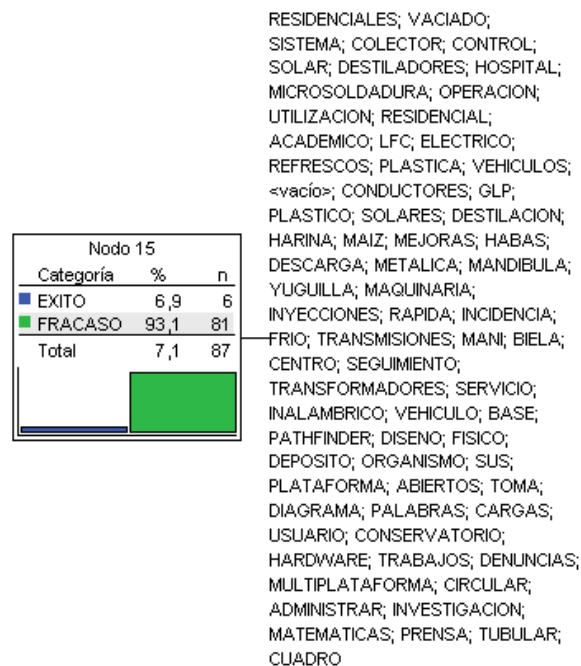


Figura 74. Nodo 15 del árbol de CHAID Exhaustivo.

El nodo 15 pertenece a la columna 5 y se observa en la Figura 74, este nodo es precedido por el nodo 9 que pertenece a la columna 2 y es presentado en la Figura 73, y este nodo es precedido por el nodo 2 que pertenece a la columna 1 y es mostrado en la Figura 65. Se determinó que si una de las palabras del nodo 2 se encuentra en la posición 1 y una de las palabras del nodo 9 se encuentra en la posición 2 y una de las palabras pertenecientes al nodo 15 se encuentra en la posición 5, entonces el proyecto de titulación tiene un 93.1% de probabilidad de fracasar. El siguiente nodo evaluado es el nodo 16 que es precedido por los nodos 2 y 9.

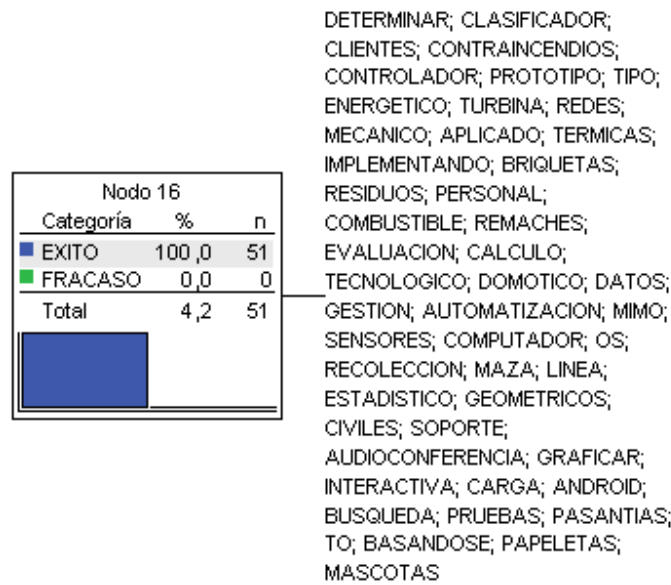


Figura 75. Nodo 16 del árbol de CHAID Exhaustivo.

La Figura 75 presenta las palabras que contiene el nodo 16, este nodo al ser precedido por los nodos 2 y 9, influye en el tema sin tomar en cuenta conectores de la siguiente manera: si una de las palabras del nodo 2 se encuentra en la posición 1 del tema y una de las palabras del nodo 9 se encuentra en la posición 2 del tema y una de las palabras del nodo 16 está en la posición 5 del tema de un proyecto, entonces el proyecto tiene un 100% de probabilidad de culminar con éxito.

3.1.6. Random Forest.

Con este algoritmo se obtuvo el árbol de la Figura 76, en la cual se observa que la columna más influyente en los temas es la 22 y las palabras que se dirigen a los estados de éxito o fracaso son las palabras que más influyen en esa columna.

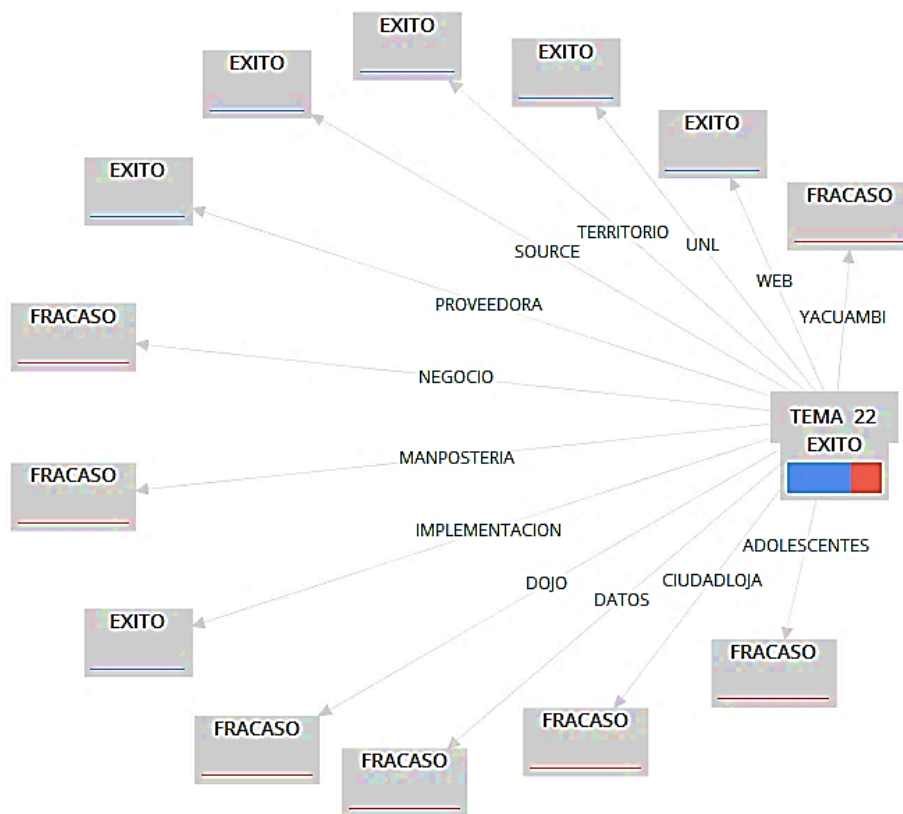


Figura 76. Árbol obtenido con el algoritmo Random Forest.

El árbol indica que el proyecto culmina con éxito si en la posición 22 del tema del proyecto se encuentra una de las siguientes palabras: WEB, UNL, TERRITORIO, SOURCE, PROVEEDORA e IMPLEMENTACION. También indica que el proyecto fracasa si en la posición 22 del tema se encuentra alguna de las siguientes palabras: YACUAMBI, NEGOCIO, MANPOSTERIA, DOJO, DATOS, CIUDADLOJA Y ADOLESCENTES.

3.2. Evaluación de los algoritmos aplicados.

Para la evaluación de los algoritmos se consideró relevante que una forma óptima de evaluar algoritmos no supervisados es mediante la interpretación humana y que de los algoritmos aplicados el único que proporcionó un porcentaje de eficiencia es el algoritmo CHAID Exhaustivo con 80.9%, por tal motivo el resto de algoritmos son evaluados respecto al algoritmo CHAID y en base a los conocimientos generados en la evaluación de resultados. Antes de proceder a la evaluación de los algoritmos, se evaluó la nube de palabras y se obtuvo lo siguiente:

- La **nube de palabras** presenta que la palabra “web” influye en el éxito de los proyectos. La palabra **web** está presente en la posición 10 con 98.8% de

probabilidad de éxito. El resultado obtenido con esta representación gráfica de los términos más usados en el conjunto de datos, coincide con los resultados del algoritmo CHAID Exhaustivo y se considera relevante el resultado.

Una vez evaluadas las nubes de palabras se realizó la evaluación de los algoritmos de clasificación, tal como se indica a continuación:

- Con el **dendograma** se obtuvo que la relación “**móvil-aplicación**” influye en el fracaso del proyecto. La relación está presente en el árbol, cuando esta la palabra **aplicación** en la posición 1 se tiene un 62.1% de probabilidad que el proyecto culmine con éxito y al estar la palabra **móvil** en la posición 14 la probabilidad de éxito aumenta a 98.8%. Se encontró la palabra **aplicación** sin estar relacionada con la palabra **móvil** en la posición 2 con 58.7% de probabilidad de fracaso y en posición 13 con 100% de posibilidad de culminar el proyecto con éxito. También se halló la palabra **móvil** sin estar relacionada con la palabra **aplicación** en la posición 14 con 100% de probabilidad de éxito. Los resultados de la palabra **aplicación** al estar sola se contradicen y de la palabra **móvil** conllevan al éxito, pero la relación “**móvil-aplicación**” que es la relación obtenida en el dendograma influye en el fracaso y no coincide con el algoritmo CHAID Exhaustivo donde la relación influye en el éxito. Tomando en cuenta lo anterior se determina que los resultados obtenidos con el dendograma son ineficientes al no coincidir los resultados.
- Las reglas de asociación **Apriori** dan como resultado la relación “**gestión-sistema**” que influye en el éxito del proyecto. Las palabras de la relación están presentes en el árbol, pero no se encuentran relacionadas entre sí. Tenemos que la palabra **sistema** está presente en las posiciones: 1 con 62.1% y 2 con 74.6% de porcentaje de éxito y en las posiciones: 11 con 100% y 5 con 93.1% de probabilidad de fracaso, y también la palabra **gestión** se encuentra en la posición 5 con 100% de probabilidad de culminar el proyecto con éxito. Ambas palabras no tienen una relación en el árbol y tienen contradicción en los porcentajes debido a que la palabra **sistema** consta de probabilidades de éxito como también de fracaso por lo que pierden relevancia los resultados la no coincidir.
- El algoritmo **K-means** presenta las siguientes relaciones de éxito como resultados: “**DISEÑO CONSTRUCCIÓN PROTOTIPO**” y “**LOJA TERRITORIO ESTUDIO**”. También se obtuvo una relación de fracaso que es la siguiente: “**APLICACION DESARROLLO MOVIL**”.

A la primera relación no se la encontró presente en el árbol, es decir se hallaron todas las palabras de la relación en el árbol, pero no relacionadas entre sí. Teniendo así que la palabra **construcción** está presente en las posiciones: 1 con 62.1% y 2 con 74.6% de probabilidades de éxito, la palabra **diseño** se encuentra en las posiciones: 1 con 70.8% y 2 con 100% de probabilidades de éxito y en la posición 2 con 93.1% de probabilidad de fracaso, y por último la palabra **prototipo** está en la posición 1 con 62.1% de porcentaje de éxito y en la posición 2 con 58.7% de probabilidades de fracaso. En las palabras **prototipo** y **diseño** se contradicen los porcentajes y al no existir una relación en el árbol entre las tres palabras de la relación planteada, se descarta los resultados al perder relevancia.

La segunda relación fue encontrada en árbol, donde la palabra **estudio** se encuentra en el nodo 1 con 62.1% de probabilidad de éxito, la palabra **territorio** está en la posición 14 con 53.3% de probabilidad de culminar con éxito y la palabra **Loja** se encuentra en la posición 11 con 58.3% de probabilidad de éxito. También se encontró las palabras sin estar relacionadas y tenemos que la palabra **Loja** está en las posiciones: 14 con 98.8% y 13 con 68.3% de probabilidad de culminar el proyecto con éxito, por otra parte, la palabra **territorio** se la encontró presente en las posiciones: 11 con 58.3% y 13 con 68.3% de probabilidades de éxito. Con los resultados anteriores se establece que las palabras pertenecientes a la relación influyen en el éxito del proyecto al coincidir los resultados.

La relación de fracaso no fue encontrada en el árbol, pero si se encontró las palabras **aplicación** y **móvil** relacionadas, la palabra **aplicación** está en la posición 1 con 62.1% de probabilidad de éxito y la palabra **móvil** se encuentra en la posición 14 con 98.8% de probabilidad de culminar el proyecto con éxito. Otras palabras relacionadas que se encontraron son **desarrollo** y **aplicación**, la palabra **desarrollo** se encuentra en la posición 1 con 70.8% de éxito y la palabra **aplicación** en la posición 2 con 58.7% de probabilidad de fracasar. También se halló la relación entre **desarrollo** y **aplicación** en otras posiciones, la palabra **desarrollo** en la posición 2 con 74.6% de éxito y la palabra **aplicación** en la posición 13 con 100% de probabilidad de culminar el proyecto con éxito. Al no encontrar la relación de fracaso en el árbol y existir contradicciones en las relaciones halladas entre las palabras pertenecientes a la relación de fracaso, se

descartan los resultados por perder relevancia al no coincidir con los resultados del algoritmo CHAID Exhaustivo.

- **K-medoids** obtuvo como resultados la palabra "LOJA". La palabra **Loja** ya se evaluó en el algoritmo k-means, donde se determinó que si influye en el éxito de los proyectos.
- Con el algoritmo **Random Forest** se obtuvo que las palabras que influyen en el éxito son: WEB, UNL, TERRITORIO, SOURCE, PROVEEDORA e IMPLEMENTACION. También se obtuvo a las palabras que influyen en el fracaso que son: YACUAMBI, NEGOCIO, MANPOSTERIA, DOJO, DATOS, CIUDADLOJA Y ADOLESCENTES. De las palabras de éxito no se encontró en el árbol a las palabras **source** y **proveedora**. El resto de palabras son evaluadas a continuación:

WEB, ya se evaluó anteriormente con los resultados de la nube de palabras en donde se determinó que si influye en el éxito de los proyectos.

UNL, se encuentra en las posiciones: 14 con 53.3%, 11 con 58.3% y 13 con 68.3% de probabilidades de éxito.

TERRITORIO, ya fue evaluada esta palabra en el algoritmo K-means en donde se determinó que si influye en el éxito.

IMPLEMENTACION, está en las posiciones: 1 con 62.1% y 2 con 74.6% de probabilidades de culminar el proyecto con éxito.

Los resultados de las palabras que influyen en el éxito coinciden, por lo que el algoritmo Random Forest es eficiente para determinar la influencia de las palabras en el éxito de los proyectos.

De las palabras que influyen en el fracaso las palabras **negocio**, **mampostería**, **dojo** y **adolescentes** no están presentes en el árbol del algoritmo CHAID Exhaustivo, por lo que las palabras restantes son evaluadas a continuación:

YACUAMBI, está en la posición con 100% de probabilidad de fracaso.

CIUDADLOJA, se encuentra en las posiciones: 11 con 58.3%, 14 con 53.3% y 13 con 100% de probabilidades de éxito.

DATOS, se encuentra en las posiciones: 11 con 100% y 5 con 100% de probabilidad de culminar el proyecto con éxito.

Debido a que solo 3 de las 7 palabras de fracaso fueron encontradas en el árbol y que dos de ellas se contradicen con los resultados del algoritmo Random Forest, en donde se obtuvo que **DATOS** y **CIUDADLOJA** influyen en el fracaso y que en el algoritmo CHAID Exhaustivo influye en el éxito, se determina que el algoritmo

Random Forest es ineficiente para identificar las palabras que influyen en el fracaso del proyecto.

3.3. Introducción de nuevos datos.

Para la introducción de nuevos datos se empleó un nuevo conjunto de datos, el cual fue proporcionado nuevamente por la secretaria de la Carrera de Ingeniería en Sistemas, **la solicitud para la obtención del conjunto de datos se puede ver en el Anexo 5**. En este conjunto de datos se encontraron los proyectos anteriormente analizados y nuevos proyectos pertenecientes a la Carrera de Ingeniería en Sistemas, además de presentar nuevas observaciones en algunos de los anteriores proyectos pertenecientes a la carrera. Debido a que este nuevo conjunto de datos cuenta con datos actualizados de solo la Carrera de Ingeniería en Sistemas, se seleccionó los proyectos pertenecientes a esa carrera para la introducción de nuevos datos.

La evaluación y categorización de los nuevos proyectos se la realizó de acuerdo a la entrevista realizada a la secretaria de la carrera y siguiendo el mismo procedimiento de los puntos 1 y 2 de esta sección, en el **Anexo 6 se visualiza la entrevista realizada a la secretaria de la carrera y en el Anexo 9 se observa la certificación del nuevo conjunto de datos obtenido**.

Una vez categorizados los datos se procedió a su limpieza usando nuevamente Openrefine y posterior a ello se aplicó los algoritmos de clasificación, los parámetros e indicaciones para limpiar el conjunto de datos y aplicar los algoritmos de clasificación son los usados anteriormente para realizar estas mismas actividades en el anterior conjunto de datos.

El nombre del nuevo conjunto de datos después de haber realizado la evaluación, categorización y depuración es "0_Nuevo_Conjunto_de_Datos.csv", donde los algoritmos de clasificación no supervisados nuevamente serán aplicados en todo el conjunto de datos, en los proyectos con estado de éxito y en los proyectos con estado de fracaso. Las representaciones gráficas de los resultados serán realizadas usando el código de la TABLA XXX al igual que en el punto 7 de esta sección.

Las palabras más utilizadas en el conjunto de datos se representaron mediante **Nubes de Palabras** como se indica a continuación:



Figura 79. Nueva Nube de Palabras de los proyectos de fracaso.

La Figura 79 muestra que las palabras más frecuentes en los proyectos con estado de fracaso son: “sistema”, “desarrollo”, “implementacion”, “unl”, “diseno”, “web”, “control”, “ciudadloja”, “aplicacion” y “gestion”.

A continuación, la TABLA XXXVIII presenta las palabras más usadas en todo el conjunto de datos y en los proyectos con estado de éxito y fracaso.

TABLA XXXVIII

COMPARATIVA DE NUEVOS RESULTADOS DE NUBES DE PALABRAS

Todo el conjunto de datos	Proyectos con estado de éxito	Proyectos con estado de fracaso
sistema	sistema	sistema
desarrollo	web	desarrollo
unl	unl	implementacion
web	desarrollo	unl
diseno	implementacion	diseno
ciudadloja	diseno	web
aplicacion	gestion	aplicacion
gestion	software	ciudadloja
loja	loja	control
implementacion	ciudadloja	gestion

Las palabras que coinciden con uno de los estados de éxito o fracaso y en todo el conjunto de datos son las palabras más influyentes o relevantes en todo el conjunto de datos. De esas palabras la que influye en el éxito del proyecto es: “loja” y la palabra que influye en el fracaso del proyecto es: “aplicacion”.

La Figura 80 presenta la representación gráfica de la palabra “loja”.

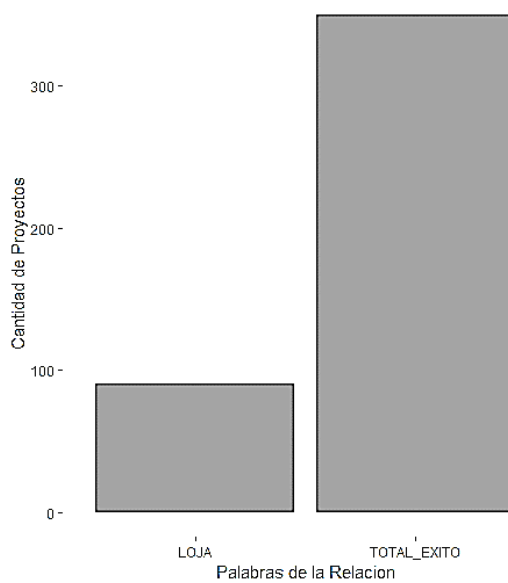


Figura 80. Nueva representación gráfica de la palabra Loja.

En la Figura 81 se observa la representación gráfica de la palabra “aplicacion”.

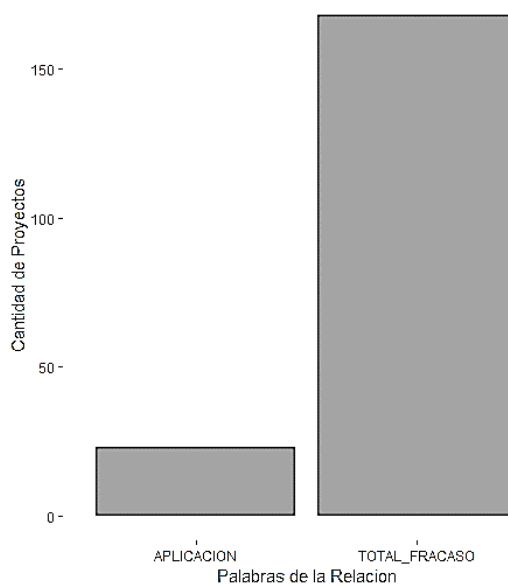


Figura 81. Nueva representación gráfica de la palabra aplicación.

Al igual que en los puntos anteriores de esta sección, una vez obtenido las representaciones de las palabras más influyentes se procede a la aplicación de los algoritmos de clasificación.

3.3.1. Aplicación de los algoritmos de Clasificación.

- **Dendograma.**

Con el Dendograma se obtuvo los siguientes resultados:

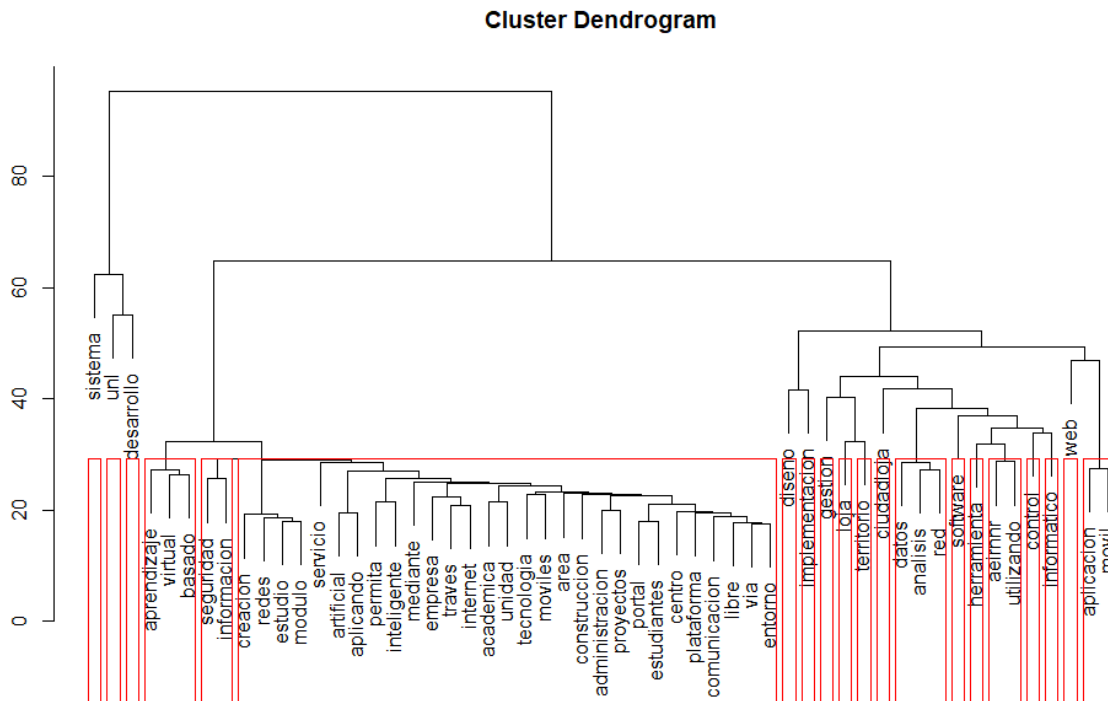


Figura 82. Dendograma del nuevo conjunto de datos.

En la Figura 82 se observa las siguientes relaciones todo el conjunto de datos: “virtual aprendizaje basado”, “seguridad informacion”, “datos análisis red”, “aieirnr utilizando” y “aplicacion movil”.

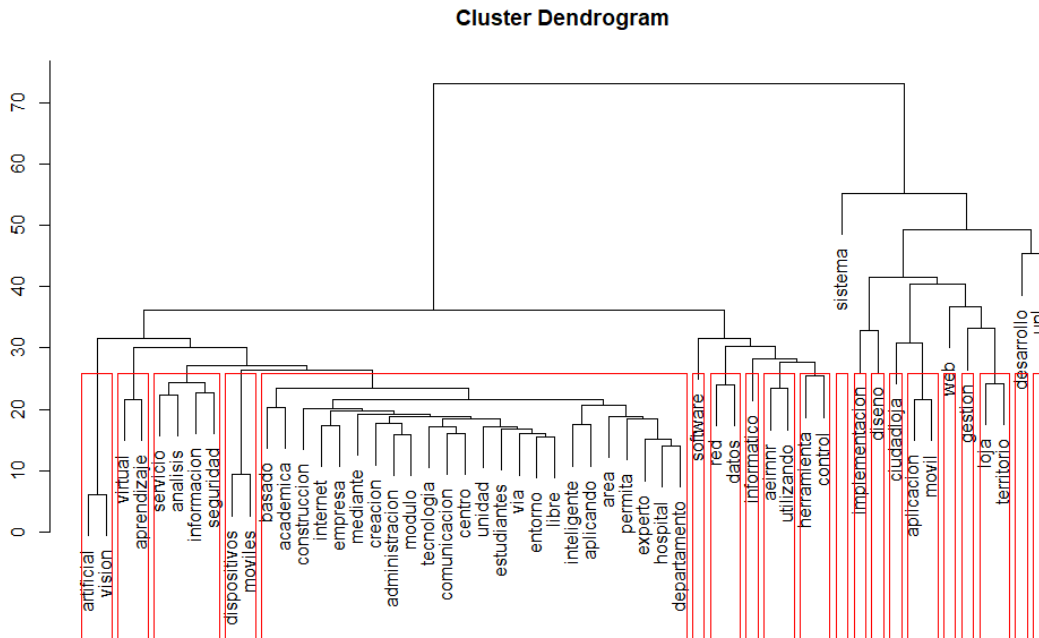


Figura 83. Dendrograma de nuevos proyectos de éxito.

Se observa en la Figura 83 que las relaciones de éxito obtenidas son: “artificial vision”, “virtual aprendizaje”, “dispositivos moviles”, “aplicación movil”, “red datos”, “servicio analisis informacion seguridad”, “aeirnr utilizando”, “herramienta control” y “loja-territorio”.

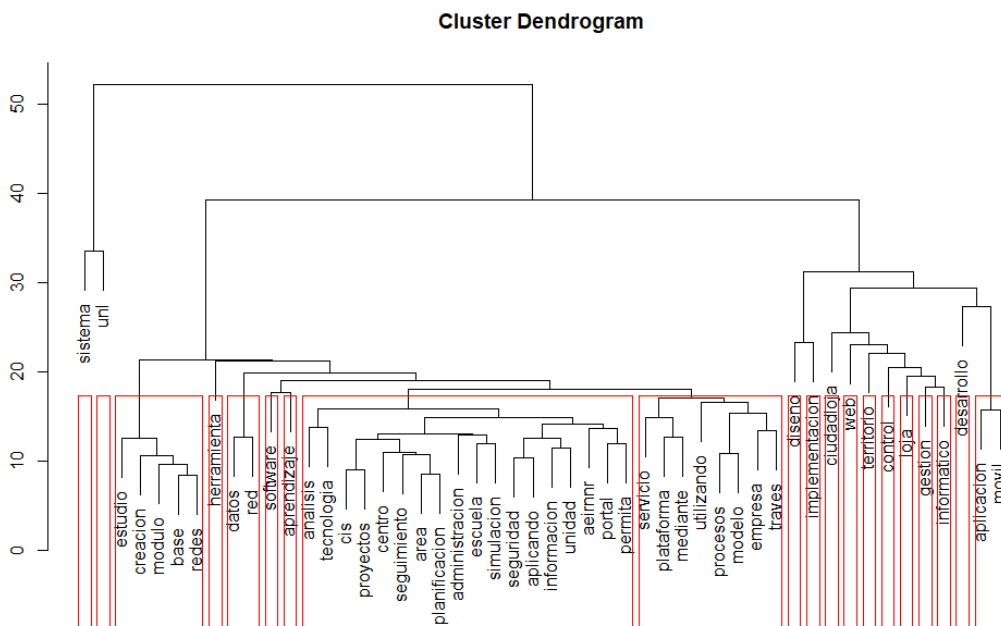


Figura 84. Dendrograma de nuevos proyectos de fracaso.

Las relaciones de fracaso se observan en la Figura 84 y son: “estudio creacion modulo base redes”, “datos red” y “aplicación movil”.

TABLA XXXIX

COMPARATIVA DE NUEVOS RESULTADOS DEL DENDOGRAMA

Todo el conjunto de datos	Proyectos con estado de éxito	Proyectos con estado de fracaso
virtual aprendizaje basado	artificial vision	estudio creacion modulo base redes
seguridad informacion	virtual aprendizaje	datos red
datos analisis red	dispositivos moviles	aplicacion movil
aeirnr utilizando	aplicacion movil	
aplicacion movil	red datos	
	servicio analisis informacion seguridad	
	aeirnr utilizando	
	herramienta control	
	loja territorio	

En la TABLA XXXIX se observa que de las relaciones obtenidas en los dendogramas una de ellas que es contenida en las relaciones de éxito coincide con una de las relaciones de todo el conjunto de datos, demostrando que esta relación tiene relevancia en el conjunto de datos, la relación es “aeirnr utilizando”. En la Figura 85 se observa la representación gráfica de la relación “aeirnr utilizando”.

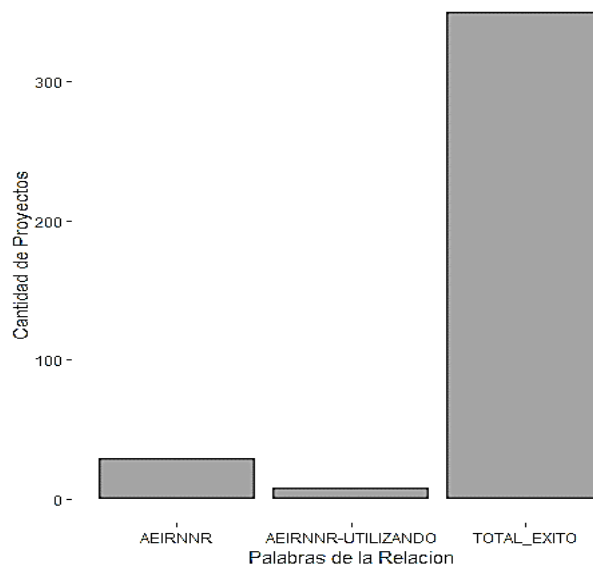


Figura 85. Nueva representación gráfica de la relación AEIRNNR-utilizando.

- **Reglas de Asociación Apriori.**

Con las **Reglas de Asociación** se obtuvieron los siguientes resultados:

	lhs	rhs	support	confidence	lift	count
[1]	{MOVIL}	=> {APLICACION}	0.05212355	0.7500000	6.070312	27
[2]	{GESTION}	=> {SISTEMA}	0.11196911	0.7532468	1.849203	58
[3]	{GESTION,WEB}	=> {SISTEMA}	0.04054054	0.7500000	1.841232	21

Figura 86. Reglas de Asociación del nuevo conjunto de datos.

En la Figura 86 se muestra las 3 reglas que se obtuvieron en todo el conjunto de datos y son: la regla 1 “MOVIL-APLICACION”, la regla 2 “GESTION-SISTEMA” y la regla 3 “GESTION-SISTEMA o WEB-SISTEMA”, en esta última regla se observa que una de las opciones coincide con la regla 2 por lo que la regla 3 queda como “WEB-SISTEMA”. La representación gráfica de todas las reglas es:

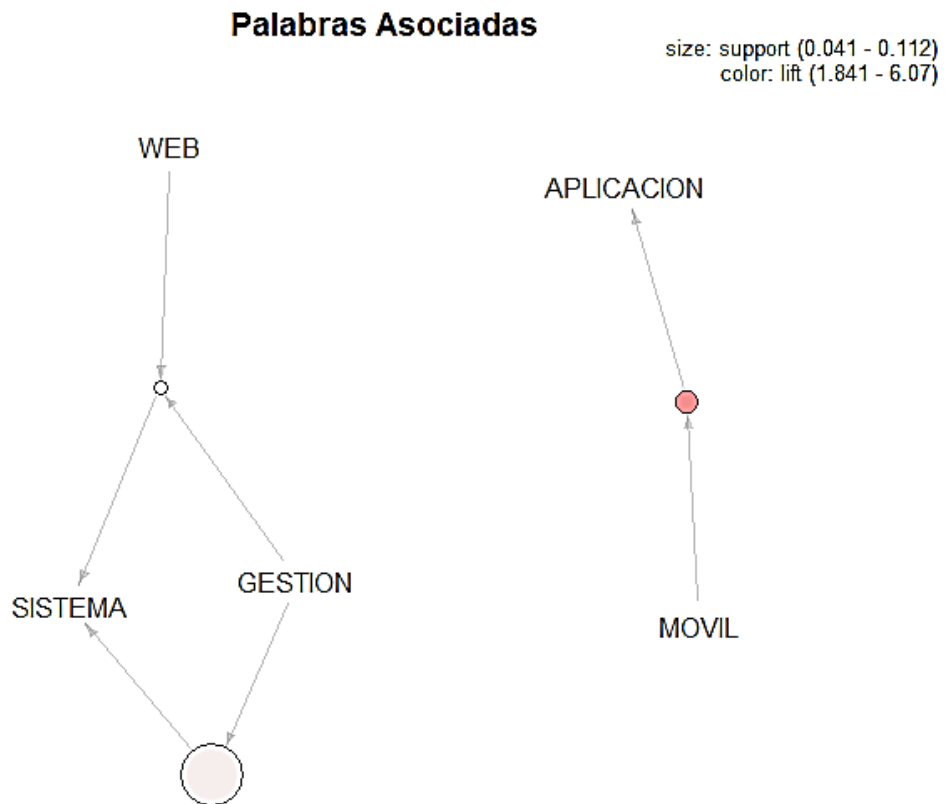


Figura 87. Reglas relacionadas del nuevo conjunto de datos.

La Figura 87 presenta las reglas de asociación de todo el conjunto de datos, en donde se indica que la palabra “APLICACION” está presente cuando la palabra “MOVIL” se encuentra en el tema. Otras relaciones que se observan son: cuando está presente en

el tema la palabra “GESTION” también se encuentra en el tema la palabra “SISTEMA”, y cuando están en el tema las palabras “GESTION” y “WEB” también se encuentra en el tema la palabra “SISTEMA”.

	lhs	rhs	support	confidence	lift	count
[1]	{LIBRE}	=> {SOFTWARE}	0.04285714	0.9375000	7.457386	15
[2]	{GESTION}	=> {SISTEMA}	0.12285714	0.7962963	1.922095	43
[3]	{GESTION,WEB}	=> {SISTEMA}	0.04857143	0.7083333	1.709770	17

Figura 88. Reglas de Asociación de los nuevos proyectos de éxito.

Se observa en la Figura 88 las 3 reglas de éxito obtenidas y son: la regla 1 “LIBRE-SOFTWARE”, la regla 2 “GESTION-SISTEMA” y la regla 3 “GESTION-SISTEMA” o “WEB-SISTEMA”. En la regla 3 hay dos relaciones, pero una de ellas es la relación de la regla 2 por lo que la regla 3 queda como “WEB-SISTEMA”. La representación gráfica de las reglas obtenidas es:

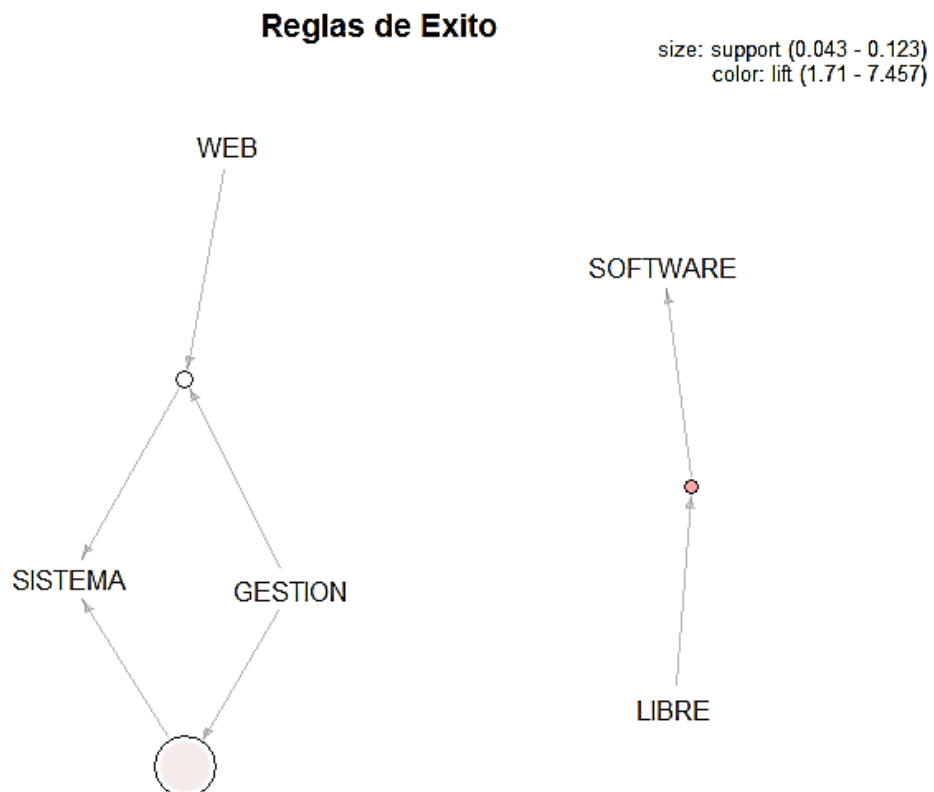


Figura 89. Reglas relacionadas de los nuevos proyectos de éxito.

En la Figura 89 se observa que: la palabra “SOFTWARE” se encuentra siempre que se encuentre la palabra “LIBRE” en el tema, la palabra “SISTEMA” está presente cuando

se encuentra la palabra “GESTION”, y también se encuentra en el tema cuando están las palabras “WEB” y “GESTION”.

	lhs	rhs	support	confidence	lift	count
[1]	{MOVIL}	=> {APLICACION}	0.06547619	0.8461538	6.769231	11
[2]	{PLATAFORMA}	=> {DESARROLLO}	0.04761905	0.8000000	3.054545	8
[3]	{INFORMATICO}	=> {SISTEMA}	0.06547619	0.7333333	1.866667	11

Figura 90. Reglas de Asociación de los nuevos proyectos de fracaso.

La Figura 90 muestra las 3 reglas de fracaso obtenidas y son: la regla 1 “MOVIL- APLICACION”, la regla 2 “PLATAFORMA-DESARROLLO” y la regla 2 “INFORMATICO- SISTEMA”. La representación gráfica de las reglas de fracaso es:

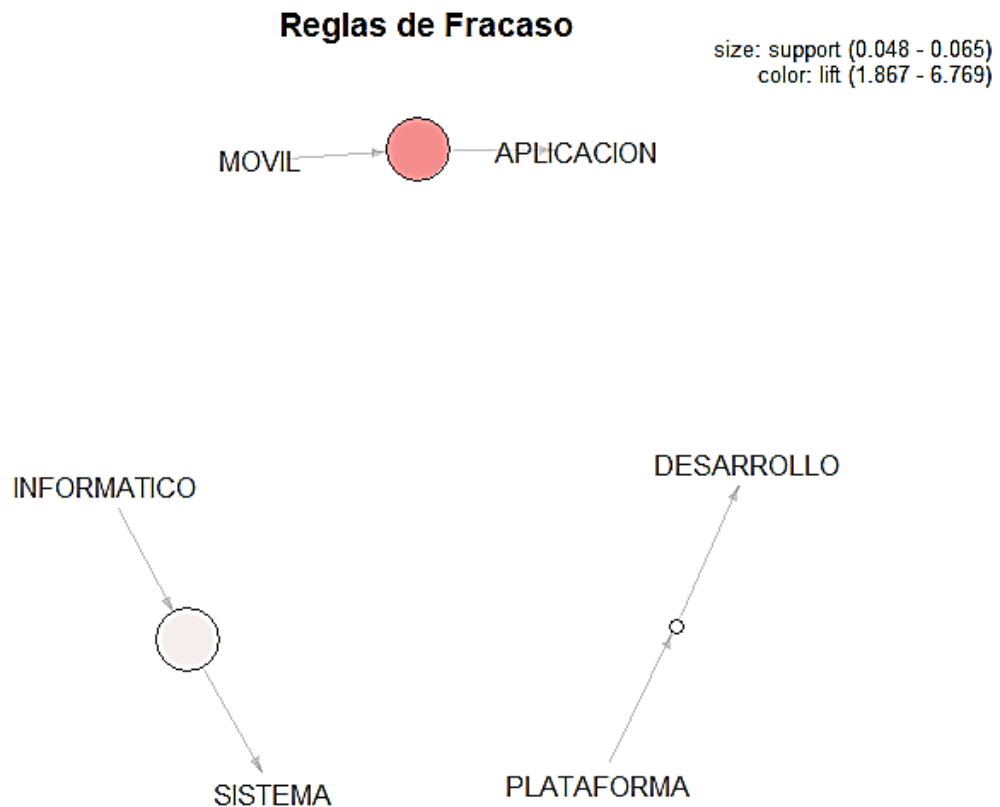


Figura 91. Reglas relacionadas de los nuevos proyectos de fracaso.

Se observa en la Figura 91 que la palabra “APLICACION” se encuentra en el tema cuando la palabra “MOVIL” también está presente en el tema. Otras relaciones que se observan son: cuando está presente en el tema la palabra “PLATAFORMA” se encuentra también la palabra “DESARROLLO”, y la palabra “SISTEMA” está presente en tema cuando también se encuentra la palabra “INFORMATICO”.

TABLA XL

COMPARATIVA DE NUEVOS RESULTADOS DE APRIORI

Todo el conjunto de datos	Proyectos con estado de éxito	Proyectos con estado de fracaso
MOVIL-APLICACION	LIBRE-SOFTWARE	MOVIL-APLICACION
GESTION-SISTEMA	GESTION-SISTEMA	PLATAFORMA-DESARROLLO
WEB-SISTEMA	WEB-SISTEMA	INFORMATICO-SISTEMA

En la TABLA XL se visualiza que de las relaciones de éxito dos de ellas coinciden con dos relaciones de todo el conjunto de datos, determinando así que las relaciones relevantes para el éxito son “GESTION-SISTEMA” y “WEB-SISTEMA”. Otra relación que coincide con una de todo el conjunto de datos es la relación de fracaso “MOVIL-APLICACION”, estableciendo que esta relación si es relevante para determinar el fracaso. La representación gráfica de la relación de éxito “GESTION-SISTEMA” se observa en la Figura 92.

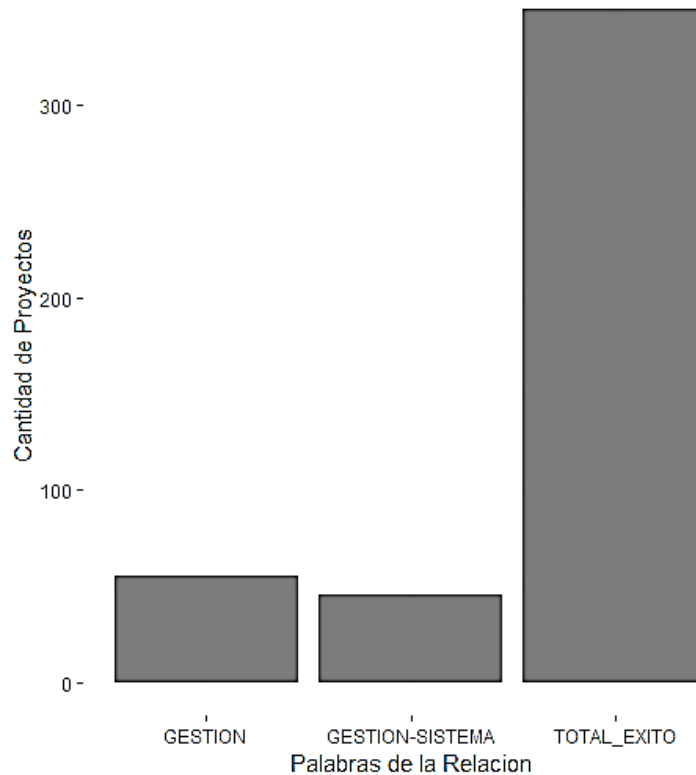


Figura 92. Nueva representación gráfica de la relación gestión-sistema.

En la Figura 93 se observa la representación gráfica de la relación de éxito “WEB-SISTEMA”.

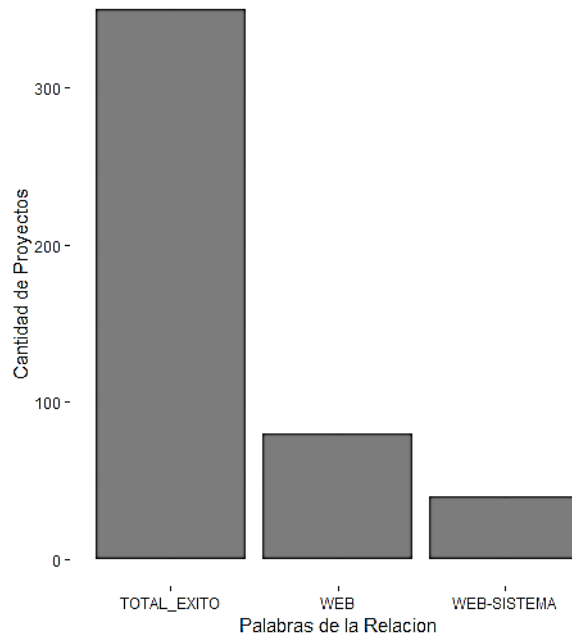


Figura 93. Nueva representación gráfica de la relación web-sistema.

La Figura 94 presenta la representación gráfica de la relación de fracaso “MOVIL-APLICACION”.

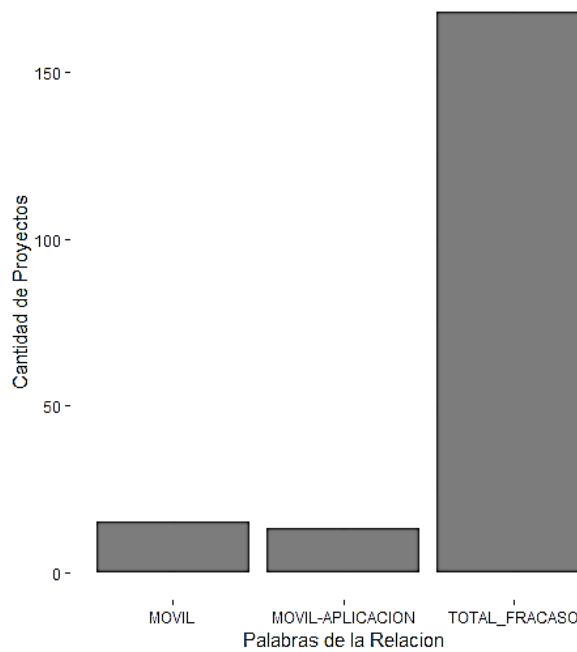


Figura 94. Nueva representación gráfica de la relación móvil-aplicación.

- **K-means.**

Con K-means se obtuvieron los siguientes resultados:

En todo el conjunto de datos:

```
kmeans_todo 1: sistema gestion web
kmeans_todo 2: informacion unl seguridad
kmeans_todo 3: empresa web sistema
kmeans_todo 4: aplicacion web movil
kmeans_todo 5: implementacion sistema desarrollo
kmeans_todo 6: territorio loja sistema
kmeans_todo 7: aprendizaje virtual entorno
kmeans_todo 8: unl implementacion diseno
kmeans_todo 9: desarrollo unl software
```

Figura 95. Resultados de k-means en el nuevo conjunto de datos.

En los proyectos con estado de éxito:

```
kmeans_exito 1: unl implementacion aeirnr
kmeans_exito 2: aprendizaje virtual entorno
kmeans_exito 3: implementacion desarrollo sistema
kmeans_exito 4: sistema gestion web
kmeans_exito 5: software desarrollo unl
kmeans_exito 6: aplicacion desarrollo diseno
kmeans_exito 7: unl sistema web
kmeans_exito 8: territorio loja sistema
kmeans_exito 9: software libre utilizando
```

Figura 96. Resultados de k-means en los nuevos proyectos de éxito.

En los proyectos con estado de fracaso:

```
kmeans_fracaso 1: aplicacion redes ciudadloja
kmeans_fracaso 2: web sistema control
kmeans_fracaso 3: desarrollo aplicacion control
kmeans_fracaso 4: unl herramienta implementacion
kmeans_fracaso 5: territorio sistema implementacion
kmeans_fracaso 6: sistema desarrollo web
kmeans_fracaso 7: sistema unl gestion
kmeans_fracaso 8: diseno implementacion sistema
kmeans_fracaso 9: loja portal web
```

Figura 97. Resultados de k-means en los nuevos proyectos de fracaso.

Como se observa en la Figura 95, Figura 96 y Figura 97 las relaciones al aplicar k-means a todo el conjunto de datos se denomina kmeans_todo, al aplicar en los proyectos con estado de éxito se denomina kmeans_exito y al aplicar en los proyectos con estado de fracaso se denomina kmeans_fracaso. Todas las relaciones obtenidas al aplicar el algoritmo son presentadas en la TABLA XLI.

TABLA XLI

COMPARATIVA DE NUEVOS RESULTADOS DE K-MEANS

No.	Kmeans_todo	Kmeans_exito	Kmeans_fracaso
1	sistema gestion web	unl implementacion aeirnr	aplicacion redes ciudadloja
2	informacion unl seguridad	aprendizaje virtual entorno	web sistema control
3	empresa web sistema	implementacion desarrollo sistema	desarrollo aplicacion control
4	aplicacion movil web	sistema gestion web	unl herramienta implementacion
5	implementacion desarrollo sistema	software desarrollo unl	territorio sistema implementacion
6	territorio loja sistema	aplicacion desarrollo diseno	sistema web desarrollo
7	aprendizaje virtual entorno	sistema unl web	sistema unl gestion
8	diseno implementacion unl	territorio loja sistema	diseno implementacion sistema
9	desarrollo unl software	software libre utilizando	loja web portal

La TABLA XLI presenta que las relaciones que coinciden son kmeans_todo 1 y kmeans_exito 4, determinando que la relación “sistema-gestion-web” influye en el éxito del proyecto. También se observa que las relaciones kmeans_todo 5 y kmeans_exito 3 coinciden, las relaciones kmeans_todo 6 y kmeans_exito 8 coinciden, las relaciones

kmeans_todo 7 y kmeans_exito 2 coinciden, y las relaciones kmeans_todo 9 y kmeans_exito 5 coinciden, determinando así que las relaciones “territorio-loja-sistema”, “implementacion-desarrollo-sistema”, “sistema-gestion-web”, “software-desarrollo-unl” y “aprendizaje-virtual-entorno” influyen en el éxito del proyecto.

La representación gráfica de la relación de éxito “territorio-loja-sistema” se observa en la Figura 98.

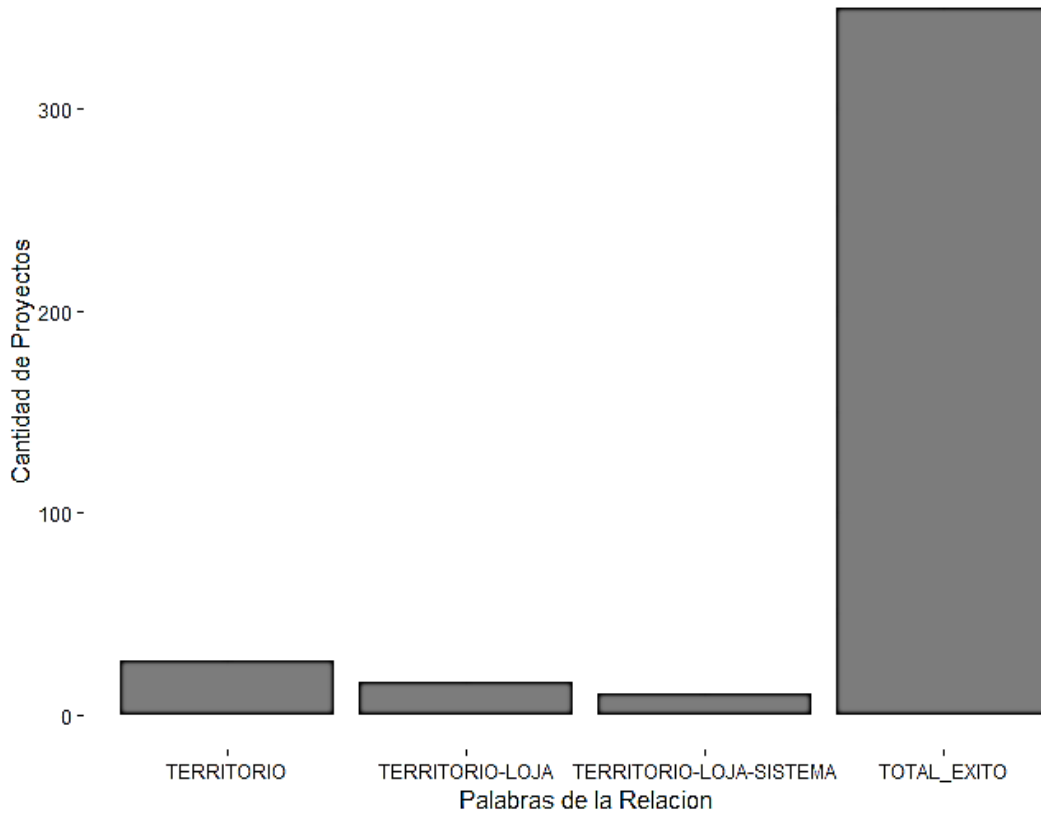


Figura 98. Nueva representación gráfica de la relación territorio-Loja-sistema.

La Figura 99 presenta la representación gráfica de la relación de éxito “aprendizaje-virtual-entorno”.

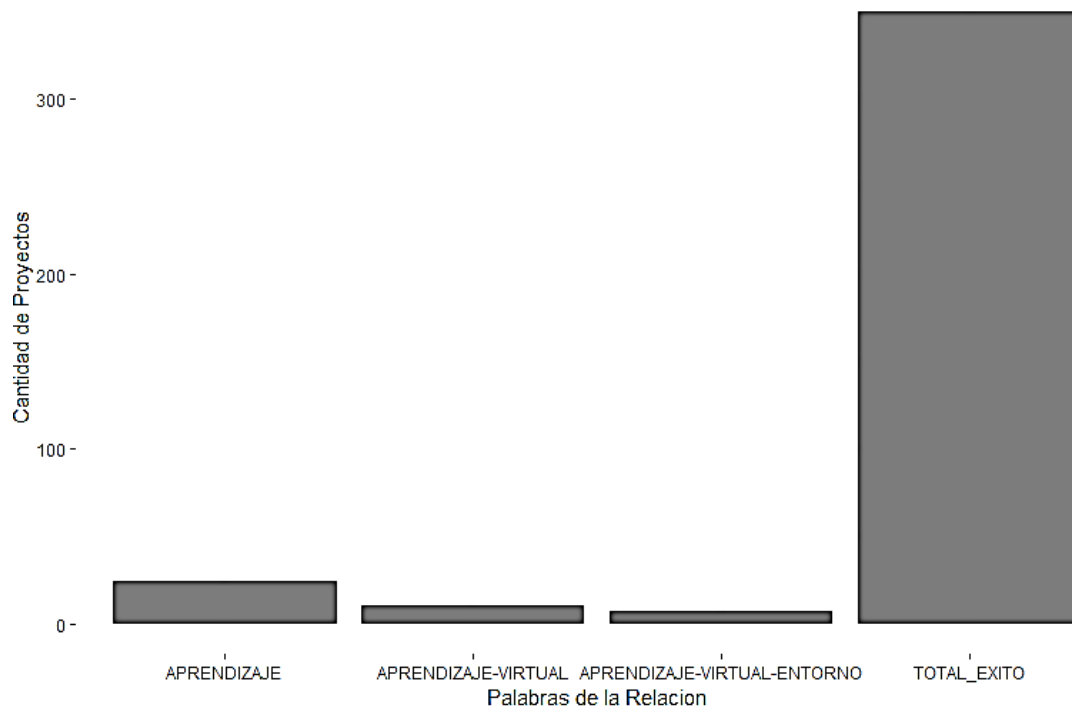


Figura 99. Nueva representación gráfica de la relación aprendizaje-virtual-entorno.
 En la Figura 100 se observa la representación gráfica de la relación de éxito “implementacion-desarrollo-sistema”.

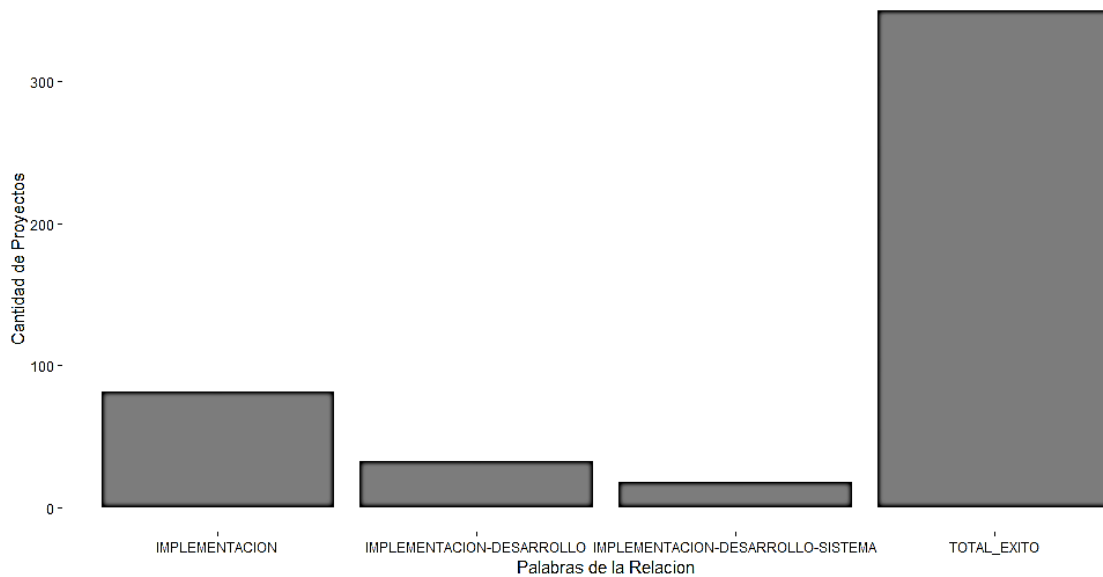


Figura 100. Nueva representación gráfica de la relación implementación-desarrollo-sistema.

La representación gráfica de la relación de éxito “sistema-gestion-web” se visualiza en la Figura 101.

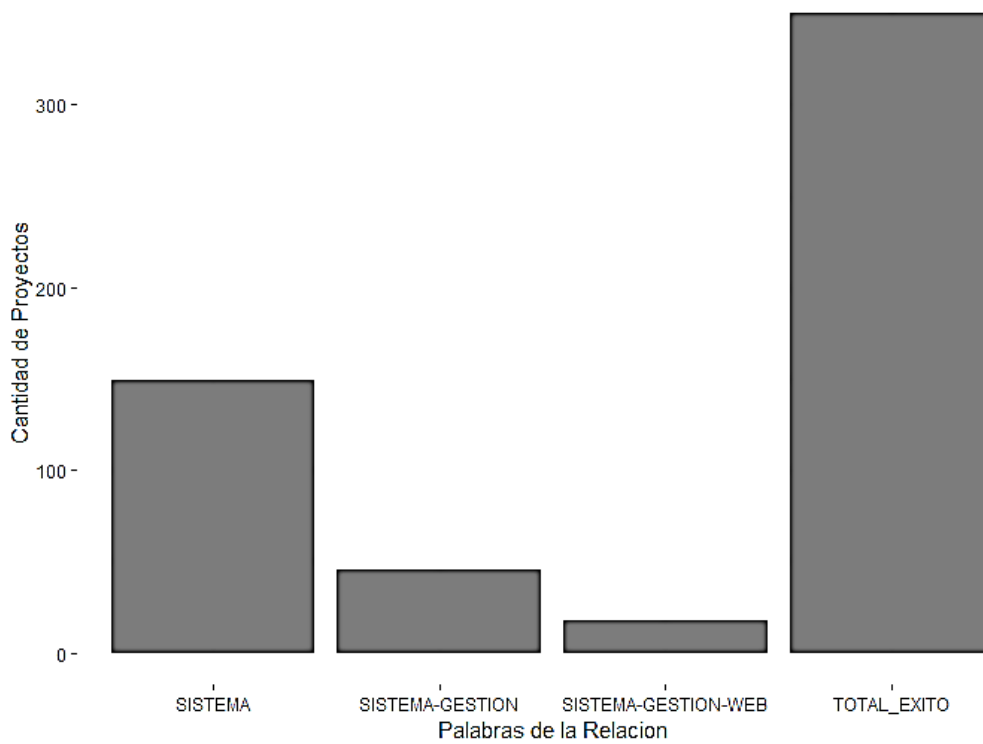


Figura 101. Nueva representación gráfica de la relación sistema-gestión-web.

La Figura 102 presenta la representación gráfica de la relación de éxito “software-desarrollo-unl”.

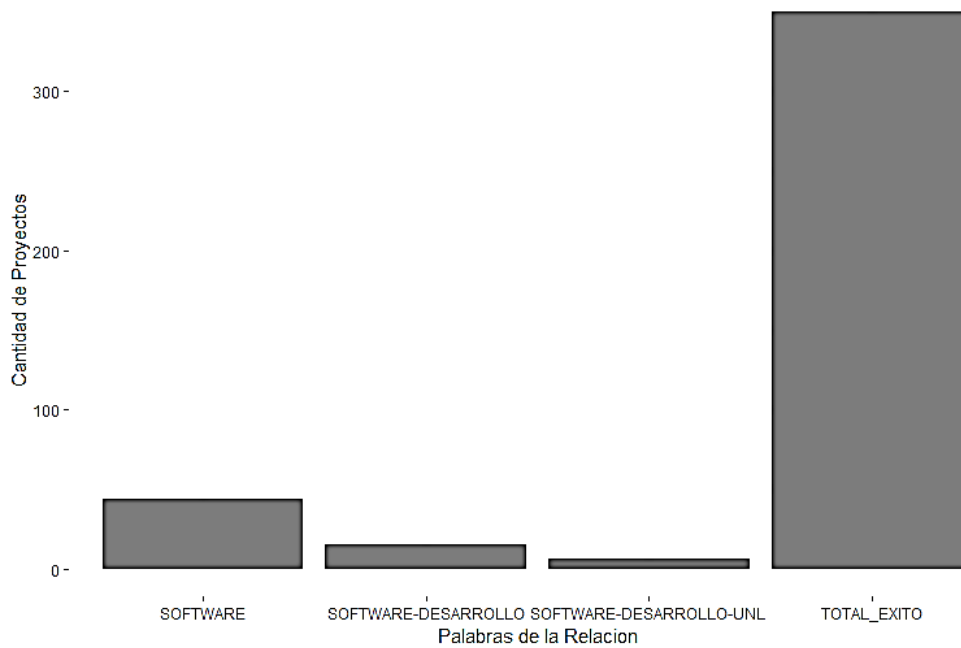


Figura 102. Nueva representación gráfica de la relación software-desarrollo-UNL.

- **K-medoids.**

Con **K-medoids** se obtuvieron los siguientes resultados:

En todo el conjunto de datos:

```
kmedoids_todo 1 : sistema
kmedoids_todo 2 :
```

Figura 103. Resultados de k-medoids en el nuevo conjunto de datos.

En los proyectos con estado de éxito:

```
kmedoids_exito 1 : sistema
kmedoids_exito 2 :
```

Figura 104. Resultados de k-medoids en los nuevos proyectos de éxito.

En los proyectos con estado de fracaso:

```
kmedoids_fracaso 1 : sistema
kmedoids_fracaso 2 : unl
kmedoids_fracaso 3 :
kmedoids_fracaso 4 : desarrollo aplicacion movil
```

Figura 105. Resultados de k-medoids en los nuevos proyectos de fracaso.

En la Figura 103 se observan las relaciones obtenidas en todo el conjunto de datos y se denominan kmedoids_todo. También en la Figura 104 se observan las relaciones obtenidas en los proyectos con estado de éxito y se denominan kmedoids_exito. Otras relaciones obtenidas son las de fracaso que se observan en la Figura 105 y se denominan kmedoids_fracaso. Se presentan todas las relaciones en la TABLA XLII.

TABLA XLII

COMPARATIVA DE NUEVOS RESULTADOS DE K-MEDOIDS

No.	Kmedoids_todo	Kmedoids_exito	Kmedoids_fracaso
1	sistema	sistema	sistema
2			unl
3			desarrollo aplicacion movil

Como se observa en la tabla anterior, la única relación de kmedoids_todo coincide con la única relación de kmedoids_exito, pero estas a su vez coinciden con la relación kmedoids_fracaso 1, determinando de esta manera que la relación de kmedoids_todo “sistema” no influye en el éxito o fracaso de los proyectos de titulación al coincidir con

relaciones tanto de éxito como de fracaso. El resto de relación de kmedoids_fracaso al no coincidir con alguna relación de kmedoids_todo pierden relevancia al no influir también en todo el conjunto de datos.

- **CHAID Exhaustivo.**

Con CHAID Exhaustivo se obtuvo un gráfico que debido a su tamaño no es muy visible su contenido por ello se separó el árbol en sus tres niveles para una más detallada visualización y explicación de los resultados. El algoritmo presenta un porcentaje de eficiencia de 85.9% y el primer nivel del árbol es el siguiente:

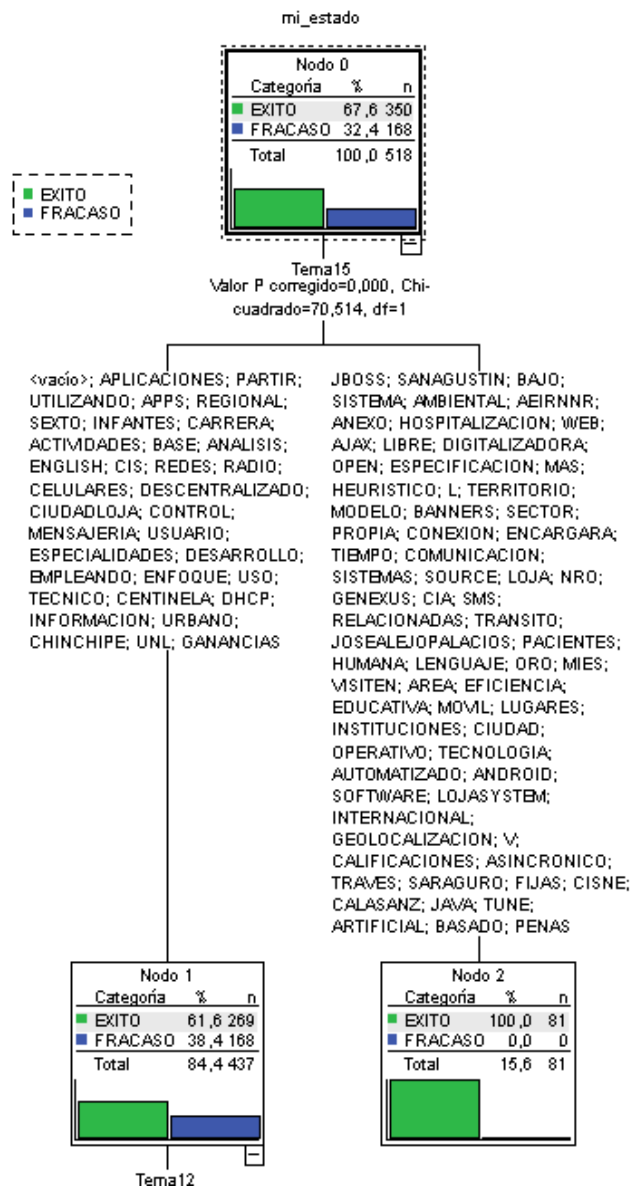


Figura 106. Primer nivel del nuevo árbol de CHAID Exhaustivo.

Se observa en la Figura 106 que la columna 15 es la más influyente en el éxito o fracaso del proyecto, es decir que las palabras ubicadas en la posición 15 del tema sin tomar en cuenta conectores son las palabras más influyentes en el estado de un proyecto de titulación. También se observa que, si una de las palabras del nodo 2 se encuentra en la posición 15 del tema, el proyecto tiene un 100% de probabilidad de culminar con éxito, y que si una de las palabras del nodo 1 se ubica en la posición 15 el proyecto tiene 61.6% de probabilidad de culminar con éxito y 38.4% de probabilidad de culminar con fracaso.

El segundo nivel del árbol es:

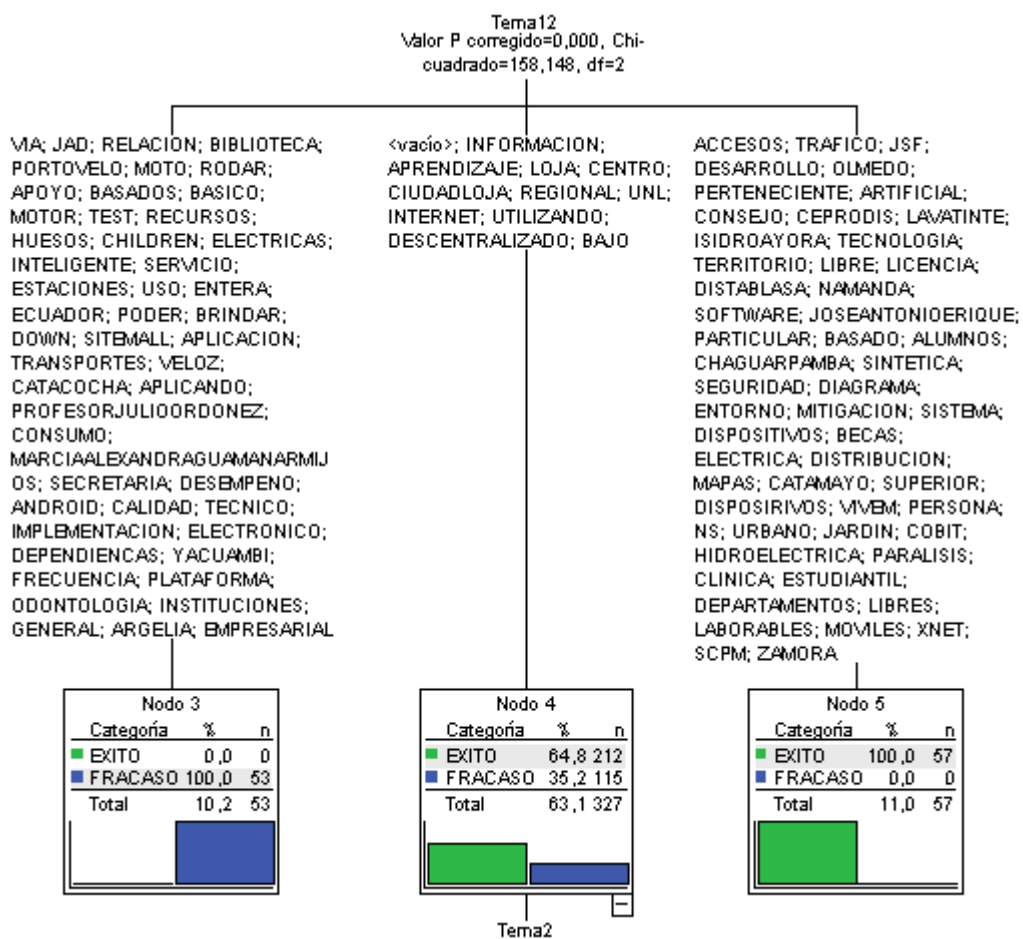


Figura 107. Segundo nivel del nuevo árbol de CHAID Exhaustivo.

En la Figura 107 se observa que, si una de las palabras del nodo 3 se encuentra en la posición 12 del tema y una de las palabras del nodo 1 se encuentra en la posición 15 del tema, el proyecto tiene un 100% de probabilidad de fracasar. También si una de las palabras del nodo 5 se encuentra en la posición 12 y una de las palabras del nodo 1 en la posición 15, entonces el proyecto tiene 100% de probabilidad de culminar con éxito.

Otro nodo presente en el nivel 2 del árbol es el nodo 4 en el que se indica que si una de las palabras de ese nodo está en la posición 12 y una de las palabras del nodo 1 en la posición 15 el proyecto tiene un 64.8% de probabilidad de culminar con éxito y 35.2% de probabilidad de fracasar. El tercer nivel del árbol se observa a continuación.

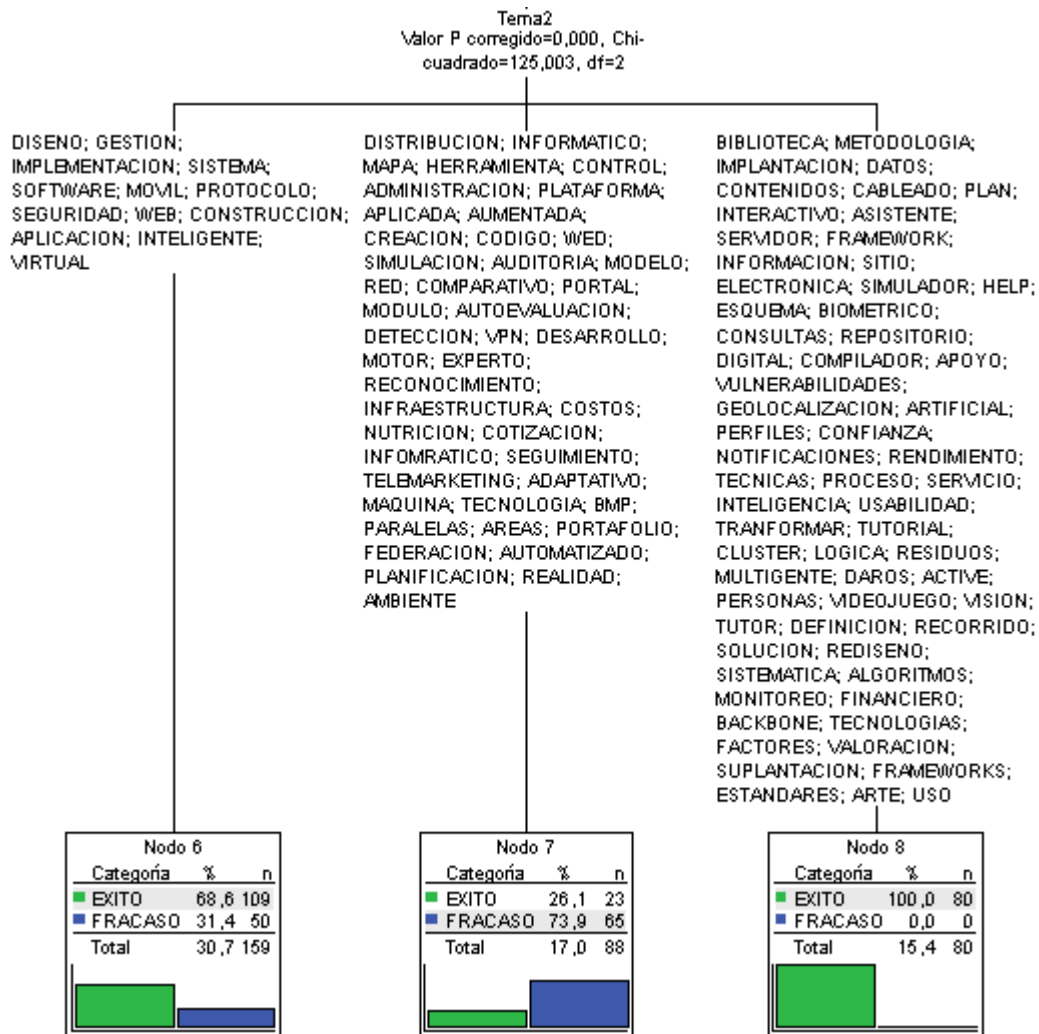


Figura 108. Tercer nivel del nuevo árbol de CHAID Exhaustivo.

La Figura 108 muestra como los nodos presentes en el nivel tres del árbol están precedidos por un solo nodo, que es el nodo 4 y este a su vez esta precedido por el nodo 1. Como se ha visto anteriormente el nodo 1 indica la posición 15 del tema de una de las palabras pertenecientes a ese nodo y el nodo 4 indica la posición 12 en el tema de una de las palabras pertenecientes a ese nodo. Los nodos en el nivel tres del árbol son los nodos 6, 7 y 8 que indican la posición 2 del tema de una de las palabras presentes en los nodos. Por lo tanto, teniendo en cuenta las consideraciones anteriores se determina que si una de las palabras pertenecientes al nodo 1, una de las palabras

pertencientes al nodo 4 y una de las palabras pertenecientes al nodo 6 se encuentran presentes en el tema en sus respectivas posiciones el proyecto tiene un 68.6% de probabilidad de culminar con éxito y un 31.4% de probabilidad de fracasar. Los porcentajes cambian cuando es el nodo 7 el precedido por los nodos 1 y 4, obteniendo así un 26.1% de probabilidad de culminar con éxito el proyecto y un 73.9% de probabilidad de que el proyecto fracase. También se obtienen otros porcentajes cuando el nodo 8 es precedido por los nodos 1 y 4, obteniendo así un 100% de probabilidad de culminar el proyecto con éxito.

- **Random Forest.**

Con Random Forest se obtuvieron los siguientes resultados:

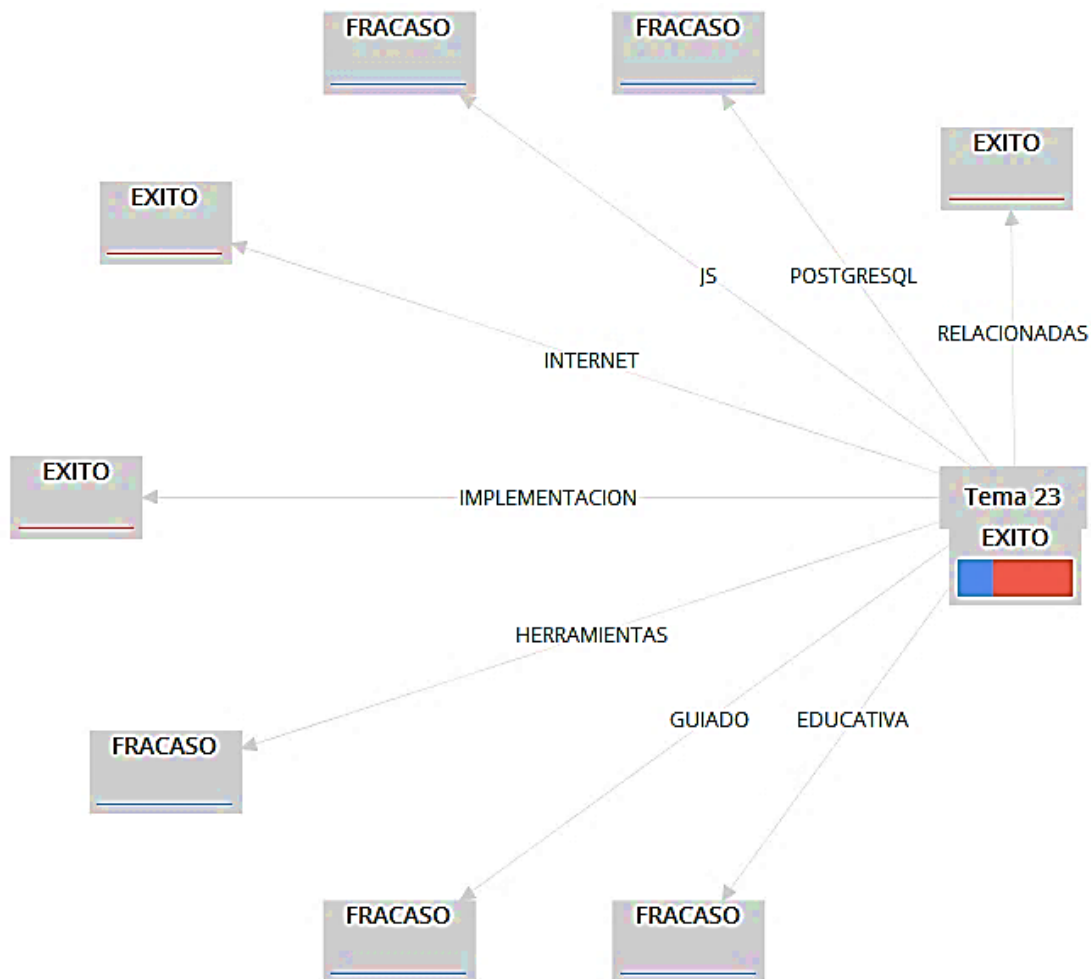


Figura 109. Nuevo árbol obtenido al aplicar Random Forest.

Se observa en la Figura 109 que la columna más influyente es la 23, esto quiere decir que la palabra presente en la posición 23 del tema de un proyecto es la que más influye en el éxito o fracaso del proyecto. Las palabras que influyen en el éxito si se encuentran

en la posición 23 del tema sin tomar en cuenta los conectores son: RELACIONADAS, INTERNET e IMPLEMENTACION. Las palabras que influyen en el fracaso son: POSTGRESQL, JS, HERRAMIENTAS, GIUADO y EDUCATIVA.

3.3.2. Evaluación de los algoritmos.

Luego de aplicar los algoritmos se procedió a evaluarlos de igual manera que en el punto 8 de esta sección, para ello usaremos la información obtenida en este punto referente a los resultados de los algoritmos. La evaluación de los algoritmos y de la representación gráfica de las palabras de los temas se hizo en base a los resultados obtenidos con el algoritmo CHAID Exhaustivo que presenta un porcentaje de eficiencia de 85.9%.

La evaluación de la representación gráfica de las palabras se detalla a continuación:

- Con la **Nube de Palabras** se obtuvo que la palabra “**loja**” influye en el éxito de los proyectos de titulación y que la palabra “**aplicacion**” influye en el fracaso de los proyectos. La palabra **Loja** está presente en el tema en las posiciones: 15 con 100% y 12 con 64.8% de probabilidad de culminar el proyecto con éxito. También se encuentra en el árbol la palabra **aplicación** en la posición 2 con 68.6% de probabilidad de culminar el proyecto con éxito y en la posición 12 con 100% de probabilidad de fracasar. Con la palabra **Loja** se coincide que influye en el éxito, mientras que con la palabra **aplicación** existe una contradicción y no se determina si influye en el éxito o fracaso.

La evaluación de los algoritmos de clasificación se describe a continuación:

- **Dendograma.**

El **Dendograma** dio como resultado que la relación “**aeirnnr-utilizando**” influye el éxito de los proyectos de titulación. Las palabras de la relación se encuentran en el árbol, pero no relacionadas entre sí. La palabra **AEIRNNR** está en la posición 15 del tema con 100% de probabilidad de éxito, y la palabra **utilizando** se encuentra en las posiciones: 15 con 61.6% y 12 con 64.8% de probabilidad de culminar el proyecto con éxito. El resultado no coincide al no estar presente la relación en el árbol, pero las palabras que forman la relación de éxito están presentes en el árbol e influyen en el éxito, por lo tanto, se determina que el algoritmo fue eficiente al obtener esta relación de éxito.

- **Apriori.**

Las reglas de asociación **Apriori** dan como resultado que las relaciones “**GESTION-SISTEMA**” y “**WEB-SISTEMA**” influyen en el éxito de los proyectos, mientras que la relación “**MOVIL-APLICACION**” influye en el fracaso de los proyectos.

A las palabras de la primera relación de éxito se las encontró presentes en el árbol, pero no relacionadas entre sí, teniendo así que: la palabra **gestión** está en la posición 2 con 68.6% de probabilidad de culminar el proyecto con y la palabra **sistema** está presente en las posiciones: 15 con 100%, 12 con 100% y 2 con 68.6% de probabilidad de éxito.

La segunda relación de éxito se encuentra formada por las palabras web y sistema, la palabra **web** está presente en las posiciones: 2 con 68.6% y 15 con 100% de probabilidad de éxito y la palabra **sistema** se encuentra en las posiciones: 15 con 100%, 12 con 100% y 2 con 68.6% de probabilidad de éxito.

La relación que influye en el fracaso está formada por las palabras móvil y aplicación, la palabra **móvil** se encuentra en las posiciones: 15 con 100% y 2 con 68.6% de probabilidad de éxito y la palabra **aplicación** está presente la posición 2 con 68.6% de probabilidad de culminar el proyecto con éxito.

Las palabras que forman parte de las relaciones de éxito tienden a influir al éxito también en el árbol, siendo así relevantes al coincidir; por otra parte, las palabras que forman parte de la relación de fracaso tienden a influir en el éxito, existiendo una contradicción entre los resultados.

- **K-means.**

El algoritmo **K-means** presenta como resultado que las relaciones “**territorio-loja-sistema**”, “**aprendizaje-virtual-entorno**”, “**desarrollo-sistema-implementacion**”, “**sistema-gestion-web**” y “**software-desarrollo-unl**” influyen en el éxito de los proyectos.

En la relación de éxito “**territorio-loja-sistema**” se encuentra la palabra **Loja** relacionada con la palabra **sistema** de la siguiente manera: la palabra **Loja** se encuentra en la posición 12 con 64.8% de éxito y la palabra **sistema** está en la posición 2 con 68.6% de probabilidad de culminar el proyecto con éxito. También están las palabras presentes en el árbol sin estar relacionadas, las palabras **sistema** y **territorio** están presentes en las posiciones: 12 con 100% y 15 con 100% de probabilidad de culminar el proyecto con éxito.

En la segunda relación de éxito “**aprendizaje-virtual-entorno**” las palabras virtual y aprendizaje se encuentran relacionadas, la palabra **aprendizaje** está en la posición 12 con 64.8% de probabilidad de éxito y la palabra **virtual** se encuentra en la posición 2 con 68.6% de probabilidad de éxito. La palabra **entorno** no se encuentra relacionada con las otras palabras que forman parte de la relación, pero está presente en la posición 12 con 100% de probabilidad de culminar el proyecto con éxito.

En la tercera relación de éxito “**desarrollo-sistema-implementacion**” las palabras que forman la relación no se encuentran relacionadas en el árbol, pero si están presentes en el árbol. La palabra **desarrollo** está en las posiciones: 15 con 61.6% y 12 con 100% de probabilidad de culminar el proyecto con éxito, y 2 con 73.9% de probabilidad de fracasar. Otra palabra de la relación es la palabra **sistema** que se encuentra en las posiciones: 15 con 100%, 12 con 100% y 2 con 68.6% de probabilidad de éxito. También se halla la palabra **implementación** que se encuentra en la posición 12 con 100% de probabilidad de fracasar y en la posición 2 con 68.6% de probabilidad de culminar el proyecto con éxito. Las palabras **desarrollo** e **implementación** tienen contradicciones en los resultados del árbol, así que esta relación pierde relevancia al no coincidir los resultados.

En la cuarta relación de éxito “**sistema-gestion-web**” las palabras de la relación no están relacionadas entre sí y se encuentran de la siguiente manera: la palabra **gestión** está en la posición 2 con 68.6% de probabilidad de culminar el proyecto con éxito, la palabra **sistema** se encuentra en las posiciones: 15 con 100%, 12 con 100% y 2 con 68.6% de probabilidad de éxito, y la palabra **web** está presente en las posiciones: 2 con 68.6% y 15 con 100% de probabilidad de éxito

En la quinta relación de éxito “**software-desarrollo-unl**” la palabra **UNL** está relacionada con la palabra **software** y la palabra **desarrollo** respectivamente, la palabra **UNL** está en la posición 12 con 64.8% de probabilidad de éxito mientras que la palabra **software** se encuentra en la posición 2 con 68.6% de probabilidad de culminar el proyecto con éxito y la palabra **desarrollo** está en la posición 2 con 73.9% de probabilidad de fracasar. También se encuentran palabras de la relación en el árbol sin estar relacionadas, así tenemos a la palabra **software** que se encuentra en las posiciones: 15 con 100% y 12 con 100% de probabilidad de éxito, y la palabra **desarrollo** que está en las posiciones: 15 con 61.6% y 12 con 100% de probabilidad de culminar el proyecto con éxito. Las palabras **UNL** y **desarrollo** están relacionadas en el

árbol e influyen en el fracaso de los proyectos, por lo que existe una contradicción en los resultados y la relación pierde relevancia.

Los resultados de las relaciones de éxito “**territorio-loja-sistema**”, “**aprendizaje-virtual-entorno**” y “**sistema-gestion-web**” coinciden con los resultados del árbol, por lo que el algoritmo fue eficiente en la obtención de estas relaciones de éxito.

- **K-medoids.**

Con el algoritmo **K-medoids** no se obtuvieron resultados relevantes.

- **Random Forest.**

El algoritmo **Random Forest** dio como resultado palabras que influyen en el éxito o en el fracaso, de esas palabras las que influyen en el éxito son: **RELACIONADAS**, **INTERNET** e **IMPLEMENTACION**; y las palabras que influyen en el fracaso son: **POSTGRESQL**, **JS**, **HERRAMIENTAS**, **GIUADO** y **EDUCATIVA**.

En las palabras de éxito tenemos que:

RELACIONADAS, se encuentra en la posición 15 con 100% de probabilidad de éxito.

INTERNET, está en la posición 12 con 64.8% de probabilidad de culminar el proyecto con éxito.

IMPLEMENTACION, se encuentra en la posición 12 con 100% de probabilidad de fracasar y en la posición 2 con 68.6% de probabilidad de culminar el proyecto con éxito.

En las palabras de fracaso no se encontró en el árbol a las palabras: **POSTGRESQL**, **JS**, **GUIADO**. Las palabras de fracaso que si están presentes en el árbol son:

HERRAMIENTAS, se encuentra en la posición 2 con 73.9% de probabilidad de fracasar.

EDUCATIVA, está en la posición 15 con 100% de probabilidad de éxito.

Las palabras que influyen en el éxito obtenidas con este algoritmo sí coinciden con el árbol, mientras que de las 5 palabras de fracaso solo 2 se hallaron en el árbol y una de estas contradice los resultados ya que en el árbol influye en el éxito. Por tal motivo el algoritmo fue eficiente para determinar que palabras influyen en el éxito, pero fue deficiente para determinar que palabras influyen en el fracaso.

g. Discusión

La definición del tema de un trabajo de titulación es el primer obstáculo para la elaboración del mismo, en muchos casos la definición del tema implica demasiado tiempo en su elaboración, y una mala definición del tema acarrea problemas en el desarrollo del proyecto que a su vez pueden ocasionar cambios en objetivos, prórrogas o bien el fracaso del trabajo de titulación. En la Secretaría General de la Facultad de Energía, las Industrias y los Recursos Naturales No Renovables de la Universidad Nacional de Loja, hay información de proyectos de titulación que fueron aprobados para su desarrollo, esta información contiene atributos del estado en el que se encuentran los proyectos, los autores de los proyectos, entre otras características.

También cabe destacar que en la biblioteca de la facultad se pudo encontrar información de los proyectos de titulación que culminaron con éxito. A toda esta información no se le da ningún uso, y tener abundancia de información no implica poseer conocimiento, por lo que se debe tener en cuenta que en esta información se encuentra oculta información de gran importancia, que sirva de orientación tanto a estudiantes como a docentes en la definición del tema de un proyecto de titulación. El presente trabajo de titulación busca determinar que influencia tiene el tema de un proyecto de titulación en su posterior éxito o fracaso, por tal motivo se planteó aplicar algoritmos de clasificación que permitan predecir patrones dentro del conjunto de datos de los proyectos de titulación aprobados para encontrar información no evidente.

1. Desarrollo de la Propuesta Alternativa

Objetivo 1: Depurar el conjunto de datos obtenido de los proyectos de titulación.

Antes de depurar el conjunto de datos se procedió a evaluarlo, para ello se realizó una entrevista a la persona encargada del manejo del conjunto de datos para comprender como se encontraban estructurados los datos y conocer que variables eran útiles para el desarrollo del proyecto, posterior a ello se aplicó la técnica del muestreo aleatorio simple para validar el conjunto de datos con los proyectos de titulación que ya se encontraban clasificados. Se identificó que 1502 proyectos formaban el conjunto de datos y 1023 de esos proyectos ya se encontraban clasificados, en estos proyectos ya clasificados se aplicó el muestreo aleatorio simple, lo que permitió obtener el número de proyectos a ser evaluados, se evaluaron 279 proyectos elegidos de forma aleatoria mediante la herramienta RStudio.

Luego de evaluar el conjunto de datos se obtuvo que de los 279 proyectos evaluados 4 de ellos estaban mal clasificados, lo que representa que un 98.56% de los proyectos se encuentran correctamente clasificados. En el conjunto de datos se encontraron proyectos de toda la FEIRNNR de la UNL, de los cuales algunos pertenecían a maestrías y tecnologías, dando un total de 1502 proyectos en el conjunto de datos. Las maestrías y tecnologías no se ofertan actualmente en la facultad por lo que son información no útil y representan ruido en el conjunto de datos al momento de aplicar los algoritmos de clasificación, por tal motivo se descartó los proyectos de las maestrías y tecnologías, dejando así 1226 proyectos en el conjunto de datos. Luego se categorizó los proyectos de titulación restantes, evidenciando que en el conjunto de datos existían proyectos que no estaban categorizados y algunos de los proyectos ya categorizados se encontraban mal categorizados, por lo que se procedió a categorizar manualmente todos los proyectos usando como guía la información obtenida de la entrevista realizada a la persona a cargo del manejo del conjunto de datos.

Para categorizar correctamente se identificó que variables del conjunto de datos determinan si el proyecto fracasó o tuvo éxito, y se estableció que relaciones entre esas variables están presentes en el conjunto de datos. Una vez categorizado el conjunto de datos se obtuvo 834 proyectos con estado de éxito y 392 proyectos con estado de fracaso. La herramienta usada para la eliminación de la información no útil fue Openrefine, esta herramienta fue desarrollada por Google para el refinamiento o depuración de información en grandes conjuntos de datos. Esta herramienta permitió estandarizar el conjunto de datos aplicando comandos para convertir los temas de los proyectos a mayúsculas y reemplazar un valor por otro, de esta manera se corrigió palabras mal escritas, se reemplazó la letra ñ por la letra n, y se eliminó tildes y otros caracteres mostrados en la TABLA II. También se reemplazó palabras semejantes utilizando algunos algoritmos presentes en la herramienta que agrupan y reemplazan términos semejantes, un ejemplo de ello son las palabras análisis y analizar que implican una misma acción y fueron estandarizadas con la palabra análisis.

Objetivo 2: Aplicar los algoritmos de clasificación al conjunto de datos de los proyectos de titulación.

En la búsqueda de algoritmos de clasificación, se empleó la metodología de la revisión de literatura. Para hallar estudios similares a este trabajo, el principal criterio para seleccionar un estudio fue que abarque al menos 3 de los 4 temas generales de la revisión de literatura, estos temas son: aprendizaje automático, minería de datos,

minería de texto y algoritmos de clasificación. Se obtuvo 12 estudios relacionados al presente trabajo, se procedió a identificar los algoritmos de clasificación más usados en los estudios seleccionados y estos algoritmos fueron establecidos como los algoritmos de clasificación encontrados. Para seleccionar los algoritmos a aplicar en este trabajo se usó características presentes en el Anexo 11 de los algoritmos de clasificación que son útiles para el desarrollo del presente trabajo. Las características se establecieron en base a los resultados y conclusiones de los estudios seleccionados anteriormente. Por último, se realizó una tabla comparativa entre los algoritmos encontrados con las características establecidas, los algoritmos que fueron seleccionados para el desarrollo de este trabajo fueron los que cumplieron con más características, los algoritmos seleccionados fueron: Agrupamiento Jerárquico o Dendograma, K-means, K-medoids, Reglas de Asociación A priori, CHAID Exhaustivo y Random Forest.

Previo a la aplicación de los algoritmos se realizó el pre procesamiento del conjunto de datos para encontrar oraciones como “Área de la Energía, las Industrias y los Recursos Naturales No Renovables”, las cuales contienen palabras que se encuentran relacionadas debido a que implican el nombre de un lugar, por tal motivo se las denominó relaciones obvias. Para realizar el pre procesamiento se usó Nubes de Palabras y Dendogramas, algunas de las relaciones obvias encontradas son: “Universidad Nacional de Loja” que fue reemplazada por sus siglas “UNL” y “Área de la Energía, las Industrias y los Recursos Naturales No Renovables” que fue reemplazada por sus siglas “AEIRNNR”. Luego de haber realizado el pre procesamiento de datos se obtuvo un nuevo conjunto de datos sin relaciones obvias, se cargó los datos en la herramienta RStudio y se controló que los conectores (de, con, para, etc.) que forman parte del tema no se carguen en la bolsa de palabras, porque estos conectores afectan a las relaciones entre las palabras.

La representación gráfica denominada Nube de Palabras permitió conocer las palabras más usadas en todo el conjunto de datos, en los proyectos con éxito y en los proyectos que fracasaron, para obtener una idea general de cuáles son las temáticas más empleadas en los proyectos. Por último, se procedió a la aplicación de los algoritmos de clasificación, los algoritmos de clasificación no supervisados fueron aplicados de igual manera que la Nube de palabras, es decir en todo el conjunto de datos y en cada clase. Los algoritmos de clasificación supervisada en cambio se aplicaron solamente en todo el conjunto de datos, debido a que este tipo de algoritmos clasifican usando las etiquetas de cada proyecto.

Objetivo 3: Evaluar los algoritmos aplicados.

Esta actividad tuvo dos etapas, la primera se evaluó los resultados obtenidos con la aplicación de los algoritmos y en la segunda se evaluó los algoritmos haciendo uso de los resultados evaluados. En la evaluación de resultados se realizó una tabla comparativa por cada algoritmo de clasificación no supervisado, en la que se comparó los resultados obtenidos en los proyectos de éxito y fracaso con los resultados obtenidos en todo el conjunto de datos, si uno de los resultados obtenidos en el éxito o fracaso no coinciden entre ellos y coincide con uno de los resultados de todo el conjunto de datos, se considera ese resultado como relevante al influir también en todo el conjunto de datos, los resultados de los algoritmos de clasificación supervisados no se evaluaron de esta manera debido a que clasifican haciendo uso de las etiquetas.

Para la evaluación de los algoritmos se tomó como base el algoritmo CHAID Exhaustivo, debido a que es el único que brinda un porcentaje de eficiencia, el proceso realizado fue buscar los resultados de los otros algoritmos en el árbol del algoritmo CHAID Exhaustivo y determinar los porcentajes de éxito o fracaso que le corresponden a cada resultado, en caso de existir una contradicción de los resultados de los algoritmos con los resultados presentes en el árbol, se determina que el algoritmo evaluado fue ineficiente en ese resultado obtenido. Con la evaluación de los resultados y los algoritmos se obtuvo lo siguiente:

- La relación de fracaso “móvil-aplicación” obtenida con el **Dendograma** no coincide con los resultados del árbol por lo que se considera ineficiente el algoritmo.
- Con el algoritmo **Apriori** se obtuvo la relación de éxito “gestión-sistema”, la cual presenta contradicciones en el árbol por lo que se considera ineficiente al algoritmo.
- Las relaciones “diseño-construcción-prototipo” de éxito y “aplicación-desarrollo-móvil” de fracaso no coinciden con las relaciones del árbol por lo que se considera ineficiente al algoritmo **K-means** en estas relaciones, pero con la relación de éxito “Loja-territorio-estudio” el algoritmo **K-means** es eficiente al coincidir con los resultados del árbol.
- El algoritmo **K-medoids** dio como resultado que la palabra “Loja” influye en el éxito, coincide con los resultados del árbol por lo que se determina eficiente el algoritmo.

- De las palabras obtenidas con el algoritmo **Random Forest** tan solo algunas palabras se encontraron presentes en el árbol, las palabras “web”, “UNL”, “territorio” e “implementación” que influyen en el éxito coinciden con los resultados del árbol por lo que se considera eficiente el algoritmo en los resultados que influyen en el éxito. La palabra “Yacuambi” que influye en el fracaso coincide con los resultados del árbol, y las palabras “ciudadloja” y “datos” que también influyen en el fracaso presentan contradicciones en los resultados del árbol, por lo que se determina ineficiente el algoritmo en los resultados que influyen en el fracaso de los proyectos.

Para la introducción de nuevos datos se obtuvo un nuevo conjunto de datos, el cual es una nueva actualización del anterior conjunto de datos, es decir cuenta con todos los datos del anterior conjunto de datos, pero algunos de los datos anteriores muestran nueva información y también cuentan con nuevos datos pertenecientes a nuevos proyectos de la Carrera de Ingeniería en Sistemas. Con este conjunto de datos se empleó la misma metodología usada con el anterior conjunto de datos, por lo que se realizaron las mismas actividades explicadas anteriormente. Con la evaluación de los resultados y los algoritmos se obtuvo lo siguiente:

- La relación “AEIRNNR-utilizando” que influye en el éxito coincide con los resultados del árbol por lo que se considera eficiente el **Dendograma**.
- Las relaciones “gestión-web” y “web-sistema” que influyen en el éxito coinciden con los resultados del árbol, por lo que el algoritmo **Apriori** es eficiente para determinar relaciones de éxito, pero la relación “móvil-aplicación” que influye en el fracaso no coincide con los resultados del árbol, por lo que se considera al algoritmo ineficiente para determinar relaciones de fracaso.
- El algoritmo **K-means** presenta que las relaciones “territorio-Loja-sistema”, “aprendizaje-virtual-entorno”, “sistema-gestión-web”, “software-desarrollo-UNL” y “desarrollo-sistema implementación” influyen en el éxito, las tres primeras relaciones coinciden con los resultados del árbol y se determina que el algoritmo es eficiente en la obtención de esas relaciones.
- Con **K-medoids** no se obtuvieron resultados relevantes.
- Con el algoritmo **Random Forest** se obtuvo que las palabras “relacionadas”, “internet” e “implementación” influyen en el éxito, estos resultados coinciden con los resultados del árbol, considerando así eficiente al algoritmo en las relaciones de éxito. Por otra parte, de las palabras de fracaso solo las palabras

“herramientas” y “educativa” se encuentran presentes en el árbol, pero existen contradicciones en los resultados, por lo que se considera que el algoritmo es ineficiente para determinar las palabras que influyen en el fracaso.

2. Valoración técnica económica ambiental

El desarrollo del presente trabajo de titulación implicó una inversión económica, puesto que se utilizó recursos para realizar los objetivos planteados en este trabajo, en la TABLA XLIII son presentados los costos de los recursos utilizados.

TABLA XLIII
VALORACIÓN ECONÓMICA DEL PROYECTO

Recurso Humano			
Equipo de trabajo	Horas	Precio/Hora	Valor Total
Pablo Valdivieso	960	\$5,00	\$4.800,00
Director	26	\$15,00	\$390,00
SUBTOTAL			\$5.190,00
Descripción	Cantidad	Valor	Valor Total
Recurso Hardware			
Computadora	1	\$1.200,00	\$1.200,00
Impresora	1	\$60,00	\$60,00
SUBTOTAL			\$1.260,00
Recurso Software			
Openrefine	1	\$0,00	\$0,00
RStudio	1	\$0,00	\$0,00
IBM SPSS Statistics	1	\$0,00	\$0,00
RapidMiner Studio	1	\$0,00	\$0,00
SUBTOTAL			\$0,00
Recurso Varios			
Internet	400	\$0,70	\$280,00
Transporte	50	\$0,30	\$15,00
SUBTOTAL			\$295,00
TOTAL			\$6.745,00

El coste del proyecto corresponde con el tiempo de ejecución del trabajo, también cabe mencionar que los costos fueron adjudicados por el autor de este trabajo.

h. Conclusiones

Las conclusiones presentadas en esta sección son obtenidas a partir de la evaluación de los resultados y algoritmos de clasificación:

- Las relaciones para el éxito de un proyecto se pueden determinar eficientemente con las reglas de asociación Apriori y las relaciones para los casos de fracaso de un proyecto con las reglas de asociación Apriori y el Dendograma.
- La relación que influye en el éxito de los proyectos está formada por las palabras “**gestión**” y “**sistema**”, esta relación es obtenida por el algoritmo Apriori en el primer conjunto de datos y en la introducción de nuevos datos, también es validada por el algoritmo CHAID Exhaustivo, lo que la hace una relación relevante.
- La relación que influye en el fracaso de los proyectos está formada por las palabras “**móvil**” y “**aplicación**”, esta relación es negada por el algoritmo CHAID Exhaustivo, pero es obtenida en el primer conjunto de datos y en la introducción de nuevos datos como una relación de fracaso, lo que la hace una relación relevante.
- Se puede excluir al algoritmo k-medoids para determinar las relaciones que influyen en el éxito o fracaso, debido a que los resultados obtenidos con este algoritmo fueron irrelevantes tanto en su aplicación en el primer conjunto de datos como también en la introducción de los nuevos datos.

i. Recomendaciones

Al finalizar el trabajo de titulación se plantearon las siguientes recomendaciones:

- Estandarizar la forma en la que se almacena la información de los proyectos de titulación e incluir una variable que señale el éxito o fracaso de los proyectos.
- Para trabajos similares usar los algoritmos Apriori y Dendograma, ya que son muy beneficiosos o eficientes para encontrar patrones o relaciones fiables entre las palabras.
- Usar Openrefine para el refinamiento o depuración de la información ya que permite reemplazar valores, estandarizar el conjunto de datos, obtener la estructura deseada del conjunto de datos y además agrupar palabras similares para eliminar el ruido y optimizar la calidad del conjunto de datos.
- El uso de la herramienta RapidMiner se recomienda para la aplicación del algoritmo Random Forest ya que proporciona diversos criterios para obtener resultados óptimos y permite modificar la forma en la que se presenta la gráfica, para obtener una mejor visualización de los resultados.
- Desarrollar un software que permita la gestión de los proyectos de titulación que se encuentran aprobados, en el que se considere como campo principal el estado actual del proyecto.

j. Bibliografía

Referencias Bibliográficas

- [1] C. De Contabilidad, Y. Auditoría, B. Torres, and Y. Mauricio, "IMPORTANCIA DE TENER UN CONOCIMIENTO PREVIO SOBRE UN TEMA ESPECÍFICO Y SU INCIDENCIA EN LA EJECUCIÓN DE UN TRABAJO INVESTIGATIVO," p. 2019, 2016.
- [2] R. Camana, "Potenciales Aplicaciones de la Minería de Datos en Ecuador," *Potenciales Apl. la Minería Datos en Ecuador*, vol. 29, no. Julio, pp. 170–183, 2016.
- [3] C. E. Riofrio, "FACEBOOK COMO HERRAMIENTA PARA EL APRENDIZAJE COLABORATIVO DE LA INTELIGENCIA ARTIFICIAL," vol. IX, pp. 27–36, 2018.
- [4] M. G. Victoria, "Investigación del problema inverso de reconstrucción tomográfica en óptica adaptativa para astronomía a través de técnicas de minería de datos e inteligencia artificial," 2014.
- [5] L. C. Blanco, O. Lidia, P. Gonzalez, Á. Mercedes, and M. Sánchez, "USO DE TÉCNICAS DE MINERÍA DE DATOS EN LA ENSEÑANZA DEL ALGEBRA LINEAL," pp. 1420–1427.
- [6] R. R. A. Santiago Leonardo Morales Cardoso, Mario Raúl Morales Morales, "Metodología para Procesos de Inteligencia de Negocios con mejoras en la extracción y transformación de fuentes de Datos," no. 11, pp. 107–119, 2017.
- [7] O. Adrián *et al.*, "CONSTRUCCIÓN DE UN MODELO DE PREDICCIÓN PARA APOYO AL DIAGNÓSTICO DE DIABETES," vol. 40, no. 130, pp. 2105–2122, 2018.
- [8] C. A. R. Romero, "ESTUDIO COMPARATIVO DE ALGORITMOS DE INTELIGENCIA ARTIFICIAL Y MINERIA DE DATOS ENFOCADOS A LA TOMA DE DECISIONES EMPRESARIALES DE SELECCION DE PERSONAL." 2018.
- [9] J. Luis *et al.*, "Análisis de deserción escolar con minería de datos," vol. 93, no. 2015, pp. 71–82, 2015.
- [10] J. C. Acosta, D. La, R. Martínez, C. Primorac, and U. Nacional, "Determinación de perfiles de rendimiento académico en la UNNE con Minería de Datos

- Educacional Facultad de Ciencias Exactas y Naturales y Agrimensura,” 2018.
- [11] A. P. H. Escobar, M. Alcivar, “Aplicaciones de Minería de Datos en Marketing,” vol. 3, no. 8, pp. 503–512, 2016.
- [12] G. P. Romero and A. P. Niz, “Análisis de la deserción estudiantil en la USB, facultad Ingeniería de Sistemas, con técnicas de minería de datos,” pp. 1–6, 2015.
- [13] U. Tecnológica, L. Red, and M. Karanik, “Perfiles de Rendimiento Académico : Un Modelo basado en Minería de datos Academic Performance Profiles : A Model based on data Mining,” vol. IV, no. 2015, pp. 12–30, 2015.
- [14] F. Cortés-martínez, A. Treviño-cansino, A. T. Espinoza-fraire, J. A. Sáenz-esqueda, J. Gerardo, and L. Vélez, “Rules for predicting compliance with the quality of wastewater in a treatment plant applying data mining Reglas para predecir el cumplimiento de la calidad del agua residual en una planta tratadora con minería de datos,” vol. 21, no. 62, pp. 13–24, 2018.
- [15] A. Consuegra, Y. M. Salazar, J. H. Garc, and D. H. Vizcaino, “MINERÍA DE TEXTO COMO UNA HERRAMIENTA PARA LA BÚSQUEDA DE ARTÍCULOS CIENTÍFICOS PARA LA INVESTIGACIÓN,” pp. 14–20, 2016.
- [16] M. R. Mounier, K. B. Acosta, F. Favret, D. A. Godoy, and J. D. D. Benítez, “POLIMORFISMOS DE NUCLEÓTIDOS SIMPLES RELACIONADOS AL RIESGO DE ENFERMEDADES: CLASIFICACIÓN CASO-CONTROL UTILIZANDO TÉCNICAS DE MINERÍA DE TEXTO,” pp. 330–334, 2018.
- [17] R. A. Cortez-reyes, “Extracción de conocimiento a partir de textos obtenidos de Twitter Extraction of knowledge from texts obtained from Twitter,” pp. 30–41, 2018.
- [18] D. A. OBANDO VELÁSQUEZ, “Estudio y análisis de entornos comerciales mediante la evaluación, comparación y experimentación de algoritmos de minería de datos,” 2017.
- [19] S. N. O. ARENAS, “COMBINACION DE METRICAS Y RASGOS LEXICO-SEMANTICOS PARA EL ANALISIS DE SIMILITUD TEXTUAL ENTRE DOS FRASES,” 2017.
- [20] A. Addati and S. Roger, “Agentes Inteligentes y Web Semántica : Preprocesamiento de Texto de Redes Sociales,” pp. 16–20, 2017.

- [21] D. E. La, D. E. Miner, and I. A. D. E. T. Aplicaciones, *NUEVAS TÉCNICAS DE MINERÍA DE TEXTOS: APLICACIONES*. 2017.
- [22] H. D. E. M. D. E. Texto, L. Andrea, and T. Samboni, "ANÁLISIS DE SENTIMIENTOS SOBRE EL POSCONFLICTO COLOMBIANO UTILIZANDO HERRAMIENTAS DE MINERÍA DE TEXTO," 2015.
- [23] A. Pastor, L. Monroy, M. Montes, H. J. Escalante, and S. M. Tonantzintla, "Categorización mediante técnicas de minería de texto," no. 1, 2016.
- [24] G. Ram and S. Christian, "Agrupamiento de textos cortos en dominios cruzados Cross-domain Clustering for Short Texts," vol. 115, pp. 133–145, 2016.
- [25] F. D. L. Castillo, C. Alberto, and S. Santos, "Nubes de palabras animadas para la visualización de información textual de Publicaciones Académicas Introducción," pp. 77–84, 2016.
- [26] F. Garelli and A. Mengascini, "Formación docente y representaciones sobre Salud : caminos para la Educación en Salud desde una mirada crítica," pp. 1–20, 2017.
- [27] H. Referentes, A. L. Barrio, and L. A. Floresta, "MINERÍA DE TEXTO DE LA WEB, DE OPINIÓN PÚBLICA Y HECHOS REFERENTES AL BARRIO LA FLORESTA," 2018.
- [28] J. L. P. A. Hernández-mendo, "Algoritmos de clasificación y redes neuronales en la observación automatizada de registros Classi cation algorithms and neural networks in automated observation records Algoritmos de classi cação e redes neurais em registros de observação automatizados," vol. 15, no. 2001, pp. 31–40, 2015.
- [29] D. D. Castillo *et al.*, "Algoritmos de aprendizaje automático para la clasificación de neuronas piramidales afectadas por el envejecimiento Machine learning algorithms for classification of pyramidal neurons affected by aging," vol. 8, no. 3, pp. 559–571, 2016.
- [30] S. Rodríguez-tapia and J. Camacho-cañamón, "LOS MÉTODOS DE APRENDIZAJE AUTOMÁTICO SUPERVISADO EN LA CLASIFICACIÓN TEXTUAL SEGÚN EL GRADO DE ESPECIALIZACIÓN," pp. 1–28, 2018.
- [31] A. N. G. Gómez and L. S. P. Chaparro, "PLUGIN PARA LA CLASIFICACIÓN

- SUPERVISADA DE IMÁGENES SATELITALES MEDIANTE EL USO DEL ALGORITMO PERCEPTRÓN MULTICAPA BASADOS EN REDES NEURONALES,” 2016.
- [32] R. R. Camacho, R. B. Fernández, and A. G. Arenas, “Modelado y propagación de valores de sentimiento en relaciones de usuario Modelling and Propagation of Sentiment Values in Relations between Users,” vol. 107, pp. 9–17, 2015.
- [33] P. Javier *et al.*, “Métodos de clasificación : Análisis de fertilidad,” no. 111, pp. 43–57, 2015.
- [34] S. R. Suarez, L. Vidal, P. Salio, and Y. G. Skabar, “Técnicas de Clasificación Supervisada para la Discriminación entre Ecos Meteorológicos y No Meteorológicos usando Información de un Radar Meteorológico de Banda C,” 2017.
- [35] J. Rosario, V. Morales, D. Omar, and R. Buenrostro, “Aplicación de algoritmos de clasificación para el análisis de tejido mamario y detección de cáncer de mama,” no. 114, pp. 260–271, 2015.
- [36] Á. Freddy and G. Viera, “Técnicas de aprendizaje de máquina utilizadas para la minería de texto,” vol. 31, pp. 103–126, 2015.
- [37] L. F. Jeri and J. Salinas, “Evaluación de la decisión de obtener el título profesional con la elaboración de la tesis mediante técnicas multivariantes : Caso Universidad Nacional Agraria La Molina Evaluation of the decision to obtain the professional title with the development of th,” vol. 78, no. 2, pp. 92–99, 2017.
- [38] J. Javier and L. Salazar, “ANÁLISIS DE PREFERENCIA DE SERVICIOS DE TELEFONÍA MÓVIL PARA SEGMENTAR A LOS CLIENTES QUE USAN SMARTPHONE,” 2016.
- [39] D. Cómputos, I. C. Cynthia, I. M. Calixto, I. P. Florencia, M. Gimena, and S. D. Matías, “Fusión de Algoritmos Bayesianos y Árboles de Clasificación como Propuesta para la Clasificación Supervisada de Fallos de Equipos en un laboratorio,” pp. 72–76.
- [40] R. I. On, “EVALUACION DE ALGORITMOS DE CLASIFICACION SUPERVISADA PARA EL MINADO DE OPINION EN TWITTER,” vol. 36, no. 3, pp. 194–205, 2015.

- [41] F. D. E. E. Y. Planificación, "SEGMENTACIÓN DE CLIENTES DE UN CASINO UTILIZANDO EL ALGORITMO PARTICIÓN ALREDEDOR DE MEDOIDES (PAM) CON DATOS MIXTOS," 2018.
- [42] R. Adalberto, C. Reyes, O. Otoniel, and F. Cortez, *Compilación de investigaciones de tecnología 2017*. 2017.
- [43] A. Flores-fuentes and R. Alejo, "Minería de datos aplicada para la identificación de factores de riesgo en alumnos," vol. 139, no. 2017, pp. 177–189, 2017.
- [44] M. B. A. ZHAÑAY, "MINERÍA DE TEXTO EN MEDIOS SOCIALES: CASO DE ESTUDIO DEL PROYECTO TRANVÍA DE CUENCA," 2016.
- [45] Y. Zhao, "R and Data Mining : Examples and Case Studies," no. December 2012, 2015.
- [46] E. Curi, "Algoritmos evolutivos para agrupar informacion biomedica en un numero desconocido de grupos," pp. 35–44.
- [47] T. T. Cis, "Minería de Datos para segmentación de clientes en la empresa tecnológica Master PC," 2015.
- [48] A. Elena and S. Ávila, "El uso de herramientas tecnológicas de minería de datos en el análisis de datos climatológicos," vol. 7, 2018.
- [49] "OpenRefine." [Online]. Available: <http://openrefine.org/>. [Accessed: 13-Jan-2019].
- [50] "RStudio - RStudio." [Online]. Available: <https://www.rstudio.com/products/RStudio/>. [Accessed: 13-Jan-2019].
- [51] "R: ¿Qué es R?" [Online]. Available: <https://www.r-project.org/about.html>. [Accessed: 13-Jan-2019].
- [52] "SPSS Statistics - Visión general - España | IBM." [Online]. Available: <https://www.ibm.com/es-es/products/spss-statistics>. [Accessed: 13-Jan-2019].
- [53] "RapidMiner Studio." [Online]. Available: <https://rapidminer.com/products/studio/>. [Accessed: 13-Jan-2019].
- [54] H. Díaz-barrios, Y. Alemán-rivas, L. Cabrera-hernández, and A. Morales-hernández, "Algoritmos de aprendizaje automático para clasificación de Splice Sites en secuencias genómicas Machine Learning algorithms for Splice Sites

- classification in genomic sequences,” vol. 9, no. 4, pp. 155–170, 2015.
- [55] K. B. Eckert, “Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos Analysis of Attrition-Retention of College Students Using Classification Technique in Data Mining,” vol. 8, pp. 3–12, 2015.
- [56] P. Santana Mansilla, R. Costaguta, and D. Missio, “Aplicación de algoritmos de clasificación de minería de textos para el reconocimiento de habilidades de E-tutores colaborativos,” *Intel. Artif.*, vol. 17, no. 53 SPEC. ISS., pp. 57–67, 2014.
- [57] O. M. C. Pineda, “CATEGORIZACIÓN AUTOMÁTICA DE TWEETS SOBRE EL TEMA POLÍTICO ELECTORAL APLICANDO ALGORITMOS DE CLASIFICACIÓN SUPERVISADA,” 2017.
- [58] A. Isabel and V. Moreno, “Técnicas estadísticas en Minería de Textos,” 2017.
- [59] M. D. Q. Villavicencio, “Aplicación de Algoritmos Genéticos en la Ingeniería del Software: Revisión Sistemática del Estado del Arte,” 2017.
- [60] J. M. I. Banegas, “Revisión Sistemática de Literatura : Seguridad en Ambientes Web Utilizando Framework,” 2016.
- [61] F. J. García peñalvo, “Revisión sistemática de literatura en los Trabajos de Final de Máster y en las Tesis Doctorales,” *Grial*, 2017.
- [62] J. M. ORTIZ-LOZANO, A. RUA VIEITES, and P. BILBAO CALABUIG, “Aplicacion De Arboles De Clasificacion a La Deteccion Precoz De Abandono En Los Estudios Universitarios De Administracion Y Direccion De Empresas,” *Rev. Electrónica Comun. y Trab. ASEPUMA*, vol. 18, no. 1, pp. 177–201, 2017.
- [63] O. Andrés and M. Medina, “Algoritmo de Clustering Basado en el Concepto de Densidad Atómica,” pp. 758–764, 2016.
- [64] y A. F. Juan Pablo Braña, Alejandra M.J. Litterio, Cristina Camós, “Modelo de Sentiment Analysis para la clasificación de noticias en tiempo real en el Mercado de Valores de Buenos Aires,” 2014.
- [65] E. B. Sanz, “Algoritmos de clustering y aprendizaje automático aplicados a Twitter,” 2016.

k. Anexos

Anexo 1

Loja, 06 de agosto del 2018

Señor Ingeniero

Edison Leonardo Coronel Romero.

GESTOR ACADEMICO DE LA CARRERA DE INGENIERIA EN SISTEMAS

Ciudad. -

De mi consideración:

Pablo Leonardo Valdivieso Orellana, estudiante del noveno ciclo paralelo "B" de la Carrera que se encuentra bajo su dirección, solicito a usted muy comedidamente se me conceda información relacionada con los Proyectos de Titulación aprobados en nuestra Facultad, información que me servirá para poder presentar la propuesta de proyecto de trabajo de titulación, asignatura de aprobación en la malla curricular en nuestra Carrera.

Por la atención que se sirva dar al presente me suscribo de usted.

Atentamente,



Pablo Leonardo Valdivieso Orellana

PETICIONARIO

RECIBIDO POR:	Elisa
FECHA:	06/08/2018
HORA:	11:00

Anexo 2

Entrevista a la secretaria de la Secretaría General de la Facultad de Energía, las Industrias y los Recursos Naturales No Renovables	
Cargo: Secretaria de la Secretaría General de la FEIRNNR.	Fecha: 18/09/2018.
Nombre: Lic. Rosalba Jaramillo Zúñiga.	
Objetivo: Obtener información relevante acerca del conjunto de datos de los proyectos de titulación aprobados con el fin de que la información obtenida permita establecer correctamente la categorización de los proyectos de titulación.	
Descripción	
Pregunta 1: ¿Cada que tiempo se actualiza la información de los proyectos?	
Respuesta: En el conjunto de datos de proyectos de tesis se añade nueva información de forma manual una vez por ciclo, esta información es correspondiente a todos los proyectos de titulación que han sido aprobados para su desarrollo.	
Pregunta 2: ¿Cuál es la herramienta usada para administrar la información de los proyectos?	
Respuesta: El conjunto de datos es un archivo Excel, el cual consta de varias hojas, de las cuales la hoja "TESIS DE GRADO, MEMORIAS TÉCNIC" es la única válida puesto que las otras eran para dar seguimiento a los graduados, pero se decidió hacer un nuevo archivo para hacer ese seguimiento, así que están incompletas las otras pestañas del archivo Excel.	
Pregunta 3: ¿De qué manera está conformada la información en la hoja de cálculo "TESIS DE GRADO, MEMORIAS TÉCNIC"?	
Respuesta: En la hoja de cálculo "TESIS DE GRADO, MEMORIAS TÉCNIC" se encuentran varias columnas, de las cuales se explican cada una: En la columna "INGRESO" se ingresa la fecha en la que se registra el proyecto de titulación aprobado. La columna "PROYECTO" muestra la numeración del proyecto en el conjunto de datos. La columna "NOMBRE DE LA TESIS" presenta el tema de cada proyecto de titulación, cabe destacar que en algunos proyectos las palabras se encuentran mal escritas o con faltas ortográficas. La columna "POSTULANTES" contiene el nombre del autor o los nombres de los autores en caso de que el proyecto de titulación lo desarrollo más de una persona. La columna" ESPECIALIDAD" indica la especialidad a la que pertenece ese proyecto de titulación, pero se debe tener en cuenta que hay proyectos de titulación de maestrías y tecnologías que son especialidades que actualmente no oferta la facultad	

de energía. También hay casos en los que el nombre vario un poco pero que se refiere a la misma especialidad, por ejemplo: ingeniería electromecánica e ingeniería en electromecánica.

La columna "FECHA DE APROBACION" que contiene las fechas en que los proyectos de titulación fueron aprobados para su desarrollo.

La columna "DIRECTOR" tiene el nombre o los nombres de los directores asignados a cada proyecto.

La columna "TIEMPO DE EJECUCION" muestra el tiempo estimado para la duración del proyecto en meses.

La columna "OBSERVACIONES" generalmente se encuentra vacía, pero hay casos en los que sirve para indicar cuando un proyecto fue anulado o abandonado, también cuando se quiere indicar cambios en objetivos o el tema y cuando el estudiante se gradúa por otra modalidad diferente a la del proyecto de titulación.

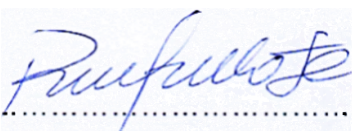
La columna "FECHA DE GRADO", contiene las fechas de grado solamente cuando el estudiante se graduó sin importar la modalidad. Si en el campo observaciones no indica que se graduó el estudiante por otra modalidad pues indica que el proyecto culminó con éxito.

Pregunta 4: ¿Qué representan los colores de las filas del conjunto de datos?

Respuesta: Las filas del conjunto de datos en su gran mayoría se encuentran pintadas, en las cuales las que tienen un color celeste fuerte son el encabezado de las columnas, las de color celeste claro son los proyectos de titulación que han terminado con éxito, las de color rojo indican cuando un proyecto ha fracasado, y las de color amarillo no ayudan a determinar un estado, ese color solo era usado para marcar proyectos de titulación en donde terminaba cada actualización del conjunto de datos, este color se dejó de usar así que no aporta mayor relevancia.

Conclusión: La información recolectada permitió conocer el contenido de cada una de las variables del conjunto de datos de los proyectos, permitiendo así identificar cuando un proyecto culminó con éxito o fracasó.

Firma:


.....
Lic. Rosalba Jaramillo Zúñiga.

Anexo 3

Loja, 29 de agosto del 2018

Tecn.

Sandra Castillo Orellana.

RESPONSABLE DE BIBLIOTECA DE LA FEIRNNR

Ciudad. -

De mi consideración:

Pablo Leonardo Valdivieso Orellana, estudiante del décimo ciclo paralelo "B" de la Carrera que se encuentra bajo su dirección, solicito a usted muy comedidamente se me conceda información relacionada con los Proyectos de Titulación de la Facultad, información que me servirá para poder continuar con la realización de mi proyecto de titulación.


Por la atención que se sirva dar al presente me suscribo de usted.

Atentamente,



Pablo Leonardo Valdivieso Orellana

PETICIONARIO

Recibido

29/12/2018.
15400

Anexo 4

Entrevista a la bibliotecaria de la Biblioteca de la Facultad de Energía, las Industrias y los Recursos Naturales No Renovables

Cargo: Bibliotecaria de la Biblioteca de la FEIRNNR.

Fecha: 28/09/2018.

Nombre: Tecn. Sandra Castillo Orellana.

Objetivo: Indagar cuáles son los proyectos de titulación que contiene el conjunto de datos de la biblioteca para conocer si son de utilidad en la evaluación de los proyectos.


Descripción

Pregunta: ¿Cómo está conformada la información contenida en el conjunto de datos de los proyectos?

Respuesta: El conjunto de datos de la biblioteca contiene información de todas las tesis que se han desarrollado con éxito, consta de siglas por cada carrera e indica si el proyecto de titulación pertenece a una tecnología, grado o postgrado. También presenta el tema de cada proyecto de titulación, el director de cada proyecto y el año en el que culminó el proyecto.

Conclusión: Los proyectos culminados con éxito son los que conforman el conjunto de datos de la Biblioteca, permitiendo evaluar si los proyectos de la Secretaría General se encuentran correctamente clasificados.

Firma:


.....
Tecn. Sandra Castillo Orellana.

Anexo 5

Loja, 11 de diciembre del 2018

Señor Ingeniero

Hernán Leonardo Torres Carrión.

GESTOR ACADEMICO DE LA CARRERA DE INGENIERIA EN SISTEMAS

Ciudad. -

De mi consideración:

Pablo Leonardo Valdivieso Orellana, estudiante del décimo ciclo paralelo "B" de la Carrera que se encuentra bajo su dirección, solicito a usted muy comedidamente se me conceda información relacionada con los Proyectos de Titulación aprobados para su desarrollo en la Carrera de Ingeniera en Sistemas, información que me servirá para poder continuar con la realización de mi proyecto de titulación.

Por la atención que se sirva dar al presente me suscribo de usted.

Atentamente,



Pablo Leonardo Valdivieso Orellana

PETICIONARIO

RECIBIDO POR:	<u>E^osa</u>
FECHA:	<u>11/12/2018</u>
HORA:	<u>09:45</u>

Anexo 6

Entrevista a la Secretaria de la Carrera de Ingeniería en Sistemas de la Facultad de Energía, las Industrias y los Recursos Naturales No Renovables

Cargo: Secretaria de la Carrera de Ingeniería en Sistemas de la FEIRNNR.

Fecha: 11/12/2018.

Nombre: Prof. Elisa Beatriz Orellana Bravo

Objetivo: Comprender la estructura del nuevo conjunto de datos para realizar correctamente la categorización de los nuevos proyectos.

Descripción

Pregunta 1: ¿Cómo es administrada la información de los proyectos?

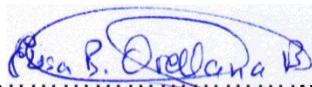
Respuesta: El conjunto de datos de los proyectos que se maneja actualmente es el obtenido previamente de la Secretaria General, a partir de ahí se trabajó con ese conjunto de datos actualizando datos e ingresando nuevos proyectos de titulación. En este nuevo conjunto de datos los nuevos proyectos ingresados no cuentan con datos en el campo “tiempo de ejecución”, pero en el campo “observaciones” se encuentran todas las indicaciones de los proyectos, es decir en caso de haber sido anulado, abandonado, renuncia o haberse graduado en modalidad de examen complejo.

Pregunta 2: ¿De qué manera es posible identificar los proyectos actualizados y los nuevos proyectos introducidos?

Respuesta: Mediante contenido del campo “observación”, en los proyectos actualizados la letra cuenta con negrita lo que permite una rápida identificación de los proyectos a los que se les ha actualizado un estado. Los últimos proyectos introducidos no cuentan con tiempo de ejecución, pero se encuentran en desarrollo debido a que los proyectos fueron aprobados en fechas muy cercanas a esta entrevista, e inclusive de los proyectos recién aprobados para su desarrollo solo existe un proyecto que fue anulado y al cual se ingresó la información de anulado en observaciones.

Conclusión: La estructura del nuevo conjunto de datos de los proyectos es la misma que el conjunto de datos obtenido en Secretaria General, por lo que es factible realizar el mismo proceso para categorizar los proyectos de titulación.

Firma:



Prof. Elisa Beatriz Orellana Bravo

Anexo 7

Lic. Rosalba Jaramillo Zúñiga

**ASISTENTE DE SECRETARIA GENERAL DE LA FACULTAD DE *ENERGÍA*,
LAS INDUSTRIAS Y LOS RECURSOS NATURALES NO RENOVABLES DE
LA UNIVERSIDAD NACIONAL DE LOJA**

CERTIFICO:

Que se otorgó al Sr. Pablo Leonardo Valdivieso Orellana con número de cédula 1150025748, estudiante de la Carrera de Ingeniería en Sistemas, toda la información requerida en relación a los proyectos de los Trabajo de Titulación que reposan en la Secretaria de General de la Facultad de Energía, información que le servirá para el desarrollo de su trabajo de titulación denominado **“APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN PARA DETERMINAR COMO INFLUYE LA DEFINICIÓN DEL TEMA DE UN PROYECTO DE TITULACIÓN EN SU ÉXITO O FRACASO”**.

Es todo cuanto puedo certificar en honor a la verdad, facultando al interesado hacer uso del presente documento en lo que se estime conveniente.


.....

Lic. Rosalba Jaramillo Zúñiga

Loja, 18 de Septiembre del 2018

Anexo 8


Tecn. Sandra Castillo Orellana

RESPONSABLE DE BIBLIOTECA DE LA FACULTAD DE *ENERGÍA*, LAS INDUSTRIAS Y LOS RECURSOS NATURALES NO RENOVABLES DE LA UNIVERSIDAD NACIONAL DE LOJA

CERTIFICO:

Que se otorgó al Sr. Pablo Leonardo Valdivieso Orellana con número de cédula 1150025748, estudiante de la Carrera de Ingeniería en Sistemas, toda la información requerida en relación a los proyectos de los Trabajo de Titulación que reposan en Biblioteca de la Facultad de Energía, información que le servirá para el desarrollo de su trabajo de titulación denominado **“APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN PARA DETERMINAR COMO INFLUYE LA DEFINICIÓN DEL TEMA DE UN PROYECTO DE TITULACIÓN EN SU ÉXITO O FRACASO”**.

Es todo cuanto puedo certificar en honor a la verdad, facultando al interesado hacer uso del presente documento en lo que se estime conveniente



Tecn. Sandra Castillo Orellana



Loja, 28 de septiembre del 2018

Anexo 9



Prof. Elisa Orellana Bravo

**SECRETARIA DE LA CARRERA DE INGENIERÍA EN SISTEMAS -
COMPUTACIÓN DE LA FACULTAD DE *ENERGÍA*, LAS INDUSTRIAS Y LOS
RECURSOS NATURALES NO RENOVABLES DE LA UNIVERSIDAD
NACIONAL DE LOJA**

CERTIFICO:

Que se otorgó al Sr. Pablo Leonardo Valdivieso Orellana con número de cédula 1150025748, estudiante de la Carrera de Ingeniería en Sistemas, toda la información requerida en relación a los proyectos de los Trabajo de Titulación que reposan en la Secretaria de la Carrera de Ingeniería en Sistemas y Computación de la Facultad de Energía, información que le servirá para el desarrollo de su trabajo de titulación denominado **“APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN PARA DETERMINAR COMO INFLUYE LA DEFINICIÓN DEL TEMA DE UN PROYECTO DE TITULACIÓN EN SU ÉXITO O FRACASO”**.

Es todo cuanto puedo certificar en honor a la verdad, facultando al interesado hacer uso del presente documento en lo que se estime conveniente


.....
 UNIVERSIDAD NACIONAL DE LOJA
FACULTAD DE ENERGÍA, LAS INDUSTRIAS Y
LOS RECURSOS NATURALES NO RENOVABLES
CARRERA DE INGENIERIA EN
SISTEMAS - COMPUTACION
SECRETARIA DE CARRERA

Prof. Elisa Beatriz Orellana Bravo

Loja, 11 de diciembre del 2018

Anexo 10

Análisis del conjunto de datos de los proyectos obtenido en Secretaría General.

Objetivo: Obtener el número de proyectos que han fracasado en la facultad haciendo uso de los proyectos que se encuentran categorizados.

Desarrollo: El conjunto de datos es un archivo Excel que consta de 10 hojas. A continuación, se muestra una captura del mismo en la Figura 110.

REGISTRO DE APROBACIÓN DE PROYECTOS DE TESIS								
INGRESO	PROYECTO	NOMBRE DE LA TESIS	POSTULANTES	ESPECIALIDAD	FECHA DE APROBACION	DIRECTOR	TIEMPO DE EJECUCIÓN	OBSERVACIONES
133		DISEÑO Y CONSTRUCCIÓN DE UN AEROGENERADOR EXPERIMENTAL MODULAR PARA APLICACIÓN RURAL	Oscar Iván Cabrera González Darwin Rigoberto Cuenca Quinde	Ingeniería Electromecánica	11/05/2006	Ing. Jorge Luis Maldonado Correa	11 meses	
134		METODOLOGÍA PARA LA APLICACIÓN DE ENESYOS NO DESTRUCTIVOS MEDIANTE ULTRASONIDO	Vicente Manuel Álvarez Vega, Carlos Manuel Jimbo Muñoz, Galo Fernando Medina Rivera, Alex Patricio Ordóñez Pico	Ingeniería Electromecánica	19/05/2006	Ing. Ramiro Gonzalo Riofrio Cruz	10 meses	
135	anulado	APLICACIÓN PARA EL CONTROL DE GARANTÍAS DE EQUIPOS Y COMPONENTES DE COMPUTACIÓN APOYADOS EN LOS PROCESOS DE FACTURACIÓN E INVENTARIO UTILIZANDO JAD COMO METODOLOGÍA INTEGRADORA DE APLICACIONES PARA UN RÁPIDO DESARROLLO Y UML COMO METODOLOGÍA DE MODELADO	Mónica María Díaz Samaniego, Pablo Oswaldo Córdova García	Ingeniería en Sistemas	30/06/2006	Franco Salcedo López	6 meses	anulado 5 ago
136	ANULADO	CONSTRUCCIÓN DE UN SISTEMA AUTOMATIZADO DE FERTIRRIGACIÓN PARA LOS TERRENOS DEL PROYECTO RECURSOS FITOGENÉTICOS DEL ÁREA AGROPECUARIA Y DE LOS RECURSOS NATURALES RENOVABLES	Hernán Diego León Abarca, Juan Carlos Rodríguez Bequerro, Juan Carlos Villamagua Armijos	Ingeniería Electromecánica	29/06/2006	Ing. Darwin Geovanny Tapia Peralta	11 meses	ANULADO 08/07/2011
137		ELABORACION DE UN PORTAL WEB PARA LA CAMARA DE LA CONSTRUCCIÓN DE LOJA	Yadira Jhoana Banegas Michay y César Leonardo Delgado Correa	Ingeniería en Sistemas	18/07/2006	Ing. Hernán Leonardo Torres Carrión	9 meses	
138		DISEÑO DE UN PLAN PILOTO DE TELEMETRÍA Y CONTROL DEL CONSUMO DE ENERGÍA.	Edgar Geovanny Alvarez chevez Jairo Patricio Cabrera Medina	Ingeniería Electromecánica	18/07/2008	Ing. Juan Carlos Solano Jiménez	10 meses	Reglame Estructu
139		DISEÑO Y CONSTRUCCIÓN DE UN TABLERO DIDÁCTICO DE TRANSFERENCIA DE ENERGÍA ELÉCTRICA PARA EL LABORATORIO DE ELECTRICIDAD	Vinicio Antonio Ortega Guamán y Milton Leonán León Aguilera	TECNOLOGÍA EN ELECTRICIDAD	19/07/2006	Ing. Juan Carlos Ochoa Alfaro HCA: 16/06/2011 cambia director a 2. Ing. José Espinosa S, Ing. Norman Augusto Jiménez León	5 meses	Reglame Estructu
140		SISTEMA INFORMÁTICO PARA LA ELABORACION Y RESOLUCIÓN DE EXÁMENES EN LÍNEA, TUTILIZANDO LA TECNOLOGÍA JAVA Y EL	María Magdalena Armijos JumboSilvia	Ingeniería en	20/07/2006	Ing. Wilman Patricio Chamba		Reglame Estructu

Figura 110. Captura del conjunto de datos de los proyectos.

La hoja "TESIS DE GRADO, MEMORIAS TÉCNIC" es la única válida, el resto contiene información incompleta debido a que fueron creadas para dar seguimiento a los graduados de las carreras de la facultad. Esta hoja contiene proyectos de las carreras que se ofertan actualmente en la Facultad y también de las tecnologías que se ofertaban anteriormente.

El total de proyectos que contiene el conjunto de datos es de 1502 proyectos de los cuales algunos se encuentran categorizados, es decir los proyectos de color celeste que se observan en la imagen anterior son aquellos que culminaron con éxito, pero sí en cambio los proyectos se encuentran pintados de color rojo esto indica que los proyectos fracasaron.

Para contabilizar el número de proyecto que se encuentran categorizados se realizó un nuevo conjunto de datos en el cual están todos los proyectos que se encuentran pintados, cabe destacar que en muchos casos los temas de los proyectos se encuentran en celdas combinadas tal como se muestra en la Figura 111.

16	CAMBIO DE NIVELES DE ILUMINACIÓN EN EL BLOQUE DE AULAS NÚMERO UNO DEL ÁREA DE ENERGÍA, LAS INDUSTRIAS Y LOS RECURSOS NATURALES NO RENOVABLES DE LA UNL		
17			
18			
19	CÁLCULO, DISEÑO Y RECONSTRUCCIÓN DE LOS CIRCUITOS DE ILUMINACIÓN EN EL EDIFICIO CENTRAL DEL AREA DE ENERGÍA.....		HCA:07/07/05 autoriza continuen los señores Díaz y Peralta; HCA: 21/06/06 incluye Juan Pablo Ponce Toledo
20			
21			
22			
23			
24	ESTUDIO DE LA INFRAESTRUCTURA NECESARIA (HARDWARE Y SOFTWARE) PARA EL DESARROLLO DE APLICACIONES WAP		cor 17/
25	DISEÑO Y CONSTRUCCIÓN DE UN TABLERO DIDÁCTICO PARA UNA CENTRALILLA TELEFÓNICA		cor 17/
26	HERRAMIENTA DE MONITOREO PARA LA RED DE LA UNIVERSIDAD NACIONAL DE LOJA		11/
27			19/
28	DISEÑO Y CONSTRUCCIÓN DE UN BANCO DE PRUEBAS HIDRAULICAS		
29			
30			
31	DSEÑO E IMPLEMENTACIÓN DE UN SISTEMA DE MONITOREO QUE PERMITA EVALUAR LA EFICIENCIA ENERGÉTICA E HIDRÁULICA DEL HOSPITAL REGIONAL ISIDRO AYORA		
32			
33	Diseño, Construcción y Montaje de un Mezclador y un sistema Elevador de Agregados Pétreos para una Planta Adoquinera Bloquera en la Universidad Nacional de Loja		
34			
35			

Figura 111. Temas de los proyectos en celdas combinadas.

Al estar los proyectos en celdas combinadas se dificulta el conteo, por lo que se extrajeron solamente los temas de los proyectos en una nueva hoja de cálculo y en esta nueva hoja de cálculo se des combinaron las celdas de los temas, luego se agregó otra columna denominada “Número de proyectos” para contabilizar los proyectos y posterior a ello en la nueva columna se usó la fórmula: “=SI(ESBLANCO(A2); B1+0;B1+1)”. La fórmula planteada en “Número de proyectos” permite aumentar el conteo de proyectos siempre que la celda en la columna de los temas no esté vacía.

1033	APLICACIÓN DE ALGORITMOS GENÉRICOS EN LA INGENIERÍA DEL SOFTWARE: REVISIÓN SISTEMÁTICA DEL ESTADO DEL ARTE	777
1034	GENERACIÓN SOBRE LA BASE SIG DEL MAPA GEOMORFOLÓGICO ESCALA 1:25 000, ORIENTADO A LA GESTIÓN TERRITORIAL DEL CANTÓN OLMEDO, PROVINCIA DE LOJA	778
1035	METODOLOGÍA PARA EJECUTAR MEDICIONES DE CALIDAD DEL SERVICIO ELÉCTRICO EN BAJA TENSIÓN	779
1036	DISEÑO DE UN ELEVADOR ELECTROMECÁNICO BI PERSONAL PARA MANTENIMIENTO INTERIOR DE ALTURA DE LA BASÍLICA DE NUESTRA SEÑORA DE EL CISNE	780
1037	GENERACIÓN SOBRE LA BASE SIG DEL MAPA GEOMORFOLÓGICO ESCALA 1:25 000, ORIENTADO A LA GESTIÓN TERRITORIAL DEL CANTÓN BALSAS, PROVINCIA DE EL ORO	781
1038	REVISIÓN SISTEMÁTICA DE LITERATURA: VULNERABILIDAD DE SISTEMAS BIOMÉTRICOS	782
1039	SUSCEPTIBILIDAD A DESLIZAMIENTOS EN LA VÍA DE INTEGRACIÓN BARRIAL CON UN TRAMO DE 1.5 KM DESDE EL REDONDEL DE CARIGAN HACIA LA AVENIDA 8 DE DICIEMBRE, UBICADA EN LA CIUDAD, CANTÓN Y PROVINCIA DE LOJA	783
1040	PROPUESTA PARA EL RECICLAJE DE ACEITES LUBRICANTES USADOS PROVENIENTES DE TALLERES AUTOMOTRICES Y LUBRICADORAS EXISTENTES EN LA CIUDAD DE CARIAMANGA	784
1041	PROPUESTA TÉCNICA PARA LA REDUCCIÓN DE LOS ACCIDENTES DE TRÁNSITO EN LA CIUDAD DE LOJA, DESDE EL PUNTO DE VISTA HUMANO-VEHÍCULO-EQUIPAMIENTO AMBIENTAL-	785
1042	ALIMENTACIÓN DE UN MOTOR MONO CILÍNDRICO CON HIDRÓGENO OBTENIDO A TRAVÉS DE LA ELECTRÓLISIS DE AGUA	786
1043	DISEÑO Y CONSTRUCCIÓN DE UN TABLERO DIDÁCTICO DE TRANSFERENCIA ELÉCTRICA AUTOMÁTICA DESDE UN GENERADOR	787
1044	Creación e Implementación del Portal Web para el Área de la Energía, las Industrias y los Recursos Naturales No Renovables	788
1045	IMPLEMENTACIÓN DE UN PROTOTIPO MANEJADO A CONTROL REMOTO	789
1046		789
1047	Desarrollo de un Sistema Distribuidor de Boletería, Entrega, Recepción de Encomiendas y la Generación de los Cuadros de Trabajo para la Cooperativa de Transportes Loja Internacional	790
1048		790
1049		790
1050	CONSTRUCCIÓN DE UN SISTEMA AUTOMATIZADO DE FERTIRRIGACIÓN PARA LOS TERRENOS DEL PROYECTO RECURSOS FITOGENÉTICOS DEL ÁREA AGROPECUARIA Y DE LOS RECURSOS NATURALES RENOVABLES	791
	DESARROLLO DE UN SISTEMA INFORMÁTICO DE CONTROL Y SEGUIMIENTO DE TRÁMITES Y EXPEDIENTES PARA LA UNIVERSIDAD NACIONAL DE	

Figura 112. Proyectos enumerados para contabilizarlos.

En la Figura 112 se observa a la izquierda como al des combinar las filas de los temas quedan celdas vacías, en la parte derecha se muestra como el valor aumenta cuando hay un tema. Luego de realizar el conteo se obtuvo que 786 proyectos de los 1023 culminaron con éxito, por ende 237 proyectos han fracasado en la Facultad. Si se

considera a 1023 como el 100% de los proyectos se obtiene que el 23.16% de los proyectos han fracasado.

Conclusión: Al estar 1023 proyecto ya categorizados de los 1502 en total, se obtuvo que el 23.16% de los proyectos han fracaso. Sin embargo, hay proyectos que no se encuentran categorizados e inclusive algunos de ellos aún están desarrollándose, por lo que no es posible contabilizar en su totalidad todos los proyectos que han fracasado. Tomando en cuenta lo mencionado anteriormente es evidente que el porcentaje de los proyectos que han fracaso puede variar, debido a esto el porcentaje obtenido es un valor aproximado, estableciendo así que el 20% de los proyectos desarrollados en la Facultad han fracasado.

Anexo 11

Búsqueda de los algoritmos de clasificación utilizando la metodología para obtención de información planteada en este trabajo de titulación.

Para seleccionar algoritmos eficientes con el caso de estudio del trabajo de titulación se procedió a realizar una búsqueda de estudios relacionados aplicando la metodología planteada en este trabajo para la obtención de información. El proceso para obtener los algoritmos se detalla a continuación:

- **Objetivo para la obtención de información.**

Conocer que algoritmos son aplicados en estudios similares al presente trabajo de titulación y sus características.

- **Criterios de inclusión.**

Los estudios se seleccionan tomando en cuenta los siguientes criterios:

- Fecha de publicación: Estudios realizados a partir del 2014.
- Motor de búsqueda: Google Académico.
- Idioma: español o inglés.
- Tipos de estudios: artículos, libros, revistas o trabajos relacionados.
- Contenido: Los estudios contengan al menos 3 de los 4 temas que abarca la revisión de literatura, estos temas son: aprendizaje automático, minería de texto, minería de datos y algoritmos de clasificación.

- **Criterio de exclusión.**

Se excluyo los estudios que no cumplen lo siguiente:

- Los criterios de inclusión previamente detallados.
- El objetivo planteado para la obtención de información.

- **Fuente de búsqueda.**

La fuente seleccionada para la realización de la búsqueda es Google Académico:
<https://scholar.google.com/>

- **Cadenas de Búsqueda.**

Es muy importante la identificación de palabras clave al momento de formular cadenas para poder realizar una eficiente búsqueda de información. Las cadenas de búsqueda planteadas son las siguientes:

Cadenas de Búsqueda
(algoritmos clasificación "aprendizaje automático" Y "minería de datos" O "minería de texto") O (minería de texto "minería de

datos" Y "algoritmos clasificación") O (minería de texto "aprendizaje automático" Y "algoritmos clasificación")

- **Ejecución de la consulta.**

Una vez establecidas las cadenas de búsqueda necesarias para realizar la consulta, se procede a su ejecución en la fuente de búsqueda definida, donde los estudios seleccionados deben:

- Apoyar el objetivo establecido de la revisión sistemática.
- Proporcionar un aporte extra al proceso de obtención de información.

Luego de ejecutar las consultas en la fuente de búsqueda se obtuvo un total de 180 estudios, de los cuales se debe seleccionar los relevantes.

- **Selección de los estudios primarios.**

Para la selección de los estudios se eliminó los estudios duplicados y se utilizó los criterios de inclusión y exclusión, también se tomó en cuenta partes importantes de los estudios como metodología, resultados y conclusiones.

De los estudios relacionados obtenidos en la búsqueda, se seleccionaron 12 estudios que cumplen los criterios de inclusión y apoyan al cumplimiento del objetivo, se detalla los estudios a continuación:

Id	Estudios analizados	Año de publicación
E1	Aplicación de árboles de clasificación a la detección precoz de abandono en los estudios universitarios de administración y dirección de empresas.	2017
E2	Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos.	2015
E3	Aplicación de Algoritmos de Clasificación de Minería de Textos para el Reconocimiento de Habilidades de E-tutores Colaborativos.	2014
E4	Minería de Datos para segmentación de clientes en la empresa tecnológica Master PC.	2015

E5	Técnicas de aprendizaje de máquina utilizadas para la minería de texto.	2015
E6	Algoritmo de clustering basado en el concepto de densidad atómica.	2016
E7	Modelo de Sentiment Analysis para la clasificación de noticias en tiempo real en el Mercado de Valores de Buenos Aires.	2014
E8	Algoritmos de clustering y aprendizaje automático aplicados a Twitter.	2016
E9	R and Data Mining Examples and Case Studies.	2015
E10	Técnicas estadísticas en Minería de textos.	2017
E11	Evaluación de la decisión de obtener el título profesional con la elaboración de la tesis mediante técnicas multivariantes Caso Universidad Nacional Agraria La Molina.	2017
E12	Minería de datos aplicada para la identificación de factores de riesgo en alumnos.	2017

- **Presentación de resultados.**

Los hallazgos de los estudios seleccionados se presentan a continuación:

- El estudio E1 presenta que los estudiantes abandonan sus estudios en el primer nivel universitario, por lo que plantea analizar si es factible determinar el abandono en tres momentos: al momento de inscribirse, al iniciar el curso académico y tras la realización de los primeros exámenes. Para resolver la problemática planteada se utiliza los algoritmos: AID, CHAID Exhaustivo, QUEST y CART, obteniendo así que los estudiantes que abandonan son los que presentan bajas calificaciones en sus exámenes. Los árboles de clasificación utilizados permiten encontrar estructuras en grandes conjuntos de datos, además se denota que los árboles de clasificación difieren respecto de otras técnicas de clasificación debido a que trabajan con una variable dependiente y facilitan la interpretación de los datos mediante su porcentaje de acierto y representación gráfica. La herramienta de software usada para la aplicación de los algoritmos en este estudio fue la SPSS.

- En el estudio E2 se analiza información para identificar factores que influyen en la deserción de estudiantes mediante técnicas de minería de datos, realiza selección y depuración de datos para aplicar algoritmos de clasificación como arboles de decisión, redes bayesianas y reglas de asociación. También se categoriza los datos en cuatro tipos para que sirva de variable dependiente. El software utilizado es Weka y las técnicas empleadas en esa herramienta son: C4.5 (J48) que genera un árbol de decisión, Naive Bayes (BayesNet) que es veloz en grandes conjuntos de datos y Reglas de asociación que son uno de los clasificadores más sencillos. Los algoritmos utilizados presentan un porcentaje de eficiencia y poseen representación gráfica.
- El estudio E3 aplica técnicas de minería de textos para identificar automáticamente habilidades manifestadas por e-tutores. También establece categorías para cada una de las instancias, obtener un conjunto de datos de cada clase, identificar las variables útiles y aplicar los algoritmos de clasificación. La herramienta utilizada es GATE y los algoritmos aplicados son: KNN (k vecino más cercano), SVM (máquinas de soporte vectorial), PAUM, Naive Bayes y C4.5.
- El estudio E4 obtiene la segmentación de clientes mediante técnicas de minería de datos, los algoritmos aplicados son: k-means, k-medoids, Self-Organizing Maps (SOM) y Apriori. Los algoritmos de agrupamiento permiten establecer la cantidad de agrupamientos y para validar estos algoritmos se utilizó como base al algoritmo de clasificación C5 y la herramienta utilizada para la aplicación de los algoritmos fue RStudio, esta herramienta fue seleccionada en base a una comparativa en la cual se determina que RStudio destaca del resto, pero también se presenta que RapidMiner permite una óptima visualización de los resultados.
- El estudio E5 identifica las principales formas de aprendizaje máquina empleadas en la minería de texto, se estableció que los algoritmos más usados son: máquinas de soporte vectorial, k-means, k-nearest neighbors y naive bayes. También se analizó 56 documentos de los cuales los algoritmos que aparecen con mayor frecuencia son los árboles de decisión aleatorios, reglas de asociación apriori y los mencionados anteriormente.
- El estudio E6 parte en búsqueda de un algoritmo más natural que los de agrupamiento, para ello en base al concepto de densidad atómica se generó uno nuevo utilizando los algoritmos k-means y k-medoids. La herramienta

utilizada para la aplicación de los algoritmos fue RapidMiner, los algoritmos permitieron establecer el número de clústeres y su centroide de manera aleatoria.

- El estudio E7 tiene como propósito mostrar el monitoreo automático de noticias en tiempo real mediante algoritmos de aprendizaje máquina. Se realiza la extracción, análisis y clasificación de opiniones de Twitter para formar la bolsa de palabras. Los algoritmos aplicados son: Random Forest, Naive Bayes y Support Vector Machines; estos algoritmos trabajan con un conjunto de datos clasificados manualmente por expertos.
- El estudio E8 realiza un estado del arte de los algoritmos de aprendizaje automático más importantes, así como también evaluar las herramientas de machine learning más populares. Las herramientas evaluadas son RStudio, Scikit-learn y Weka. Los algoritmos encontrados son: k-means, k-NN, SVM, Naive Bayes y CHAID Exhaustivo. De los algoritmos mencionados anteriormente los utilizados en las herramientas son k-means y Naive Bayes.
- El estudio E9 introduce el uso de RStudio para la minería de datos, presenta diversos ejemplos de casos de estudio de aplicaciones del mundo real. Los algoritmos aplicados en este estudio son las principales técnicas de minería de datos, incluyen clasificación, agrupación, reglas de asociación y minería de texto. En el apartado de minería de texto los algoritmos aplicados son: k-means, k-medoids, Dendograma y Random Forest.
- El estudio E10 explica técnicas y modelos de minería de datos, análisis de sentimientos, extracción de información y de clasificación de documentos. También presenta un listado de las herramientas de software para analizar datos en forma de texto. En los casos presentados en este estudio se obtiene la bolsa de palabras, se usa como representación gráfica la nube de palabras y posterior a ello se aplica el algoritmo Dendograma (Agrupación Jerárquica). La herramienta utilizada es RStudio y las que se analizaron son: SAS Text Miner, SPSS, Megaputer y Google Cloud Plataform. También se obtiene que la aplicación del Dendograma resulta una tarea efectiva en el conjunto de datos a pesar de su gran tamaño, debido a que elimina términos dispersos de la bolsa de palabras.
- El estudio E11 evalúa factores que explican la decisión de elaborar la tesis o no con el fin de obtener el título profesional. Se categorizo los datos en dos categorías: si decide o no realizar la tesis y se utilizó el modelo de regresión

logística. Los algoritmos aplicados son: Regresión Logística y CHAID Exhaustivo; estos algoritmos fueron aplicados en la herramienta SPSS por lo que presentan un porcentaje de eficiencia.

- El estudio E12 implementa un sistema de tutorías en una plataforma digital y realiza un procesamiento en línea para que sea más eficiente el proceso de tutorías. Los algoritmos aplicados son Apriori y k-means, finalmente esto permitió obtener patrones entre los datos y así optimizar el sistema de tutorías. La herramienta utilizada para la aplicación de los algoritmos fue Weka.

- **Conclusión.**

Luego de analizar los estudios seleccionados se obtuvo los algoritmos de clasificación utilizados, algunos de los algoritmos encontrados son: J48, Naive Bayes, K-means, K-NN, K-medoids, Apriori, Random Forest, SVM, entre otros.

Algunos estudios establecen características para los algoritmos de clasificación, por ejemplo: los estudios E1 y E2 indica que los algoritmos utilizados en ellos presentan porcentaje de eficiencia, son eficientes en grandes conjuntos de datos y poseen representación gráfica. Por otra parte, los estudios E4 y E6 se indica que los algoritmos utilizados en estos estudios son efectivos en grandes conjuntos de datos y permiten establecer la cantidad de agrupamientos.

Al analizar los estudios se obtuvo que algunos algoritmos son utilizados en más de un estudio como es el caso del algoritmo k-means en los estudios: E4, E5, E6, E8 y E9. Tomando en cuenta lo anterior algunos los algoritmos van a compartir características, es decir los algoritmos presentes en un estudio va a tener la misma característica establecida de ese algoritmo en otro estudio.

Las características de los algoritmos que se obtuvieron al analizar los estudios son: presenta porcentaje de eficiencia, posee representación gráfica, es efectivo en grandes conjuntos de datos y permite especificar la cantidad de agrupamientos.

Las herramientas más utilizadas en los estudios son Weka, RStudio, RapidMiner y SPSS. De estas herramientas, Weka no permite tener una óptima visibilidad de los resultados cuando se trabaja con grandes conjuntos de datos.

Licencia Creative Commons del Normativo



Normativo para la presentación del informe final del Proyecto Fin de Carrera por [Carrera de Ingeniería en Sistemas](#) se encuentra bajo una [Licencia Creative Commons Atribución-NoComercial-CompartirIgual 3.0 Unported](#).