UNIVERSIDAD TÉCNICA DEL NORTE

Facultad de Ingeniería en Ciencias Aplicadas

Carrera de Ingeniería en Sistemas Computacionales

IMPLEMENTACIÓN DE MINERÍA DE DATOS PARA EL ANÁLISIS DEL DESEMPEÑO ESTUDIANTIL BASADO EN LOS RECURSOS Y ACTIVIDADES DEL ENTORNO VIRTUAL DE APRENDIZAJE DEL SIIU-UTN.

Trabajo de grado previo a la obtención del título de Ingeniero en Sistemas Computacionales.

Autor:

Leonardo Moises Aguagallo Aigaje

Director:

PhD. Iván Danilo García Santillán

Ibarra – Ecuador

2022

Autorización de uso y publicación a favor de la Universidad

UNIVERSIDAD TÉCNICA DEL NORTE
BIBLIOTECA UNIVERSITARIA

AUTORIZACIÓN DE USO Y PUBLICACIÓN A FAVOR DE LA UNIVERSIDAD TÉCNICA DEL NORTE

1. IDENTIFICACIÓN DE LA OBRA

En cumplimiento del Art. 144 de la Ley de Educación Superior, hago la entrega del presente trabajo a la Universidad Técnica del Norte para que sea publicado en el Repositorio Digital Institucional, para lo cual pongo a disposición la siguiente información:

DATOS DE CONTACTO			
CÉDULA DE IDENTIDAD:	1726117581		
APELLIDOS Y NOMBRES:	Aguagallo Aigaje Leonardo Moises		
DIRECCIÓN:	Cardenal de la Torre 3-72 – Cayambe - Pichincha		
EMAIL:	lmaguagalloa@utn.edu.ec, lmmoyses.sh@gmail.com		
TELÉFONO FIJO:	-	**TELÉFONO MÓVIL:**	+593 99 473 6136

DATOS DE LA OBRA	
TÍTULO:	IMPLEMENTACIÓN DE MINERÍA DE DATOS PARA EL ANÁLISIS DEL DESEMPEÑO ESTUDIANTIL BASADO EN LOS RECURSOS Y ACTIVIDADES DEL ENTORNO VIRTUAL DE APRENDIZAJE DEL SIIU-UTN
AUTOR (ES):	Aguagallo Aigaje Leonardo Moises
FECHA: DD/MM/AAAA	30/09/2022
SOLO PARA TRABAJOS DE GRADO	
PROGRAMA:	■ **PREGRADO** ☐ **POSGRADO**
TITULO POR EL QUE OPTA:	**INGENIERIA EN SISTEMAS COMPUTACIONALES**
ASESOR /DIRECTOR:	**PhD. Iván Danilo García Santillán**

2. CONSTANCIAS

El autor (es) manifiesta (n) que la obra objeto de la presente autorización es original y se la desarrolló, sin violar derechos de autor de terceros, por lo tanto la obra es original y que es (son) el (los) titular (es) de los derechos patrimoniales, por lo que asume (n) la responsabilidad sobre el contenido de la misma y saldrá (n) en defensa de la Universidad en caso de reclamación por parte de terceros.

Ibarra, a los 10 días del mes de octubre de 2022

EL AUTOR:

(Firma).....
Nombre: Leonardo Aguagallo

Certificación del director de Trabajo de Grado

UNIVERSIDAD TÉCNICA DEL NORTE
FACULTAD DE INGENIERÍA EN CIENCIAS APLICADAS
CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

CERTIFICADO DEL DIRECTOR

En mi calidad de Tutor de Trabajo de Grado presentado por el egresado, **Leonardo Moises Aguagallo Aigaje** para optar por el Título de Ingeniero en Sistemas Computacionales, cuyo tema es: **IMPLEMENTACIÓN DE MINERÍA DE DATOS PARA EL ANÁLISIS DEL DESEMPEÑO ESTUDIANTIL BASADO EN LOS RECURSOS Y ACTIVIDADES DEL ENTORNO VIRTUAL DE APRENDIZAJE DEL SIIU-UTN**. Considero que le presente trabajo reúne los requisitos y méritos suficientes para ser sometido a la presentación pública y evaluación por parte del tribunal examinador.

En la ciudad de Ibarra, a los 30 días del mes de septiembre del 2022.

1002292603
IVAN DANILO
GARCIA
SANTILLAN

Firmado digitalmente por 1002292603 IVAN DANILO GARCIA SANTILLAN
Fecha: 2022.10.03 07:40:12 -05'00'

PhD. Iván Danilo García Santillán
DIRECTOR DE TRABAJO DE GRADO

Dedicatoria

Al que me permitió lograr lo que parecía imposible, me dio fuerza y sabiduría para lograr llegar a este punto, al Espíritu Santo. El me permitió levantarme y lograra superar la depresión y me enseño que se puede tener sueños y luchar para cumplirlos.

A mi madre, una mujer fuerte, mujer aguerrida, que no le peso trabajar cada día en la finca bajo el ardiente sol y el trabajo pesado para darme las condiciones de estudiar, a ella no le importo quitarse el pan de su boca o dejar de comprar sus cosas para darme los pasajes, que se levantó antes que saliera el sol para enviarme el almuerzo, a ti muchas gracias, madre, no se aun como lograré pagar todo lo que haces por mí, muchas gracias mama.

Leonardo Aguagallo

Agradecimiento

A todas las personas que estuvieron presentes y me ayudaron en cada una de las etapas de mi vida estudiantil, agradezco aquella palabra de aliento, aquel gesto amable, aquella mano amiga muchas gracias.

A la institución que me alojo y sus profesionales quienes compartieron sus conocimientos y experiencias. A todos lo que forman parte de la carrera de Ingeniería en Sistemas Computacionales y mi docente tutor el Ing. Iván García.

Leonardo Aguagallo

Tabla de Contenido

Autorización de uso y publicación a favor de la Universidad ii
Certificación del director de Trabajo de Grado iii
Dedicatoria iv
Agradecimiento v
Tabla de Contenido vi
Índice de Figuras x
Índice de Cuadros xii
Resumen xiv
INTRODUCCIÓN xv
Antecedentes xv
Problema xv
Objetivos xvi
Objetivo General xvi
Objetivos Específicos xvi
Justificación xvi
Alcance xvii
CAPÍTULO 1 19
Marco Teórico 19
1.1. Entornos Virtuales de Aprendizaje 19
1.1.1. Componentes 19
1.1.2. Dimensiones 19
Dimensión Tecnológica 20
Dimensión Educativa 20
1.1.3. Recursos 20
1.1.4. Actividades 20
1.2. Rendimiento académico 20
1.3. Minería de Datos 21
1.3.1. Características de la Minería de Datos 21
1.3.2. Relación de la Minería de Datos con otras áreas 21
1.4. Técnicas de Minería de Datos 23
1.4.1. Técnicas Predictivas 23
1.4.1.1. Clasificación 24
Bosques aleatorios (Random Forest) y Arboles de clasificación (Decisión Tree) 24
K Vecinos más Cercanos (K Nearest Neighbours) 25
Clasificación Bayesiana (Naive Bayes) 26
Máquinas de Vectores de Soporte (Support Vector Machine) 27

1.4.1.2. Regresión 28
1.4.2. Técnicas Descriptivas 28
1.4.2.1. Clustering o agrupación 28
K-Means 29
K-Prototype 29
K-Modes 29
Algoritmo Esperanza-Maximización (Expectation–maximization algorithm) 30
1.4.2.2. Asociación 30
Apriori 30
1.4.2.3. Correlación 30
Pearson 32
Spearman 32
Kendal 33
1.5. Proceso de descubrimiento de conocimiento (KDD) 34
1.5.1. Recopilación y Integración 35
1.5.2. Preprocesamiento 36
1.5.2.1. Selección 36
1.5.2.2. Eliminación 36
1.5.2.3. Transformación 37
1.5.3. Minería de Datos 39
1.5.4. Interpretación y Validación 39
1.6. Herramientas para Minería de Datos 40
1.6.1. Inteligencia de Negocios 40
Pentaho Data Integración 40
1.6.2. Herramientas de minería de datos 40
Weka 40
1.7. Trabajos Relacionados 40
1.8. Propuesta de Minería de Datos 42
CAPÍTULO 2 44
Proceso de Descubrimiento de Conocimientos 44
2.1. Visión General del Proyecto 44
2.2. Entregables del proyecto 44
2.3. Organización del Proyecto 45
2.3.1. Participantes del Proyecto 45
2.3.2. Roles y Responsabilidades 45
2.4. Gestión del Proceso 46
2.4.1. Estimaciones 46

2.4.2. Plan del Proyecto 47
2.5. Fase: Recopilación y Integración de Datos 47
2.5.1. Recopilación de datos 47
2.5.2. Integración de Datos 53
2.5.3. Data Warehouse 57
2.6. Fase: Preprocesamiento 57
2.6.1. Selección 57
2.6.2. Transformación 59
a. Discretización 59
b. Normalización 63
2.6.3. Eliminación 65
2.6.4. Vista Minable 67
2.7. Fase: Minería de Datos 69
2.7.1. Correlación 69
a. Correlación de atributos continuos 70
b. Correlación de atributos ordinales 74
2.7.2. Modelo de Clasificación 76
Selección de variables independientes 77
Random Forest 79
Support Vector Machine 80
2.7.3. Agrupación 81
K-Means 82
K-Prototype 82
K-Modes 83
2.7.4. Asociación 83
CAPÍTULO 3 84
Validación de Resultados 84
3.1. Fase: Validación y Interpretación 84
3.1.1. Validación y Análisis 84
3.1.1.2. Validación e análisis de Correlación 84
3.1.1.2. Validación e análisis del modelo de predicción 86
3.1.1.3. Validación e análisis de agrupación 89
3.1.1.4. Validación e análisis de asociación 91
3.1.2. Interpretación de Resultados 92
Interpretación e análisis de Coeficientes de Correlación 92
Interpretación e análisis de Clasificación 96
Interpretación e análisis de Agrupación 97

Interpretación e análisis de Asociación 99
3.3. Obtención de conocimiento 100
3.4. Discusión 101
CONCLUSIONES 104
RECOMENDACIONES 105
BIBLIOGRAFÍA 106

Índice de Figuras

Fig. 1. Árbol de problemas xvi
Fig. 2. Alcance de la propuesta xviii
Fig. 3. Relación de la ciencia de datos con otras áreas según Karoussi 22
Fig. 4. Relación de la ciencia de datos con otras áreas 22
Fig. 5. Técnicas de Minería de datos 23
Fig. 6. Bosque aleatorio 25
Fig. 7. Estructura del árbol de clasificación. 25
Fig. 8. K Vecinos más cercanos. 26
Fig. 9. Estructura de Naive Bayes. 27
Fig. 10. Support Vector Machine 27
Fig. 11. Aumento de dimensión en SVM 28
Fig. 12. Agrupación K-Means. 29
Fig. 13. Proceso KDD 35
Fig. 14. Matriz de Propuesta 43
Fig. 15. Integración datos socioeconómicos 54
Fig. 16. Integración datos académicos 55
Fig. 17. Integración datos de interacciones 56
Fig. 18. Construcion del Data Warehouse 57
Fig. 19. Porción del Data Warehouse 57
Fig. 20. Filtrado de atributos 58
Fig. 21. Filtrado de registros 59
Fig. 22. Uso Number Ranges 63
Fig. 23. Uso Replace in string 63
Fig. 24. Eliminación de campos vacíos – atributo nota final 66
Fig. 25. Eliminación de campos vacíos – atributo actividades y recursos 66
Fig. 26. Eliminación de campos vacíos – atributos socioeconómicos 66
Fig. 27. Atributos de la vista minable 67
Fig. 28. Vista minable cualitativa 67
Fig. 29. Vista minable cuantitativa 68
Fig. 30. Fase de Procesamiento de datos 68
Fig. 31. Dataset continuo 70
Fig. 32. Histograma de los datos continuos 71
Fig. 33. Resultados de las pruebas de normalidad 72
Fig. 34. Resultados de la correlación de Pearson 72
Fig. 35. Resultados de la correlación de Spearman 72
Fig. 36. Dataset de ordinales 73

Fig. 37. Resultados de la correlación de Kendall 73
Fig. 38. Dataset de nominales 74
Fig. 39. Algoritmo punto biserial 74
Fig. 40. Dataset de ordinales 74
Fig. 41. Resultados de correlación de Kendall 75
Fig. 42. Resultados de correlación de Spearman 75
Fig. 43. Dataset de nominales 76
Fig. 44. Resultados de correlación de Spearman 76
Fig. 45. Preselección de variables independientes 77
Fig. 46. Grado de correlación de variables independientes 78
Fig. 47. Modelo XGBoost 79
Fig. 48. Nivel de importancia de las variables independientes 79
Fig. 49. Modelo Random Forest 80
Fig. 50. Resultados del entrenamiento del modelo RF 80
Fig. 51. Modelo Support Vector Machine 81
Fig. 52. Resultados de entrenamiento del modelo SVM 81
Fig. 53. Algoritmo Simple K-Means 82
Fig. 54. Algoritmo K-Prototype 82
Fig. 55. Algoritmo K-Modes 83
Fig. 56. Algoritmo de Apriori para reglas de asociación 83
Fig. 57. Matrix de confusión del modelo Random Forest 87
Fig. 58. Curva ROC del modelo Random Forest 87
Fig. 59. Matrix de confusión del modelo Support Vector Machine 88
Fig. 60. Curva ROC del modelo Support Vector Machine 88
Fig. 61. Grado de correlación de las variables – análisis 1 93
Fig. 62. Grado de correlación de las variables – análisis 2 95
Fig. 63. Clusters con K-Means 97
Fig. 64. Clusters con K-Prototype 97
Fig. 65. Clusters con K-Modes 98
Fig. 66. Reglas de decisión del uso de recursos 100
Fig. 67. Reglas de decisión de porcentaje en faltas 100

Índice de Cuadros

TABLA 1.1 INTERPRETACIÓN DE CORRELACIÓN 31
TABLA 1.2 ELECCIÓN DE MEDIDAS DE ASOCIACIÓN 32
TABLA 1.3 SIMBOLOGÍA DE CORRELACIÓN PEARSON 32
TABLA 1.4 SIMBOLOGÍA DE CORRELACIÓN SPEARMAN 33
TABLA 1.5 SIMBOLOGÍA DE CORRELACIÓN KENDALL 34
TABLA 1.6 SIMBOLOGÍA NORMALIZACIÓN MIN-MAX 37
TABLA 1.7 SIMBOLOGÍA NORMALIZACIÓN POR EL MÁXIMO 38
TABLA 1.8 SIMBOLOGÍA DISCRETIZACIÓN MISMO ANCHO 38
TABLA 2.1 ENTREGABLES DE PROYECTO 44
TABLA 2.2 DIRECTORES DE LAS ÁREAS 45
TABLA 2.3 PARTICIPANTES DIRECTOS 45
TABLA 2.4 ROLES Y RESPONSABILIDADES 45
TABLA 2.5 TALENTO HUMANO DEL PROYECTO 46
TABLA 2.6 RECURSOS MATERIALES 46
TABLA 2.7 COSTO TOTAL DEL PROYECTO 47
TABLA 2.8 FASES DEL PROYECTO Y DISTRIBUCIÓN DE HORAS 47
TABLA 2.9 ESTRUCTURA PERSONA 48
TABLA 2.10 ESTRUCTURA ETNIA 48
TABLA 2.11 ESTRUCTURA DISCAPACIDAD 48
TABLA 2.12 ESTRUCTURA SOCIOECONOMICO 48
TABLA 2.13 ESTRUCTURA INGRESOS 49
TABLA 2.14 ESTRUCTURA MATRICULA 49
TABLA 2.15 ESTRUCTURA DETALLE_MATRICULA 50
TABLA 2.16 ESTRUCTURA NOTAS 50
TABLA 2.17 ESTRUCTURA DEPENDENCIA 51
TABLA 2.18 ESTRUCTURA CICLO_ACADEMICO 51
TABLA 2.19 ESTRUCTURA NUMERO_ACTIVIDADES 52
TABLA 2.20 ESTRUCTURA NUMERO_RECURSOS 52
TABLA 2.21 ESTRUCTURA RECURSOS_SIIU 52
TABLA 2.22 ESTRUCTURA ACTIVIDADES_SIIU 52
TABLA 2.23 DISCRETIZACIÓN ATRIBUTO EDAD 59
TABLA 2.24 SBU Y CBF DE LOS ÚLTIMOS 5 AÑOS 60
TABLA 2.25 DISCRETIZACIÓN ATRIBUTO INGRESOS 60
TABLA 2.26 DISCRETIZACIÓN ATRIBUTO ETNIA 60
TABLA 2.27 DISCRETIZACIÓN ATRIBUTO DISCAPACIDAD 61
TABLA 2.28 DISCRETIZACIÓN ATRIBUTO PORCENTAJE FALTAS 61

TABLA 2.29 DISCRETIZACIÓN LOS ATRIBUTOS DE LAS NOTAS 61
TABLA 2.30 DISCRETIZACIÓN NÚMERO DE ACTIVIDADES 62
TABLA 2.31 DISCRETIZACIÓN NÚMERO DE RECURSOS 62
TABLA 2.32 NUMERIZACIÓN ATRIBUTO CARRERA 63
TABLA 2.33 NUMERIZACIÓN ATRIBUTO NUMERO ACTIVIDADES Y RECURSOS 63
TABLA 2.34 NUMERIZACIÓN ATRIBUTO NOTA FINAL 64
TABLA 2.35 NUMERIZACIÓN DE ATRIBUTOS NOMINALES 64
TABLA 2.36 NUMERIZACIÓN ATRIBUTO EDAD 64
TABLA 2.37 NUMERIZACIÓN DE ATRIBUTOS GENERO 64
TABLA 2.38 NUMERIZACIÓN ATRIBUTO ESTADO CIVIL 64
TABLA 2.39 NUMERIZACIÓN ATRIBUTO DISCAPCIDAD 65
TABLA 2.40 NUMERIZACIÓN ATRIBUTO ETNIA 65
TABLA 2.41 NUMERIZACIÓN ATRIBUTO INGRESOS 65
TABLA 2.42 VARIABLES INDEPENDIENTES 77
TABLA 3.1 NIVEL DE SIGNIFICANCIA EN ATRIBUTOS CONTINUOS 84
TABLA 3.2 NIVEL DE SIGNIFICANCIA EN ATRIBUTOS ORDINARIOS 85
TABLA 3.3 MEDIDAS ESTADÍSTICAS DEL MODELO RF 87
TABLA 3.4 MEDIDAS ESTADÍSTICAS DEL MODELO RF 88
TABLA 3.5 CLÚSTERS CON K-MEANS 89
TABLA 3.6 CLÚSTERS CON K-MEANS 90
TABLA 3.7 CLÚSTERS CON K-MEANS 90
TABLA 3.8 REGLAS DE ASOCIACIÓN – SET DE DATOS I 91
TABLA 3.9 REGLAS DE ASOCIACIÓN – SET DE DATOS II 91
TABLA 3.10 COEFICIENTE DE CORRELACIÓN DE LOS ATRIBUTOS CONTINUOS 92
TABLA 3.11 COEFICIENTE DE CORRELACIÓN DE LOS ATRIBUTOS ORDINALES 94
TABLA 3.12 ERRORES DE LOS MODELOS 96

Resumen

El desempeño académico de los estudiantes forma parte de los aspectos relevantes a tratar en relación con la calidad de la Educación de una institución, además constituye un indicador de la realidad educativa del establecimiento. El presente trabajo permitió identificar factores que influyen en el rendimiento académico, patrones de uso-rendimiento, reglas de asociación y generar un modelo de predicción del éxito académico, por medio de la aplicación de técnicas descriptivas y predictivas de minería de datos. Se analizó un conjunto de datos conformado por datos socioeconómicos, académicos y interacciones con el Entorno Virtual de Aprendizaje (EVA) de los años 2017 a 2018, con total de 26 atributos y 57 115 instancias. Aplicando la metodología de descubrimiento de conocimientos en bases de datos (KDD) se desarrolló cada una de las fases de la minería de datos. Los resultados más relevantes mostraron que las notas de las parciales o evaluaciones, porcentaje de faltas, la entrega de pruebas y trabajos son factores que influyen en el desempeño académico.

Palabras clave: rendimiento académico, minería de datos, técnicas predictivas, técnicas descriptivas, correlación, agrupación, clasificación, predicción, KDD.

INTRODUCCIÓN

Antecedentes

El rendimiento académico de los estudiantes universitarios constituye un aspecto importante al momento de tratar el tema de la calidad de la Educación Superior, en vista que se trata de un indicador que aproxima la realidad educativa de una institución. (De Miguel Díaz, Apocada, Arias, Escudero, Rodríguez & Vidal, 2002).

Los diferentes cuestionamientos sociales en la relación al costo – beneficio de la educación ha producido en las autoridades universitarias un especial interés por los resultados académicos de los estudiantes, en vista que un estudio o análisis constituye una sólida herramienta para la construcción de indicadores que guíen en la toma de decisiones en la educación superior (Garbanzo, 2007).

Tal es el caso de la Universidad Técnica del Norte (UTN) que es una institución de educación superior que desarrolla su labor académica e investigativa para la Zona 1 (Esmeraldas, Carchi, Imbabura y Sucumbíos) del Ecuador, esta cuenta con un Sistema Integrado Institucional Universitario (SIIU) que integra y relaciona los procesos de las diferentes áreas, entre ellas la académica (Chamorro, 2018), pero se ha observado la carencia de una herramienta que permita analizar el rendimiento académico de los estudiantes.

En SIIU ha venido recolectando la información de los estudiantes, docentes, periodos académicos, calificaciones, actividades y recursos entre otras. Esta información puede ser útil para ser analizada con técnicas de minería de datos en busca de factores que influyan en el rendimiento académico de los estudiantes.

Problema

El rendimiento académico es un punto fundamental a la hora de abordar la Calidad de la Educación superior. El estudio del rendimiento académico es una herramienta útil para el descubrimiento de indicadore para la toma de decisiones académicas, además se constituye un tema de particular interés para las autoridades de la institución. En la UTN, se ha observado la carencia de una herramienta que permita analizar el rendimiento académico de los estudiantes aprovechando los datos provenientes del SIIU, esto ha generado un desconocimiento de los factores que

influyen (actividades y recursos) en el rendimiento académico, provocando de alguna manera la baja calidad en la educación y sus posibles efectos.

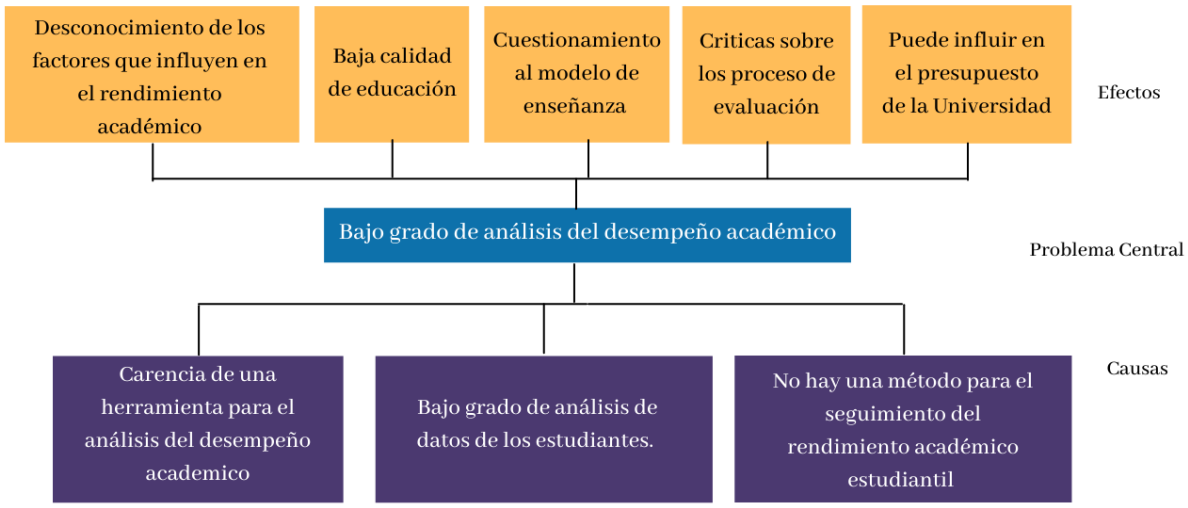


Fig. 1. Árbol de problemas

Objetivos

Objetivo General

Implementar minería de datos para el análisis del desempeño estudiantil basado en los recursos y actividades del Entorno virtual de aprendizaje del SIIU-UTN.

Objetivos Específicos

Elaborar un marco teórico que sustente las técnicas de minería de datos en entornos virtuales de aprendizaje.

Implementar algunas técnicas de minería de datos para el análisis descriptivo y/o predictivo del rendimiento académico en el entorno virtual de aprendizaje del SIIU-UTN.

Validar los resultados obtenidos utilizando métricas cuantitativas utilizadas en la ciencia de datos.

Justificación

Dentro de los Objetivos y metas de desarrollo sostenible (ODS) el objetivo 4 cita a la Educación de Calidad, especificando que los alumnos adquieran conocimientos necesarios para promover su desarrollo sostenible en la meta 4.7 (Naciones Unidas,

2015). Por esa razón es importante realizar un análisis del rendimiento académico con el fin de contribuir a la mejora de la educación de la UTN.

Justificación Tecnológica

La minería de datos se ha tornado una herramienta frecuentemente empleada en la visualización de resultados, la educación en Ecuador con el uso de la analítica de datos pretende tomar decisiones relevantes basándose en la gestión y el crecimiento educativo del país. (Tejada, Murrieta, Villao & Garzón, 2018).

El análisis en base a técnicas de minería de datos y sus resultados proporcionará conocimientos necesarios para la toma de decisiones y de esa forma contribuir a la mejora de la Calidad de Educación en la UTN.

Justificación Social

El rendimiento académico de un alumno es un valor que se le atribuye a los logros del estudiante en las diferentes actividades académicas, la cual es medida de forma cuantitativa. (Pérez, 2005, como se citó en Garbanzo, 2007).

El presente trabajo se basará en técnicas de minería de datos y sus conceptos, buscando brindar soluciones que mitiguen la problemática, por medio del proceso de descubrimiento de conocimientos en las bases de datos recolectadas por el SIIU. De esta forma se pretende contribuir al proceso de toma de decisiones mediante el análisis del rendimiento académico, de esa forma se espera minimizar los posibles efectos que genera el bajo rendimiento académico.

Alcance

El presente trabajo de grado tiene la finalidad identificar factores que influyen en desempeño académico, creación de un modelo de predicción, obtener patrones de uso-rendimiento académico y descubrir reglas de asociación existentes. Para la siguiente propuesta se contará con una fuente de datos proporcionada por el Departamento de Informática, correspondiente a los últimos 5 años, pertenecientes a la Facultad de Ingeniería en Ciencias Aplicadas (FICA) de la Universidad Técnica del Norte.

El desarrollo del mismo se hará uso de la metodología de descubrimiento de conocimientos en bases de datos o KDD. Para la extracción, transformación, y

limpieza de datos (ETL) se pretende utilizar la Spoon de la Suite Pentaho en un ambiente local. Posterior al proceso al proceso ETL se pretende obtener una vista minable con las variables de interés, misma en la que se aplicarán técnicas descriptivas y predictivas de minería de datos utilizando las herramientas de Python, R Studio y Weka. Luego se validaría los resultados obtenidos con las métricas cuantitativas de las respetivas técnicas. Finalmente se procederá a interpretar los resultados obtenidos. La Fig. 2 muestra el alcance de la propuesta.

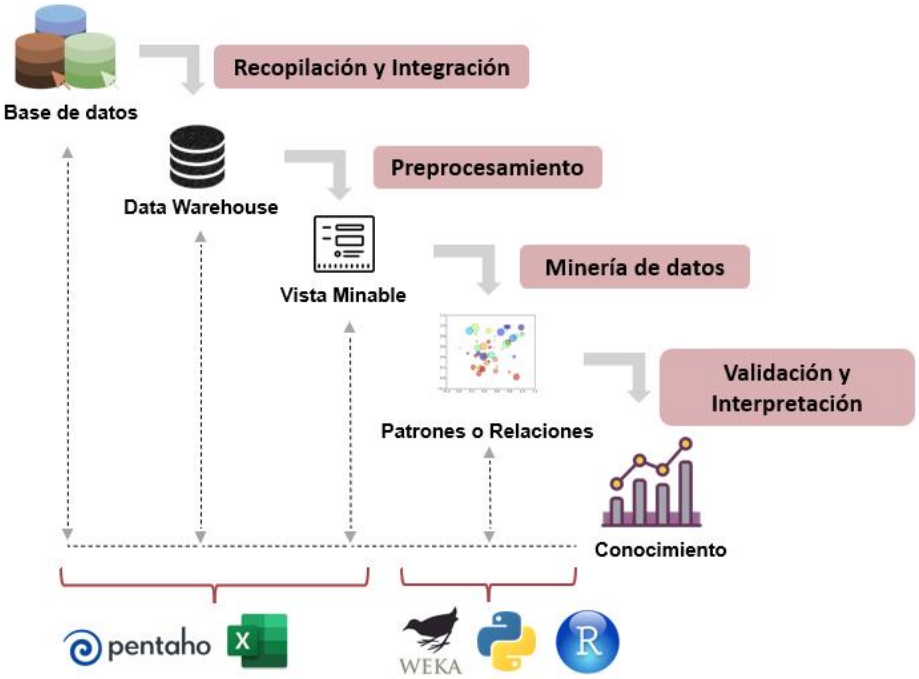


Fig. 2. Alcance de la propuesta

CAPÍTULO 1

Marco Teórico

1.1. Entornos Virtuales de Aprendizaje

Un Entorno Virtual de Aprendizaje (EVA) es un conjunto de medios para una interacción sincrónica y asincrónica, que permite efectuar el proceso enseñanza y aprendizaje, por medio de un sistema de administración de aprendizaje (Trejo,2013).

Salinas (2011) señala que un EVA es un espacio educativo hospedado en internet, que se conforma por un conjunto de herramientas informáticas, posibilitando la interacción didáctica con los estudiantes. El mismo que posee cuatro características básicas:

a) Un entorno virtual construido por tecnologías digitales.
b) Se encuentra alojado en la red por lo que se puede acceder remotamente.
c) Sirve como un complemento en las actividades académicas y gestión administrativa.
d) Desarrollo de actividades en tiempo real y posterior a su lanzamiento.

1.1.1. Componentes

Según Vintimilla (2015) manifiesta que los componentes de un EVA son:

a) Gestión y distribución de contenido
b) Comunicación
c) Recursos personales
d) Trabajo colaborativo
e) Evaluación y seguimiento

1.1.2. Dimensiones

La composición de estos entornos nos da a entender que poseen una dimensión tecnológica y una educativa, y cada una se interrelacionan y potencia entre sí (Salinas, 2011).

Dimensión Tecnológica

Está representada por herramientas o aplicaciones informáticas que se usaron en la construcción del entorno (Salinas, 2011). Estos artefactos que conforman el entorno sirven como soporte de las diversas propuestas educativas a desarrollarse. De forma independiente un entorno debe brindan las siguientes acciones básicas: publicación de actividades, comunicación, colaboración y organización.

Dimensión Educativa

La dimensión educativa es un componente que se desarrolla al interior del entorno, el cual está dado por el proceso de enseñanza – aprendizaje (Salinas, 2011). Se caracteriza por ser un espacio humano y social, donde existe una interrelación entre el docente y el alumno, basado en el planteo y la resolución de actividades.

1.1.3. Recursos

Los recursos son todos los componentes que permiten acceder al contenido educativo disponible dentro del Entorno de virtual de aprendizaje.

Según Marqués (2000), se considera un recurso a todo tipo de materiales que tiene la finalidad ya sea didáctica o sirve de apoyo en el desarrollo de las diversas actividades académicas.

1.1.4. Actividades

Se considera actividades todos los trabajos propuestos a los estudiantes con la finalidad de contribuir a la comprensión, análisis, síntesis y valoración de los contenidos que conforman las diferentes materias, y el desarrollo de la actividad permite transformar la información en conocimiento, habilidades y actitudes referente al área de estudio (Cabero & Román, 2006).

1.2. Rendimiento académico

El rendimiento académico considerado también un sinónimo del desempeño académico se relaciona tanto con una métrica para valorar las evaluaciones, así como una forma de medir las habilidades y actitudes de un estudiante (Coello & Cachón, 2017). Se puede definir también como el grado de los logros obtenidos en los hitos de una planificación educativa.

1.3. Minería de Datos

El incesante crecimiento de datos juntamente con el uso de dispositivos electrónicos en tareas empresariales y cotidianas ha impulsado el almacenamiento de información, esto ha abierto un mundo de posibilidades en la búsqueda de conocimiento en los grandes volúmenes de datos.

Según Witten, Frank & Hall (2011) la minería de datos busca solucionar problemas mediante el análisis de los datos que se encuentran presentes en una base de datos.

Buczak & Guven (2016) mencionan que la minería de datos se enfoca en el descubrimiento de propiedades desconocidas, centrándose en encontrar nuevos e interesantes conocimientos y no solo un objetivo en particular.

La minería de datos se define como un proceso para encontrar patrones significativos que contribuyen con alguna ventaja. Los patrones descubiertos nos dan la posibilidad de hacer predicciones no triviales sobre nuevos datos (Witten et al., 2011).

1.3.1. Características de la Minería de Datos

Jiménez (2015) sostiene que la característica principal de la minería de datos es el descubrimiento inductivo de información y patrones ocultos en los datos. Añade también otra característica como es la exploración de datos almacenada durante años en bases de datos y repositorios. Según Pesantez (2016) detalla las siguientes características:

- Permite la recolección a gran escala, unificando los datos de todas las bases de datos disponibles, internas o externas.
- Procesamiento de grandes volúmenes de datos por medio de la combinación de procesadores y recursos informáticos.
- Búsqueda de información oculta aplicando herramientas algorítmicas.

1.3.2. Relación de la Minería de Datos con otras áreas

Karoussi (2012) establece una relación en base a las diferentes técnicas y métodos que se fundamenta la minería de datos como se muestra en la Fig. 3:

- Estadística: resaltando el muestreo, la estimación y probar hipótesis a partir de la estadística.

- Inteligencia artificial: uso y aplicación de algoritmos de búsqueda, técnicas de modelado, técnicas de machine learning.

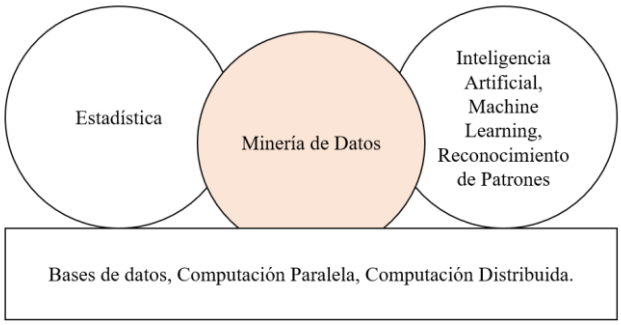


Fig. 3. Relación de la ciencia de datos con otras áreas según Karoussi (Elaboración propia)

Lara (2014) muestra una relación más amplia como se puede observar en la Fig.4.

- Estadística: Un gran número de técnicas que se usa en la minería de datos nacen de los conceptos y técnicas estadísticas.
- Bases de Datos: El proceso KDD habitualmente parte de los datos almacenados en las diversas bases de datos disponibles.
- Visualización: La minería de datos tiene como objetivo la búsqueda de conocimiento, el cual debe ser representado para ser de utilidad. Para lo cual se usa diagramas, gráficos, etc.
- Aprendizaje automático: Su relación se basa en la obtención y la aplicación de modelos del aprendizaje automático.
- Otras áreas: la minería de datos es muy versátil y se relacionan con áreas como los sistemas de apoyo de decisión, recuperación de información tratamiento y procesamientos de datos.

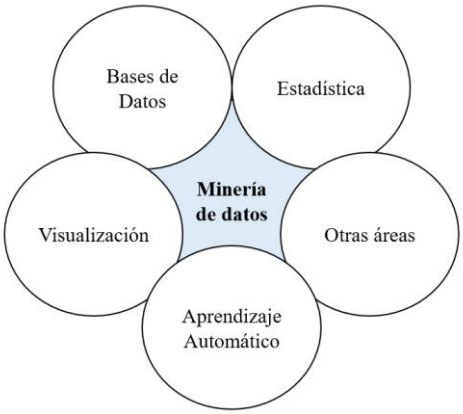


Fig. 4. Relación de la ciencia de datos con otras áreas (Elaboración propia)

1.4. Técnicas de Minería de Datos

Según Pérez & Santín (2007), una clasificación dentro de las técnicas de minería de datos se diferencia entre las técnicas predictivas donde las variables al inicio se pueden clasificar en dependientes e independientes, las técnicas descriptivas las cuales al inicio tienen el mismo estatus. La Fig.5 podemos visualizar las técnicas de minería de datos y sus subclasificaciones.

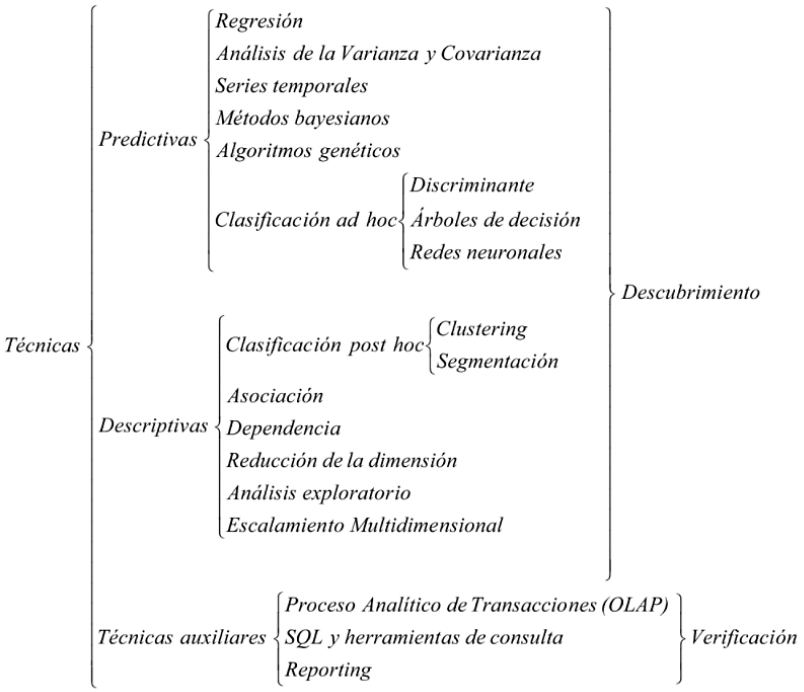


Fig. 5. Técnicas de Minería de datos (Pérez & Santín, 2007)

Las técnicas de clasificación pueden pertenecer al grupo de predictivas cuando clasifican dentro de grupos ya establecidos, y al grupo de descriptivas cuando clasifican sin una definición previa de los grupos.

1.4.1. Técnicas Predictivas

El propósito es predecir un valor desconocido de una variable que por lo general es un valor futuro (Jain & Srivastava, 2013), esta variable puede ser llamada como variable dependiente u objetivo, mientras que la variable independiente la que se usa para la predicción.

Según Tamilselvi & Kalaiselvi (2013), las técnicas predictivas establecen el modelo de datos a partir del conocimiento previo obtenido. Este modelo requiere ser

entrenado para lo cual se usa en set de datos de esa forma proporcionara resultados en base al aprendizaje.

En resumen, las técnicas predictivas inducen los datos disponibles con el fin de hacer predicciones en bases a estos datos.

1.4.1.1. Clasificación

La clasificación consiste en definir a que grupo o conjunto de datos pertenece un elemento, dicho de otra forma, se encarga de identificar las características y atributos que hacen que un elemento pertenezca a un grupo (Martínez, 2012).

La clasificación es una de las técnicas más comúnmente aplicadas, esta hace uso de un conjunto de datos y determina el modelo en base a los resultados que se va obteniendo. Es sumamente adecuada para detección de fraudes y riesgos crediticios, ya que emplea arboles de decisión o algoritmos de clasificación basados en redes neuronales.

Según Ramageri (2010), el proceso de clasificación consta de aprendizaje y clasificación:

- En el aprendizaje, por medio de un algoritmo de clasificación se analizan los datos de entrenamiento.
- En la clasificación, se utilizan los datos para estimar la precisión de las reglas de clasificación.

En el caso que la exactitud es aceptable, las reglas podrán ser usadas para nuevos sets de datos.

Entre las principales técnicas de clasificación tenemos: Bosques aleatorios, Arboles de clasificación, K vecinos más cercanos, clasificación bayesiana.

Bosques aleatorios (Random Forest) y Arboles de clasificación (Decisión Tree)

Los bosques aleatorios o Random Forest es una técnica sumamente versátil capaz de realizar tareas de regresión como de clasificación. De la misma forma es usado para tareas de reducción de dimensión y tratamientos de valores atípicos. Sus ventajas incluyen un alto grado de precisión, identificación de las variables más significativas, recupera de forma eficiente valores perdidos (Cutler D., Edwards T., Beard K., Cutler A., Hess K., Gibson J. & Lawler J., 2007)

El algoritmo es un clasificador que se compone de un conjunto de árboles de clasificación $\{h(x,k),\ k=1,\dots\}$ donde k corresponde a vectores independientes distribuidos de forma uniforme a partir del conjunto de datos de entrada, y cada árbol arroja un voto a una clase de la entrada x (Breiman, 2001), donde el resultado es dado por la clase con mayor número de votos dentro del bosque de árboles de clasificación. La Fig.6 muestra la estructura del bosque aleatorio.

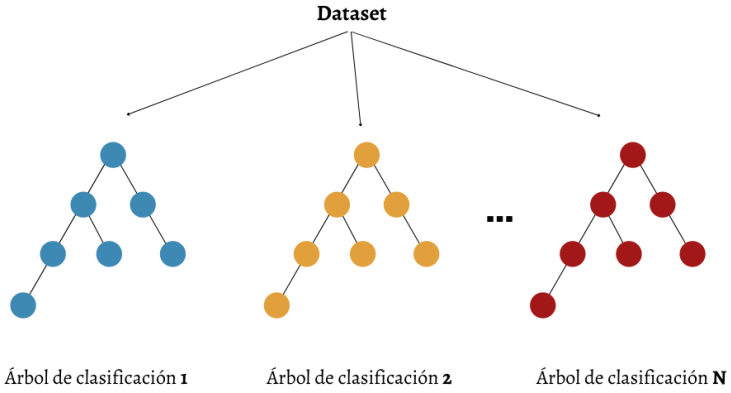


Fig. 6. Bosque aleatorio.

Los bosques aleatorios tienen su base en los árboles de clasificación, que es un conjunto de datos estructurados en forma de árbol, y están compuestos por un nodo raíz, nodos de prueba y nodos internos o hojas (Franco, 2010), donde las variables independientes conforman las condiciones o nodos de prueba y las variables dependientes representaría la predicción o las hojas. La Fig. 7 muestra la estructura de un árbol de clasificación.

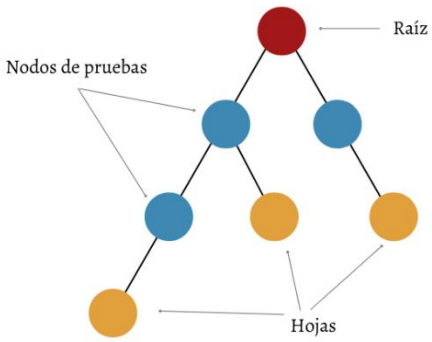


Fig. 7. Estructura del árbol de clasificación.

K Vecinos más Cercanos (K Nearest Neighbours)

K vecinos más cercanos (KNN) es una técnica que puede ser empleada en tareas de clasificación, así como de predicción, es una técnica no paramétrica (Fukunaga K.

& Narendra P., 1975), lo que significa que pueden tener su punto de partida con muestras compuestas por parámetros no específicos, dichos de otra forma no siguen un orden lógico o una distribución normal para hacer inferencias. KNN también se ha denominado un clasificador basado en instancias (Álvarez & Mayo, 2019), esto se refiere a que no utiliza un modelo de forma explícita si no que utiliza estancias de formación de una base de datos para realizar una predicción.

El objetivo del algoritmo de KNN, es la clasificación en función al mayor número de K vecinos, que son un conjunto de datos más próximos al objetivo en un espacio de datos, para hacer uso de KNN se requiere encontrar el valor óptimo de K, la validación cruzada es uno de los métodos más usados para determinar K (Izurieta & Moyano, 2019). La Fig. 8 muestra el algoritmo de KNN.

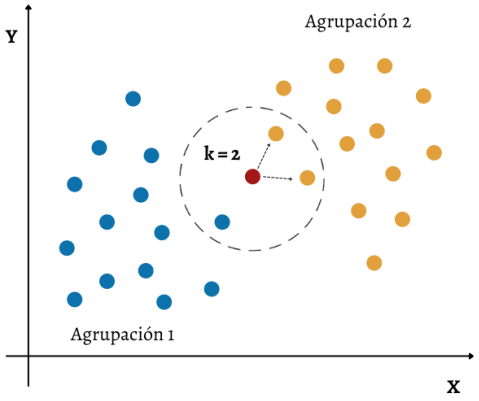


Fig. 8. K Vecinos más cercanos.

Clasificación Bayesiana (Naive Bayes)

El falsificador bayesiano o Naive Bayes es un algoritmo de clasificación que se fundamenta en el teorema de Bayes, la cual es una técnica de clasificación estadística. La forma en que funciona es por medio de la creación de un modelo grafico probabilístico compuesto por nodos y arcos. En el cual las variables están representadas por los nodos y los arcos es la dependencia existente entre las variables (Rodríguez, Pedro & Cruz, 2013). En la Fig. 9 se puede comprender la definición.

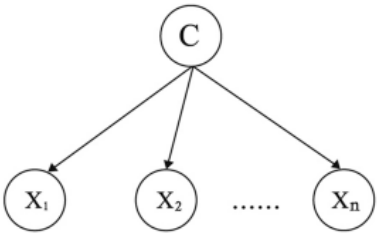


Fig. 9. Estructura de Naive Bayes.

La clasificación se da cuando se aplica la regla de bayes para calcular la probabilidad del nodo raíz en base a las probabilidades de cada una de las variables individual y posteriormente prediciendo la variable con mayor probabilidad (Friedman, Geiger & Goldszmidt, 1997). Una de sus ventajas, no requiere un aprendizaje estructural, sino aprende las probabilidades de sus variables y asume que las características son independientes entre sí.

Máquinas de Vectores de Soporte (Support Vector Machine)

Es un clasificador discriminativo que se basa en encontrar un hiperplano de separación entre dos categorías de forma que la distancia entre el hiperplano y los puntos de datos de la clase sea la máxima (Buczak A., & Guven E., 2016), la Fig. 10 muestra su descripción grafica.

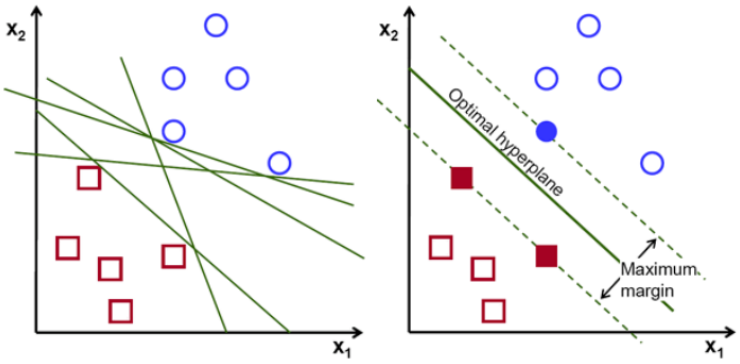


Fig. 10. Support Vector Machine

El hiperplano es una recta que divide al plano en dos subespacios donde cada categoría se encuentra en uno, el algoritmo calcula el hiperplano optimo a fin de que cuando haya una nueva entrada de datos se pueda identificar a que categoría pertenece, esto se aplica cuando los datos son linealmente separables.

Cuando los datos no son linealmente separables se puede optar por extender las dimensiones, convirtiendo al plano en uno de tres dimensiones para la separación de

las clases o categorías existentes (Amat J., 2020). La Fig. 11 muestra el aumento de dimensiones.

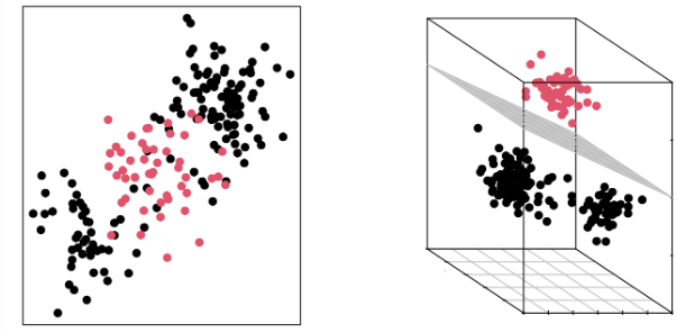


Fig. 11. Aumento de dimensión en SVM (Amat J., 2020)

1.4.1.2. Regresión

Es una técnica predictiva frecuentemente usada para determinar la relación existente entre las variables dependientes y las variables independientes (Aldowah, Al-Samarraie & Fauzy, 2019). Es una función a la cual se le asigna un valor real a un elemento para la predicción de datos futuros, hace usos de técnicas estadísticas como es la regresión lineal (Martínez, 2012). Dentro de las técnicas de regresión podemos considerar las siguientes:

1.4.2. Técnicas Descriptivas

Se enfoca en descubrir patrones que representen la información que puede ser interpretada en base al conjunto de datos asignados (Tamilselvi et al, 2013). No se contempla la existencia de variables dependientes o independientes y el modelo de datos se crea automáticamente a partir de los distintos patrones obtenidos (Perez et al., 2007). En conclusión, podemos afirmar que las técnicas descriptivas obtienen las propiedades de los datos con el fin de obtener patrones. Entra las principales técnicas descriptivas tenemos el clustering y asociación.

1.4.2.1. Clustering o agrupación

Consiste en identificar clases similares de objetos, a través del uso de técnicas de agrupamiento se hace posible identificar más regiones densas y dispersas en el espacio de objetos y de esa forma descubrir patrones de distribución general, así como correlaciones entre los atributos de los datos (Ramageri, 2010). Y los datos que no pertenecen a estas clases de objetos se los considera como valores atípicos.

K-Means

K-means es un algoritmo perteneciente al análisis de conglomerados que tiene por función la agrupación de los datos teniendo por criterio de agrupación la similitud entre ellos. Dado un conjunto de datos el algoritmo agrupa en K grupos distintos, siendo K un numero entero determinado por el analista (Cáceres, 2019).

Para la agrupación el algoritmo selecciona k puntos de forma aleatoria dentro del conjunto de datos los cuales se llaman centroides, el conjunto restante es asignado al clúster del centroide más cercano, este proceso se repite hasta que las agrupaciones sean similares o se cumpla el número de repeticiones establecidas (Esteves, Hacker & Rong, 2013).

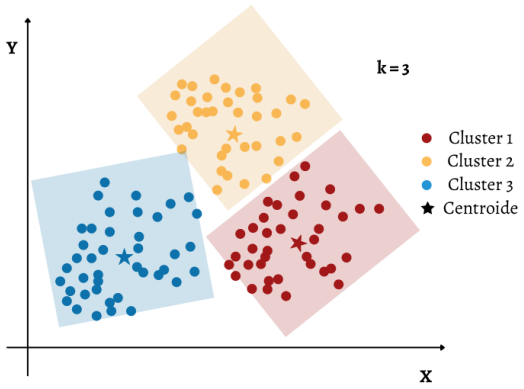


Fig. 12. Agrupación K-Means.

K-Prototype

K-prototype es un algoritmo de agrupamiento particional donde cada objeto formara parte de un clúster, este algoritmo es adecuado para conjunto de datos mixtos formados por atributos continuos y categóricos. La búsqueda de centroides para datos continuos se realiza por medio de la distancia euclidiana y medida de disimilitud de concordancia simple para datos categóricos (Szepannek G., 2018). Es considerado una combinación de K-Means y K-Modes.

K-Modes

Es un algoritmo de agrupamiento no paramétrico usado en especial para datos categóricos, se basa en el algoritmo de K-Means pero con distinta manera de calcular la distancia entre dos puntos, además remplaza los centroides por la moda. La distancia entre dos puntos se calcula por medio de la medida de disimilitud de

concordancia simple que toma el valor de cero si es similar y 1 si es distinto para cada atributo (Ng M., Li M., Huang J. & He Z., 2007)

Algoritmo Esperanza-Maximización (Expectation–maximization algorithm)

El algoritmo Esperanza-Maximización (EM) funciona de forma iterativa con la finalidad obtener estimadores de máxima verosimilitud. En el algoritmo EM se presenta la idea que dentro del conjunto de datos además de los datos observables existe datos ocultos o perdidos (Ávila, 2018). La máxima verosimilitud hace referencia a un método que permite ajustar un modelo y estimar los parámetros.

El proceso de EM consiste primero, calcula la probabilidad logarítmica del conjunto de datos y los parámetros del modelo. Segundo, los parámetros del modelo son actualizados con los resultados anteriores. Estos pasos se repiten hasta la convergencia (Tomas & Sousa, 2007).

1.4.2.2. Asociación

Esta técnica es frecuentemente utilizada para descubrir relaciones entre variables y grupos de atributos de un determinado patrón de entrada (Ramageri, 2010). Se conforma de un grupo de atributos y reglas de asociación, las cuales se usan para cuantificar la relación existente de entre dos o más atributos (Larose & Larose, 2014)

Apriori

El Apriori es un algoritmo con un enfoque iterativo que implementa la búsqueda por niveles con el fin buscar reglas de asociación, este parte de un k-conjunto de elementos para explorar (k+1) conjunto de elementos (Sumithra & Paul, 2010).

Este algoritmo hace su ejecución en dos fases, en la primera busca los ítemsets frecuentes dentro del conjunto de datos. En la segunda fase transforma los itemset previamente obtenidos en reglas de asociación (Amat, 2018). El resultado son reglas de asociación estas den ser evaluadas para eliminar reglas redundantes.

1.4.2.3. Correlación

La correlación busca saber la asociación o que tan relacionadas se encuentran dos variables, para esto es necesario conocer el grado de correlación o el coeficiente de correlación que se representa por la letra *r*.

Según Chok N. (2010) dos variables *x* e *y* están asociadas cuando el valor que toma una variable afecta la distribución de la otra variable. Szmidt E. & Kacprzyk J. (2011) mencionan que la correlación muestra el grado en el que dos variables con distribución normal se mueven juntas de forma lineal. El coeficiente de correlación tiene un límite superior de +1 que indica una correlación positiva, un límite inferior de -1 indicando una correlación negativa y si r es igual 0 no hay correlación, la Tabla 1.1 muestra la interpretación de los coeficientes.

TABLA 1.1
INTERPRETACIÓN DE CORRELACIÓN

r	Interpretación de relación
-1	Correlación negativa perfecta
-0.9 a -0.99	Correlación negativa muy alta
-0.7 a -0.89	Correlación negativa alta
-0.04 a -0.69	Correlación negativa moderada
-0.2 a -0.39	Correlación negativa baja
-0.01 a -0.19	Correlación negativa muy baja
0	No existe correlación
0.01 a 0.19	Correlación positiva muy baja
0.2 a 0.39	Correlación positiva baja
0.4 a 0.69	Correlación positiva moderada
0.7 a 0.89	Correlación positiva alta
0.9 a 0.99	Correlación positiva muy alta
1	Correlación positiva perfecta

Fuente: Adaptación de Khamis H. (2008)

Para que una correlación sea estadísticamente significativa el nivel de significancia debe ser menor a 0.05 para aceptar como valida la asociación.

La selección de la correcta medida de asociación es vital para no obtener resultados erróneos, la selección de estas medidas se basa en las características de las variables de estudio, es decir en su nivel de medición (continua, ordinal y nominal). Khamis H. (2008) propone un marco de selección de las medidas de asociación en base a las variables de estudio en la Tabla 1.2.

TABLA 1.2
ELECCIÓN DE MEDIDAS DE ASOCIACIÓN

	VARIABLE X		
VARIABLE Y	NOMINAL	ORDINAL	CONTINUA
NOMINAL	φ o λ	Rango Biserial	Biserial Puntual
ORDINAL	Rango Biserial	τ_b o Spearman	τ_b o Spearman
CONTINUA	Biserial Puntual	τ_b o Spearman	Pearson o Spearman

φ = coeficiente de phi, λ = lambda de Goodman y Kruskal, τ_b = Tau-b de Kendall

Fuente: Khamis H. (2008)

Entre los principales métodos para medir asociaciones esta la correlación de Pearson, Spearman y Kendall.

Pearson

Es unos de los métodos más utilizados y conocidos, pero para determina la asociación entre dos variables por este método es necesario que las variables sean continuas y tenga una distribución normal. Pearson calcula la relación entre la covarianza de las dos variables y el producto de sus desviaciones estándar (Chok, 2010) y el coeficiente de correlación se denota con la letra r. La Ec.1 muestra la fórmula de la correlación de Pearson y la Tabla 1.3 su simbología.

$$r = \frac{n\,\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[n\,\Sigma x^2 - (\Sigma x)^2][n\,\Sigma y^2 - (\Sigma \mathrm{y})^2]}} \quad \text{Ec. 1}$$

Donde:

TABLA 1.3
SIMBOLOGÍA DE CORRELACIÓN PEARSON

Símbolo	Descripción
n	Tamaño de muestra
x, y	Variables de estudio
Σ	Sumatoria

Fuente: Elaboración Propia

Spearman

Se usa la correlación de Spearman cuando las variables no tiene distribución normal en variables continuas. También, podemos usar cuando las variables son ordinales con 5 o más categorías. El coeficiente de correlación se denota como *rho* o r_s. La Ec. 2 muestra la fórmula de Spearman y la Tabla 1.4 muestra su simbología.

$$r_s = 1 - \frac{6\,\Sigma\,D^2}{n\,(n^2-1)} \qquad \text{Ec. 2}$$

Donde:

TABLA 1.4
SIMBOLOGÍA DE CORRELACIÓN SPEARMAN

Símbolo	Descripción
n	Tamaño de muestra
D^2	Diferencia de los rangos de las variables

Fuente: Elaboración Propia

Kendal

La correlación de Kendall o Tau de Kendall, este método se basa en intervalos de observación llamados pares concordantes o discordantes. El coeficiente de correlación se denota con la letra τ (tau). Este método se conforma por tres fórmulas matemáticas para varias situaciones las Ec. 3 – 5 muestran cada una de ellas y la Tabla 1.5 muestra su simbología.

- Tau-a es la forma más sencilla ya que no toma en cuenta valores repetidos o categorías (niveles)

$$\tau_a = \frac{2(n_c - n_d)}{n(n-1)} \qquad \text{Ec. 3}$$

- Tau-b se usa cuando los dos atributos a estudiar tienen el mismo número de categorías.

$$\tau_b = \frac{(n_c - n_d)}{\sqrt{\frac{1}{2}n(n-1) - T_x} * \sqrt{\frac{1}{2}n(n-1) - T_y}} \qquad \text{Ec. 4}$$

- Tau-c cuando tengo diferentes números de categorías.

$$\tau_c = \frac{m(n_c - n_d)}{n^2(m-1)} \qquad \text{Ec. 5}$$

Donde:

TABLA 1.5
SIMBOLOGÍA DE CORRELACIÓN KENDALL

Símbolo	Descripción
n	Tamaño de muestra
n_c	Sumatoria de pares concordantes
n_d	Sumatoria de pares discordantes
T_x , T_y	Ajuste valores repetidos en x
m	Valor mínimo de categorías en x y en y

Fuente: Elaboración Propia

Además, hay otras medidas como Point-biserial cuando tengo una variable continua y otra nominal. De la misma forma Rank-biserial nos sirve cuando tengo una variable ordinal y otra nominal.

1.5. Proceso de descubrimiento de conocimiento (KDD)

El proceso de descubrimiento de conocimiento en bases de datos o conocido en inglés como Knowledge Discovery in Databases (KDD), es el proceso de extracción de información implícita en los datos disponibles. Según Martinez (2012), KDD es una metodología genérica para el descubrimiento de conocimiento implícito, anteriormente desconocido y de gran utilidad.

El proceso KDD implica el uso de algoritmos de minería de datos para identificar lo que se contempla como conocimiento conforme a ciertos parámetros específicos, lo cual se aplica sobre una base de datos juntamente con el preprocesamiento y postprocesamiento.

Este proceso con frecuencia es confundido con la minería de datos en sí, más involucra diferentes pasos y técnicas (Santos, 2009), por lo cual se puede agregar que KDD hace referencia al proceso en general, mientras que la minería de datos hace referencia a un paso en particular dentro del proceso de descubrimiento de conocimiento (Karoussi, 2012), como muestra Fig. 13.

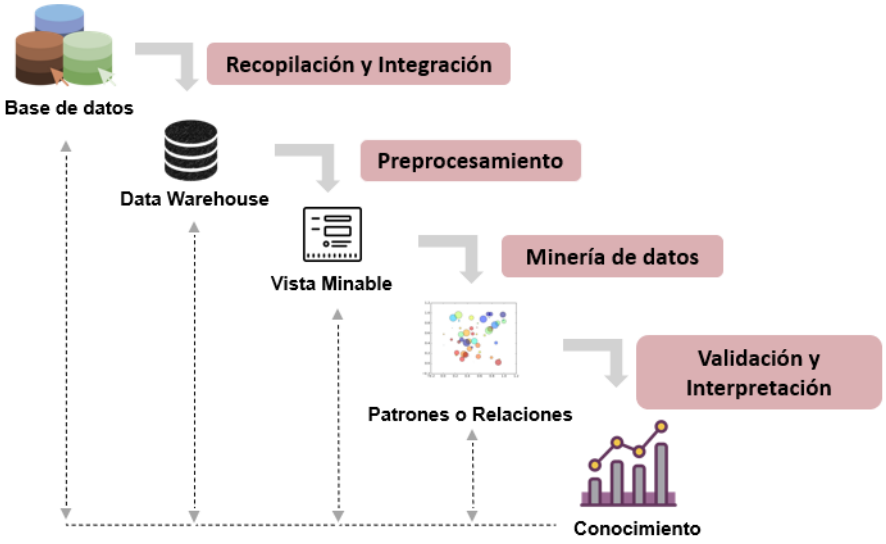


Fig. 13. Proceso KDD (Adaptación Santos 2009)

El proceso KDD inicia con la fase de Recopilación y Integración, en la cual se recopila información de diferentes orígenes (base de datos, registros, etc.) para posteriormente integrar los datos y así crear un Data Warehouse. En fase de Procesamiento se realiza la selección, la limpieza y la transformación de datos. La selección se hace en base al interés del estudio (Perez & Santin, 2007), luego en las fases de limpieza y transformación se obtiene una vista minable. La Fase Minería de Datos el DataSet es sometida técnicas descriptivas o predictivas con fin de identificar patrones o modelos. Finalmente se concluye con la validación y interpretación de resultados.

Es necesario tomar en cuenta que las fases iniciales son cruciales para que en las fases posteriores se pueda extraer conocimientos útiles de la información existente.

1.5.1. Recopilación y Integración

Esta fase inicial con la recopilación de información, la cual se recolecta de diferentes fuentes para ser utilizados en la fase de procesamiento. En la recopilación, los datos pueden provenir de base de datos relacionales o no relacionales, registros, backups, etc. En los días actuales los datos para una futura minería provienen de base datos, aquí es necesario conocer el esquema de base de datos para la etapa de integración.

En la integración de datos consiste en unir la información recolectada de forma que se encuentre en un estado lineal, por ejemplo, si tenemos las tablas clientes, ventas y productos la integración consiste en unir la información de las tres tablas de forma

que tengamos registros que contengan la información del cliente con las compras realizadas y cuáles fueron los productos que adquirió. El resultado de estas dos etapas es un Data Warehouse listo para las etapas del preprocesamiento de datos.

1.5.2. Preprocesamiento

1.5.2.1. Selección

La selección de atributos o también conocido como selección de características tiene el objetivo de identificar atributos relevantes para la etapa de minería de datos, visto de otra forma es la identificación de atributos irrelevantes que no aportan al análisis (Gironés J., Casas J., Minguillón J. & Caihuelas R., 2017).

Previo a la selección es necesario conocer lo que se va a investigar y definir las metas del proceso KDD, esto se debe a que la selección de los atributos puede afectar directamente los resultados (Rodríguez F., 2018).

La selección se puede hacer por niveles como es el filtrado por atributos y el filtrado de registros:

- El filtrado por atributos es la selección solamente de los atributos relevantes para el estudio.
- Filtrado de registros se presenta cuando los atributos están en un estado que lo excluye de del análisis, por ejemplo, en el análisis del rendimiento académico se va a excluir a los estudiantes que se retiraron o anularon matricula porque sus registros estarán vacíos.

1.5.2.2. Eliminación

Según Larose (2014) una gran parte de los datos alojado en bases de datos están sin preprocesamiento, incompletos o ruidos. Por ejemplo, pueden contener campos obsoletos o redundantes, valores atípicos, datos incompresibles para los algoritmos, estos deben someterse a una limpieza y transformación para ser usados en la minería de datos.

En la presente etapa se requiere la eliminación de ruido, tomar decisiones sobre atributos vacíos o incompletos, o corrección si fuere necesario. La toma de estas decisiones se puede basar en la forma de transformación los datos en la próxima etapa o el tipo de algoritmos que se usaran en el análisis.

1.5.2.3. Transformación

La transformación tiene como objetivo proveer de una vista minable lista para analizarlos por los algoritmos de minería de datos. En etapa consiste en la construcción de nuevos atributos mediante la aplicación de alguna operación o función a los atributos existentes con el objetivo de facilitar el proceso de minería (Calvache-Fernandez, Álvarez-Vallejo & Triviño-Arbelaez, 2018).

Dentro de esta etapa se debe considerar métodos como la normalización y discretización, que consiste en reducir conjuntos de datos para estructurarlos en atributos precisos y manejables (Rosero, 2021).

- **Normalización**

La normalización consiste en convertir los valores numéricos para que se ubiquen dentro un rango especifico (Roiger, 2017, p. 251). La normalización beneficia especialmente a clasificadores, redes neuronales y algoritmos que funcionan con medidas de distancias. Es recomendable usar cuando los atributos están en diferentes escalas.

- Normalización Min-Max: compensa el efecto causado por la distancia del valor que se está trabajando con respecto a al máximo valor del atributo (Gironés J. et al, 2017). La Ec. 6 muestra su fórmula y la Tabla 1.6 muestra su simbología:

$$z_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \qquad \text{Ec. 6}$$

Donde:

TABLA 1.6
SIMBOLOGÍA NORMALIZACIÓN MIN-MAX

Símbolo	Descripción
z_i	Resultado de normalización
x_i	Valor para normalizar
x_{max}	Máximo valor del atributo
x_{min}	Mínimo valor del atributo

Fuente: Elaboración Propia

- Normalización por el máximo: identifica el valor máximo del atributo que se va a normalizar y lo divide por todos los valores, de forma que el máximo

valor se le asigna uno y los demás un rango entre 0 a 1. La Ec. 7 muestra su fórmula y la Tabla 1.7 muestra su simbología:

$$z_i = \frac{x_i}{x_{max}} \qquad \text{Ec. 7}$$

Donde:

TABLA 1.7
SIMBOLOGÍA NORMALIZACIÓN POR EL MÁXIMO

Símbolo	Descripción
z_i	Resultado de normalización
x_i	Valor para normalizar
x_{max}	Máximo valor del atributo

Fuente: Elaboración Propia

- **Discretización**

La discretización es un método de conversión o reducción de datos, consiste en transformar gran cantidad de valores continuos a grupos reducido con sus respectivos limites, facilitando el manejo y sin perdidas de datos. Según Hacibeyoglu & Ibrahim (2016) la discretización es un proceso de reducción de variables continuas a un conjunto finito de intervalos. Además, se considera un método muy útil al momento de transformar atributos numéricos a continuos.

- Discretización supervisada que usa datos previamente etiquetados los principales métodos son entrópica, arboles de decisión y Naive Bayes.
- Discretización no supervisada subdivide los datos en intervalos en base a anchura o frecuencia y numero de intervalos proporcionada por el usuario.
  - ❖ Mismo Ancho: consiste en dividir el rango de valores en k número de intervalos, es necesario ordenar los valores e identificar mínimos y máximos. La Ec. 8 muestra el cálculo del rango del intervalo y la Ec. 9 el cálculo del límite. La Tabla 1.8 muestra su simbología.

$$intervalo\ (i) = \frac{\max - min}{k} \qquad \text{Ec.8}$$

$$limite = min + i, min + 2i, .., min + (k-1)i \qquad \text{Ec.9}$$

Donde:

TABLA 1.8
SIMBOLOGÍA DISCRETIZACIÓN MISMO ANCHO

Símbolo	Descripción
i	Rango del intervalo
k	Número de intervalos
min	Máximo valor del atributo
max	Mínimo valor del atributo

Fuente: Elaboración Propia

1.5.3. Minería de Datos

La etapa de minería de datos tiene por objetivo la búsqueda y descubrimiento de patrones, a través de tareas de minería de datos como lo son la clasificación, clustering, patrones secuenciales y asociaciones (Timarán, Hernández, Caicedo, Hidalgo & Alvarado, 2016)

De acuerdo con Rodríguez (2018) la minería de datos suele estar compuesto de tres elementos:

- Modelo, hace referencia a los parámetros a fijarse a partir de los datos d entrada.
- Criterio de referencia, que nos pueden ser de utilidad para comprar modelos alternativos.
- Algoritmo de búsqueda, para la búsqueda de patrones.

1.5.4. Interpretación y Validación

Según Calvache-Fernandez et al., (2018) en esta etapa se evalúa la calidad de los patrones resultantes, los cuales deben tener tres cualidades, como:

- Ser precisos
- Ser comprensibles
- Ser interesantes (útiles y novedosos)

Caso los resultados no sea satisfactores es posible retornar a las anteriores etapas para una posterior iteración. Las tareas más comunes en esta etapa suelen ser visualización de patrones, eliminación de los patrones redundantes e interpretación de patrones útiles. Además, aquí es donde se reúne el conocimiento descubierto para acciones posteriores, documentarlo, resolver conflictos, etc. (Timarán et al., 2016).

1.6. Herramientas para Minería de Datos

1.6.1. Inteligencia de Negocios

Estas son herramientas que permite la transformación de datos en conocimientos el cual es aprovechado por las empresas, pymes y negocios para la toma de decisiones entre las cuales tenemos a:

Pentaho Data Integración

Pentaho es una herramienta que dispone de funcionalidades para realizar el proceso ETL que corresponde a las actividades de extracción y análisis de datos. Conocida también como Kettle nos permite la extracción de datos de fuentes diferentes para su posterior integración (Brinquis, 2020).

1.6.2. Herramientas de minería de datos

Estas herramientas tienen el objetivo de aplicar las técnicas de minería de datos sobre un conjunto de datos previamente procesado en búsqueda de patrones o conocimiento para la toma de decisiones.

Weka

Waikato Environment for Knowledge Analysis o más conocido por sus siglas Weka, es un software open source basado en Java, permite el procesamiento de datos con técnicas de aprendizaje automático, análisis de clústeres, correlación, regresión y clasificación. Además, da acceso a Redes neuronales artificiales, arboles de decisión, algoritmo ID3 y C4.5 (Ionos, 2018).

1.7. Trabajos Relacionados

Cerezo R., Sánchez-Santillán M., Paule-Ruiz M. & Núñez J. (2016) realizaron un análisis para identificar patrones de interacciones usando atributos relacionadas con el tiempo (tiempo en tareas, tiempo en teoría, tiempo en foros), esfuerzo (accesos, tiempo de entrega de la tarea, numero palabras en foros) y notas finales extraídos de Moodle. Primero, por medio de K-means y Expectation-Maximization (EM) formaron clústeres en base al comportamiento. Después, aplicaron ANOVA a los clústeres contra las notas finales, concluyendo que los atributos tiempo en tareas, tiempo de entrega y numero de palabras en foros están más relacionadas con las notas finales.

Helal S., Li J., Liu L., Ebrahimie E., Dawson S. & Murray J. (2017) mediante métodos de descubrimiento de subgrupos realizaron un análisis para la identificación de factores que influyen en el desempeño estudiantil, donde se contempló tres conjuntos de datos; datos de matriculación, datos de Moodle y la combinación de las dos anteriores. Usando SNS, DSSD, NMEEF-SD, BSD, SD-Map y APRIORI-SD determinaron que la situación económica, nivel educativo de los padres, participación en las actividades, foros, visualización de recursos son factores clave en el rendimiento académico.

Conijn R., Snijders C., Kleingeld A. & Matzat U. (2017) analizaron modelos de regresión y clasificación con los datos de Moodle, los resultados mostraron que las notas intermedias tienen una fuerte correlación y foros, wikis una correlación baja. Los análisis de regresión lineal y binaria mostraron diferentes variables predictoras para cada curso.

Hasan R., Palaniappan S., Abdul R., Mahmood S. & Sarker K. (2018) usaron arboles de decisión J48, REP Tree, Random Forest, Logistic Model Tree, Hoeffding Tree, Decisión Stump, Naive Bayes y SMO para la predicción del éxito académico. Para ello se utilizó un conjunto de datos compuestos por información académica (notas y promedios, métrica de riesgo y recuento de plagio) y actividades (accesos y tiempo de permanencia) del estudiante. Los resultados mostraron que Random Forest, Naive Bayes y SMO son los más eficientes para esta tarea.

Bharara, Sabitha & Bansal (2018) por medio del algoritmo K-means realizaron una búsqueda de características relevantes que afectan el rendimiento académico usando información demográfica, datos académicos, participación de los padres (Encuestas, test de satisfacción) y interacciones de los estudiantes. Identificando que participaciones como levantar la mano, debates, ver anuncios y visita de recursos influyen en el rendimiento académico y las notas.

Bravo, Romero & Pamplona (2021) usando correlación, regresiones lineales y agrupación realizaron un análisis para la predicción temprana del rendimiento académico con datos de Moodle. El modelo mostro las siguientes variables predictoras: tarea, accesos (inicios de sesión, participación en foros y acceso a material educativo, foros y glosarios), cuestionarios y edad. Estos resultados coinciden con el análisis de correlación y agrupación.

Calderon-Valenzuela, Payihuanca-Mamani & Bedregal-Alpaca (2022) realizaron un análisis sobre el uso de los recursos y actividades con la finalidad de identificar factores asociados al éxito estudiantil. Utilizaron K-means para agrupar los docentes asociados al uso de estas herramientas y A-priori para identificar asociaciones de uso. Se determino que los recursos como archivos, url y las actividades como tareas, cuestionarios son mayormente usados, además de patrones de comportamiento de docentes con respecto al uso de los recursos y actividades.

Ouatik F., Erritali M., Ouatik F., & Jourhmane M. (2022) usando datos personales, notas de las evaluaciones, interacciones en el EVA, datos psicológicos y ambiente estudiantil experimentaron con tres modelos de clasificación como SVM, C4.5 y KNN para realizar predicciones del éxito académico (aprueba o reprueba) de los alumnos, los resultados mostraron que las notas de las evaluaciones, estatus económico, nivel educativos de los padres, la distancia del hogar a la institución, el grado de interés del alumno, problemas psicológicos y el número de accesos al EVA son los factores que más influyen el éxito académico. Asimismo, el algoritmo de KNN arrojo mejores resultados con respecto a los otros algoritmos.

1.8. Propuesta de Minería de Datos

En base a la revisión de los diferentes estudios realizados, se creó un matriz que sirve como base para la propuesta de minería de datos. La Fig. 14 muestra las técnicas de la minería de datos, los algoritmos, los datos y el estudio que realizo los diversos autores. Para el presente trabajo se pretende utilizar datos socioeconómicos, académicos y interacciones, para identificar factores relacionados al rendimiento académico con correlación, se construirá un modelo de predicción, con agrupación se identificará patrones de uso-rendimiento y descubrir las reglas de asociación existentes.

Autor	Técnicas				Algoritmos																Actividades																		Recursos							Conexión			Datos Académicos				Resultados del estudio
	Predictivas		Descriptivas																		Foros				Tareas				Questionarios				Wiki		Pruebas																		
	Clasificación	Regresión	Agrupamiento	Asociación	K-means	Expectation–maximization	Regresion lineal	Regresion multiple	Decision Tree J48	REP Tree	Random Forest	Logistic Model Tree	Hoeffding Decision Trees	Naive Bayes	Support Vector Machine (SVM)	C4.5	K-Nearest-Neighbor (KNN)	SMO	Apriori-SD	Acceso a foros	Tiempos de permanencia	Num palabras y Num Comentarios	Frecuencia de consulta	Tiempo de Entrega	Num. Tareas Enviadas	Total Tareas entregadas	Total Tareas	Num. de Vistas a questionarios	Frecuencia de acceso a questionarios	Intentos de resolver questionario	Questionarios de satisfacción	Num. Vistas	Num. Wikis Editadas	Num. Test iniciados	Num. Test Vistos	Num. Intentos	Num. Tests aprobados	Participaciones	Frecuencia acceso a material didáctico	Frecuencia de acceso a todos los glosarios	Num links vistos	Num Total clicks	Num paginas vistas	Num recursos vistos	Tiempo en contenidos	Num inicios de sesion	Total tiempo online	Max tiempo inactivo	Datos sociodemográficos	Nota Actividades y Examen final	Nota Final del curso	Nota final de los semestres anteriores	
Cerezo et al., 2016			X		X	X															X	X		X																			X				X	X		Determinar la relacion entre los clusteres y notas finales usando ANOVA
Helal et al., 2017				X															X	X				X												X	X			X	X						X			Realacion entre acceso a recursos-actividades y la tasa de éxito
Conijn et al., 2017		X					X															X				X	X				X	X	X	X				X	X	X	X		X	X	X		X	X		Prediccion de la nota final
Hasan et al., 2018	X								X	X	X	X	X	X				X																		X		X		X				X				X	X		Definió una categoría dentro de una escala de (Reprobado, Bueno, Promedio, Aprobado y Excelente)
Bharara et al., 2018			X		X															X									X							X	X				X				X			X		Determinar variables que afectan el rendimiento academico y agrupacion.
Bravo et al., 2021		X						X														X	X		X	X		X	X	X							X	X						X			X	X	Prediccion de la nota final y determinacion de factores de rendimiento academico.
			X		X																	X	X		X	X		X	X	X							X	X						X			X	X	Agrupación según comportamientos academicos
Calderon-Valenzuela etal., 2022			X		X															X						X											X	X	X								X	Agrupación de docentes asociados al uso de los recursos y actividades
				X															X	X						X											X	X	X								X	Identificar asociaciones de uso de las actividades y recursos
Ouatik F. et al,. 2022	X														X	X	X					X																					X			X	X	X		Modelo declasificacion para predecir el exit académico.

Fig. 14. Matriz de Propuesta

CAPÍTULO 2

Proceso de Descubrimiento de Conocimientos

El desarrollo del presente trabajo se pretende analizar los datos socioeconómicos, académicos y interacciones del estudiante, por medios de la metodología de KDD se busca identificar factores relacionados al rendimiento académico con correlación, se construirá un modelo de predicción, con agrupación se identificará patrones de uso-rendimiento y descubrir las reglas de asociación existentes de patrones entre el rendimiento académico y las interacciones.

2.1. Visión General del Proyecto

- **Suposiciones**

Para el desarrollo de proyecto, se dispone de datos socioeconómicos, académicos y interacciones con el SIIU, recopilada de la Base de Datos de la UTN, esta fuente de datos en bruto será el insumo de entrada del proceso KDD y del cual se obtendrá como resultado conocimiento para determinar las acciones a tomar con el fin de mitigar la problemática.

- **Restricciones**

Para el desarrollo del proyecto se ha estimado un periodo de tiempo de 6 meses, considerando el tiempo de cada actividad dentro de la planificación, así como la disponibilidad de la información a ser analizada.

El proyecto se desarrollará acorde con el tiempo disponible de los encargados en cada una de las áreas, con el fin de obtener insumos e asesoría y proporcionar resultados relevantes.

2.2. Entregables del proyecto

La Tabla 2.1 menciona los artefactos a ser generados y utilizados en el desarrollo del proyecto, los mismos que pueden cambiar conforme transcurra el proceso KDD.

TABLA 2.1
ENTREGABLES DE PROYECTO

Entregables	Detalle
Vista Minable	Producto de la fase de procesamiento (selección, transformación y limpieza) de datos.

Factores que influyen en el rendimiento	Análisis de correlación.
Modelos de predicción	Modelos de clasificación binaria
Patrones de uso – rendimiento	Análisis de agrupación (Clústeres).
Reglas de asociación	Análisis de asociación.
Conocimiento	Interpretación y validación, resultados.

2.3. Organización del Proyecto

2.3.1. Participantes del Proyecto

En la Tabla 2.2 detalla a los directores de áreas involucrados

TABLA 2.2
DIRECTORES DE LAS ÁREAS

Dependencia	Encargado	Función
Coordinación de la Carrera en Sistemas Computacionales	Mgs. Pedro Granda	Asignar especialistas en Minería de Datos
Dirección de Desarrollo Tecnológico e Informático	Mgs. Juan Carlos García	Asignar especialistas en Bases de Datos

La Tabla 2.3 muestra los participantes directos en el proyecto

TABLA 2.3
PARTICIPANTES DIRECTOS

Rol	Dependencia	Nombre
Jefe de proyecto	Carrera de Ingeniería en Sistemas Computacionales	PhD. Iván García
Administrador de bases de datos	Dirección de Desarrollo Tecnológico e Informático	Ing. Eveling Enríquez Ing. Carlos Guevara
Analista de Sistemas	Carrera de Ingeniería en Sistemas Computacionales	Sr. Leonardo Aguagallo

2.3.2. Roles y Responsabilidades

Los roles y responsabilidades de los participantes directos se detallan en la Tabla 2.4.

TABLA 2.4
ROLES Y RESPONSABILIDADES

Rol	Dependencia
Jefe de proyecto	Persona responsable de la planificación, designación, ejecución y se encarga de hacer cumplir las actividades fijadas dentro de la planificación (Rosero, 2021).

Administrador de bases de datos	Proporcionar los datos (datos socioeconómicos, académicos y interacciones) necesaria para el análisis y aclarar posibles dudas de la composición de los datos.
Analista de Sistemas	Ejecución del proceso KDD, validación y presentación de resultados, documentación.

2.4. Gestión del Proceso

2.4.1. Estimaciones

La Tabla 2.5 detalla las horas y costo por hora que compone el talento humano y el correspondiente presupuesto.

TABLA 2.5
TALENTO HUMANO DEL PROYECTO

Descripción	N. de horas	Costo por hora ($)	Costo total ($)
Horas de investigación del proyecto	210	20.00	4200.00
Horas de desarrollo del proyecto	210	20.00	4200.00
		TOTAL	8400.00

Fuente: Elaboración propia

En la Tabla 2.6 detalla los recursos materiales y su costo, los cuales fueron usado para el desarrollo de proyecto.

TABLA 2.6
RECURSOS MATERIALES

Descripción	Costo Real ($)	Costo Actual ($)
HARDWARE		
Laptop	1200.00	00.00
SOFTWARE		
Microsoft Word	00.00	00.00
Microsoft Excel	00.00	00.00
EndNote	00.00	00.00
Pentaho Data Integration	00.00	00.00
Jupyter Notebook	00.00	00.00
R Studio	00.00	00.00
Weka	00.00	00.00
MATERIAS DE OFICINA		
Impresiones	70.00	70.00
CDs	00.30	00.30
Internet	138.00	138.00
Flash Memory	16.00	16.00
Cuaderno	01.50	01.50

Esfero		01.05	01.05
INVESTIGACIÓN			
Libros		80.00	00.00
	TOTAL	1506.85	226.85

La Tabla 2.7 se evidencia la estimación total del proyecto.

TABLA 2.7
COSTO TOTAL DEL PROYECTO

Descripción		Costo ($)
Talento Humano		8400.00
Recursos Materiales		1506.85
	TOTAL	9906.85

Fuente: Elaboración propia

2.4.2. Plan del Proyecto

Para el desarrollo del presente trabajo se hace uso de la metodología KDD en la Tabla 2.8 muestra las fases y el tiempo que se destinara para cada fase.

TABLA 2.8
FASES DEL PROYECTO Y DISTRIBUCIÓN DE HORAS

Fase		Tiempo en Horas
Fase de Recopilación e Integración		40
Fase de Selección, Limpieza y Transformación		40
Fase de Minería de Datos		40
Fase de Evaluación e Interpretación		50
Investigación y Documentación		180
Análisis de Resultados		35
Presentación de Resultados		35
	TOTAL	420

2.5. Fase: Recopilación y Integración de Datos

2.5.1. Recopilación de datos

Para el desarrollo del proyecto se utilizó el proceso KDD, los datos empleados para el estudio son datos socioeconómicos, académicos y interacciones con el SIIU provenientes de la base de datos de la UTN. Estos fueron entregados al Analista de Sistemas en documentos de formato Microsoft Excel. Los tipos de datos que componen los registros son: enteros, decimales, cadenas de caracteres y fechas.

- **Datos socioeconómicos**

Las Tablas 2.9 – 2.14 detalla la estructura de los datos socioeconómicos

TABLA 2.9
ESTRUCTURA PERSONA

ATRIBUTO	TIPO DE DATO
CEDULA	Cadena de caracteres
LUGAR DE NACIMIENTO	Cadena de caracteres
LUGAR_RESIDENCIA	Cadena de caracteres
NACIONALIDAD	Cadena de caracteres
LUGAR_PROCEDENCIA	Cadena de caracteres
TIPO_IDENTIFICACION	Carácter
FECHA_NACIMIENTO	Fecha
GENERO	Carácter
ESTADO_CIVIL	Carácter
ESTADO	Carácter
TIPO_SANGRE	Cadena de caracteres
ID_SUBGRUPO_DISCAPACIDAD	Entero
CARNET_CONADIS	Entero
PORCENTAJE_DISCAPACIDAD	Decimal
COD_ETNIA	Cadena de caracteres
IDENTIFICACIÓN	Cadena de caracteres
DISCAPACIDAD_SIIES	Cadena de caracteres
COD_NACIONALIDAD_INDIGENA	Entero

Fuente: DDTI-UTN

TABLA 2.10
ESTRUCTURA ETNIA

ATRIBUTO	TIPO DE DATO
COD_ETNIA	Cadena de caracteres
ETNIA	Cadena de caracteres

Fuente: DDTI-UTN

TABLA 2.11
ESTRUCTURA DISCAPACIDAD

ATRIBUTO	TIPO DE DATO
ID_SUBGRUPO_DISCAPACIDAD	Entero
ID_GRUPO_DISCAPACIDAD	Entero
DESCRIP_SUBGRUPO_DISCAPACIDAD	Cadena de caracteres
DISCAPACIDAD_SIIES	Cadena de caracteres

Fuente: DDTI-UTN

TABLA 2.12
ESTRUCTURA SOCIOECONOMICO

ATRIBUTO	TIPO DE DATO
CI_PASAPORTE	Cadena de caracteres
COD_MATRICULA	Entero
INST_CODIGO	Entero

FECHA_INGRESO	Fecha
COD_VIVIENDA	Entero
COD_ESTUDIOS_FINAN	Entero
COD_DEP_ECONOMICA	Entero
COD_ING_MENSUAL	Entero
USUARIO	Cadena de caracteres
SERVICIO_AGUA_POTABLE	Cadena de caracteres
SERVICIO_ALCANTARILLADO	Cadena de caracteres
SERVICIO_ENERGIA	Cadena de caracteres
SERVICIO_TELEFONO	Cadena de caracteres
SERVICIO_INTERNET	Cadena de caracteres
SERVICIO_CABLE	Cadena de caracteres
TIPO_BECA	Cadena de caracteres
ESTADO_BECA	Cadena de caracteres
NECESIDAD_PEDAGOGICA	Cadena de caracteres

Fuente: DDTI-UTN

TABLA 2.13
ESTRUCTURA INGRESOS

ATRIBUTO	TIPO DE DATO
COD_ING_MENSUAL	Entero
DESCRIPCION	Cadena de caracteres
RANGO1	Decimal
RANGO2	Decimal
GRUPO	Cadena de caracteres

Fuente: DDTI-UTN

- **Datos académicos**

Las Tablas 2.14 – 2.18 detallan la estructura de los datos académicos.

TABLA 2.14
ESTRUCTURA MATRICULA

ATRIBUTO	TIPO DE DATO
ESTUDIANTE_CEDULA	Cadena de caracteres
CODIGO	Entero
INST_CODIGO	Entero
SIST_ESTUD_CODIGO	Entero
TCICLOACAD_CODIGO	Entero
TFINANCIA_CODIGO	Entero
CICLO_ACAD_CODIGO	Cadena de caracteres
DEPEN_CODIGO	Entero
TMATRICULA_CODIGO	Cadena de caracteres
ESTADO	Carácter
NUMERO_MATRICULA	Entero
FECHA_MATRICULA	Fecha

NIVEL_CODIGO	Entero
LEGALIZADO	Carácter
FECHA_LEGALIZACION	Fecha
ESTUDIANTE_REGULAR	Cadena de caracteres
PORCENTAJE_REGULAR	Decimal

Fuente: DDTI-UTN

TABLA 2.15
ESTRUCTURA DETALLE_MATRICULA

ATRIBUTO	TIPO DE DATO
PARALELO_CODIGO	Carácter
MATERIA_CODIGO	Cadena de caracteres
DOCENTE_CEDULA	Cadena de caracteres
INST_CODIGO	Entero
MODA_ESTUD_CODIGO	Entero
SIST_ESTUD_CODIGO	Entero
TCICLOACAD_CODIGO	Entero
DEPEN_CODIGO	Entero
CICLO_ACAD_CODIGO	Cadena de caracteres
NIVEL_CODIGO	Entero
MATRICULA_CODIGO	Entero
ESTUDIANTE_CEDULA	Cadena de caracteres
NUMERO_MATRICULA	Entero
ANULACIÓN	Carácter
FECHA_ANULACION	Fecha
PENSUM_CODIGO	Carácter
OBSERVACIÓN	Cadena de caracteres

Fuente: DDTI-UTN

TABLA 2.16
ESTRUCTURA NOTAS

ATRIBUTO	TIPO DE DATO
MATERIA_CODIGO	Cadena de caracteres
PARALELO_CODIGO	Carácter
MATRICULA_CODIGO	Entero
DOCENTE_CEDULA	Cadena de caracteres
INST_CODIGO	Entero
MODA_ESTUD_CODIGO	Entero
SIST_ESTUD_CODIGO	Entero
TCICLOACAD_CODIGO	Entero
TFINANCIA_CODIGO	Entero
DEPEN_CODIGO	Entero
CICLO_ACAD_CODIGO	Cadena de caracteres
NIVEL_CODIGO	Entero
ESTUDIANTE_CEDULA	Cadena de caracteres
APROBO	Carácter

NOTA1	Entero
NOTA2	Entero
NOTA3	Entero
NOTA4	Entero
NOTA5	Entero
RESULTADO1	Entero
RESULTADO2	Entero
RESULTADO3	Entero
FINAL1	Decimal
FINAL2	Decimal
FINAL3	Decimal
NOTA_FINAL	Decimal
OBSERVACIÓN	Cadena de caracteres
PORCENTAJE_FALTAS	Decimal
PIERDE_POR_FALTAS	Carácter

Fuente: DDTI-UTN

TABLA 2.17
ESTRUCTURA DEPENDENCIA

ATRIBUTO	TIPO DE DATO
INST_CODIGO	Entero
CODIGO	Entero
NOMBRE	Cadena de caracteres
FUNCIÓN	Carácter
DEPEN_CODIGO	Entero
DEPEN_INST_CODIGO	Entero
DESCRIPCIÓN	Cadena de caracteres
SIGLAS	Cadena de caracteres
OBSERVACIÓN	Cadena de caracteres
TDEPEN_CODIGO	Entero
ESTADO	Carácter

Fuente: DDTI-UTN

TABLA 2.18
ESTRUCTURA CICLO_ACADEMICO

ATRIBUTO	TIPO DE DATO
CÓDIGO	Cadena de caracteres
PER_ACAD_CODIGO	Cadena de caracteres
FECHA_INICIO	Fecha
FECHA_FIN	Fecha
ESTADO	Carácter
ORDEN	Entero
TCICLOACAD_CODIGO	Entero
AÑO	Entero

Fuente: DDTI-UTN

- **Datos de interacciones con el SIIU**

Las Tablas 2.19 – 2.22 detalla la estructura de los datos de interacciones

TABLA 2.19
ESTRUCTURA NUMERO_ACTIVIDADES

ATRIBUTO	TIPO DE DATO
CÓDIGO	Cadena de caracteres
CICLO_ACAD_CODIGO	Cadena de caracteres
MATERIA_CODIGO	Cadena de caracteres
NUM_ACTIVIDADES	Entero

Fuente: DDTI-UTN

TABLA 2.20
ESTRUCTURA NUMERO_RECURSOS

ATRIBUTO	TIPO DE DATO
CÓDIGO	Cadena de caracteres
CICLO_ACAD_CODIGO	Cadena de caracteres
MATERIA_CODIGO	Cadena de caracteres
NUM_RECURSOS	Entero

Fuente: DDTI-UTN

TABLA 2.21
ESTRUCTURA RECURSOS_SIIU

ATRIBUTO	TIPO DE DATO
ID_RECURSO	Cadena de caracteres
DEPENDENCIA_CODIGO	Cadena de caracteres
CICLO_ACAD_CODIGO	Cadena de caracteres
MATERIA_CODIGO	Cadena de caracteres
NIVEL_CODIGO	Entero
DOCUMENTO	Cadena de caracteres
ENLACE	Cadena de caracteres
ARCHIVO	Cadena de caracteres
AUDIO	Cadena de caracteres
IMAGEN	Cadena de caracteres
VIDEO	Cadena de caracteres

Fuente: DDTI-UTN

TABLA 2.22
ESTRUCTURA ACTIVIDADES_SIIU

ATRIBUTO	TIPO DE DATO
ID_ACTIVIDAD	Cadena de caracteres
DEPENDENCIA_CODIGO	Cadena de caracteres
CICLO_ACAD_CODIGO	Cadena de caracteres
MATERIA_CODIGO	Cadena de caracteres

CEDULA_INTEGRANTE	Cadena de caracteres
NIVEL_CODIGO	Cadena de caracteres
ACTIVIDADES_GRUPALES	Cadena de caracteres
ACTIVIDAD_INDIVIDUAL	Cadena de caracteres
DEBERES	Cadena de caracteres
EXÁMENES	Cadena de caracteres
INFORMES	Cadena de caracteres
LECCIONES	Cadena de caracteres
PARTICIPACIÓN	Cadena de caracteres
FOROS_DEBATES_OTROS	Cadena de caracteres
PROYECTO	Cadena de caracteres
PRUEBAS	Cadena de caracteres
REPORTE_LABORATORIO	Cadena de caracteres
TAREAS	Cadena de caracteres
TRABAJO_AUTONOMO	Cadena de caracteres
TRABAJOS	Cadena de caracteres

Fuente: DDTI-UTN

2.5.2. Integración de Datos

Para la integración se utilizó la herramienta Pentaho Data Integration (PDI), esta herramienta permite realizar tareas del proceso ETL. El objetivo de la integración es consolidar los datos provenientes de diferentes fuentes para obtener un Data Warehouse, en esta fase de usa funciones de JOIN para relacionar la información, la Fig. 15 – 17 muestra el proceso de integración.

- **Datos socioeconómicos**

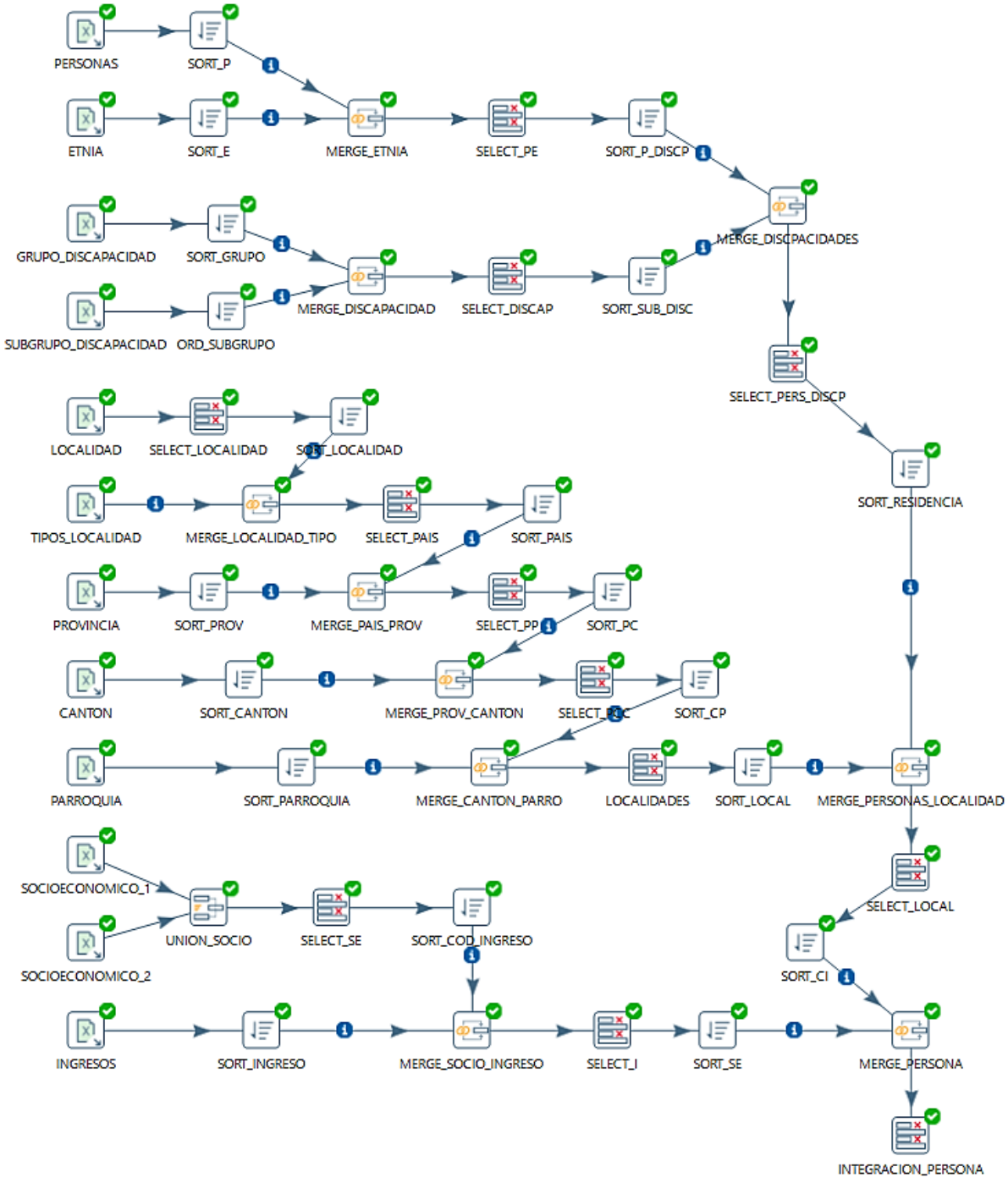


Fig. 15. Integración datos socioeconómicos

- **Datos Académicos**

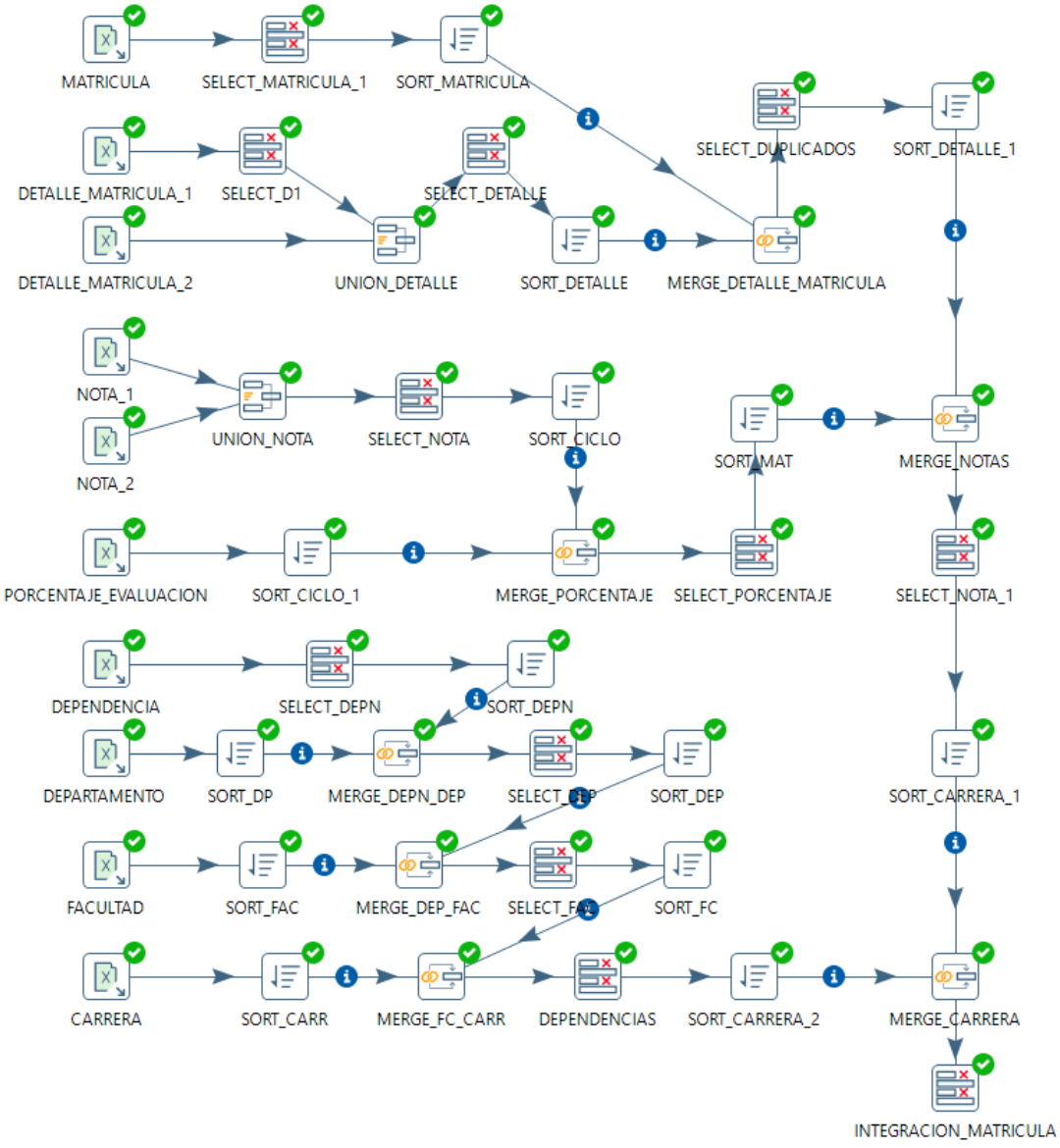


Fig. 16. Integración datos académicos

- **Datos de Interacciones**

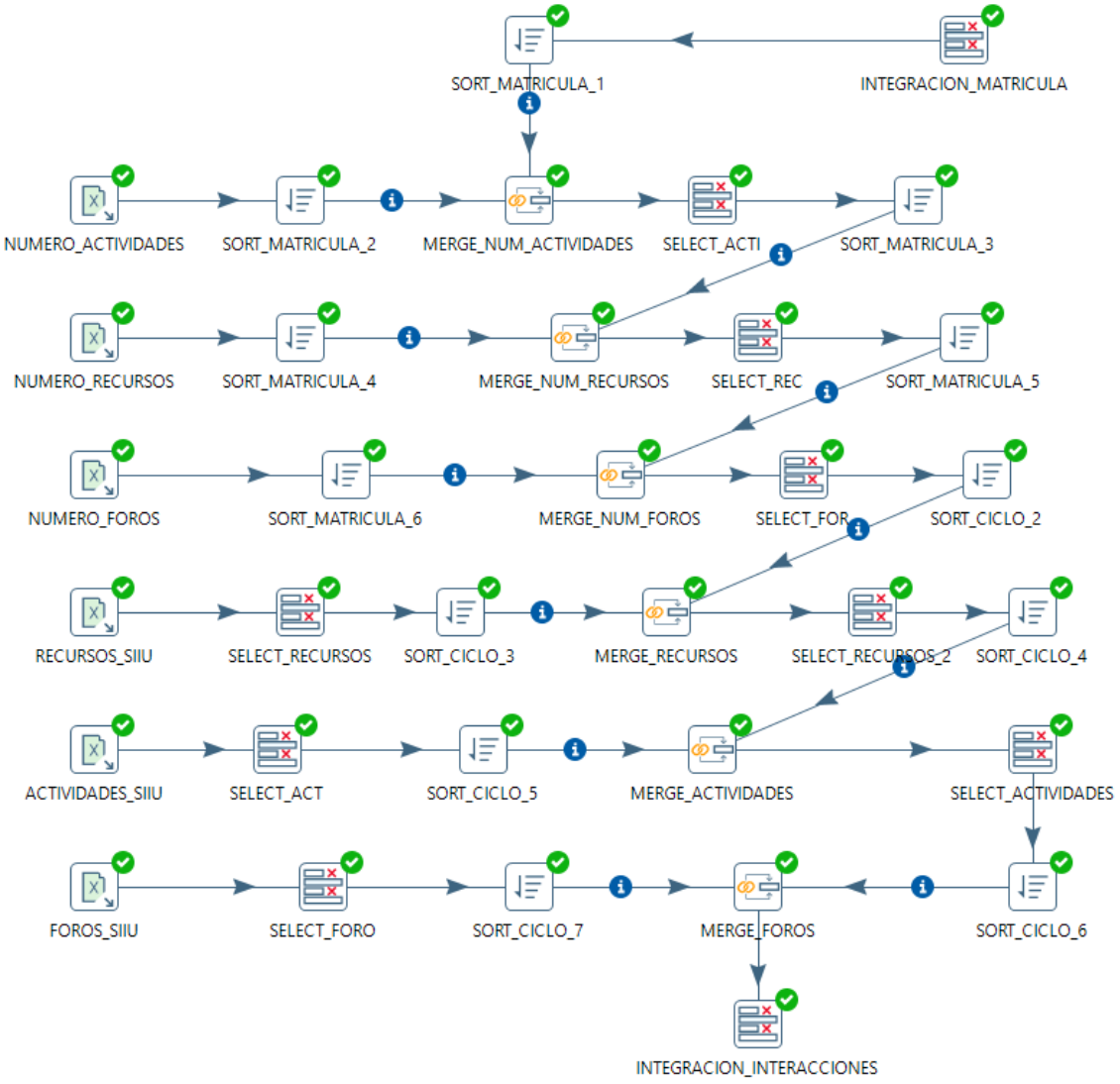


Fig. 17. Integración datos de interacciones

Debido a la gran cantidad de información se hizo una integración por partes y al final se integró todas las partes como se muestra en la Fig. 18.

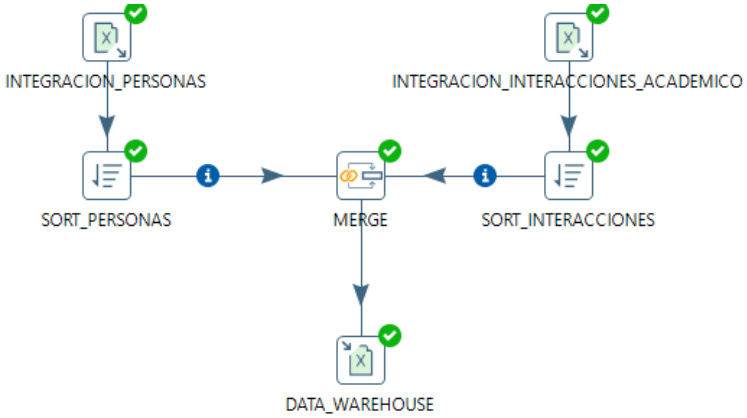


Fig. 18. Construcion del Data Warehouse

2.5.3. Data Warehouse

El resultado de la etapa de recopilación y integración de datos es un Data Warehouse listo para ser procesado. La Fig. 19 muestra una porción de los datos.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	MATERIA_CODIGO	MATRICULA_CODIG	DEPEN_CODIG	CICLO_ACAD_CODI	NIVEL_CODIG	E	APRC	SUMA	ACTIV	ACTIV	AUTO	DEBE	ENCU	EXAM	INFOR	LECC
2	EISIC-AI1	297396	00165	0917-0218	07	100	S	100.0	10	0	0	0	0	30	0	20
3	EISIC-AS2	297396	00165	0917-0218	07	100	S	100.0	25	0	0	0	0	25	0	0
4	EISIC-COM	297396	00165	0917-0218	07	100	N	100.0	25	0	0	0	0	25	0	0
5	EISIC-DEI	297396	00165	0917-0218	07	100	S	100.0	20	0	0	10	0	30	0	20
6	EISIC-HSA	297396	00165	0917-0218	7	100	S	100.0	20	0	0	0	0	30	0	30
7	EISIC-IS2	297396	00165	0917-0218	07	100	S	100.0	25	0	0	0	0	30	0	0
8	FS-A702	297396	00165	0917-0218	07	100	S	100.0	0	0	0	20	0	30	0	30
9	CISIC-COMOV	297398	00165	0917-0218	09	100	S	100.0	0	0	0	20	0	30	0	0
10	CISIC-OPT1	297398	00165	0917-0218	08	100	S	100.0	0	0	0	0	0	30	0	30
11	EISIC-INA	297398	00165	0917-0218	08	100	S	100.0	20	0	0	20	0	30	0	0
12	EISIC-TRAGRA2	297399	00165	0917-0218	10	040	S	100.0	0	0	0	0	0	30	10	0
13	EISIC-AI1	297400	00165	0917-0218	07	100	S	100.0	10	0	0	0	0	30	0	20
14	EISIC-AS2	297400	00165	0917-0218	07	100	S	100.0	25	0	0	0	0	25	0	0
15	EISIC-COM	297400	00165	0917-0218	07	100	N	100.0	25	0	0	0	0	25	0	0
16	EISIC-DEI	297400	00165	0917-0218	07	100	S	100.0	20	0	0	10	0	30	0	20
17	EISIC-HSA	297400	00165	0917-0218	07	100	S	100.0	20	0	0	0	0	30	0	30
18	EISIC-IS2	297400	00165	0917-0218	07	100	S	100.0	25	0	0	0	0	30	0	0
19	FS-A702	297400	00165	0917-0218	07	100	S	100.0	0	0	0	20	0	30	0	30
20	CISIC-COMOV	297401	00165	0917-0218	09	040	S	100.0	0	0	0	20	0	30	0	0
21	CISIC-OPT2	297401	00165	0917-0218	09	040	S	100.0	0	0	0	0	0	20	0	20
22	CISIC-OPT3	297401	00165	0917-0218	09	040	S	100.0	10	0	0	0	0	30	0	20
23	CISIC-OPT4	297401	00165	0917-0218	09	040	S	100.0	20	0	0	20	0	30	0	30
24	CISIC-SIG	297401	00165	0917-0218	09	040	S	100.0	0	0	0	0	0	30	0	30
25	EISIC-TRAGRA1	297401	00165	0917-0218	09	040	S	100.0	0	0	0	0	0	0	25	0
26	EISIC-AI1	297402	00165	0917-0218	07	100	S	100.0	10	0	0	0	0	30	0	20
27	EISIC-AS2	297402	00165	0917-0218	07	100	S	100.0	25	0	0	0	0	25	0	0
28	EISIC-COM	297402	00165	0917-0218	07	100	S	100.0	25	0	0	0	0	25	0	0
29	EISIC-DEI	297402	00165	0917-0218	07	100	S	100.0	20	0	0	10	0	30	0	20
30	EISIC-HSA	297402	00165	0917-0218	07	100	S	100.0	20	0	0	0	0	30	0	30
31	EISIC-IS2	297402	00165	0917-0218	07	100	S	100.0	25	0	0	0	0	30	0	0
32	FS-A702	297402	00165	0917-0218	07	100	S	100.0	0	0	0	20	0	30	0	30
33	EISIC-AI1	297403	00165	0917-0218	07	040	S	100.0	10	0	0	0	0	30	0	20
34	EISIC-COM	297403	00165	0917-0218	07	040	S	100.0	25	0	0	0	0	25	0	0
35	EISIC-DEI	297403	00165	0917-0218	07	040	S	100.0	20	0	0	10	0	30	0	20
36	EISIC-HSA	297403	00165	0917-0218	07	040	S	100.0	20	0	0	0	0	30	0	30
37	EISIC-IS2	297403	00165	0917-0218	07	040	S	100.0	25	0	0	0	0	30	0	0
38	FS-A702	297403	00165	0917-0218	07	040	S	100.0	0	0	0	20	0	30	0	30
39	CISIC-COMOV	297405	00165	0917-0218	09	040	S	100.0	0	0	0	20	0	30	0	0
40	CISIC-OPT2	297405	00165	0917-0218	09	040	S	100.0	0	0	0	0	0	20	0	20

Fig. 19. Porción del Data Warehouse

2.6. Fase: Preprocesamiento

Una vez concluido la consolidación de los datos y obtenido un data warehouse, en esta sección se hace la selección, transformación y limpieza.

2.6.1. Selección

- **Filtrado de atributos**

En esta etapa se seleccionó los atributos relevantes para el análisis, el criterio de selección se hiso en base a los trabajos relacionados y la matriz de trabajos relacionados. Dentro de los atributos seleccionados se encuentran atributos que se utilizarán directamente en el análisis y atributos que servirán para calcular otros atributos. La Fig. 20 muestra la primera selección de atributos.

#	Fieldname	Rename to	Length	Precision
1	MATERIA_CODIGO			
2	NIVEL_CODIGO			
3	APROBO			
4	NOTA1			
5	NOTA2			
6	NOTA_FINAL			
7	PIERDE_POR_FALTAS			
8	NUMERO_MATRICULA			
9	ANULACION			
10	FECHA_MATRICULA			
11	SIGLAS_DEP			
12	NUM_ACTIVIDADES			
13	NUM_RECURSOS			
14	NUM_FOROS			
15	DOCUMENTO			
16	ENLACE			
17	ARCHIVO			
18	AUDIO			
19	IMAGEN			
20	VIDEO			
21	ACTIVIDADES_GRUPALES			
22	ACT_INDIVIDUAL			
23	DEBERES			
24	EXAMENES			
25	INFORMES			
26	LECCIONES			
27	PARTICIPACION			
28	FOROS_DEBATES_OTROS			
29	PROYECTO			
30	PRUEBAS			
31	REPORTES_LAB			
32	TAREAS			
33	TRAB_AUTONOMO			
34	TRABAJOS			
35	FOROS			
36	FECHA_NACIMIENTO			
37	GENERO			
38	ESTADO_CIVIL			
39	PORCENTAJE_DISCAPACIDAD			
40	ETNIA			
41	SERVICIO_INTERNET			
42	DESCRIPCION			
43	RANGO2			

Fig. 20. Filtrado de atributos

- **Filtrado de registros**

Para el análisis que se va a realizar es necesario todos los registros de notas y interacciones disponibles, razón por la cual se filtró a los estudiantes que anularon la matricula. La Fig. 21 muestra la filtración de los anulados.

NOTA_FINAL	PORCENTAJE_FALTAS	PIERDE_POR_FALTAS	NUMERO_MAT	ANULACION
7.8	0	N	2	N
4.9	0	N	1	N
5.3	0	N	3	N
5.3	0	N	3	N
1.5	2.5	N	1	S
3	1.56	N	1	S
2.5	34.37	S	1	S
0.5	0	N	1	S
4	12.5	N	1	S

Fig. 21. Filtrado de registros

2.6.2. Transformación

a. Discretización

En esta sección presentamos una discretización simple, donde no se empleó ningún método estadístico apara identificar los intervalos, esto debido a que los límites de los intervalos están previamente establecidos o se establecieron en base a los siguientes criterios.

- **Edad**

Para el atributo edad se calculó en base a la fecha de nacimiento y la fecha de matriculación con el objetivo de identificar la edad exacta del estudiante al momento de realizar las interacciones. La Tabla 2.23 muestra las categorías establecidas, donde los estudiantes de 17 a 25 años se ubicaron en una edad baja porque entre esas edades inician y finalizan sus estudios superiores, los estudiantes de 26 a 30 años se ubicaron en una media edad y de 31 años en adelante en una edad alta.

TABLA 2.23
DISCRETIZACIÓN ATRIBUTO EDAD

EDAD	CATEGORÍA
17 – 25	BAJA
26 – 30	MEDIA
31 – Adelante	ALTA

Fuente: Elaboración Propia

- **Ingresos**

Para el atributo Ingreso se tomó en cuenta el sueldo básico unificado (SBU) y la canasta básica familiar (CBF) de los años últimos 5 años, debido que los registros contemplan información socioeconómica desde el 2017 al 2021 se hiso un promedio del SBU y de la CBF para tomar como base y discretizar los ingresos. La Tabla 2.24 muestra el promedio del SBU y la CBF.

TABLA 2.24
SBU Y CBF DE LOS ÚLTIMOS 5 AÑOS

AÑO	SBU	CBF
2017	375	708.98
2018	386	715.16
2019	394	715.08
2020	400	710.08
2021	400	719.65
PROMEDIO	**391**	**713.79**

Fuente: INEC

La Tabla 2.25 muestra la discretización de los ingresos.

TABLA 2.25
DISCRETIZACIÓN ATRIBUTO INGRESOS

INGRESOS ($)	CATEGORÍA
0 – 390.99	MUY BAJO
391.00 – 713.79	BAJO
713.80 – 1427.58	MEDIO
1427.59 – 2000.00	ALTO
2000.01 – 10 000.00	MUL ALTO

Fuente: Elaboración Propia

- **Etnia**

El atributo etnia se categorizo en base a las categorías identificadas en el Ceso del 2010 (Vila, 2019). La Tabla 2.26 muestra las categorías del atributo etnia.

TABLA 2.26
DISCRETIZACIÓN ATRIBUTO ETNIA

ETNIA	CATEGORÍA
AFROECUATORIANO/A	AF
MULATO/A	AF
NEGRO/A	AF
MESTIZO/A	ME
BLANCO/A	ME
MONTUBIO/A	MO
INDIGENA	IN
NO REGISTRA	NR

Fuente: Elaboración Propia

- **Porcentaje de discapacidad**

El atributo discapacidad se discretizo en base a los porcentajes establecidos en el Consejo Nacional para la Igualdad de Discapacidades (CONADIS), como se muestra en la Tabla 2.27.

TABLA 2.27
DISCRETIZACIÓN ATRIBUTO DISCAPACIDAD

PORCENTAJES (%)	CATEGORÍA
0 – 29	NINGUNA
30 – 49	LEVE
50 – 74	MODERADA
75 – 84	SEVERO
85 – 100	GRAVE

Fuente: CONADIS

- **Porcentaje de Faltas**

El atributo porcentaje faltas se discretizo en base al porcentaje establecido por la institución del 30 % reprueba la materia. La Tabla 2.28 muestra las categorías.

TABLA 2.28
DISCRETIZACIÓN ATRIBUTO PORCENTAJE FALTAS

PORCENTAJES (%)	CATEGORÍA
0 – 7.99	MUY BAJO
8.00 – 15.99	BAJO
16.00 – 23.99	MEDIO
24.00 – 29.99	ALTO
30.00 – 100	MUL ALTO

Fuente: Elaboración Propia

- **Nota 1, Nota 2 y Nota Final**

Los atributos de las Notas se discretizo basado en si la nota era suficiente para aprobar la materia. La Tabla 2.29 muestra las categorías.

TABLA 2.29
DISCRETIZACIÓN LOS ATRIBUTOS DE LAS NOTAS

PORCENTAJES (%)	CATEGORÍA
0 – 6.99	INSUFICIENTE
7.00 – 7.99	SUFICIENTE
8.00 – 8.99	BUENO
9.00 – 10	EXCELENTE

Fuente: Elaboración Propia

En esta sección se usó eso el método de discretización de Mismo Ancho para discretizar el atributo número de actividades y recursos usadas por el estudiante.

- **Número de actividades**

Para el atributo número de actividades se apoyó de la herramienta de análisis de datos de Excel para identificar el valor mínimo y máximo de actividades. Donde el mínimo fue de 1 y el máximo fue de 57 actividades durante un semestre. Además, se definió una k igual a 5 para categorizar el uso como *muy bajo, bajo, medio, alto y muy alto*. La Ec. 5 muestra el cálculo del rango de los intervalos por método de Mismo Ancho.

$$intervalo = \frac{\max - min}{k} = \frac{57 - 1}{5} = 11.2 \qquad \text{Ec.5}$$

La Tabla 2.30 muestra los rangos y categorías de las actividades

TABLA 2.30
DISCRETIZACIÓN NÚMERO DE ACTIVIDADES

INTERVALO	CATEGORÍA DE USO
1 – 12.20	MUY BAJO
12.21 – 23.40	BAJO
23.41 – 34.60	MEDIO
34.61 – 45.80	ALTO
45.81 – 57.00	MUY ALTO

Fuente: Elaboración Propia

- **Numero de Recursos**

Para el atributo número de recursos se siguió el mismo procedimiento que en el atributo número de actividades. Para este atributo su mínimo fue 1 y su máximo 1652, y para k se definió el valor de 5. La Tabla 2.31 muestra los intervalos y las categorías de uso de los recursos.

TABLA 2.31
DISCRETIZACIÓN NÚMERO DE RECURSOS

INTERVALO	CATEGORÍA DE USO
1 – 330.20	MUY BAJO
330.21 – 660.40	BAJO
660.41 – 990.60	MEDIO
990.61 – 1320.80	ALTO
1320.81 – 1652.00	MUY ALTO

Fuente: Elaboración Propia

Para la discretización se usó las opciones de Number Ranges y Replace in string de acuerdo con la naturaleza de la información. Para rangos de valores que contienen un mínimo y máximo se usó la opción Number Ranges como se muestra en la Fig. 22 y Replace in string para remplazar cadenas como se muestra en la Fig. 23.

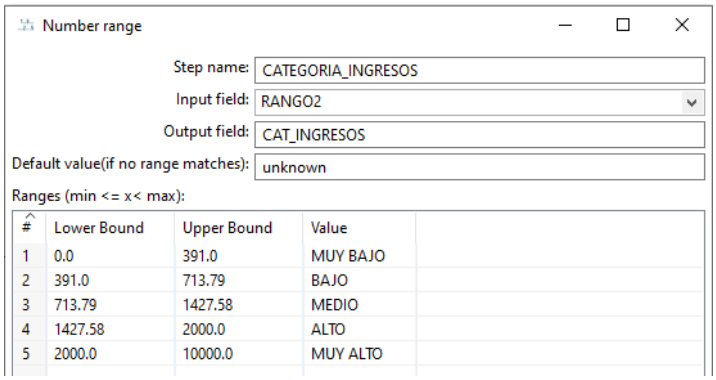


Fig. 22. Uso Number Ranges

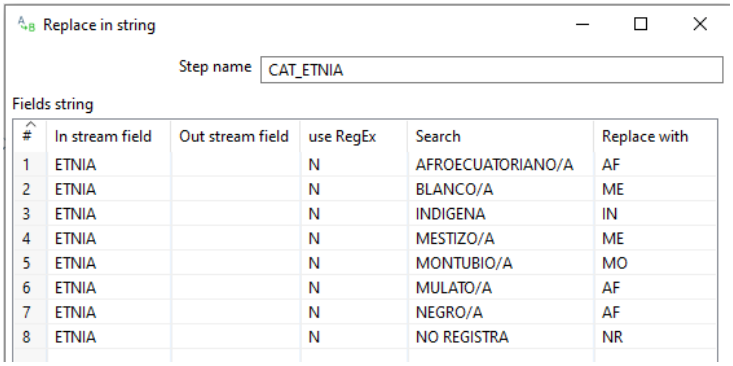


Fig. 23. Uso Replace in string

b. Normalización

Debido a limitación de algunos algoritmos y la exigencia de técnicas como la correlación de Pearson que requiere atributos numéricos se procedió a la numerización de los atributos de la vista minable cualitativa. La numerización es un tipo de normalización simple y consiste en asignar un numero a las categorías de los atributos.

La Tabla 2.32 – 2.34 muestra la normalización con este método.

TABLA 2.32
NUMERIZACIÓN ATRIBUTO CARRERA

CATEGORÍAS	VALOR ASIGNADO
CIAUTO	1
CIERCOM	2
CIME	3
CINDU	4
CISIC	5
CSOFT	6

Fuente: Elaboración Propia

TABLA 2.33
NUMERIZACIÓN ATRIBUTO NUMERO ACTIVIDADES Y RECURSOS

CATEGORÍAS	VALOR ASIGNADO

MUY BAJO	1
BAJO	2
MEDIO	3
ALTO	4
MUY ALTO	5

Fuente: Elaboración Propia

TABLA 2.34
NUMERIZACIÓN ATRIBUTO NOTA FINAL

CATEGORÍAS	VALOR ASIGNADO
INSUFICIENTE	1
SUFICIENTE	2
BUENO	3
EXCELENTE	4

Fuente: Elaboración Propia

Los atributos como aprobó de los datos académicos y atributos como documento, enlace, archivo, exámenes, foros debates y otros, proyecto, pruebas, tareas y trabajos de los datos de interacciones, así como el atributo servicio de internet, son nominales y tienen la misma estructura con dos categorías (SI, NO). La Tabla 2.35 muestra la normalización de estos atributos.

TABLA 2.35
NUMERIZACIÓN DE ATRIBUTOS NOMINALES

CATEGORÍAS	VALOR ASIGNADO
SI	1
NO	0

Fuente: Elaboración Propia

La Tabla 2.36 – 2.3 muestra las normalizaciones restantes.

TABLA 2.36
NUMERIZACIÓN ATRIBUTO EDAD

CATEGORÍAS	VALOR ASIGNADO
TEMPRANA	1
MEDIA	2
ALTA	3

Fuente: Elaboración Propia

TABLA 2.37
NUMERIZACIÓN DE ATRIBUTOS GENERO

CATEGORÍAS	VALOR ASIGNADO
F	1
M	2

Fuente: Elaboración Propia

TABLA 2.38
NUMERIZACIÓN ATRIBUTO ESTADO CIVIL

CATEGORÍAS	VALOR ASIGNADO
C	1
D	2
P	3
S	4
U	5

Fuente: Elaboración Propia

TABLA 2.39
NUMERIZACIÓN ATRIBUTO DISCAPCIDAD

CATEGORÍAS	VALOR ASIGNADO
NINGUNA	1
LEVE	2
MODERADA	3
SEVERO	4
GRAVE	5

Fuente: Elaboración Propia

TABLA 2.40
NUMERIZACIÓN ATRIBUTO ETNIA

CATEGORÍAS	VALOR ASIGNADO
AF	1
ME	2
MO	3
IN	4
NR	5

Fuente: Elaboración Propia

TABLA 2.41
NUMERIZACIÓN ATRIBUTO INGRESOS

CATEGORÍAS	VALOR ASIGNADO
MUY BAJO	1
BAJO	2
MEDIO	3
ALTO	4
MUY ALTO	5

Fuente: Elaboración Propia

2.6.3. Eliminación

Se tomó la decisión que esta fase se realizaría después de la fase de transformación porque campos que contenían cero no necesariamente representaba ausencia, sino que se incluían dentro de un rango de categorías en la transformación, ejemplo: los estudiantes sin faltas tienen el porcentaje de 0%.

Dado que los datos provinieron de departamentos diferentes y se conforma de gran cantidad de tablas se evidencio muchos campos vacíos en cada uno de los atributos, se inició la limpieza con la eliminación de campos vacíos en el atributo nota final, esto debido a que es un atributo de alta importancia para el análisis de correlación. La Fig. 24 muestra estos campos.

SIGLAS_DEP	MATERIA_CODIGO	NIVEL_CODIGO	NOTA_FINAL	CAT_NOTA_FINAL
FICA	CINDU-TRAGRA2	10	10	EXCELENTE
FICA	CINDU-TRAGRA2	10	10	EXCELENTE
FICA	CINDU-TRAGRA2	10	10	EXCELENTE
FICA	CINDU-TRAGRA2	10	10	EXCELENTE
FICA	CINDU-TRAGRA2	10	10	EXCELENTE
FICA	CINDU-TRAGRA2	10	10	EXCELENTE
FICA	CIMANELE-LABVIRAC	AC	10	EXCELENTE
FICA	CIDES-ALLOG	01	10	EXCELENTE
FICA	CIDES-ALLOG	01	10	EXCELENTE
FICA	CIDES-ALLOG	01	10	EXCELENTE
FICA	FC-CA-802	08		unknown
FICA	IME0410	04		unknown

Fig. 24. Eliminación de campos vacíos – atributo nota final

Posteriormente, se eliminaron los campos vacíos en los atributos números de actividades y recursos. La Fig. 25 muestra estos campos.

NUMERO_MATRICULA	NUM_ACTIVIDADES	NUM_RECURSOS
1	9.0	10.0
1	9.0	10.0
1	9.0	10.0
	9.0	16.0
1	9.0	10.0
1		

Fig. 25. Eliminación de campos vacíos – atributo actividades y recursos

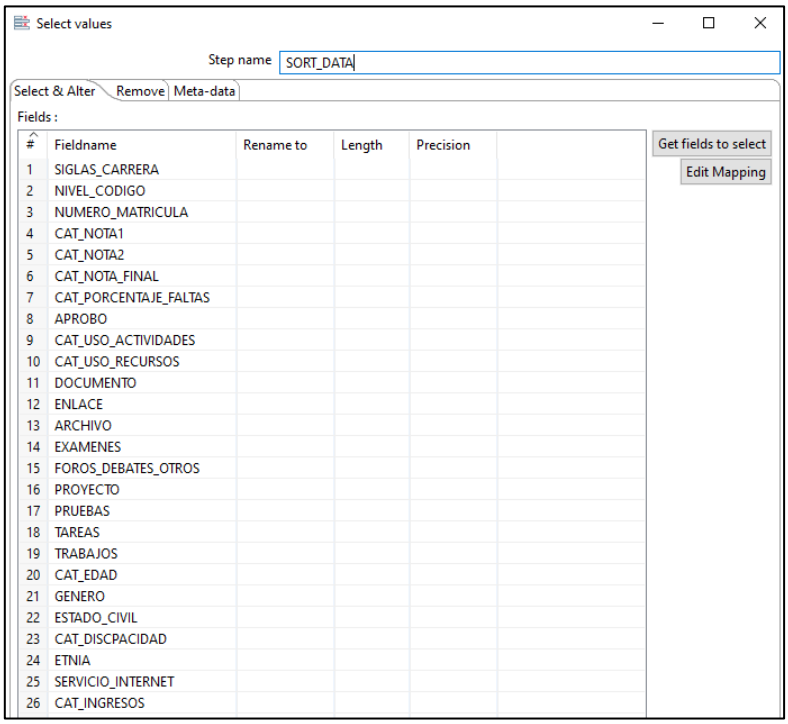
Para concluir, se eliminaron los campos vacíos en el atributo categoría discapacidad y coincidió que los mismos registros en los atributos socioeconómicos estaban vacíos, la Fig. 26 muestra el estado de los atributos.

CAT_EDAD	GENERO	ESTADO_CIVIL	CAT_DISCPACIDAD	ETNIA	SERVICIO_INTERNET	CAT_INGRESOS
ALTA	M	S	NINGUNA	IN	NO	BAJO
ALTA	M	S	NINGUNA	IN	NO	BAJO
ALTA	M	S	NINGUNA	IN	NO	BAJO
ALTA	M	S	NINGUNA	IN	NO	BAJO
ALTA	M	S	NINGUNA	IN	NO	BAJO
BAJA	F	S	SEVERA	ME	SI	BAJO
BAJA	F	S	SEVERA	ME	SI	BAJO
BAJA	F	S	SEVERA	ME	SI	BAJO
BAJA	F	S	SEVERA	ME	SI	BAJO
BAJA	F	S	SEVERA	ME	SI	BAJO
BAJA	F	S	SEVERA	ME	SI	BAJO
unknown			unknown			unknown
unknown			unknown			unknown
unknown			unknown			unknown

Fig. 26. Eliminación de campos vacíos – atributos socioeconómicos

2.6.4. Vista Minable

El resultado de la fase de procesamiento es una vista minable con 26 atributos y 57 115 registros. La Fig. 27 muestra los atributos de la vista minable.



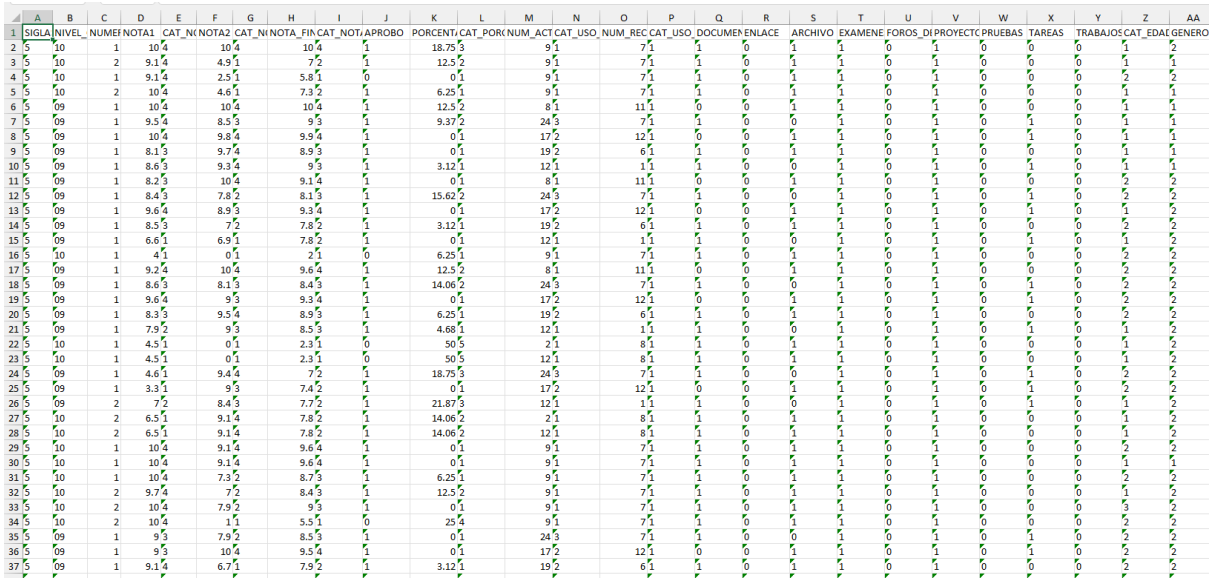
#	Fieldname	Rename to	Length	Precision
1	SIGLAS_CARRERA			
2	NIVEL_CODIGO			
3	NUMERO_MATRICULA			
4	CAT_NOTA1			
5	CAT_NOTA2			
6	CAT_NOTA_FINAL			
7	CAT_PORCENTAJE_FALTAS			
8	APROBO			
9	CAT_USO_ACTIVIDADES			
10	CAT_USO_RECURSOS			
11	DOCUMENTO			
12	ENLACE			
13	ARCHIVO			
14	EXAMENES			
15	FOROS_DEBATES_OTROS			
16	PROYECTO			
17	PRUEBAS			
18	TAREAS			
19	TRABAJOS			
20	CAT_EDAD			
21	GENERO			
22	ESTADO_CIVIL			
23	CAT_DISCPACIDAD			
24	ETNIA			
25	SERVICIO_INTERNET			
26	CAT_INGRESOS			

Fig. 27. Atributos de la vista minable

En el presente proyecto se realizá asociación, predicción, agrupación y asociación, para abastecer la exigencia de cada uno de los algoritmos se generó dos vistas minables. La Fig. 28 muestra una vista cualitativa y la Fig. 29 muestra una vista que contiene los atributos en tipo continuo, ordinal y nominal.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	SIGLAS_CA	NIVEL_CO	NUMERO_	CAT_NOTA	CAT_NOTA	CAT_POR(	APROBO		CAT_USO_	CAT_USO_	DOCUMEN	ENLACE	ARCHIVO	EXAMENE	FOROS_DI	PROYECTO	PRUEBAS	TAREAS	TRABAJOS	CAT_EDAD	GENERO	ESTADO_C	CAT_DISC	ETNIA	SERVICIO_	CAT_INGRESOS	
2	CISIC	10	1	EXCELENT	EXCELENT	EXCELENT	MEDIO	SI	MUY BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	TEMPRAN	M	S	NINGUNA	ME	SI	MUY BAJO	
3	CISIC	10	2	EXCELENT	INSUFICIE	SUFICIENT	BAJO	SI	MUY BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	TEMPRAN	F	S	NINGUNA	ME	NO	BAJO	
4	CISIC	10	1	EXCELENT	INSUFICIE	INSUFICIE	MUY BAJO	NO	MUY BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	MEDIA	M	S	NINGUNA	ME	SI	BAJO	
5	CISIC	10	2	EXCELENT	INSUFICIE	SUFICIENT	MUY BAJO	SI	MUY BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	TEMPRAN	F	S	NINGUNA	ME	NO	BAJO	
6	CISIC	09	1	EXCELENT	EXCELENT	EXCELENT	BAJO	SI	MUY BAJO	MUY BAJO	NO	NO	SI	SI	NO	SI	NO	NO	NO	TEMPRAN	F	S	NINGUNA	ME	SI	MUY BAJO	
7	CISIC	09	1	EXCELENT	BUENO	BUENO	BAJO	SI	MEDIO	MUY BAJO	SI	NO	NO	SI	NO	SI	NO	SI	NO	TEMPRAN	F	S	NINGUNA	ME	SI	MUY BAJO	
8	CISIC	09	1	EXCELENT	EXCELENT	EXCELENT	MUY BAJO	SI	BAJO	MUY BAJO	NO	NO	SI	SI	NO	SI	NO	SI	NO	TEMPRAN	F	S	NINGUNA	ME	SI	MUY BAJO	
9	CISIC	09	1	BUENO	EXCELENT	BUENO	MUY BAJO	SI	BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	TEMPRAN	F	S	NINGUNA	ME	SI	MUY BAJO	
10	CISIC	09	1	BUENO	EXCELENT	BUENO	MUY BAJO	SI	MUY BAJO	MUY BAJO	SI	NO	NO	SI	NO	SI	NO	SI	NO	TEMPRAN	F	S	NINGUNA	ME	SI	MUY BAJO	
11	CISIC	09	1	BUENO	EXCELENT	EXCELENT	MUY BAJO	SI	MUY BAJO	MUY BAJO	NO	NO	SI	SI	NO	SI	NO	NO	NO	MEDIA	M	S	NINGUNA	ME	SI	MEDIO	
12	CISIC	09	1	BUENO	SUFICIENT	BUENO	BAJO	SI	MEDIO	MUY BAJO	SI	NO	NO	SI	NO	SI	NO	SI	NO	MEDIA	M	S	NINGUNA	ME	SI	MEDIO	
13	CISIC	09	1	EXCELENT	BUENO	EXCELENT	MUY BAJO	SI	BAJO	MUY BAJO	NO	NO	SI	SI	NO	SI	NO	SI	NO	TEMPRAN	M	S	NINGUNA	ME	SI	MEDIO	
14	CISIC	09	1	BUENO	SUFICIENT	SUFICIENT	MUY BAJO	SI	BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	MEDIA	M	S	NINGUNA	ME	SI	MEDIO	
15	CISIC	09	1	INSUFICIE	INSUFICIE	SUFICIENT	MUY BAJO	SI	MUY BAJO	MUY BAJO	SI	NO	NO	SI	NO	SI	NO	SI	NO	TEMPRAN	M	S	NINGUNA	ME	SI	MEDIO	
16	CISIC	10	1	INSUFICIE	INSUFICIE	INSUFICIE	MUY BAJO	NO	MUY BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	MEDIA	M	S	NINGUNA	ME	SI	MEDIO	
17	CISIC	09	1	EXCELENT	EXCELENT	EXCELENT	BAJO	SI	MUY BAJO	MUY BAJO	NO	NO	SI	SI	NO	SI	NO	NO	NO	MEDIA	M	S	NINGUNA	ME	SI	BAJO	
18	CISIC	09	1	BUENO	BUENO	BUENO	BAJO	SI	MEDIO	MUY BAJO	SI	NO	NO	SI	NO	SI	NO	SI	NO	MEDIA	M	S	NINGUNA	ME	SI	BAJO	
19	CISIC	09	1	EXCELENT	BUENO	EXCELENT	MUY BAJO	SI	BAJO	MUY BAJO	NO	NO	SI	SI	NO	SI	NO	SI	NO	MEDIA	M	S	NINGUNA	ME	SI	BAJO	
20	CISIC	09	1	BUENO	EXCELENT	BUENO	MUY BAJO	SI	BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	MEDIA	M	S	NINGUNA	ME	SI	BAJO	
21	CISIC	09	1	SUFICIENT	BUENO	BUENO	MUY BAJO	SI	MUY BAJO	MUY BAJO	SI	NO	NO	SI	NO	SI	NO	SI	NO	TEMPRAN	M	S	NINGUNA	ME	SI	BAJO	
22	CISIC	10	1	INSUFICIE	INSUFICIE	INSUFICIE	MUY ALTO	NO	MUY BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	TEMPRAN	M	S	NINGUNA	ME	SI	MEDIO	
23	CISIC	10	1	INSUFICIE	INSUFICIE	INSUFICIE	MUY ALTO	NO	MUY BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	TEMPRAN	M	S	NINGUNA	ME	SI	MEDIO	
24	CISIC	09	1	INSUFICIE	EXCELENT	SUFICIENT	MEDIO	SI	MEDIO	MUY BAJO	SI	NO	NO	SI	NO	SI	NO	SI	NO	MEDIA	M	S	NINGUNA	ME	SI	MEDIO	
25	CISIC	09	1	INSUFICIE	BUENO	SUFICIENT	MUY BAJO	SI	BAJO	MUY BAJO	NO	NO	SI	SI	NO	SI	NO	SI	NO	MEDIA	M	S	NINGUNA	ME	SI	MEDIO	
26	CISIC	09	2	SUFICIENT	BUENO	SUFICIENT	MEDIO	SI	MUY BAJO	MUY BAJO	SI	NO	NO	SI	NO	SI	NO	SI	NO	TEMPRAN	M	S	NINGUNA	ME	SI	MEDIO	
27	CISIC	10	2	INSUFICIE	EXCELENT	SUFICIENT	BAJO	SI	MUY BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	TEMPRAN	M	S	NINGUNA	ME	NO	MUY BAJO	
28	CISIC	10	2	INSUFICIE	EXCELENT	SUFICIENT	BAJO	SI	MUY BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	TEMPRAN	M	S	NINGUNA	ME	NO	MUY BAJO	
29	CISIC	10	1	EXCELENT	EXCELENT	EXCELENT	MUY BAJO	SI	MUY BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	MEDIA	M	S	NINGUNA	ME	SI	MUY ALTO	
30	CISIC	10	1	EXCELENT	EXCELENT	EXCELENT	MUY BAJO	SI	MUY BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	TEMPRAN	F	S	NINGUNA	AF	SI	BAJO	
31	CISIC	10	1	EXCELENT	SUFICIENT	BUENO	MUY BAJO	SI	MUY BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	MEDIA	M	C	NINGUNA	ME	SI	MEDIO	
32	CISIC	10	2	EXCELENT	SUFICIENT	BUENO	BAJO	SI	MUY BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	TEMPRAN	M	U	NINGUNA	ME	SI	BAJO	
33	CISIC	10	2	EXCELENT	SUFICIENT	BUENO	MUY BAJO	SI	MUY BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	ALTA	M	S	NINGUNA	ME	NO	MUY ALTO	
34	CISIC	10	2	EXCELENT	INSUFICIE	INSUFICIE	ALTO	NO	MUY BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	MEDIA	M	S	NINGUNA	ME	SI	MUY BAJO	
35	CISIC	09	1	BUENO	SUFICIENT	BUENO	MUY BAJO	SI	MEDIO	MUY BAJO	SI	NO	NO	SI	NO	SI	NO	SI	NO	MEDIA	M	S	NINGUNA	ME	SI	MUY BAJO	
36	CISIC	09	1	BUENO	EXCELENT	EXCELENT	MUY BAJO	SI	BAJO	MUY BAJO	NO	NO	SI	SI	NO	SI	NO	SI	NO	MEDIA	M	S	NINGUNA	ME	SI	MUY BAJO	
37	CISIC	09	1	EXCELENT	INSUFICIE	SUFICIENT	MUY BAJO	SI	BAJO	MUY BAJO	SI	NO	SI	SI	NO	SI	NO	NO	NO	MEDIA	M	S	NINGUNA	ME	SI	MUY BAJO	

Fig. 28. Vista minable cualitativa



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	SIGLA	NIVEL_	NUMEF	NOTA1	CAT_N(	NOTA2	CAT_N(	NOTA_FIN	CAT_NOTA	APROBO	PORCENTA	CAT_POR(	NUM_ACT	CAT_USO_	NUM_REC	CAT_USO_	DOCUMEN	ENLACE	ARCHIVO	EXAMENE	FOROS_DI	PROYECTO	PRUEBAS	TAREAS	TRABAJOS	CAT_EDAD	GENERO
2	5	10	1	10	4	10	4	10	4	1	18.75	3	9	1	7	1	1	0	1	1	0	1	0	0	0	1	2
3	5	10	2	9.1	4	4.9	1	7	2	1	12.5	2	9	1	7	1	1	0	1	1	0	1	0	0	0	1	1
4	5	10	1	9.1	4	2.5	1	5.8	1	0	0	1	9	1	7	1	1	0	1	1	0	1	0	0	0	2	2
5	5	10	2	10	4	4.6	1	7.3	2	1	6.25	1	9	1	7	1	1	0	1	1	0	1	0	0	0	1	1
6	5	09	1	10	4	10	4	10	4	1	12.5	2	8	1	11	1	0	0	1	1	0	1	0	0	0	1	1
7	5	09	1	9.5	4	8.5	3	9	3	1	9.37	2	24	3	7	1	1	0	0	1	0	1	0	1	0	1	1
8	5	09	1	10	4	9.8	4	9.9	4	1	0	1	17	2	12	1	0	0	1	1	0	1	0	1	0	1	1
9	5	09	1	8.1	3	9.7	4	8.9	3	1	0	1	19	2	6	1	1	0	1	1	0	1	0	0	0	1	1
10	5	09	1	8.6	3	9.3	4	9	3	1	3.12	1	12	1	1	1	1	0	0	1	0	1	0	1	0	1	1
11	5	09	1	8.2	3	10	4	9.1	4	1	0	1	8	1	11	1	0	0	1	1	0	1	0	0	0	2	2
12	5	09	1	8.4	3	7.8	2	8.1	3	1	15.62	2	24	3	7	1	1	0	0	1	0	1	0	1	0	2	2
13	5	09	1	9.6	4	8.9	3	9.3	4	1	0	1	17	2	12	1	0	0	1	1	0	1	0	1	0	1	2
14	5	09	1	8.5	3	7	2	7.8	2	1	3.12	1	19	2	6	1	1	0	1	1	0	1	0	0	0	2	2
15	5	09	1	6.6	1	6.9	1	7.8	2	1	0	1	12	1	1	1	1	0	1	1	0	1	0	1	0	1	2
16	5	10	1	4	1	0	1	2	1	0	6.25	1	9	1	7	1	1	0	1	1	0	1	0	0	0	2	2
17	5	09	1	9.2	4	10	4	9.6	4	1	12.5	2	8	1	11	1	0	0	1	1	0	1	0	0	0	2	2
18	5	09	1	8.6	3	8.1	3	8.4	3	1	14.06	2	24	3	7	1	1	0	0	1	0	1	0	1	0	2	2
19	5	09	1	9.6	4	9	3	9.3	4	1	0	1	17	2	12	1	0	0	1	1	0	1	0	1	0	2	2
20	5	09	1	8.3	3	9.5	4	8.9	3	1	6.25	1	19	2	6	1	1	0	1	1	0	1	0	0	0	2	2
21	5	09	1	7.9	2	9	3	8.5	3	1	4.68	1	12	1	1	1	1	0	0	1	0	1	0	1	0	1	2
22	5	10	1	4.5	1	0	1	2.3	1	0	50	5	2	1	8	1	1	0	1	1	0	1	0	0	0	1	2
23	5	10	1	4.5	1	0	1	2.3	1	0	50	5	12	1	8	1	1	0	1	1	0	1	0	0	0	1	2
24	5	09	1	4.6	1	9.4	4	7	2	1	18.75	3	24	3	7	1	1	0	0	1	0	1	0	1	0	2	2
25	5	09	1	3.3	1	9	3	7.4	2	1	0	1	17	2	12	1	0	0	1	1	0	1	0	1	0	2	2
26	5	09	2	7	2	8.4	3	7.7	2	1	21.87	3	12	1	1	1	0	0	1	1	0	1	0	1	0	1	2
27	5	10	2	6.5	1	9.1	4	7.8	2	1	14.06	2	2	1	8	1	1	0	1	1	0	1	0	0	0	1	2
28	5	10	2	6.5	1	9.1	4	7.8	2	1	14.06	2	12	1	8	1	1	0	1	1	0	1	0	0	0	1	2
29	5	10	1	10	4	9.1	4	9.6	4	1	0	1	9	1	7	1	1	0	1	1	0	1	0	0	0	2	2
30	5	10	1	10	4	9.1	4	9.6	4	1	0	1	9	1	7	1	1	0	1	1	0	1	0	0	0	1	1
31	5	10	1	10	4	7.3	2	8.7	3	1	6.25	1	9	1	7	1	1	0	1	1	0	1	0	0	0	2	2
32	5	10	2	9.7	4	7	2	8.4	3	1	12.5	2	9	1	7	1	1	0	1	1	0	1	0	0	0	1	2
33	5	10	2	10	4	7.9	2	9	3	1	0	1	9	1	7	1	1	0	1	1	0	1	0	0	0	3	2
34	5	10	2	10	4	1	1	5.5	1	0	25	4	9	1	7	1	1	0	1	1	0	1	0	0	0	2	2
35	5	09	1	9	3	7.9	2	8.5	3	1	0	1	24	3	7	1	1	0	0	1	0	1	0	1	0	2	2
36	5	09	1	9	3	10	4	9.5	4	1	0	1	17	2	12	1	0	0	1	1	0	1	0	1	0	2	2
37	5	09	1	9.1	4	6.7	1	7.9	2	1	3.12	1	19	2	6	1	1	0	1	1	0	1	0	0	0	2	2

Fig. 29. Vista minable cuantitativa

La Fig. 30 muestra la fase procesamiento en la herramienta Spoon de Pentaho, para la generación de las vistas se usó la opción de reemplazar s y se realizó en otra hoja de transformación de la herramienta.

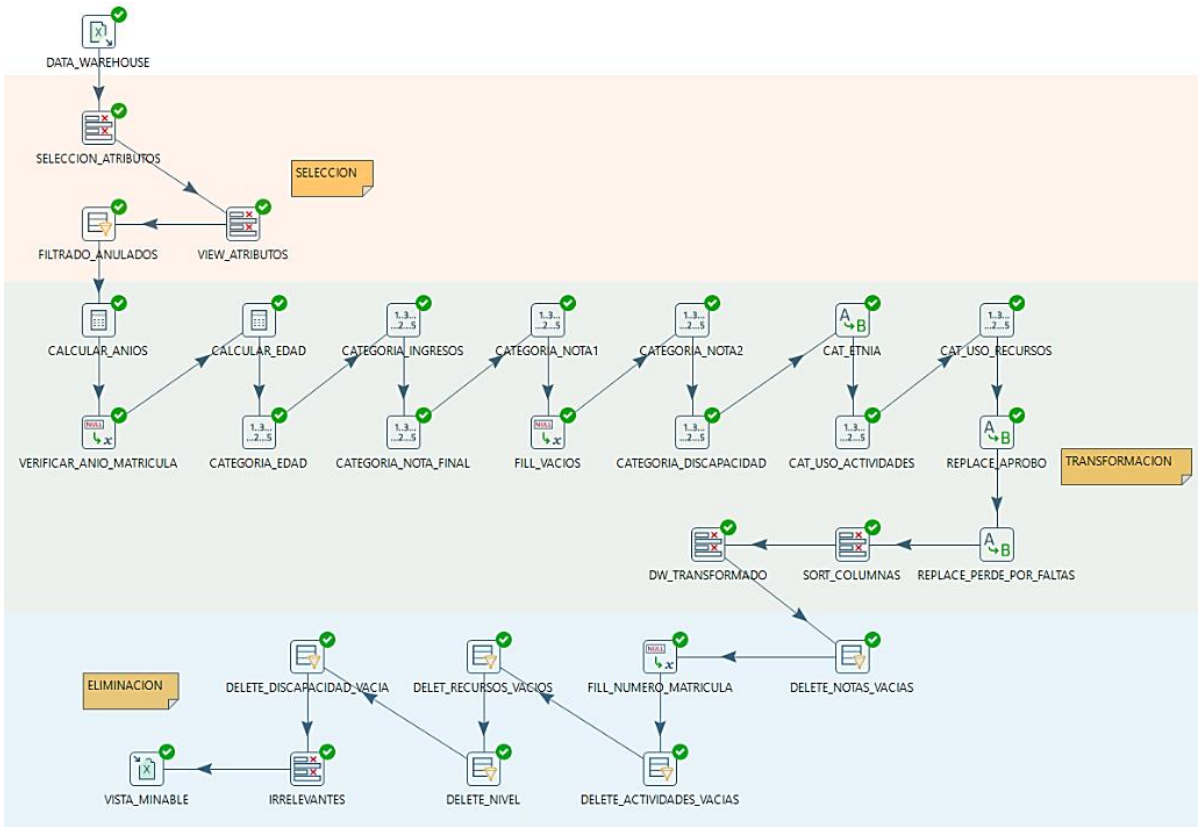


Fig. 30. Fase de Procesamiento de datos

2.7. Fase: Minería de Datos

El presente estudio tiene como objetivo el análisis del desempeño estudiantil en base a los recursos y actividades del SIIU (interacciones), además se está usando información académica y socioeconómica de los estudiantes.

Para lograr el objetivo número dos se pretende, primero hacer un análisis de correlación entre el atributo nota final y los atributos académicos, interacciones y socioeconómicos, segundo se creará un modelo de clasificación, tercero se realizará un análisis de conglomerados para identificar patrones de uso – desempeño académico y por último se aplicará asociación para descubrir reglas de asociación.

2.7.1. Correlación

La correlación mide el grado de correlación entre dos variables (x,y), en la presente sección se analizará la correlación entre la nota final (promedio) del semestre contra los datos académicos, interacciones y socioeconómico de la vista minable, de esa forma identificar cuales atributos influyen en el desempeño académico. El análisis se realiza en Jupyter Notebook con el lenguaje de Python y las librerías para ciencia de datos.

Para el presente análisis se usa la vista minable cuantitativa donde los atributos NOTA1, NOTA2, NOTA_FINAL, PORCENTAJE_FALTAS, NUM_ACTIVIDADES y NUM_RECURSOS se analizan en tipo de datos continuo y categórico (Tablas 2.28 – 2.31) los atributos restantes permanecerán como se detalla en la etapa de transformación, lo que significa que análisis se hará en dos partes de esa forma obtener resultados más sólidos.

La selección de las medidas de asociación se hace en base de la Tabla 1.2 y el análisis se organiza de la siguiente manera:

a. Correlación de atributos continuos: cuando el Atributo *NOTA_FINAL* es de tipo continuo

- Atributo continuo vs Atributo continuo:
  - *Correlación de Pearson*
  - *Correlación de Spearman*
- Atributo continuo vs Atributo ordinal:

- Correlación de Kendall
- Atributo continuo vs Atributo nominal:
  - Correlación biserial puntual

b. Correlación de atributos ordinales: cuando el Atributo *NOTA_FINAL* es de tipo ordinal resultado de la normalización de la fase de Transformación de datos:

- Atributo ordinal vs Atributo ordinal:
  - *Correlación de Kendall (Tau – b)*
  - *Correlación de Spearman*
- Atributo Ordinal vs Atributo nominal
  - *Correlación de Spearman*

a. Correlación de atributos continuos

En este paso los atributos NOTA1, NOTA2, NOTA_FINAL, PORCENTAJE_FALTAS, NUM_ACTIVIDADES y NUM_RECURSOS se analizan en tipo de datos continuo y para aplicar las medidas de asociación correcta se crea un conjunto con cada tipo de dato y se aplica correlación con la nota final.

Atributo continuo vs Atributo continuo

Para los atributos continuos se recomienda la correlación de Pearson o Spearman, crea un Dataframe con los atributos de tipo continuo. La Fig. 31 muestra estos datos.

```
In [8]: # Crear un DataFrame con ciertas columnas
        mv_continua = co_mining_view.loc[:,('NOTA1','NOTA2','PORCENTAJE_FALTAS',
                                            'NUM_ACTIVIDADES','NUM_RECURSOS','NOTA_FINAL')]
        #type(vm_continua)
        mv_continua.head()
```

Out[8]:

	NOTA1	NOTA2	PORCENTAJE_FALTAS	NUM_ACTIVIDADES	NUM_RECURSOS	NOTA_FINAL
0	10.0	10.0	18.75	9	7	10.0
1	9.1	4.9	12.50	9	7	7.0
2	9.1	2.5	0.00	9	7	5.8
3	10.0	4.6	6.25	9	7	7.3
4	10.0	10.0	12.50	8	11	10.0

Fig. 31. Dataset continuo

Correlación de Pearson

Este método mide el coeficiente de correlación en datos cuantitativos con distribución normal, es decir los valores de los atributos deben ajustarse a la campana de gauss, por eso es necesario someterse a pruebas de normalidad. Para probar la

normalidad se puede hacer por medio de la graficación de histogramas, pruebas de contraste como Kolmogorov – Smirnov o Shapiro – Wilk o también por medio de un gráfico Quantile – Quantile. En contra parte, mismo que los datos no tengan una distribución normal es posible aplicar estos algoritmos.

La Fig. 32 muestra la graficación de histogramas.

Fig. 32. Histograma de los datos continuos

Los histogramas mostraron que los atributos no tienen una distribución normal sino una distribución libre y para corroborar se aplicó las pruebas de contraste de Kolmogorov – Smirnov que define que un atributo tiene una distribución norma cuando $p > 0.05$, la Fig. 33 muestra los resultados de la prueba.

```
In [26]: # Kolmogorov-Smirnov
for column in normalize_mv:
    atributo = normalize_mv[column]
    stadistic_s, p_value_s = stats.kstest(atributo, 'norm')
    print(column)
    print('Estadistico = %.3f, p_value = %.3f' % (stadistic_s, p_value_s))
```

```
NOTA1
Estadistico = 0.658, p_value = 0.000
NOTA2
Estadistico = 0.644, p_value = 0.000
NOTA_FINAL
Estadistico = 0.677, p_value = 0.000
PORCENTAJE_FALTAS
Estadistico = 0.500, p_value = 0.000
NUM_ACTIVIDADES
Estadistico = 0.500, p_value = 0.000
NUM_RECURSOS
Estadistico = 0.500, p_value = 0.000
```

Fig. 33. Resultados de las pruebas de normalidad

Tras el análisis se evidencio que los atributos continuos no tienen una distribución normal. La Fig. 34 muestra la aplicación de la correlación de Pearson.

```
In [19]: # Grado de significancia entre variables
nivel_significancia = pg.pairwise_corr(normalize_mv,
                                       columns=[
                                           ['NOTA_FINAL'],
                                           ['NOTA1','NOTA2','PORCENTAJE_FALTAS','NUM_ACTIVIDADES','NUM_RECURSOS']],
                                       method='pearson')

# Ordenar en funcion del grado de significancia & Redondear decimales
nivel_significancia.sort_values(by = ['p-unc'])[['X', 'Y', 'n', 'r', 'p-unc']].round(9)
```

Out[19]:

	X	Y	n	r	p-unc
0	NOTA_FINAL	NOTA1	57114	0.848858	0.0
1	NOTA_FINAL	NOTA2	57114	0.910309	0.0
2	NOTA_FINAL	PORCENTAJE_FALTAS	57114	-0.289449	0.0
3	NOTA_FINAL	NUM_ACTIVIDADES	57114	-0.099376	0.0
4	NOTA_FINAL	NUM_RECURSOS	57114	0.060584	0.0

Fig. 34. Resultados de la correlación de Pearson

Correlación de Spearman

Spearman mide el coeficiente de correlación en atributos cuantitativos sin una distribución normal, la Fig. 35 muestra el algoritmo.

```
In [23]: # Grado de significancia entre variables
ns_spearman = pg.pairwise_corr(normalize_mv,
                               columns=[['NOTA_FINAL'],
                                        ['NOTA1','NOTA2','PORCENTAJE_FALTAS','NUM_ACTIVIDADES','NUM_RECURSOS']],
                               method='spearman')

# Ordenar en funcion del grado de significancia & Redondear decimales
ns_spearman.sort_values(by = ['p-unc'])[['X', 'Y', 'n', 'r', 'p-unc']].round(8)
```

Out[23]:

	X	Y	n	r	p-unc
0	NOTA_FINAL	NOTA1	57114	0.855510	0.0
1	NOTA_FINAL	NOTA2	57114	0.873549	0.0
2	NOTA_FINAL	PORCENTAJE_FALTAS	57114	-0.225410	0.0
4	NOTA_FINAL	NUM_RECURSOS	57114	0.118170	0.0
3	NOTA_FINAL	NUM_ACTIVIDADES	57114	-0.114879	0.0

Fig. 35. Resultados de la correlación de Spearman

Atributo continuo vs Atributo ordinal

Para los datos de tipo ordinal se aplica la correlación de Kendall, es necesario mencionar que los datos de tipo ordinal y nominal tiene una distribución libre y no es necesario probar la normalidad en los atributos.

Al igual que en el parte anterior se creó un DataFrame con los atributos ordinales. La Fig. 36 muestra estos datos.

```
In [28]: # Crear un DataFrame con continuo-ordianl
         ordinal_mv = co_mining_view.loc[:,('NIVEL_CODIGO','NUMERO_MATRICULA','CAT_EDAD',
                                            'CAT_DISCPACIDAD','CAT_INGRESOS','NOTA_FINAL')]
         #type(vm_continua)
         ordinal_mv.head()
```

Out[28]:

	NIVEL_CODIGO	NUMERO_MATRICULA	CAT_EDAD	CAT_DISCPACIDAD	CAT_INGRESOS	NOTA_FINAL
0	10	1	1	1	1	10.0
1	10	2	1	1	2	7.0
2	10	1	2	1	2	5.8
3	10	2	1	1	2	7.3
4	9	1	1	1	1	10.0

Fig. 36. Dataset de ordinales

Correlación de Kendall

La correlación de Kendall mide la correlación en base a pares discordantes o concordantes, para este análisis se seleccionó el Tau-b de kendall la Fig. 37 muestra el algoritmo.

```
In [36]: # Grado de significancia entre variables
         or_taub = pg.pairwise_corr(ordinal_mv,
                                    columns=[['NOTA_FINAL'],
                                             ['NIVEL_CODIGO','NUMERO_MATRICULA','CAT_EDAD',
                                              'CAT_DISCPACIDAD','CAT_INGRESOS']],
                                    method='kendall')

         # Ordenar en funcion del grado de significancia & Redondear decimales
         or_taub.sort_values(by = ['p-unc'])[['X', 'Y', 'n', 'r', 'p-unc']].round(6)
```

Out[36]:

	X	Y	n	r	p-unc
0	NOTA_FINAL	NIVEL_CODIGO	57114	0.107709	0.0
1	NOTA_FINAL	NUMERO_MATRICULA	57114	-0.106106	0.0
2	NOTA_FINAL	CAT_EDAD	57114	-0.056913	0.0
3	NOTA_FINAL	CAT_DISCPACIDAD	57114	-0.032946	0.0
4	NOTA_FINAL	CAT_INGRESOS	57114	0.017231	0.0

Fig. 37. Resultados de la correlación de Kendall

Atributo Continuo vs Atributo Nominal

Cuando se tiene un atributo de tipo continuo y otro nominal es recomendable usar la correlación Biserial Puntual. La Fig. 38 muestra la creación de un DataFrame con estos datos y la Fig. 39 el algoritmo de la correlación.

```
In [39]: # Crear un DataFrame con continuo-nominal
nominal_mv = co_mining_view.loc[:,('SIGLAS_CARRERA','APROBO','DOCUMENTO','ENLACE','ARCHIVO','EXAMENES',
                                   'FOROS_DEBATES_OTROS','PROYECTO','PRUEBAS','TAREAS','TRABAJOS',
                                   'GENERO','ESTADO_CIVIL','ETNIA','SERVICIO_INTERNET','NOTA_FINAL',)]
nominal_mv.head()
```

Out[39]:

	SIGLAS_CARRERA	APROBO	DOCUMENTO	ENLACE	ARCHIVO	EXAMENES	FOROS_DEBATES_OTROS	PROYECTO	PRUEBAS	TAREAS
0	5	1	1	0	1	1	0	1	0	0
1	5	1	1	0	1	1	0	1	0	0
2	5	0	1	0	1	1	0	1	0	0
3	5	1	1	0	1	1	0	1	0	0
4	5	1	0	0	1	1	0	1	0	0

Fig. 38. Dataset de nominales

```
In [40]: # Correlacion Biserial Puntual

# atributo x
xc = nominal_mv['NOTA_FINAL']

for column in nominal_mv:
    # variable y
    yc = nominal_mv[column]
    correlation, p_value = stats.pointbiserialr(list(xc), list(yc))
    print("NOTA_FINAL vs " + column)
    print('r = %.3f, p_value = %.3f' % (correlation, p_value) + '\n')
```

```
NOTA_FINAL vs SIGLAS_CARRERA
r = -0.001, p_value = 0.760
```

Fig. 39. Algoritmo punto biserial

b. Correlación de atributos ordinales

En este paso los atributos NOTA1, NOTA2, NOTA_FINAL, PORCENTAJE_FALTAS, NUM_ACTIVIDADES y NUM_RECURSOS se analizan en tipo de datos ordinal resultado de la fase de transformación de datos.

Atributo ordinal vs Atributo ordinal

Para los atributos de tipo ordinal se recomienda el uso de la correlación de Spearman o Kendall, se crea un conjunto de datos de tipo ordinal como se puede observar en la Fig. 40.

```
In [6]: # Crear un DataFrame con variables ordinales
ord_miningview = or_mining_view.loc[:,('NIVEL_CODIGO','NUMERO_MATRICULA','CAT_NOTA1','CAT_NOTA2',
                                       'CAT_PORCENTAJE_FALTAS','CAT_USO_ACTIVIDADES','CAT_USO_RECURSOS',
                                       'CAT_EDAD','CAT_DISCPACIDAD','CAT_INGRESOS','CAT_NOTA_FINAL')]
#type(vm_continua)
ord_miningview.head()
```

Out[6]:

	NIVEL_CODIGO	NUMERO_MATRICULA	CAT_NOTA1	CAT_NOTA2	CAT_PORCENTAJE_FALTAS	CAT_USO_ACTIVIDADES	CAT_USO_RECU
0	10	1	4	4	3	1	
1	10	2	4	1	2	1	
2	10	1	4	1	1	1	
3	10	2	4	1	1	1	
4	9	1	4	4	2	1	

Fig. 40. Dataset de ordinales

Correlación de Kendall

La Fig. 41 muestra la aplicación de esta correlación al conjunto de datos.

```
In [8]: # Grado de significancia entre variables
or_ns_taub = pg.pairwise_corr(ord_miningview,
                              columns=[['CAT_NOTA_FINAL'],
                                       ['NIVEL_CODIGO','NUMERO_MATRICULA','CAT_NOTA1','CAT_NOTA2',
                                        'CAT_PORCENTAJE_FALTAS','CAT_USO_ACTIVIDADES','CAT_USO_RECURSOS',
                                        'CAT_EDAD','CAT_DISCPACIDAD','CAT_INGRESOS']],
                              method='kendall')

# Ordenar en funcion del grado de significancia & Redondear decimales
or_ns_taub.sort_values(by = ['p-unc'])[['X', 'Y', 'n', 'r', 'p-unc']].round(6)
```

Out[8]:

	X	Y	n	r	p-unc
2	CAT_NOTA_FINAL	CAT_NOTA1	57114	0.727850	0.0
3	CAT_NOTA_FINAL	CAT_NOTA2	57114	0.750709	0.0
4	CAT_NOTA_FINAL	CAT_PORCENTAJE_FALTAS	57114	-0.144441	0.0
0	CAT_NOTA_FINAL	NIVEL_CODIGO	57114	0.109400	0.0
5	CAT_NOTA_FINAL	CAT_USO_ACTIVIDADES	57114	-0.119241	0.0
1	CAT_NOTA_FINAL	NUMERO_MATRICULA	57114	-0.109763	0.0
6	CAT_NOTA_FINAL	CAT_USO_RECURSOS	57114	0.072379	0.0
7	CAT_NOTA_FINAL	CAT_EDAD	57114	-0.056675	0.0
8	CAT_NOTA_FINAL	CAT_DISCPACIDAD	57114	-0.037312	0.0
9	CAT_NOTA_FINAL	CAT_INGRESOS	57114	0.018782	0.0

Fig. 41. Resultados de correlación de Kendall

Correlación de Spearman

Al mismo conjunto de datos se aplicó la correlación de Spearman como se muestra en la Fig. 42.

```
In [11]: # Grado de significancia entre variables
od_ns_spearman = pg.pairwise_corr(ord_miningview,
                                  columns=[['CAT_NOTA_FINAL'],
                                           ['NIVEL_CODIGO','NUMERO_MATRICULA','CAT_NOTA1','CAT_NOTA2',
                                            'CAT_PORCENTAJE_FALTAS','CAT_USO_ACTIVIDADES','CAT_USO_RECURSOS',
                                            'CAT_EDAD','CAT_DISCPACIDAD','CAT_INGRESOS']],
                                  method='spearman')

# Ordenar en funcion del grado de significancia & Redondear decimales
od_ns_spearman.sort_values(by = ['p-unc'])[['X', 'Y', 'n', 'r', 'p-unc']].round(6)
```

Out[11]:

	X	Y	n	r	p-unc
2	CAT_NOTA_FINAL	CAT_NOTA1	57114	0.790149	0.0
3	CAT_NOTA_FINAL	CAT_NOTA2	57114	0.811961	0.0
4	CAT_NOTA_FINAL	CAT_PORCENTAJE_FALTAS	57114	-0.156144	0.0
0	CAT_NOTA_FINAL	NIVEL_CODIGO	57114	0.134545	0.0
5	CAT_NOTA_FINAL	CAT_USO_ACTIVIDADES	57114	-0.131684	0.0
1	CAT_NOTA_FINAL	NUMERO_MATRICULA	57114	-0.118736	0.0
6	CAT_NOTA_FINAL	CAT_USO_RECURSOS	57114	0.078085	0.0
7	CAT_NOTA_FINAL	CAT_EDAD	57114	-0.061721	0.0
8	CAT_NOTA_FINAL	CAT_DISCPACIDAD	57114	-0.040229	0.0
9	CAT_NOTA_FINAL	CAT_INGRESOS	57114	0.021903	0.0

Fig. 42. Resultados de correlación de Spearman

Atributo ordinal vs Atributo nominal

Para los atributos de tipo nominal se recomienda la correlación de Spearman, para eso se creó un dataset con este tipo de atributos como se muestra en la Fig. 43.

```
In [14]: # Crear un DataFrame con ordinal-nominal
nom_mimingview = or_mining_view.loc[:,('SIGLAS_CARRERA','APROBO','DOCUMENTO','ENLACE','ARCHIVO','EXAMENES',
                                       'FOROS_DEBATES_OTROS','PROYECTO','PRUEBAS','TAREAS','TRABAJOS',
                                       'GENERO','ESTADO_CIVIL','ETNIA','SERVICIO_INTERNET','CAT_NOTA_FINAL',)]
nom_mimingview.head()
```

Out[14]:

	SIGLAS_CARRERA	APROBO	DOCUMENTO	ENLACE	ARCHIVO	EXAMENES	FOROS_DEBATES_OTROS	PROYECTO	PRUEBAS	TARI
0	5	1	1	0	1	1	0	1	0	
1	5	1	1	0	1	1	0	1	0	
2	5	0	1	0	1	1	0	1	0	
3	5	1	1	0	1	1	0	1	0	
4	5	1	0	0	1	1	0	1	0	

Fig. 43. Dataset de nominales

Correlación de Spearman

Se aplico esta correlación al conjunto de datos como se observa en la Fig. 44.

```
In [16]: # Grado de significancia entre variables
no_ns_spearman = pg.pairwise_corr(nom_mimingview,
                                  columns=[['CAT_NOTA_FINAL'],
                                           ['SIGLAS_CARRERA','APROBO','DOCUMENTO','ENLACE','ARCHIVO','EXAMENES',
                                           'FOROS_DEBATES_OTROS','PROYECTO','PRUEBAS','TAREAS','TRABAJOS',
                                           'GENERO','ESTADO_CIVIL','ETNIA','SERVICIO_INTERNET']],
                                  method='spearman')

# Ordenar en funcion del grado de significancia & Redondear decimales
no_ns_spearman.sort_values(by = ['p-unc'])[['X', 'Y', 'n', 'r', 'p-unc']].round(6)
```

Out[16]:

	X	Y	n	r	p-unc

Fig. 44. Resultados de correlación de Spearman

2.7.2. Modelo de Clasificación

Generalmente un análisis dentro de la minería de datos educacional se realiza una vez finalizado el curso y varias de las acciones correctivas son imposibles de aplicarlas. Helal (2017) menciona que la detección de riesgos en una fase temprana es vital para las instituciones.

En esta sección se construirá un modelo de clasificación con el fin de predecir el desempeño académico, específicamente si el estudiante aprobara o no aprobará. Para que el modelo tenga la capacidad de predecir en la mitad del curso o periodo académico, se lo entrena con la nota de la primera parcial y se excluirá la nota del segundo parcial y el promedio del semestre del conjunto de datos.

La clasificación es semejante a la regresión, la única diferencia es que la variable objetivo a predecir es categórica (Larose D. & Larose C., 2014), el entrenamiento de un modelo de clasificación se hace con atributos asociados o que contenga información sobre la variable dependiente los cuales se denomina variables independientes.

Selección de variables independientes

La selección las variables independientes para el modelo se realizó por medio de correlación, el proceso de selección se hace en dos pasos, primero, se creó un conjunto de datos excluyendo la nota del segundo parcial y la nota final, la Fig. 45 muestra este paso.

```
In [3]: # Datos preseleccionados para modelo
model_mining_view = mining_view.loc[:,('SIGLAS_CARRERA','NIVEL_CODIGO','NUMERO_MATRICULA','NOTA1','APROBO',
                                       'PORCENTAJE_FALTAS','CAT_USO_ACTIVIDADES','CAT_USO_RECURSOS','DOCUMENTO',
                                       'ENLACE','ARCHIVO','EXAMENES','FOROS_DEBATES_OTROS','PROYECTO',
                                       'PRUEBAS','TAREAS','TRABAJOS','CAT_EDAD','GENERO','ESTADO_CIVIL',
                                       'CAT_DISCPACIDAD','ETNIA','SERVICIO_INTERNET','CAT_INGRESOS')]

model_mining_view.head()
```

Out[3]:

	SIGLAS_CARRERA	NIVEL_CODIGO	NUMERO_MATRICULA	NOTA1	APROBO	PORCENTAJE_FALTAS	CAT_USO_ACTIVIDADES	CAT_USC
0	5	10	1	10.0	1	18.75	1	
1	5	10	2	9.1	1	12.50	1	
2	5	10	1	9.1	0	0.00	1	
3	5	10	2	10.0	1	6.25	1	

Fig. 45. Preselección de variables independientes

Luego, para identificar las variables relevantes se aplicó correlación entre las variables del paso anterior y la variable dependiente “APROBO” (SI, NO) de tipo nominal.

En el proceso se usó correlación Biserial-puntual, Spearman y coeficientes de Phi. La Tabla 2.42 se puede observar los coeficientes de correlación.

TABLA 2.42
VARIABLES INDEPENDIENTES

x	y	Biserial-Puntual	Spearman	Phi
APROBO	SIGLAS_CARRERA			0.2620
APROBO	NIVEL_CODIGO		0.0340	
APROBO	NUMERO_MATRICULA		-0.0910	
APROBO	NOTA1	0.5390		
APROBO	PORCENTAJE_FALTAS	-0.2240		
APROBO	CAT_USO_ACTIVIDADES		-0.061295	
APROBO	CAT_USO_RECURSOS		0.022612	
APROBO	DOCUMENTO			0.0434

APROBO	ENLACE		0.0516
APROBO	ARCHIVO		0.0184
APROBO	EXAMENES		0.0000
APROBO	FOROS_DEBATES_OTROS		0.1585
APROBO	PROYECTO		0.1095
APROBO	PRUEBAS		0.1463
APROBO	TAREAS		0.1308
APROBO	TRABAJOS		0.1445
APROBO	CAT_EDAD	-0.0648	
APROBO	GENERO		0.0420
APROBO	ESTADO_CIVIL		0.0112
APROBO	CAT_DISCPACIDAD	-0.0410	
APROBO	ETNIA		0.0154
APROBO	SERVICIO_INTERNET		0.0218
APROBO	CAT_INGRESOS	-0.0040	

Los atributos *EXAMENES* y *CAT_INGRESOS* mostraron una correlación nula con un nivel de significancia menos a 0.05 mostrando que no ninguna relación, los demás atributos tienen una correlación que va desde muy baja a moderada como se observa en la Fig. 46.

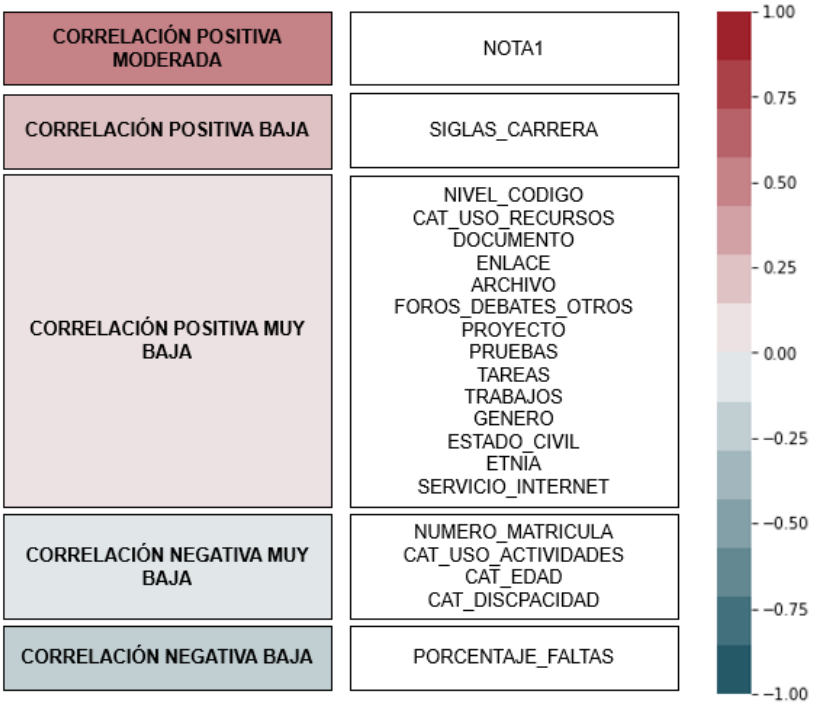


Fig. 46. Grado de correlación de variables independientes

También se usó el modelo XGBoost para identificar las variables relevantes. El algoritmo de XGBoost por medio de árboles de decisión y proceso de boosting busca

generar modelos precisos, donde crea varios modelos y cada nuevo modelo intenta corregir las deficiencias del modelo anterior (Mitchell R. & Frank E., 2017). La Fig. 47 muestra la estructura del modelo con parámetros obtenidos con Grid Search.

```
In [22]: # Estructura del modelo
         xgb_model = XGBClassifier(nestimators=10,
                                   nthreads=1,
                                   objective='binary:logistic',
                                   learning_rate = 0.1)
```

Fig. 47. Modelo XGBoost

Los resultados obtenidos mostraron una similitud a la correlación donde el atributo *EXAMENES* tiene el ultimo nivel de importancia como se observa en la Fig. 48.

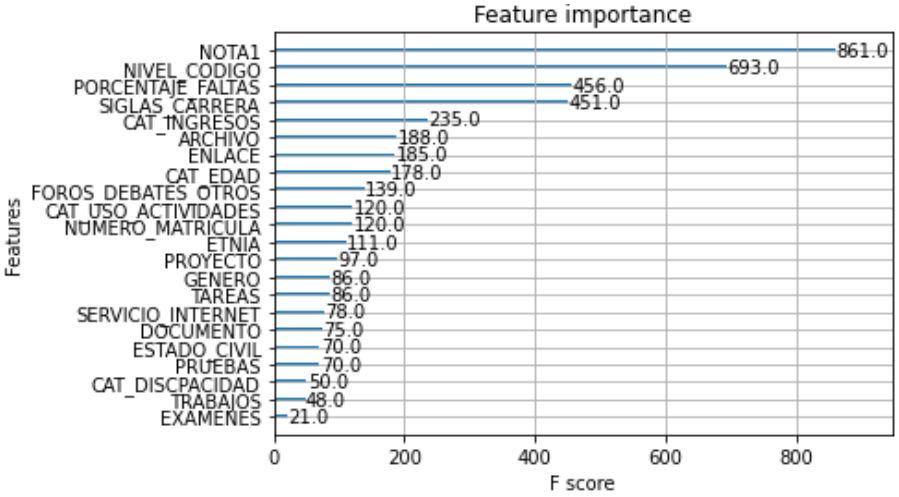


Fig. 48. Nivel de importancia de las variables independientes

Dado que los resultados no tuvieron una gran variación en los dos análisis, las variables independientes para el modelo de clasificación son los atributos analizados a excepción de *EXAMENES* y *CAT_INGRESOS.*

Finalmente, previo al entrenamiento y su posterior validación de los modelos de clasificación se dividió en datos de entrenamiento y testeo con porcentajes de 70% - 30% respectivamente.

Random Forest

Se selecciono Random Forest (RF) porque ha mostrado óptimos resultados en minería de datos academicos (Hasan, 2018), este algoritmo está compuesto de un conjunto de árboles de decisión (Decision Tree) donde cada árbol predice una categoría, y se selecciona la categoría con más votos.

La construcción del modelo se hiso mediante Grid Search basado en validación cruzada para la obtención de los mejores parámetros, de esa forma obtener un modelo optimo. Grid Search consiste en probar todas las combinaciones posibles de los parámetros de interés (Müller A. & Guido S., 2016), la Fig. 49 muestra la estructura del modelo.

```
In [25]: # Estructura del modelo
         rf_modeli = RandomForestClassifier(n_estimators=100,
                                            n_jobs=-1,
                                            random_state=1)

In [26]: # Entrenamiento
         # ============================================================================
         rf_modeli.fit(Xi_train, y_train)

Out[26]: RandomForestClassifier
         RandomForestClassifier(n_jobs=-1, random_state=1)
```

Fig. 49. Modelo Random Forest

La Fig. 50 muestra los resultados del entrenamiento del modelo.

```
In [27]: # Resultados del entrenamiento del modelo
         print_score(rf_modeli, Xi_train, y_train, Xi_test, y_test, train=True)

================================================
TRAIN RESULTS:
================================================
Accuracy Score: 99.35%
________________________________________________
CLASSIFICATION REPORT:
                     0            1  accuracy     macro avg  weighted avg
precision     0.978226     0.994743  0.993472      0.986484      0.993419
recall        0.939451     0.998178  0.993472      0.968814      0.993472
f1-score      0.958446     0.996458  0.993472      0.977452      0.993411
support    3204.000000  36775.000000  0.993472  39979.000000  39979.000000
________________________________________________
```

Fig. 50. Resultados del entrenamiento del modelo RF

El accuracy del entrenamiento nos muestra una eficiencia bastante alta en el aprendizaje.

Support Vector Machine

Support Vector Machine (SVM) es un clasificador basado en la búsqueda de un óptimo hiperplano con el fin de definir a que categoría pertenece los datos de entrada. Para la construcción del modelo también se usó Grid Search basado en validación cruzada para identificar los mejores parámetros. La Fig. 51 muestra la estructura y el entrenamiento del modelo.

```
In [39]: # Estructura del modelo
svm_optimo_model = SVC(
    C = 1,
    kernel = 'rbf',
    gamma = 1
)

In [40]: # Entrenamiento
svm_optimo_model.fit(Xi_train, y_train)

Out[40]: SVC
SVC(C=1, gamma=1)
```

Fig. 51. Modelo Support Vector Machine

El modelo está configurado con un *kernel* gaussiano, una *C* baja y una *gamma* baja. Müller A. & Guido S. (2016) menciona que el parámetro *gamma* controla la anchura del kernel gaussiano y el parámetro C regula la importancia de los puntos. Lo que significa, en un gamma bajo el límite de decisión varia lentamente (produce un modelo de baja complejidad) y una *C* pequeña hace que los puntos tengan influencia limitada (modelo más restringido).

En la Fig. 52 se puede observar el reporte del entrenamiento, el accuracy muestra una eficiencia alta en el aprendizaje.

```
In [48]: # Resultados del entrenamiento del modelo
print_score(svm_optimo_model, Xi_train, y_train, Xi_test, y_test, train=True)

================================================
TRAIN RESULTS:
================================================
Accuracy Score: 98.31%
________________________________________________
CLASSIFICATION REPORT:
                     0             1  accuracy     macro avg  weighted avg
precision     0.965035      0.984381  0.983066      0.974708      0.982830
recall        0.818352      0.997417  0.983066      0.907884      0.983066
f1-score      0.885661      0.990856  0.983066      0.938259      0.982425
support    3204.000000  36775.000000  0.983066  39979.000000  39979.000000
________________________________________________
```

Fig. 52. Resultados de entrenamiento del modelo SVM

2.7.3. Agrupación

En esta sección se realiza un análisis de conglomerados para identificar los patrones de rendimiento – uso. El conjunto de datos se conforma por atributos que tienen correlación con los atributos nota final y aprobó resultado de las etapas de correlación y clasificación.

El agrupamiento es un método no supervisado donde forma grupos o clusters en base a las características de los atributos. Se estableció un valor de $k = 4$ en base al número de categorías del atributo nota final que refleja el estado académico del

estudiante, haciendo posible identificar las características de cada uno de los 4 clusters (clusters de los estudiantes con notas insuficiente, suficiente, bueno y excelente).

K-Means

El algoritmo de K-Means calcula el centroide como una media aproximada de los datos y asocia cada centroide a un cluster, los puntos cercanos al centroide se asignan al cluster. Una de las desventajas es la elección de *k* que corresponde a número de clusters a formarse. La Fig. 53 muestra los parámetros del algoritmo en Weka.

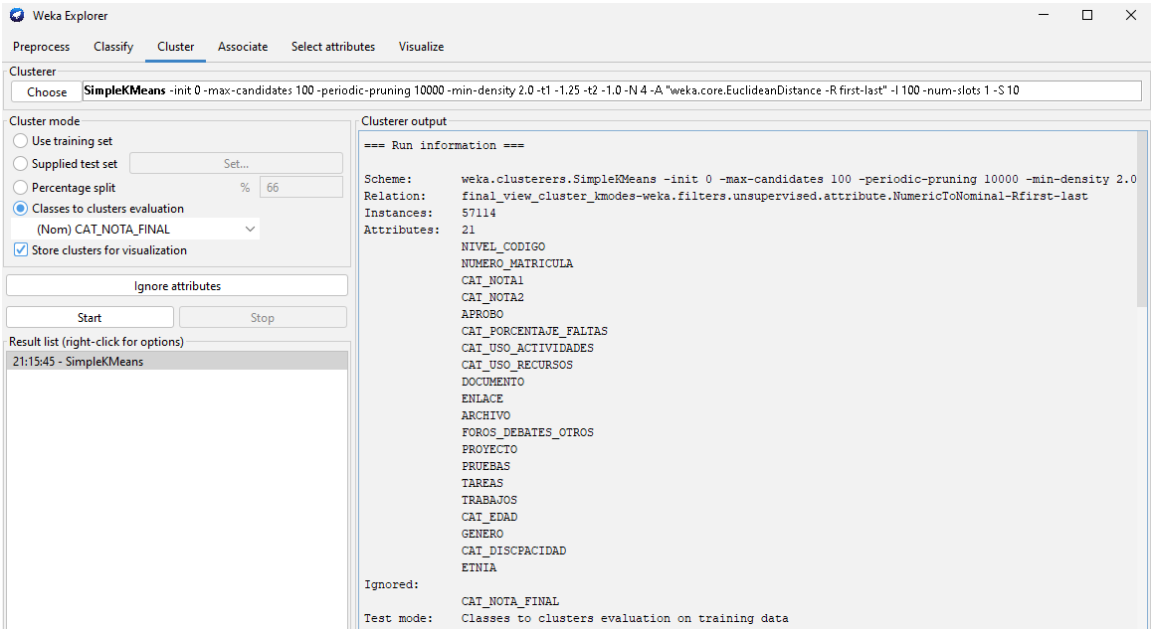


Fig. 53. Algoritmo Simple K-Means

K-Prototype

Es un algoritmo que tiene la capacidad de agrupamiento cuando un conjunto de datos es mixto es decir datos continuos y categóricos. Para la formación de clústeres usa la distancia euclidiana para datos continuos y la distancia de concordancia simple para datos categóricos. La Fig. 54 muestra la implementación en R.

```
81 # Agrupacion
82 #=============================================================
83 # Clusters con k-prototype
84 kproto_model <- kproto(agrupaminto_mv,
85                        4,
86                        iter.max = 100)
87 # resultado de agrupacion
88 kproto_model
89 # resumen de agrupacion
90 summary(kproto_model)
```

Fig. 54. Algoritmo K-Prototype

K-Modes

K-Modes es un algoritmo de agrupación para datos categóricos. Para aplicar este algoritmo se usó el mismo conjunto de datos que él le paso anterior, pero con atributos categóricos. La Fig. 55 muestra la implantación en R.

```
# Agrupacion
#=========================================================
#Clusters con k-modes
kmodes_model <- kmodes(agrupaminto_km,
                       4,
                       iter.max = 200,
                       weighted = FALSE,
                       fast = TRUE)
# agrupaciones
kmodes_model
# resumen de agrupacion
summary(kmodes_model)
```

Fig. 55. Algoritmo K-Modes

2.7.4. Asociación

En esta sección se realiza un análisis de asociación al conjunto de datos usado el algoritmo de Apriori, la asociación se aplicó a dos conjuntos de datos, el primer conjunto de datos tiene correlación con la nota final y el segundo al atributo aprueba.

La Fig. 56 muestra la aplicación de este algoritmo.

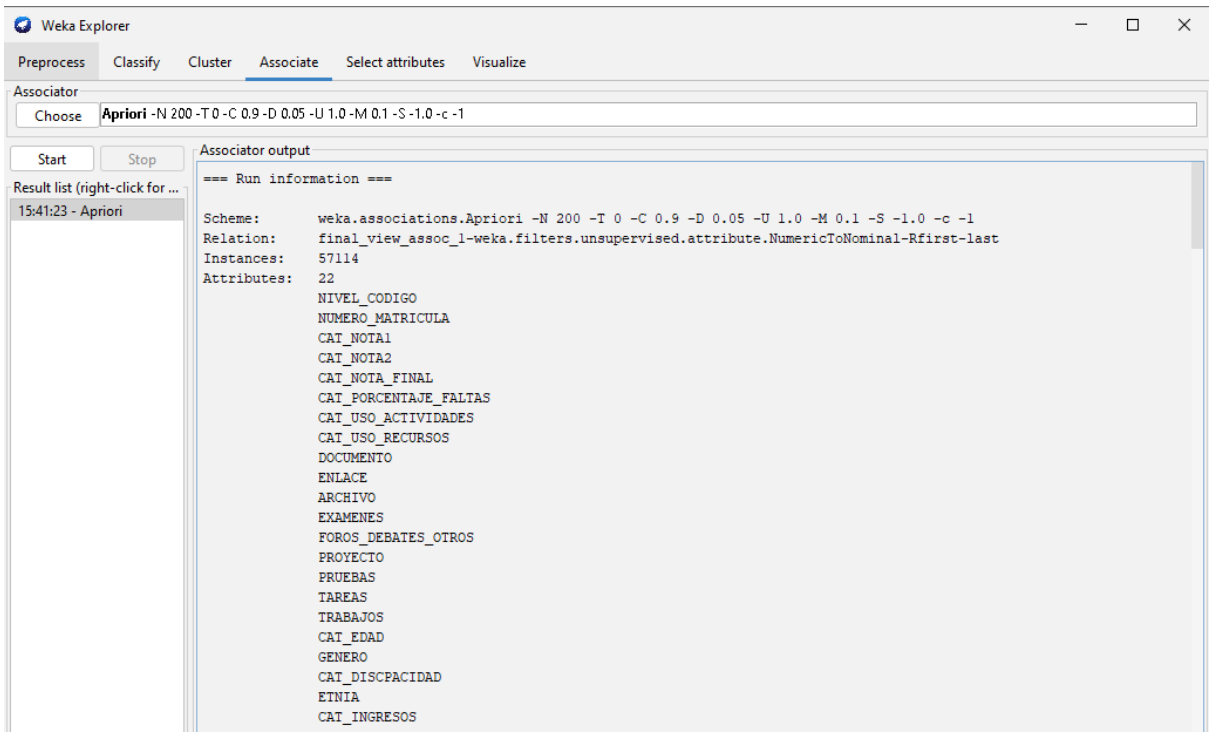


Fig. 56. Algoritmo de Apriori para reglas de asociación

CAPÍTULO 3

Validación de Resultados

3.1. Fase: Validación y Interpretación

3.1.1. Validación y Análisis

3.1.1.2. Validación e análisis de Correlación

Para identificar que atributos influyen en el desempeño académico se realizó un análisis de correlación entre la nota fina y los atributos académicos, interacciones y socioeconómicos. La correlación se realizó en dos partes, cuando el atributo *NOTA_FINAL* es continuo y ordinal.

Dado que el nivel de significancia estable una correlación entre dos variables como los menciona McDonald J.:

> El nivel de significancia o de confianza se denota con la letra *p* y establece que una correlación es estadísticamente significativa si se cumple que $p < 0.05$, lo cual representaría un 5% de posibilidad de error (McDonald J., 2014) y un 95% de confianza. Una correlación es muy significativa si $p < 0.01$ lo que significa que hay 99% de confianza.

Para la validación se calculó la significancia en cada una de las correlaciones y posteriormente su coeficiente de correlación el cual se presenta en forma de matriz.

Validación de los atributos continuos

En la Tabla 3.1 se puede observar el nivel de significancia para cada variable.

TABLA 3.1
NIVEL DE SIGNIFICANCIA EN ATRIBUTOS CONTINUOS

x	y	Pearson	Spearman	Kendall	Biserial-Puntual
NOTA_FINAL	SIGLAS_CARRERA				0.760
NOTA_FINAL	NIVEL_CODIGO			0.000	
NOTA_FINAL	NUMERO_MATRICULA			0.000	
NOTA_FINAL	NOTA1	0.000	0.000		
NOTA_FINAL	NOTA2	0.000	0.000		
NOTA_FINAL	APROBO				0.000
NOTA_FINAL	PORCENTAJE_FALTAS	0.000	0.000		
NOTA_FINAL	NUM_ACTIVIDADES	0.000	0.000		
NOTA_FINAL	NUM_RECURSOS	0.000	0.000		

NOTA_FINAL	DOCUMENTO		0.000
NOTA_FINAL	ENLACE		0.000
NOTA_FINAL	ARCHIVO		0.000
NOTA_FINAL	EXAMENES		0.000
NOTA_FINAL	FOROS_DEBATES_OTROS		0.000
NOTA_FINAL	PROYECTO		0.000
NOTA_FINAL	PRUEBAS		0.000
NOTA_FINAL	TAREAS		0.000
NOTA_FINAL	TRABAJOS		0.000
NOTA_FINAL	CAT_EDAD	0.000	
NOTA_FINAL	GENERO		0.000
NOTA_FINAL	ESTADO_CIVIL		0.298
NOTA_FINAL	CAT_DISCPACIDAD	0.000	
NOTA_FINAL	ETNIA		0.000
NOTA_FINAL	SERVICIO_INTERNET		0.053
NOTA_FINAL	CAT_INGRESOS	0.000	

Se puede constatar que todos los atributos tienen una correlación estadísticamente significativa lo que significa que si hay correlación entre los atributos a excepción de los atributos SIGLAS_CARRERA, ESTADO_CIVIL y SERVICIO_INTERNET que mostraron una $p > 0.05$.

Validación de los atributos ordinales

Para los atributos ordinales se usó el método de Kendall y Spearman, donde primero se calculó el nivel de significancia y luego le coeficiente de correlación. La Tabla 3.2 el nivel de significancia.

TABLA 3.2
NIVEL DE SIGNIFICANCIA EN ATRIBUTOS ORDINARIOS

x	y	Spearman	Kendall
NOTA_FINAL	SIGLAS_CARRERA	0.595	
NOTA_FINAL	NIVEL_CODIGO	0.000	0.000
NOTA_FINAL	NUMERO_MATRICULA	0.000	0.000
NOTA_FINAL	CAT_NOTA1	0.000	0.000
NOTA_FINAL	CAT_NOTA2	0.000	0.000
NOTA_FINAL	APROBO	0.000	
NOTA_FINAL	CAT_PORCENTAJE_FALTAS	0.000	0.000
NOTA_FINAL	CAT_USO_ACTIVIDADES	0.000	0.000
NOTA_FINAL	CAT_USO_RECURSOS	0.000	0.000
NOTA_FINAL	DOCUMENTO	0.000	
NOTA_FINAL	ENLACE	0.000	

NOTA_FINAL	ARCHIVO	0.002	
NOTA_FINAL	EXAMENES	0.000	
NOTA_FINAL	FOROS_DEBATES_OTROS	0.000	
NOTA_FINAL	PROYECTO	0.000	
NOTA_FINAL	PRUEBAS	0.000	
NOTA_FINAL	TAREAS	0.000	
NOTA_FINAL	TRABAJOS	0.000	
NOTA_FINAL	CAT_EDAD	0.000	0.000
NOTA_FINAL	GENERO	0.000	
NOTA_FINAL	ESTADO_CIVIL	0.000	
NOTA_FINAL	CAT_DISCPACIDAD	0.000	0.000
NOTA_FINAL	ETNIA	0.000	
NOTA_FINAL	SERVICIO_INTERNET	0.163	
NOTA_FINAL	CAT_INGRESOS	0.000	0.000

Se puede observar que los atributos SIGLAS_CARRERA y SERVICIO_INTERNET no tienen una correlación estadísticamente significativa, es decir no hay correlación entre estos atributos y la NOTA_FINAL.

3.1.1.2. Validación e análisis del modelo de predicción

Para analizar el desempeño académico de los estudiantes se construyó dos modelos de predicción usando clasificación a fin de determinar si el estudiante aprobara o reprobará el curso. La validación de los modelos se hizo por medio de métricas cuantitativas basada en la matriz de confusión, exactitud, precisión, sensibilidad, puntaje de f1, coeficiente Kappa y área bajo la curva. El entrenamiento se realizó con validación cruzada para observar el comportamiento del modelo en un ambiente de producción.

La variable por predecir está compuesta de dos categorías:

- 1 = Aprobó
- 0 = No aprobó

Random Forest

En la Fig. 57 se observa la matriz de confusión que resulto de la ejecución del modelo con el algoritmo de Random Forest y el conjunto de datos de prueba con 17 135 registros.

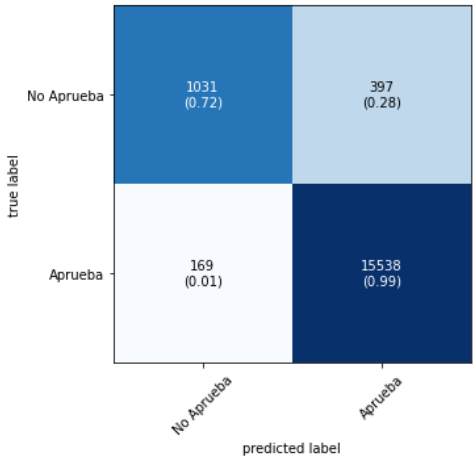


Fig. 57. Matrix de confusión del modelo Random Forest

Se puede observar, 1 031 reales negativos, 397 falsos negativos, 15 538 reales positivos y 169 falsos positivos, lo que significa que el modelo de Random Forest clasifico de forma correcta un total de 16 569 registros y 566 registros de forma incorrecta, la Fig. 58 muestra la curva ROC del modelo.

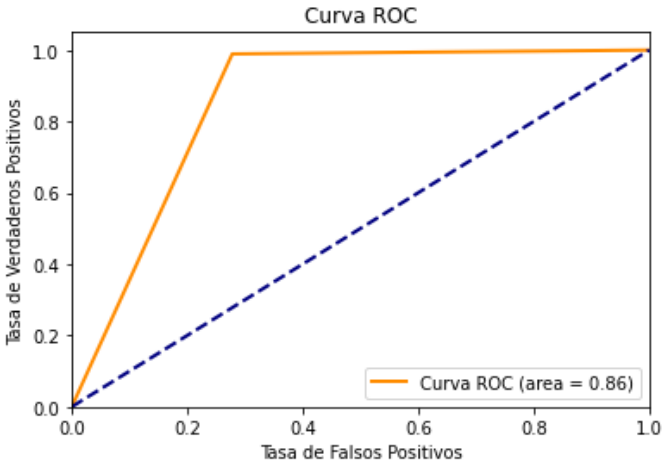


Fig. 58. Curva ROC del modelo Random Forest

Además de la matriz de confusión la Tabla 3.3 muestra las métricas estadísticas de validación:

TABLA 3.3
MEDIDAS ESTADÍSTICAS DEL MODELO RF

Medida	Valor
Accuracy	96.70%
Sensibilidad	98.92%
Precisión	97.51%
F1-score	98.21%
Área bajo la curva	0.86
Coeficiente Kappa	0.77

Support Vector Machine

La Fig. 59 muestra la matriz de confusión del modelo SVM después de la ejecución con los datos de prueba.

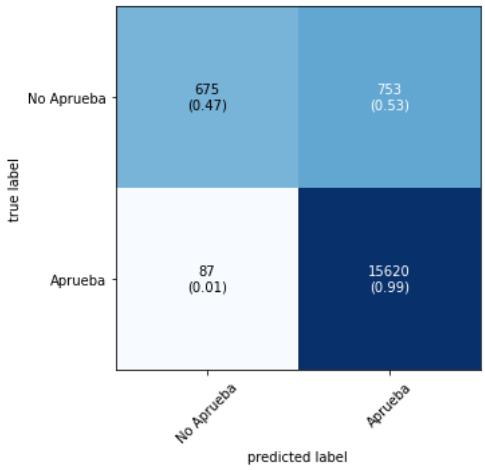


Fig. 59. Matrix de confusión del modelo Support Vector Machine

Se puede constatar, 675 reales negativos, 753 falsos negativos, 15 620 reales positivos y 87 falsos positivos, lo que significa que el modelo de SVM clasifico de forma correcta un total de 16 295 registros y 840 registros de forma incorrecta, la Fig. 60 muestra la curva ROC.

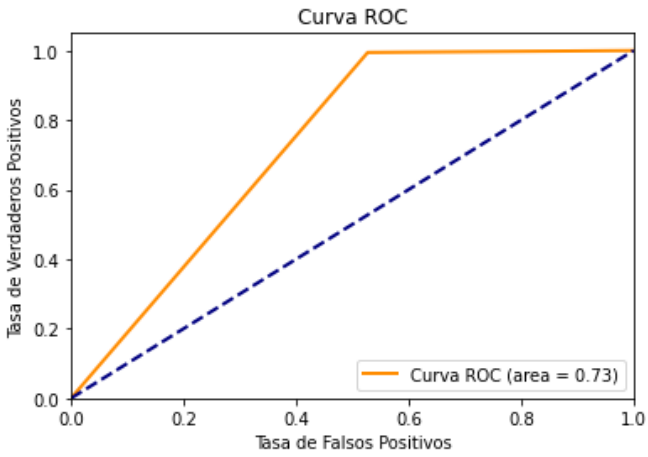


Fig. 60. Curva ROC del modelo Support Vector Machine

Además de la matriz de confusión la Tabla 3.4 muestras las métricas estadísticas de validación.

TABLA 3.4
MEDIDAS ESTADÍSTICAS DEL MODELO RF

Medida	Valor
Accuracy	95.10%
Sensibilidad	99.45%
Precisión	95.40%

F1-score	97.38%
Área bajo la curva	0.73
Coeficiente Kappa	0.59

3.1.1.3. Validación e análisis de agrupación

Para descubrir patrones relacionados al uso-desempeño se realizó un análisis de agrupación con tres algoritmos, el conjunto de datos conformado por los atributos que tiene correlación con la nota final y el atributo aprobó, la validación para tareas de agrupación es la selección del valor de *k*, en este análisis se tomó el valor de 4 para *k* correspondiente a las categorías de los atributos notas.

K-Means

La Tabla 3.5 muestra los clústeres con el algoritmo de k-means

TABLA 3.5
CLÚSTERS CON K-MEANS

Atributos	Cluster 1	Cluster 2	Cluster 3	Cluster 4
CAT_NOTA_FINAL	SUFICIENTE	INSUFICIENTE	BUENO	EXCELENTE
NIVEL_CODIGO	6	8	8	7
NUMERO_MATRICULA	1	1	1	1
NOTA1	BUENO	BUENO	BUENO	EXCELENTE
NOTA2	SUFICIENTE	EXCELENTE	EXCELENTE	EXCELENTE
APROBO	SI	SI	SI	SI
CAT_PORCENTAJE_FALTAS	MUY BAJO	MUY BAJO	MUY BAJO	MUY BAJO
CAT_USO_ACTIVIDADES	MUY BAJO	BAJO	MUY BAJO	MUY BAJO
CAT_USO_RECURSOS	MUY BAJO	MUY BAJO	MUY BAJO	MUY BAJO
DOCUMENTO	SI	SI	SI	SI
ENLACE	NO	NO	SI	NO
ARCHIVO	SI	SI	NO	SI
FOROS_DEBATES_OTROS	NO	NO	SI	SI
PROYECTO	SI	SI	NO	NO
PRUEBAS	NO	NO	SI	SI
TAREAS	NO	SI	NO	NO
TRABAJOS	NO	NO	SI	SI
CAT_EDAD	TEMPRANA	TEMPRANA	TEMPRANA	TEMPRANA
GENERO	M	M	M	M
CAT_DISCPACIDAD	NINGUNA	NINGUNA	NINGUNA	NINGUNA
ETNIA	MESTIZO	MESTIZO	MESTIZO	MESTIZO

K-Prototype

La Tabla 3.6 muestra los clústeres con el algoritmo de k-prototype.

TABLA 3.6
CLÚSTERS CON K-MEANS

Atributos	Cluster 1	Cluster 2	Cluster 3	Cluster 4
CAT_NOTA_FINAL	SUFICIENTE	BUENO	BUENO	SUFICIENTE
NIVEL_CODIGO	8	9	7	1
NUMERO_MATRICULA	1	1	1	1
NOTA1	SUFICIENTE	BUENO	SUFICIENTE	SUFICIENTE
NOTA2	SUFICIENTE	BUENO	SUFICIENTE	SUFICIENTE
APROBO	SI	SI	SI	SI
CAT_PORCENTAJE_FALTAS	MUY BAJO	MUY BAJO	MUY BAJO	MUY BAJO
CAT_USO_ACTIVIDADES	MUY BAJO	MUY BAJO	BAJO	BAJO
CAT_USO_RECURSOS	MUY BAJO	MUY BAJO	MUY BAJO	MUY BAJO
DOCUMENTO	SI	SI	SI	SI
ENLACE	NO	SI	NO	NO
ARCHIVO	NO	SI	SI	SI
FOROS_DEBATES_OTROS	SI	SI	NO	NO
PROYECTO	NO	NO	SI	NO
PRUEBAS	SI	SI	NO	NO
TAREAS	NO	NO	SI	NO
TRABAJOS	SI	SI	NO	NO
CAT_EDAD	TEMPRANA	TEMPRANA	TEMPRANA	TEMPRANA
GENERO	M	M	M	M
CAT_DISCPACIDAD	NINGUNA	NINGUNA	NINGUNA	NINGUNA
ETNIA	MESTIZO	MESTIZO	MESTIZO	MESTIZO

K-Modes

La Tabla 3.7 muestra los clústeres con el algoritmo k-modes

TABLA 3.7
CLÚSTERS CON K-MEANS

Atributos	Cluster 1	Cluster 2	Cluster 3	Cluster 4
CAT_NOTA_FINAL	SUFICIENTE	SUFICIENTE	EXCELENTE	INSUFICIENTE
NIVEL_CODIGO	6	8	9	10
NUMERO_MATRICULA	1	1	1	1
NOTA1	SUFICIENTE	BUENO	EXCELENTE	INSUFICIENTE
NOTA2	SUFICIENTE	BUENO	EXCELENTE	INSUFICIENTE
APROBO	SI	SI	SI	NO
CAT_PORCENTAJE_FALTAS	MUY BAJO	MUY BAJO	MUY BAJO	MUY BAJO
CAT_USO_ACTIVIDADES	MUY BAJO	MUY BAJO	MUY BAJO	MUY BAJO
CAT_USO_RECURSOS	MUY BAJO	MUY BAJO	MUY BAJO	MUY BAJO
DOCUMENTO	SI	SI	SI	SI
ENLACE	NO	SI	NO	NO
ARCHIVO	SI	NO	SI	NO
FOROS_DEBATES_OTROS	NO	SI	SI	SI

PROYECTO	SI	NO	NO	NO
PRUEBAS	NO	SI	SI	SI
TAREAS	SI	NO	NO	NO
TRABAJOS	NO	SI	SI	SI
CAT_EDAD	TEMPRANA	TEMPRANA	TEMPRANA	TEMPRANA
GENERO	M	M	M	M
CAT_DISCPACIDAD	NINGUNA	NINGUNA	NINGUNA	NINGUNA
ETNIA	MESTIZO	MESTIZO	MESTIZO	MESTIZO

3.1.1.4. Validación e análisis de asociación

La validación de las reglas de asociación se realiza por medio de la confianza que mide que tan confiable es cada suposición hecha, la confianza abarca un rango desde 0 a 1, siendo 1 el 100% de confianza. La Tabla 3.8 muestra las reglas del conjunto de datos correlacionados a la nota final.

TABLA 3.8
REGLAS DE ASOCIACIÓN – SET DE DATOS I

N	Antecedente	Consecuente	Confianza
1	CAT_PORCENTAJE_FALTAS=MUY BAJO	CAT_DISCPACIDAD=NINGUNA	1.00
2	DOCUMENTO=SI, EXAMENES=SI	CAT_DISCPACIDAD=NINGUNA	1.00
3	CAT_USO_RECURSOS=MUY BAJO	CAT_DISCPACIDAD=NINGUNA	1.00
4	CAT_PORCENTAJE_FALTAS=MUY BAJO	CAT_USO_RECURSOS=MUY BAJO	0.99
5	NUMERO_MATRICULA=1 CAT_PORCENTAJE_FALTAS=MUY BAJO EXAMENES=SI CAT_DISCPACIDAD=NINGUNA	CAT_USO_RECURSOS=MUY BAJO	0.99
6	CAT_PORCENTAJE_FALTAS=MUY BAJO CAT_DISCPACIDAD=NINGUNA	NUMERO_MATRICULA=1	0.96
7	CAT_PORCENTAJE_FALTAS=MUY BAJO CAT_USO_RECURSOS=MUY BAJO CAT_DISCPACIDAD=NINGUNA	NUMERO_MATRICULA=1	0.96
8	NUMERO_MATRICULA=1 CAT_USO_RECURSOS=MUY BAJO EXAMENES=SI CAT_DISCPACIDAD=NINGUNA	CAT_PORCENTAJE_FALTAS=MUY BAJO	0.96

La Tabla 3.9 muestra las reglas de los datos con correlación al atributo aprobó.

TABLA 3.9
REGLAS DE ASOCIACIÓN – SET DE DATOS II

N	Antecedente	Consecuente	Confianza
1	APROBO=SI, CAT_USO_RECURSOS=MUY BAJO, ESTADO_CIVIL=S	CAT_DISCPACIDAD=NINGUNA	1.00
2	CAT_PORCENTAJE_FALTAS=MUY BAJO CAT_USO_RECURSOS=MUY BAJO	CAT_DISCPACIDAD=NINGUNA	1.00

3	. NUMERO_MATRICULA=1 CAT_PORCENTAJE_FALTAS=MUY BAJO ESTADO_CIVIL=S CAT_DISCPACIDAD=NINGUNA	CAT_USO_RECURSOS=MUY BAJO	0.99
4	APROBO=SI, CAT_DISCPACIDAD=NINGUNA	CAT_USO_RECURSOS=MUY BAJO	0.99
5	NUMERO_MATRICULA=1 CAT_PORCENTAJE_FALTAS=MUY BAJO CAT_USO_RECURSOS=MUY BAJO CAT_DISCPACIDAD=NINGUNA	ESTADO_CIVIL=S	0.97
6	APROBO=SI, ESTADO_CIVIL=S	CAT_PORCENTAJE_FALTAS=MUY BAJO	0.97
7	NUMERO_MATRICULA=1, ESTADO_CIVIL=S	CAT_PORCENTAJE_FALTAS=MUY BAJO CAT_DISCPACIDAD=NINGUNA	0.96
8	CAT_PORCENTAJE_FALTAS=MUY BAJO CAT_USO_RECURSOS=MUY BAJO ESTADO_CIVIL=S CAT_DISCPACIDAD=NINGUNA	NUMERO_MATRICULA=1	0.96

3.1.2. Interpretación de Resultados

Interpretación e análisis de Coeficientes de Correlación

La correlación es un método que mide el grado de asociación entre dos variables, el resultado de la asociación va desde -1 a 1, donde un coeficiente de correlación negativo muestra una correlación negativa y un coeficiente de correlación positivo muestra una correlación positiva, en caso de que el coeficiente de correlación es cero significa que no hay correlación entre las variables de estudio, la Tabla 1.1 muestra las interpretaciones de forma más detallada.

Coeficientes de correlación de los atributos continuos

La Tabla 3.10 muestra los coeficientes de correlación de las variables que tienen una correlación estadísticamente significativa.

TABLA 3.10
COEFICIENTE DE CORRELACIÓN DE LOS ATRIBUTOS CONTINUOS

x	y	Pearson	Spearman	Kendall	Biserial-Puntual
NOTA_FINAL	NIVEL_CODIGO			0.1077	
NOTA_FINAL	NUMERO_MATRICULA			-0.1061	
NOTA_FINAL	NOTA1	0.8489	0.8555		
NOTA_FINAL	NOTA2	0.9103	0.8735		
NOTA_FINAL	APROBO				0.7450
NOTA_FINAL	PORCENTAJE_FALTAS	-0.2894	-0.2254		
NOTA_FINAL	NUM_ACTIVIDADES	-0.0994	-0.1149		

NOTA_FINAL	NUM_RECURSOS	0.0606	0.1182		
NOTA_FINAL	DOCUMENTO				-0.0560
NOTA_FINAL	ENLACE				0.0350
NOTA_FINAL	ARCHIVO				0.0230
NOTA_FINAL	EXAMENES				-0.0280
NOTA_FINAL	FOROS_DEBATES_OTROS				0.1850
NOTA_FINAL	PROYECTO				-0.1280
NOTA_FINAL	PRUEBAS				0.1810
NOTA_FINAL	TAREAS				-0.1270
NOTA_FINAL	TRABAJOS				0.1820
NOTA_FINAL	CAT_EDAD			-0.0569	
NOTA_FINAL	GENERO				-0.0650
NOTA_FINAL	CAT_DISCPACIDAD			-0.0329	
NOTA_FINAL	ETNIA				-0.0250
NOTA_FINAL	CAT_INGRESOS			0.0172	

La interpretación de la Tabla 3.10 muestra a cada atributo con su respectivo grado de correlación como se puede observar en la Fig. 61.

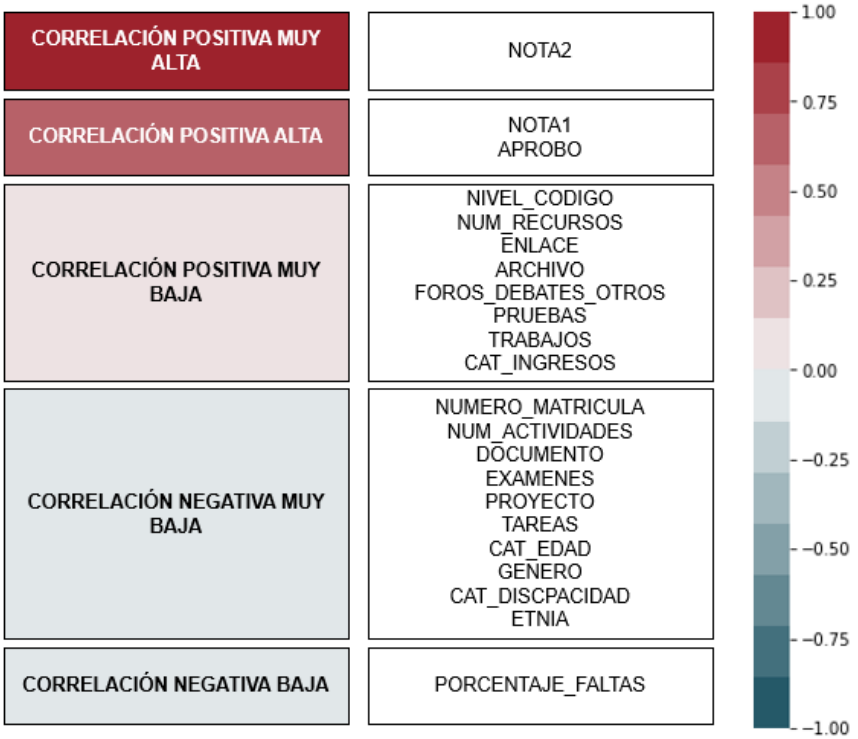


Fig. 61. Grado de correlación de las variables – análisis 1

Coeficientes de correlación de los atributos ordinales

La Tabla 3.11 muestra los coeficientes de correlación con Spearman y Kendall (Tau-b) de los atributos estadísticamente significativos.

TABLA 3.11
COEFICIENTE DE CORRELACIÓN DE LOS ATRIBUTOS ORDINALES

x	y	Spearman	Kendall
NOTA_FINAL	NIVEL_CODIGO	0.1345	0.1094
NOTA_FINAL	NUMERO_MATRICULA	-0.1187	-0.1098
NOTA_FINAL	CAT_NOTA1	0.7901	0.7279
NOTA_FINAL	CAT_NOTA2	0.8120	0.7507
NOTA_FINAL	APROBO	0.4983	
NOTA_FINAL	CAT_PORCENTAJE_FALTAS	-0.1561	-0.1444
NOTA_FINAL	CAT_USO_ACTIVIDADES	-0.1317	-0.1192
NOTA_FINAL	CAT_USO_RECURSOS	0.0781	0.0724
NOTA_FINAL	DOCUMENTO	-0.0552	
NOTA_FINAL	ENLACE	0.0428	
NOTA_FINAL	ARCHIVO	0.0127	
NOTA_FINAL	EXAMENES	-0.0477	
NOTA_FINAL	FOROS_DEBATES_OTROS	0.2334	
NOTA_FINAL	PROYECTO	-0.1649	
NOTA_FINAL	PRUEBAS	0.2322	
NOTA_FINAL	TAREAS	-0.1498	
NOTA_FINAL	TRABAJOS	0.2344	
NOTA_FINAL	CAT_EDAD	-0.0617	-0.0567
NOTA_FINAL	GENERO	-0.0703	
NOTA_FINAL	ESTADO_CIVIL	0.0202	
NOTA_FINAL	CAT_DISCPACIDAD	-0.0402	-0.0373
NOTA_FINAL	ETNIA	-0.0362	
NOTA_FINAL	CAT_INGRESOS	0.0219	0.0188

La Fig. 62 muestra el grado de correlación de los atributos de la Tabla 3.11.

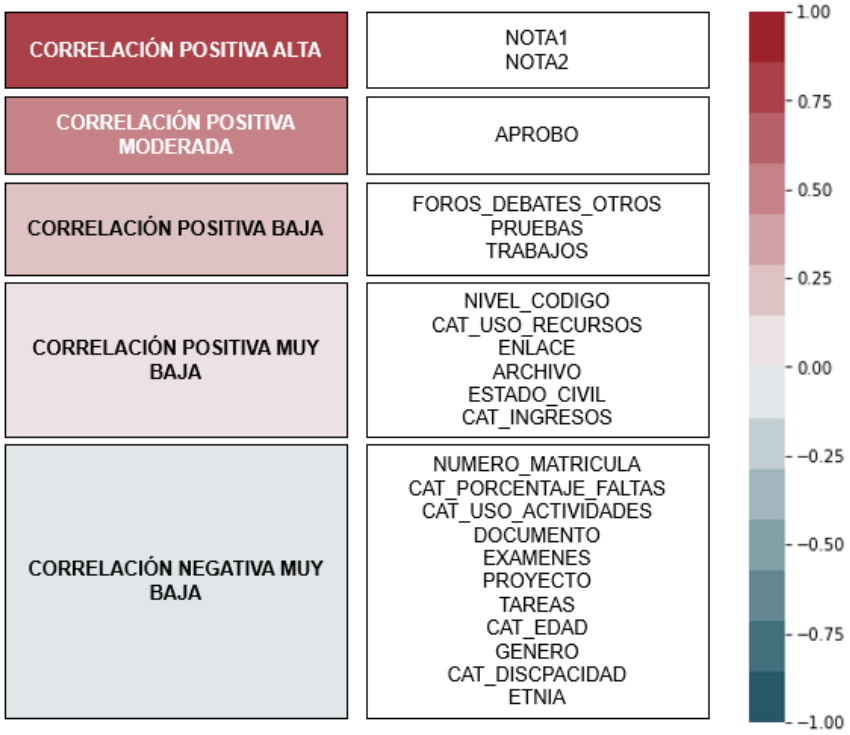


Fig. 62. Grado de correlación de las variables – análisis 2

En base a los dos análisis se tomó como atributos influyentes en el rendimiento académico a los atributos con un grado de correlación de baja a superior, a continuación su respetiva interpretación:

- Los dos análisis mostraron que los atributos NOTA1 tiene una correlación positiva alta lo que significa que a medida que la nota va subiendo la nota final también subirá, lo mismo sucede con la NOTA2. La NOTA1 corresponde a la primera parcial del semestre y la NOTA2 a la segunda parcial.
- El atributo PORCENTAJE_FALTAS mostro una correlación negativa baja es decir que mientras el porcentaje de faltas baja la nota final tiende a subir.
- Los atributos FOROS_DEBATES_OTROS, PRUEBAS y TRABAJOS mostraron una correlación positiva baja es decir mientras mayor número de intervenciones en foros, entrega de pruebas y trabajos, y mayor participación en diferentes actividades, la nota final tiende a subir.

Además, se descubrió que analizar un atributo en tipo de dato continuo y posteriormente transformar este atributo a categórico para analizarlo, los coeficientes de correlación en los dos casos no tienen una gran variación, esto se puede observar en las Tablas 3.10 y 3.11.

Interpretación e análisis de Clasificación

a. Modelo con Random Forest

En base a la matriz de confusión (Fig. 57) el modelo de clasificación hizo las siguientes predicciones:

- Categoría 1 (*aprueba*): una clasificación de 15 538 registros correctos y 169 registros incorrectos.
- Categoría 0 (*no aprueba*): una clasificación de 1031 registros correctos y 397 registros incorrectos.

b. Modelo con Support Vector Machine

En base a la matriz de confusión (Fig. 59) el modelo de clasificación con el algoritmo de SVM predijo de la siguiente manera:

- Categoría 1 (*aprueba*): una clasificación de 15 620 registros correctos y 87 registros incorrectos.
- Categoría 0 (*no aprueba*): una clasificación de 675 registros correctos y 753 registros incorrectos.

En la Tabla 3.12 se puede apreciar los errores de tipo I y II en base a las matrices de confusión de los modelos (Fig. 57 y Fig.59).

TABLA 3.12
ERRORES DE LOS MODELOS

Modelo	Tipo I	Tipo II
Random Forest	169	397
SVM	87	753

Los errores de tipo I o falsos positivos son aquellos que el modelo lo clasifico como positivos (aprobó) cuando eran negativos (reprobó). Y los errores de tipo II son aquellos se clasificaron como negativos cuando en realidad eran positivos. En vista que los errores de tipo II son los más graves errores que pueda cometer un modelo, el modelo con Random Forest se considera el mejor clasificador para la tarea de predecir si el alumno aprobara o reprobara.

Interpretación e análisis de Agrupación

Las tareas de agrupación se hicieron con k-means, k-prototype y k-modes con un valor de $k = 4$ para formar agrupaciones que reflejen el rendimiento académico. Primero se analiza el atributo nota final para identificar grupos según su rendimiento para posteriormente identificar los patrones de cada grupo en base a la Tabla 3.5, Tabla 3.6 y Tabla 3.7.

Identificación de grupos

Con el algorimo de k-means se obtuvo los siguientes clústeres: clúster 1 categoría suficiente, clúster 2 categoría insuficiente, clúster 3 categoría bueno y clúster 4 categoría excelente, como se puede observar en la Fig. 63. Con este algoritmo cada clúster pertenece a un grupo específico, donde los clústeres 1, 3 y 4 son grupos con estudiantes que aprobaron y clúster 2 con reprobados. Asimismo, dentro del grupo de estudiantes aprobados se divide en estudiantes con notas suficiente para aprobar, buena y excelente.

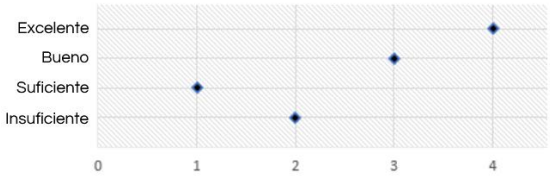


Fig. 63. Clusters con K-Means

El algoritmo de k-prototype determinó los siguientes grupos, clúster 1 y 4 categoría suficiente, clúster 2 y 3 categoría bueno como se observa en la Fig. 64.

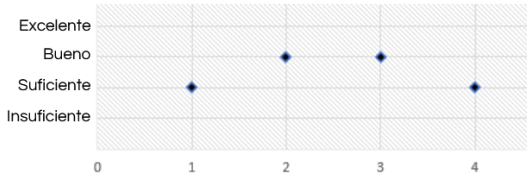


Fig. 64. Clusters con K-Prototype

Con k-modes se obtuvo los siguientes grupos, clúster 1 y 2 categoría suficiente, clúster 3 categoría excelente y clúster 4 insuficiente insuficientes como se observa en la Fig. 65.

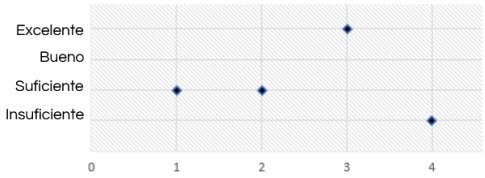


Fig. 65. Clusters con K-Modes

Patrones con K-Means

El clúster 1 correspondiente a estudiantes con una nota suficiente (7.00 a 7.99) para aprobar el semestre, este grupo muestra, una primera parcial buena y una segunda parcial suficiente, un bajo porcentaje de faltas, un bajo uso de actividades y recursos. En las interacciones no se hace uso de enlaces, foros, pruebas, tareas y trabajos.

Clúster 2, corresponde a estudiantes que reprobaron (nota entre 0 a 6.99) muestra una nota en la primera parcial de categoría buena, bajo porcentaje de faltas, bajo uso de recursos y actividades. En las interacciones no se utiliza enlaces, foros, pruebas y trabajos.

Clúster 3, grupo de estudiantes con notas de categoría buena (8.00 a 8.99) muestra un bajo porcentaje de faltas, bajo uso de recursos y actividades. En las interacciones se registra el uso de documentos, enlaces, participación en foros, pruebas y trabajos.

Clúster 4, corresponde a estudiantes con nota excelente (9.00 a 10.00) muestra en las dos parciales notas excelentes, bajo porcentaje de faltas, bajo uso de recursos y actividades. En las interacciones el uso de recursos de documentos, archivos, participación en foros, entrega de pruebas y trabajos.

Patrones con K-Prototype

Los clústeres 1 y 4 corresponden a estudiantes con notas de categoría suficiente y muestra una nota suficiente en las dos parciales, un bajo porcentaje en faltas y bajos uso de recursos y actividades, dentro de las interacciones no usan recursos como enlaces, archivos, baja participación en foros, no entrega de proyectos, pruebas y tareas.

Los clústeres 2 y 3 pertenece al grupo con notas de la categoría bueno y muestra notas de bueno y suficiente en sus parciales, bajo uso de recursos actividades, usan recursos de tipo documento, archivos, y baja participación en foros.

Patrones con K-Modes

Los clústeres 1 y 2 corresponden a estudiantes con notas de categoría suficiente y muestran notas suficientes a bueno en las dos parciales, un bajo porcentaje en faltas y bajos uso de recursos y actividades, en las interacciones no usan recursos como enlaces, archivos, baja participación en foros, no entrega de proyectos, pruebas y trabajos.

El clúster 3 es el grupo de estudiantes con notas excelentes, muestras notas excelentes en sus dos parciales, bajo porcentaje de faltas y hacen uso de recursos de tipo documento, archivos, registra una participación en foros, entrega de pruebas y trabajos.

El clúster 4 es el grupo de estudiantes reprobados, muestra notas insuficientes en sus dos parciales, bajos uso de actividades y recursos. Tiene un porcentaje de faltas bajo. No hacen uso de recursos de tipo enlace, archivos, no entregan proyectos y tareas.

Interpretación e análisis de Asociación

En el análisis de asociación las reglas resultantes mostraron una relación hacia el uso de recursos, porcentaje de faltas, numero de matrícula, discapacidad y estado civil como se observa en la Tabla 3.8 y 3.9. A continuación se interpretan las reglas relacionadas al uso de recursos y porcentaje de faltas:

- De las reglas 4 y 5 de la Tabla 3.8 se puede deducir si el alumno tiene primera matricula, tiene un bajo porcentaje en faltas, ha presentado exámenes y no tiene discapacidad entonces el uso de recursos será bajo, lo mismo sucede en las reglas 3 y 4 de la Tabla 3.9. La Fig. 66 muestra las reglas 4 y 5.

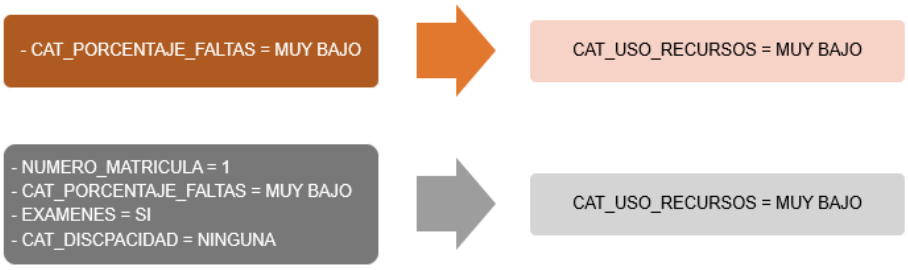


Fig. 66. Reglas de decisión del uso de recursos

- De la regla 8 de la Tabla 3.8 y la regla 6 y 7 de la Tabla 3.9 se puede deducir que si el alumno tiene primera matricula, el uso de recursos es bajo, ha presentado exámenes, no tiene discapacidad y su estado civil es soltero entonces su porcentaje de faltas será bajo. La Fig. 67 muestra la regla 8.

Fig. 67. Reglas de decisión de porcentaje en faltas

3.3. Obtención de conocimiento

Los análisis de correlación mostraron que los atributos que tienen una influencia en rendimiento académico (nota final del semestre) son la NOTA1, NOTA2, PORCENTAJE_FALTAS, FOROS_DEBATES_OTROS, PRUEBAS y TRABAJOS. En base a esto se puede afirmar que el puntaje obtenido en las parciales, el número de faltas o porcentaje en faltas, la entrega de pruebas y trabajos son factores que influyen en el desempeño dentro de SIIU-UTN.

Para predecir si el alumno aprobara o reprobará el algoritmo de Random Forest mostro mejores resultados en comparación al algoritmo de Support Vector Machine.

En agrupación se identificó 4 grupos que reflejan el rendimiento académico de estudiantes basado en su nota final:

- Grupo con nota insuficiente para aprobar (0.00 a 6.99), se caracteriza por haber reprobado, tienen notas dentro de la categoría buena o insuficiente en sus parciales, bajo usos de recursos y actividades. No utilizan los recursos de tipo enlaces, archivos, foros, no entregan pruebas, trabajos o proyectos.

- Grupo con nota suficiente (7.00 a 7.99), se caracterizan por tener notas de categoría suficiente o buena en sus parciales, un bajo uso de recursos y actividades, dentro de las interacciones no usan recursos de tipo enlaces, archivos, baja participación en foros, no entrega de proyectos, pruebas y tareas.
- Grupo con nota buena (8.00 a 8.99), se caracteriza por tener un bajo uso de recursos y actividades. También se muestra el uso de recursos de tipo documento, enlace, archivos, una baja participación en foros, así como entrega de pruebas y trabajos.
- Grupo con nota excelente (9.00 a 10), se caracteriza por tener notas de categoría excelente en sus parciales, bajo usos de recursos y actividades. Hacen uso de recursos de tipo documento, archivos, registra una participación en foros, entrega de pruebas y trabajos.

El análisis de asociación mostro que el uso de recursos y el porcentaje en faltas están directamente relacionados con el número de matrícula, los exámenes, numero de recursos usados, numero de faltas y la discapacidad.

3.4. Discusión

Los resultados obtenidos en el análisis de correlación mostraron que los factores que influyen en el desempeño académicos son las notas de las parciales, porcentaje de faltas, la entrega de pruebas y trabajos. Con agrupación pudimos identificar patrones de uso – rendimiento que mostraron que el uso de recursos de tipo documento, archivos, participación en foros, entrega de pruebas y trabajos son factores de influyen en el desempeño académico. El análisis de asociación mostros que el bajo uso de recursos estaría relacionado con un bajo porcentaje en faltas, que el alumno no tenga discapacidad o que tenga primera matricula en la asignatura, así como el bajo porcentaje en faltas estaría asociado al bajo uso de recursos, la discapacidad, el número de matrícula. Por último, el modelo de clasificación para predicción del éxito académico arrojo mejores resultados con el algoritmo de Random Forest.

Dentro del análisis de correlación se experimentó con los atributos en tipo continuo, ordinal y nominal, es decir, se creó atributos de tipo ordinal al discretizar los atributos continuos, se evidencio que los coeficientes de correlación resultantes para los datos

ordinales disminuían en 0.05 lo que no representa una gran diferencia en la interpretación del grado de correlación, por ejemplo el coeficiente con Spearman del atributo PORCETAJE_FALTAS (continuo) es -0.2254 y coeficientes de CAT_PORCENTAJE_FALTAS (ordinal) es -0.1561 como muestra en la Tabla 3.10 y 3.11.

En el presente trabajo se obtuvo varios puntos en común, las notas intermedias mostraron una correlación alta con respecto a la nota final, este resultado coincide con Conijn et al. (2017) y Ouatik et al. (2022) para identificar factores que influyen en el desempeño, esto ocurre porque las notas finales están conformadas por las notas intermedias. En los trabajos de Helal et al. (2017) y Bharara et al. (2018) se puede observar que la participación en foros y el uso de recursos son factores clave para el rendimiento académico, estos hallazgos coincidieron con nuestros resultados.

Los patrones de uso – rendimiento mostraron que el uso de recursos de tipo documento, archivos, participación en foros, entrega de pruebas y trabajos son claves para el desempeño académico, así como los estudiantes reprobados se caracterizan por no utilizan los recursos de tipo enlaces, archivos, foros, no entregan pruebas, trabajos o proyectos. Además, el bajo uso de recursos y actividades fue una constante en cada uno de los clústeres.

En las tareas de clasificación para predecir el éxito académico de los estudiantes, los resultados mostraron que el algoritmo de Random Forest arrojo mejores resultados al igual que en Hasan et al. (2018), aun cuando los atributos del conjunto de datos son distintos se obtuvieron óptimos resultados. Para la selección de variables independientes se realizó un análisis de correlación y XGBoost mostrando a los atributos como notas de evaluaciones, porcentaje en faltas, nivel que se encuentra cursando y la carrera como las principales variables predictoras.

El análisis de asociación mostros, primero el bajo uso de recursos estaría dado porque hay un bajo porcentaje en faltas, el alumno es de primera matricula, presenta exámenes y no tiene discapacidad. Segundo, el bajo porcentaje en faltas se da porque el alumno tiene primera matricula, muestra un bajo uso de recursos, presenta exámenes y no tiene discapacidad.

Con respecto a las discordancias, Ouatik et al. (2022) señala al estatus económico como factores que influyen en el rendimiento académico, en el presente estudio se

obtuvo una correlación muy baja excluyéndolo del grupo de factores relevantes. De la misma forma no se encontró alguna relación con el trabajo de Vila (2019) sobre patrones de deserción realizado con los datos socioeconómicos y académicos del SIIU-UTN.

Al observar los resultados de los análisis se puede afirmar que le SIIU-UTN estaría siendo usado como un repositorio de notas y para la toma de asistencias, con especto a las actividades y recursos, se estaría dando muy poco uso. Como trabajos futuros se puede realizar: (i) realizar análisis con datos de todas las facultades de la institución para identificar factores claves en el desempeño, patrones y relaciones generales; (ii) identificar variables predictoras a nivel general y para cada facultadas para luego establecer más modelos de predicción del éxito académico; (iii) incorporar los modelos de predicción generados en el SIIU-UTN como una herramienta para que los docentes puedan detectar alumnos vulnerables en una etapa temprana y puedan aplicar acciones correctivas.

Las principales limitaciones que se experimentaron son que no disponemos de suficientes datos de interacciones con el EVA, por ejemplo, estudios previos mostraron, que el nivel educativo de los padres, entono educativo, datos psicológicos, el número de accesos al EVA (Ouatik et al., 2022), tiempo en tareas, tiempo de entrega, numero de palabras en foros (Cerezo, 2016) están relacionadas con el rendimiento académico, estos resultados no se pudieron validar en el presente estudio.

CONCLUSIONES

Con el desarrollo de este trabajo se analizó el desempeño académico usando datos de interacciones con el EVA (recursos y actividades), académicos y socioeconómicos de los alumnos. Se logró identificar los factores que influyen en el rendimiento académico, patrones de uso-rendimiento, reglas de asociación y generar un modelo de predicción del éxito académico.

El resultado del análisis de correlación identifico a las notas de las parciales, porcentaje de faltas, la entrega de pruebas y trabajos como factores que influyen en el desempeño académicos en el SIIU-UTN. Del análisis de agrupación se obtuvo que el uso de recursos de tipo documento, archivos, participación en foros, entrega de pruebas y trabajos son factores clave en el desempeño académico, además que el bajo uso de actividades y recursos fue una constante en cada clúster.

Así mismo, se obtuvo reglas de asociación que mencionan que la primera matricula, bajo porcentaje en faltas, presentar exámenes y no tener discapacidad estaría relacionado con el bajo uso de recursos, además que el bajo porcentaje en faltas se daría por el estudiante tiene primera matricula, bajo uso de recursos, presentar exámenes y no tener discapacidad.

El modelo de predicción del éxito académico arrojo mejores resultados con el algoritmo de Random Forest. El análisis de correlación y el modelo XGBoost mostraron a las notas de evaluaciones, porcentaje en faltas, nivel que se encuentra cursando y la carrera como las principales variables predictoras.

Las principales limitaciones fue no disponer de datos de interacciones relacionados con los accesos, tiempos de permanecía, datos psicológicos y otros datos mencionados en la discusión, y entre los principales trabajos a futuro se podría hacer un análisis con datos de toda la universidad e implementar los modelos de predicción para una temprana detección de alumnos en riesgo.

RECOMENDACIONES

En función a los resultados obtenidos de los diferentes análisis se formulan algunas recomendaciones con el fin de contribuir a mejorar la calidad en la educación, incentivar el uso del SIIU-UTN, mejorar el manejo de información y contribuir a la mejora del desempeño estudiantil:

- Aumentar el número de recursos que se sube al SIIU-UTN por parte del docente, esto además de incrementar las interacciones del estudiante con el EVA contribuirá a mejorar el desempeño de los alumnos.

- Hacer uso de recursos de tipo enlace, foros y debates en el SIIU-UTN, estas actividades resultaron ser factores que influyen en el desempeño académico.

- Incorporar los modelos de predicción como una herramienta en el portafolio de los docentes del SIIU.

- Con respecto al almacenamiento de las interacciones de los estudiantes en el EVA, se recomienda crear campos que registren el número de cada tipo de actividades y recursos generados y realizados por el estudiante, además de registrar tiempos de entrega de actividades y tiempos de permanecías en los recursos, de esa forma disponer datos robustos para futuros análisis.

- Mejorar el SIIU-UTN por medio de UX (Experiencia de Usuario) para que el EVA sea más amigable con el usuario, de esa forma incentivar el uso por parte de los estudiantes.

- Generar un esquema de base de datos con la información de las tablas referentes a las interacciones en el EVA y datos académicos, esto debido que en la etapa de integración de datos nos encontramos con una gran variedad de tablas de bases de datos provenientes de diferentes departamentos y sin un esquema para realizar la integración.

BIBLIOGRAFÍA

Álvarez, F., & Mayo, L. (2019). *Algoritmo para la clasificación de aspectos de lenguaje natural basados en web semántica.* Universidad Técnica de Cotopaxi. Latacunga - Ecuador. http://repositorio.utc.edu.ec/handle/27000/5339

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, *37*, 13-49. https://doi.org/https://doi.org/10.1016/j.tele.2019.01.007

Amat, J. (2018). *Reglas de asociación y algoritmo Apriori con R* https://www.cienciadedatos.net/documentos/43_reglas_de_asociacion

Amat, J. (2020). *Máquinas de Vector Soporte (SVM) con Python*. https://www.cienciadedatos.net/documentos/py24-svm-python.html

Ang, K. L.-M., Ge, F. L., & Seng, K. P. (2020). Big Educational Data & Analytics: Survey, Architecture and Challenges. *IEEE Access*, *8*, 116392-116414. https://doi.org/10.1109/ACCESS.2020.2994561

Bharara, S., Sabitha, S., & Bansal, A. (2018). Application of learning analytics using clustering data Mining for Students' disposition analysis. *Education and Information Technologies*, *23*(2), 957-984. https://doi.org/10.1007/s10639-017-9645-7

Bohórquez, J., & Cortez, F. (2013). *Factores que determinan según su entorno el rendimiento académico de los estudiantes de la Facultad de Ciencias Económicas y Administrativas de la Universidad Católica de Santiago de Guayaquil* Universidad Católica de Santiago de Guayaquil. http://repositorio.ucsg.edu.ec/handle/3317/513

Bravo-Agapito, J., Romero, S., & Pamplona, S. (2021). Early prediction of undergraduate Student's academic performance in completely online learning: A five-year study. *Computers in Human Behavior*, *115*, 106595. https://doi.org/https://doi.org/10.1016/j.chb.2020.106595

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brinquis, C. (2020). *Qué es Pentaho - Sus Productos y Ventajas*. https://www.incentro.com/es-es/blog/stories/que-es-pentaho/

Buczak, A., & Guven, E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials*, *18*(2), 1153-1176. https://doi.org/10.1109/COMST.2015.2494502

Cabero, J., & Román, P. (2006). *E-actividades: un referente básico para la formación en Internet* (E. MAD, Ed.).

Calderon-Valenzuela, J., Payihuanca-Mamani, K., & Bedregal-Alpaca, N. (2022). Educational Data Mining to Identify the Patterns of Use made by the University Professors of the Moodle Platform. *International Journal of Advanced Computer Science and Applications*, *13*(1), 321-328. https://doi.org/10.14569/IJACSA.2022.0130140

Calvache-Fernandez, L., Álvarez-Vallejo, V., & Triviño-Arbelaez, J. (2018). Proceso KDD como apoyo a las estrategias del proyecto SARA. *Revista Educación En Ingeniería*, *Vol. 13*(Núm. 26). https://doi.org/https://doi.org/10.26507/rei.v13n26.916

Campbell, J. P. (2007). Utilizing student data within the course management system to determine undergraduate student academic success: An exploratory study. In: Purdue University. https://docs.lib.purdue.edu/dissertations/AAI3287222/

Casalino, G., Castellano, G., & Mencar, C. (2019, 2-5 July 2019). Incremental and Adaptive Fuzzy Clustering for Virtual Learning Environments Data Analysis. 2019 23rd International Conference Information Visualisation (IV). https://doi.org/https://doi.org/10.1109/IV.2019.00071

Centeno, H., Doffourt, G., Garcia, N., Gómez, G., González, E., Granado, L., Pérez, D. (2011). MINERIA DE DATOS: El arte de sacar conocimiento de grandes volúmenes de datos. In.

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., & Núñez, J. C. (2016). Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Computers & Education*, *96*, 42-54. https://doi.org/https://doi.org/10.1016/j.compedu.2016.02.006

Chok, N. S. (2010). *Pearson's Versus Spearman's and Kendall's Correlation Coefficients for Continuous Data* [Master's Thesis, University of Pittsburgh. http://d-scholarship.pitt.edu/8056/

Cisneros, S. (2019). *Detección de patrones de deserción estudiantil utilizando técnicas descriptivas de agrupamiento, asociación y atípicos en minería de datos para la gestión académica en la Universidad Técnica Norte.* http://repositorio.utn.edu.ec/handle/123456789/9516

Coello, Y., & Cachón, C. (2017). El Desempeño Académico a partir de la Implicación de los Estudiantes. In (pp. 11).

CONADIS. (2020). *Estadísticas de Discapacidad*. https://www.consejodiscapacidades.gob.ec/estadisticas-de-discapacidad/

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting Student Performance from LMS Data: A Comparison of 17 Blended Courses Using Moodle LMS. *IEEE Transactions on Learning Technologies*, *10*(1), 17-29. https://doi.org/10.1109/TLT.2016.2616312

Cutler, D. R., Edwards Jr., T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random Forests for Classification in Ecology. *Ecology*, *88*(11), 2783-2792. https://doi.org/https://doi.org/10.1890/07-0539.1

Cáceres, D. (2019). *Análisis de exactitud de los algoritmos de clustering aplicados en la base de datos del sistema académico de la UNACH* Universidad Nacional de Chimnborazo. http://dspace.unach.edu.ec/handle/51000/6114

De Miguel Díaz, M., Apocada Urquijo, P., Arias Blanco, J. M., Escudero Escorza, T., Rodriguez Espinar, S., & Vidal García, J. (2002). Evaluación del rendimiento en la enseñanza superior. Comparación de resultados entre alumnos procedentes de la LOGSE y del COU. *Revista de Investigación Educativa*, *20*(2), 357-383. https://revistas.um.es/rie/article/view/98971

Esteves, R. M., Hacker, T., & Rong, C. (2013). Competitive K-Means, a New Accurate and Distributed K-Means Algorithm for Large Datasets. *2013 IEEE 5th International Conference on Cloud Computing Technology and Science*, *1*, 17-24. https://doi.org/10.1109/CloudCom.2013.89

Franco Arcega, A. (2010). *Árboles de decisión para grandes conjuntos de datos* Instituto Nacional de Astrofísica, Óptica y Electrónica. https://inaoe.repositorioinstitucional.mx/jspui/handle/1009/506

Franco, E., & López Martínez, R. (2019). Minería de datos educativos para el análisis de rendimiento académico en una carrera de computación.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine Learning*, *29*(2), 131-163. https://doi.org/10.1023/A:1007465528199

Fukunaga, K., & Narendra, P. (1975). A Branch and Bound Algorithm for Computing k-Nearest Neighbors. *IEEE Transactions on Computers*, *C-24*(7), 750-753. https://doi.org/10.1109/T-C.1975.224297

Gandhi, R. (2018). *Support Vector Machine - Introduction to Machine Learning Algorithms*. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Garbanzo Vargas, G. M. (2007). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. *Revista Educación*, *31*(1), 43-63. https://www.redalyc.org/articulo.oa?id=44031103

Gibaja, E., Zafra, A., Luque, M., Arauzo-Azofra, A., Ramírez, A., & Olmo Ortiz, J. L. (2017). Minería de datos educativos para la detección de recursos clave. *Revista de innovación y buenas prácticas docentes*, *3*, 18. https://doi.org/10.21071/ripadoc.v3i0.9960

Gironés, J., Casas, J., Minguillón, J., & Caihuelas, R. (2017). *Minería de datos: Modelos y Algoritmos*. Editorial UOC.

HacibeyoĞLu, M., & Ibrahim, M. (2016). Comparison of the effect of unsupervised and supervised discretization methods on classification process. *International Journal of Intelligent Systems and Applications in Engineering*, 105-108. https://doi.org/10.18201/ijisae.267490

Hasan, R., Palaniappan, S., Raziff, A., Mahmood, S., & Sarker, K. U. (2018, 13-14 Aug. 2018). Student Academic Performance Prediction by using Decision Tree Algorithm. 2018 4th International Conference on Computer and Information Sciences (ICCOINS). https://doi.org/10.1109/ICCOINS.2018.8510600

Hassan, S.-U., Waheed, H., Aljohani, N., Ali, M., Ventura, S., & Herrera, F. (2019). Virtual learning environment to predict withdrawal by leveraging deep learning. *International Journal of Intelligent Systems*, *34*(8), 1935-1952. https://doi.org/10.1002/int.22129

Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., & Murray, D. J. (2019). Identifying key factors of student academic performance by subgroup discovery. *International Journal of Data Science and Analytics*, *7*(3), 227-245. https://doi.org/10.1007/s41060-018-0141-y

Hernández, R., Fernández, C., & Baptista, M. d. P. (2014). *Metodología de la Investigación* (McGRAW-HILL, Ed. Sexta ed.).

Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., & Navarro-Colorado, B. (2019). A Systematic Review of Deep Learning Approaches to Educational Data Mining. *Complexity*, *2019*. https://doi.org/10.1155/2019/1306039

Immitzer, M., Vuolo, F., & Atzberger, C. (2016). First Experience with Sentinel-2 Data for Crop and Tree Species Classifications in Central Europe. *Remote Sensing*, *8*(3). https://doi.org/10.3390/rs8030166

INEC. (2022). *Indice de precios al consumidor*. Retrieved 18/04 from https://www.ecuadorencifras.gob.ec/canasta/

Ionos. (2018). *Software de data mining: realiza análisis de datos más efectivos*. https://www.ionos.es/digitalguide/online-marketing/analisis-web/software-de-data-mining-las-mejores-herramientas/

Izurieta, G., & Moyano, R. (2019). *Predicción de clientes potenciales utilizando el algoritmo k-vecino más cercano en el área de negocios de la COAC*

"Riobamba" Ltda. Universidad Nacional de Chimborazo. http://dspace.unach.edu.ec/handle/51000/6043

Jain, N., & Srivastava, V. (2013). *Data Mining Techniques: A Survey Paper.* https://ijret.org/volumes/2013v02/i11/IJRET20130211019.pdf

Jiménez, L. (2015). *Aplicación de un sistema de alerta temprana basada en la minería de datos para identificar patrones delictivos en la ciudad de Chiclayo* Universidad Católica Santo Toribio de Mogrovejo. http://hdl.handle.net/20.500.12423/543

Karoussi, E. (2012). *Data mining K-clustering problem* University of Agder. http://hdl.handle.net/11250/137565

Khamis, H. (2008). Measures of Association: How to Choose? *Journal of Diagnostic Medical Sonography*, *24*(3), 155-162. https://doi.org/10.1177/8756479308317006

Lara Torralbo, J. A. (2014). *Fundamentos y Aplicaciones Prácticas del Descubrimiento de Conocimiento en Bases de Datos* http://repositorio.cedia.org.ec/handle/123456789/965

Larose, D., & Larose, C. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining* (J. W. Sons, Second ed.). https://doi.org/10.1002/9781118874059

Marqués, P. (2000). Los medios didácticos y los recursos educativos. http://www.peremarques.net/medios.htm

Martínez, C. (2012). *Aplicación de técnicas de minería de datos para mejorar el proceso de control de gestión en ENTEL* https://repositorio.uchile.cl/handle/2250/112065

McDonald, J. H. (2014). *Handbook of Biological Statistics* (Third ed.). Sparky House.

Mitchell, R., & Frank, E. (2017). Accelerating the XGBoost algorithm using GPU computing. *PeerJ Computer Science*, *3*, e127. https://doi.org/10.7717/peerj-cs.127

Müller, A., & Guido, S. (2016). *Introduction to Machine Learning with Python - A Guide for Data Scientists* (First Edition ed.). O'Reilly Media, Inc.

Ng, M. K., Li, M. J., Huang, J. Z., & He, Z. (2007). On the Impact of Dissimilarity Measure in k-Modes Clustering Algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 503-507. https://doi.org/10.1109/TPAMI.2007.53

Ouatik, F., Erritali, M., Ouatik, F., & Jourhmane, M. (2022). Predicting Student Success Using Big Data and Machine Learning Algorithms. *International*

Journal of Emerging Technologies in Learning (iJET), *17*(12), pp. 236-251. https://doi.org/https://doi.org/10.3991/ijet.v17i12.30259

Pezantes, J. C. (2016). *Análisis, diseño y construcción de un almacén de datos de colocaciones de crédito y captaciones de los bancos privados para aplicar algoritmos de minería de datos* Universidad Politécnica Salesiana. http://dspace.ups.edu.ec/handle/123456789/13713

Pérez López, C., & Santín Gonzales, D. (2007). *Minería de Datos: Técnicas y herramientas* (Paraninfo, Ed. 1 ed.).

Ramageri, B. M. (2010). *Data Mining Techniques and Applications* http://www.ijcse.com/docs/IJCSE10-01-04-51.pdf

Rodriguez, F. (2018). *Implementación de KDD para mejorar el proceso de identificación de estilos de aprendizaje en la Universidad Autónoma del Perú.* Universidad Autónoma del Perú. https://hdl.handle.net/20.500.13067/635

Rodríguez, V., Pedro, R., & Cruz, R. (2013). *Redes Bayesianas para la clasificación de masas en mamografías* Universidad Tecnológica de la Mixteca. http://repositorio.utm.mx:8080/jspui/handle/123456789/224

Roiger, R. J. (2017). *Data Mining: A Tutorial-Based Primer* (C. a. Hall/CRC, Ed. 2nd ed.). https://doi.org/https://doi.org/10.1201/9781315382586

Rosero, D. (2021). *Detección de patrones de contrabando para la gestión de información de aprehensiones y retenciones utilizando técnicas descriptivas de agrupamiento, asociación y atípicos en minería de datos* Universidad Técnica del Norte. http://repositorio.utn.edu.ec/handle/123456789/10851

Salinas, M. I. (2011). Entornos virtuales de aprendizaje en la escuela: tipos, modelo didáctico y rol del docente. In (pp. 12).

Sarra, A., Fontanella, L., & Di Zio, S. (2019). Identifying Students at Risk of Academic Failure Within the Educational Data Mining Framework. *Social Indicators Research*, *146*(1), 41-60. https://doi.org/10.1007/s11205-018-1901-8

Sumithra, R., & Paul, S. (2010, 29-31 July 2010). Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery. 2010 Second International conference on Computing, Communication and Networking Technologies,

Szepannek, G. (2018). clustMixType: User-Friendly Clustering of Mixed-Type Data in R. *R J.*, *10*, 200. https://doi.org/10.32614/RJ-2018-048

Szmidt, E., & Kacprzyk, J. (2011). The Spearman and Kendall rank correlation coefficients between intuitionistic fuzzy sets.

Tamilselvi, R., & Kalaiselvi, S. (2013). An Overview of Data Mining Techniques and Applications. In (Second Volumen ed.): International Journal of Science and Research. https://www.ijsr.net/archive/v2i2/IJSROFF2013059.pdf

Tejada-Escobar, F., Murrieta-Marcillo, R., Villao-Santos, F., & Garzón-Balcázar, J. (2018). Big Data en la Educación: Beneficios e Impacto de la Analítica de Datos. In (Revista Científica Y Tecnológica UPSE ed.). https://doi.org/10.26423/rctu.v5i2.424

Timaran Pereira, R., Hidalgo Troya, A., Zambrano, J., Hernández Arteaga, I., & Alvarado Pérez, J. (2016). *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional* (F. E. E. U. C. d. Colombia, Ed.). https://doi.org/http://dx.doi.org/10.16925/9789587600490

Tomas, P., & Sousa, L. (2007, 1-4 July 2007). An Efficient Expectation-Maximisation Algorithm for Spike Classification. 2007 15th International Conference on Digital Signal Processing,

Trejo, R. H. (2013). *Uso de los Entornos Virtuales de Aprendizaje en la Educación a Distancia* EDUTEC - Costa Rica 2013, https://www.uned.ac.cr/academica/edutec/memoria/ponencias/hiraldo_162.pdf

Vila, D. (2019). *Detección de patrones de deserción estudiantil utilizando técnicas predictivas de clasificación y regresión de minería de datos, para la gestión académica de la Universidad Técnica del Norte* Universidad Técnica del Norte. http://repositorio.utn.edu.ec/handle/123456789/9095

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (Third Edition ed.). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/C2009-0-19715-5