



**UNIVERSIDAD ESTATAL PENÍNSULA  
DE SANTA ELENA  
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES  
INSTITUTO DE POSTGRADO**

**TITULO DEL TRABAJO DE TITULACIÓN**

Evaluación del rendimiento académico de la U.E John F. Kennedy situada en la parroquia José Luis Tamayo del cantón Salinas aplicando técnicas de minería de datos.

**AUTOR**

**Gonzabay Panchana, Miguel Ángel**

**TRABAJO DE TITULACIÓN**

Previo a la obtención del grado académico en  
**MAGISTER EN TECNOLOGÍAS DE LA INFORMACIÓN**

**TUTOR**

PhD. Rodas Silva, Jorge

**Santa Elena, Ecuador**

**Año 2023**





**UNIVERSIDAD ESTATAL PENÍNSULA  
DE SANTA ELENA  
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES  
INSTITUTO DE POSTGRADO**

**CERTIFICACIÓN**

Certifico que luego de haber dirigido científica y técnicamente el desarrollo y estructura final del trabajo, este cumple y se ajusta a los estándares académicos, razón por el cual apruebo en todas sus partes el presente trabajo de titulación que fue realizado en su totalidad por MIGUEL ANGEL GONZABAY PANCHANA, como requerimiento para la obtención del título de Magister en Tecnologías de la Información.

**TUTOR**



---

PhD. Jorge Rodas Silva

**22 días del mes de febrero del año 2023**



**UNIVERSIDAD ESTATAL PENÍNSULA  
DE SANTA ELENA  
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES  
INSTITUTO DE POSTGRADO**

**DECLARACIÓN DE RESPONSABILIDAD**

Yo, MIGUEL ANGEL GONZABAY PANCHANA

**DECLARO QUE:**

El trabajo de Titulación, “Evaluación del rendimiento académico de la U.E John F. Kennedy situada en la parroquia José Luis Tamayo del cantón Salinas aplicando técnicas de minería de datos”, previo a la obtención del título en Magister en Tecnologías de la Información, ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.

Santa Elena, a los 22 días del mes de febrero del año 2023

A handwritten signature in blue ink, appearing to read "MGP", is written over a horizontal line.

Miguel Angel Gonzabay Panchana



**UNIVERSIDAD ESTATAL PENÍNSULA  
DE SANTA ELENA  
FACULTAD DE CIENCIAS DE LA INGENIERÍA  
INSTITUTO DE POSTGRADO**

**CERTIFICACIÓN DE ANTIPLAGIO**

Certifico que después de revisar el documento final del trabajo de titulación denominado “Evaluación del rendimiento académico de la U.E John F. Kennedy situada en la parroquia José Luis Tamayo del cantón Salinas aplicando técnicas de minería de datos”, presentado por el estudiante, GONZABAY PANCHANA MIGUEL ANGEL fue enviado al Sistema Antiplagio URKUND, presentando un porcentaje de similitud correspondiente al 2%, por lo que se aprueba el trabajo para que continúe con el proceso de titulación.

Documento	<a href="#">TT_Gonzalez_John.docx</a> (D96390558)
Presentado	2021-02-23 22:35 (-05:00)
Presentado por	T.W.D@hotmail.es
Recibido	luis.alban01.ucsg@analysis.arkund.com
Mensaje	TT_ González John <a href="#">Mostrar el mensaje completo</a> 2% de estas 42 páginas, se componen de texto presente en 6 fuentes.

**TUTOR**



Creado digitalmente por  
JORGE LUIS RODAS  
SILVA

---

PhD. Jorge Rodas Silva



**UNIVERSIDAD ESTATAL PENÍNSULA  
DE SANTA ELENA  
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES  
INSTITUTO DE POSTGRADO  
AUTORIZACIÓN**

Yo, MIGUEL ANGEL GONZABAY PANCHANA

Autorizo a la Universidad Estatal Península de Santa Elena, para que haga de este trabajo de titulación o parte de él, un documento disponible para su lectura consulta y procesos de investigación, según las normas de la Institución.

Cedo los derechos en línea patrimoniales del trabajo de titulación con fines de difusión pública, además apruebo la reproducción de este trabajo de titulación dentro de las regulaciones de la Universidad, siempre y cuando esta reproducción no suponga una ganancia económica y se realice respetando mis derechos de autor

Santa Elena, a los 22 días del mes de febrero del año 2023

A handwritten signature in blue ink, consisting of stylized letters, is written over a horizontal line.

Miguel Angel Gonzabay Panchana

## **AGRADECIMIENTO**

Agradezco primero a Dios, a mi esposa, a mi madre, a mi familia que día a día me apoyo en esta etapa de estudio.

Agradezco a los docentes que me brindaron el conocimiento necesario para lograr el objetivo de cada materia.

Agradecer a mi tutor que en la última etapa me guio para desarrollar mi proyecto final de estudio.

Agradezco a la Universidad Estatal Península de Santa Elena que me dio la oportunidad de seguir superándome.

*Miguel Angel, Gonzabay Panchana*

## **DEDICATORIA**

Dedico este logro a mis dos ángeles en el cielo, mi padre y mi hermano, quienes en vida fueron mis ejemplos y mentores como grandes profesionales en cada una de sus ramas, dedico también este logro a mi hijo y esposa que con su apoyo incondicional supo darme ánimos y motivación para seguir adelante, por último, se lo dedico a mi madre, que por ella tomé fuerzas y fue mi inspiración para seguir creciendo y dar lo mejor por mi familia.

*Miguel Angel, Gonzabay Panchana*



# ÍNDICE GENERAL

RESUMEN .....	1
ABSTRACT.....	2
INTRODUCCIÓN.....	3
Situación de la problemática.....	4
Objetivo general.....	5
Objetivos específicos .....	6
Alcance del proyecto. ....	6
JUSTIFICACIÓN .....	7
Eje social.....	7
DESARROLLO .....	8
Metodología CRISP – DM. ....	8
Fase 1: comprensión del negocio.....	8
Determinar los objetivos del negocio. ....	8
Objetivos del negocio. ....	8
Criterios del éxito del negocio. ....	9
Evaluación de la Situación.....	9
Inventario de recursos .....	9
Costos.....	9
Establecer el plan de Proyecto. ....	10

Evaluar herramientas y técnicas de minería.....	11
Fase 2: Comprensión de los datos.....	12
Datos iniciales.....	12
Descripción de los datos .....	13
Exploración de los datos .....	14
Verificar la calidad de los datos .....	16
Fase 3. Preparación de los datos .....	17
Seleccionar los datos.....	18
Limpiar los datos .....	18
Construir los datos .....	20
Integrar los datos.....	21
Fase 4. Modelado.....	21
Selección de la técnica de modelado.....	21
Generar el plan de prueba .....	22
Construir el modelo.....	23
Descripción del modelo.....	24
Fase 5. Evaluación.....	25
Evaluar el modelo.....	25
Evaluar resultados.....	26
Fase 6. Implantación.....	26

Presentar resultados de predicción en el dashboard.....	27
CONCLUSIONES .....	30
RECOMENDACIONES.....	32
ANEXOS .....	33
REFERENCIAS.....	34

## ÍNDICE DE TABLAS

Tabla 1.	Plan del Proyecto .....	10
Tabla 2.	Ficha técnica de un atributo .....	14
Tabla 3.	Tipos de datos .....	14
Tabla 4.	Escalas de aprendizaje .....	20
Tabla 5.	Cuadro comparativo entre los modelos.....	25

## ÍNDICE DE FIGURAS

Figura 1.	Base datos en Excel .....	12
Figura 2.	Datos estadísticos de los atributos .....	13
Figura 3.	Cifras estadísticas .....	15
Figura 4.	Cantidad de estudiantes en etapa supletorio .....	15
Figura 5.	Cantidad de estudiantes en etapa remedial .....	16
Figura 6.	Valores con inconsistencias .....	17
Figura 7.	Valores nulos .....	17
Figura 8.	Selección de atributos .....	18
Figura 9.	Asignación de valores a datos nulos .....	19
Figura 10.	Reemplazar valores cualitativos en materias optativas.....	19
Figura 11.	Proceso para generar nuevos atributos.....	20
Figura 12.	Mostar nuevos registros .....	21
Figura 13.	Técnicas de predicción propuestas por Rapidminer .....	22
Figura 14.	Características de los modelos .....	23
Figura 15.	Detalle del modelo .....	23
Figura 16.	Matriz de confusión .....	24
Figura 17.	Estructura del modelo .....	26
Figura 18.	Valores de predicción del modelo.....	27
Figura 19.	Dashboard con valores predictivos .....	29

## RESUMEN

En la Unidad Educativa John F. Kennedy al finalizar un parcial se desarrollan las juntas de curso, proceso donde se analiza el rendimiento académico de los estudiantes utilizando técnicas básicas de estadísticas y se observa la cantidad de estudiantes que han aprobado o reprobado, no se refleja un estudio a fondo de la causa de la deficiencia, esto no ayuda a tener una visión para dar solución al problema que afecta al grupo de estudiantes. Por tal razón se aplicó técnicas de minería de datos que mediante un análisis prescriptivo lograron determinar varios factores que estaban causando dicha anomalía. Para el proceso de minería se utilizó la herramienta Rapidminer que facilitó las tareas mediante el uso de técnicas de modelado. El proyecto siguió la estructura de la metodología CRISP-DM, que consta de seis fases basadas en comprender el negocio dando a conocer primero el objetivo general que fue establecer mejoras en las estrategias de aprendizaje aplicadas a los estudiantes, luego se establecieron los criterios de éxito basados en la calidad del aprendizaje, seguido de la comprensión de los datos donde la principal tarea fue explorar las características de los atributos, luego se prepararon los datos donde se crearon nuevos atributos, se eliminaron y sustituyeron valores, en la fase de modelado se aplicaron las técnicas apropiadas para el proyecto, luego se valoró el rendimiento de los modelos según las métricas establecidas en cada uno y posteriormente se eligió el modelo de mejor rendimiento considerando el tiempo de ejecución y la precisión. Por último, se mostró los resultados de predicción en un dashboard interactivo usando la herramienta POWER BI donde los interesados hicieron un monitoreo de la información e identificaron factores que influyen en el rendimiento académico y que sirvió para la búsqueda de nuevas estrategias de aprendizaje.

**Palabras claves:** Minería de datos, Metodología CRISP DM, Operadores.

## **ABSTRACT**

In the John F. Kennedy Educational Unit, at the end of a term, the course meetings are held, a process where the academic performance of the students is analyzed using basic statistical techniques and the number of students who have passed or failed is observed, it is not reflected an in-depth study of the cause of the deficiency, this does not help to have a vision to solve the problem that affects the group of students. For this reason, data mining techniques were applied that, through a prescriptive analysis, were able to determine several factors that were causing said anomaly. For the mining process, the Rapidminer tool was used, which facilitated the tasks through the use of modeling techniques. The project followed the structure of the CRISP-DM methodology, which consists of six phases based on understanding the business, first revealing the general objective, which was to establish improvements in the learning strategies applied to students, then the success criteria were established. based on the quality of learning, followed by understanding the data where the main task was to explore the characteristics of the attributes, then the data was prepared where new attributes were created, values were removed and substituted, in the modeling phase they were applied the appropriate techniques for the project, then the performance of the models was assessed according to the metrics established in each one and later the model with the best performance was chosen considering the execution time and precision. Finally, the prediction results were shown in an interactive dashboard using the POWER BI tool where the interested parties monitored the information and identified factors that influence academic performance and that served to search for new learning strategies.

**Keywords:** Data mining, CRISP DM Methodology, Operators.

## INTRODUCCIÓN

En la Unidad Educativa John F. Kennedy se realizan las respectivas juntas de curso al terminar un parcial, quimestre o año lectivo, estas permite analizar el rendimiento académico de los estudiantes, dicho estudio se basa en conocer quiénes aprobaron y cuántos reprobaron en una materia, esto conlleva a una conclusión que indica que se espera llegar a esta etapa para conocer el número de estudiantes inmersos en grupo con bajo rendimiento, para este proceso se usan técnicas estadísticas clásicas, obteniendo como resultado un análisis tipo descriptivo que no permite saber el motivo por el cual los estudiantes tienden a reprobado, es ahí donde nace el problema que deriva a consecuencias negativas en el proceso educativo, dificultando además la tarea de encontrar estrategias que ayuden a mejorar el nivel de aprendizaje de los alumnos.

Por este motivo se propone la aplicación de técnicas de minería de datos, que mediante un análisis prescriptivo favorezca a encontrar algún tipo de anomalía que este afectando el rendimiento de los estudiantes, para luego con los resultados obtenidos se facilite la búsqueda de estrategias novedosas de aprendizaje y que sirvan para mejorar el nivel de estudio de la comunidad estudiantil. Con esto se garantizará que los estudiantes que presente algún riesgo durante el proceso de estudio puedan terminar el año lectivo sin novedad o que en su defecto se reduzca el porcentaje de reprobados.

Para la tarea minería de datos se cuenta con la herramienta Rapidminer en donde se ejecutarán los diferentes procesos que se establecen en cada una de las etapas del proyecto, el programa cuenta con operadores que trabajan en encadenamiento a través de un entorno gráfico agradable para interactuar con el usuario, logrando hallar anomalías, patrones, correlaciones que puedan generarse de los atributos de la base de datos en un menor tiempo.

El proyecto se desarrollará basado en la metodología CRISP-DM, la misma que servirá de orientación para cumplir con cada una de las fases. La primera fase consiste en comprender el negocio, considerando los objetivos del negocio brindar educación de calidad, detectar dificultades de aprendizaje, lograr el bienestar de los integrantes de la comunidad educativa, se plantea buscar diversas estrategias de aprendizaje para mejorar el rendimiento académico en unión de uno de los criterios enfocados en las ofertas de la



institución que se basa en determinar patrones de predicción que ayuden a mejorar la actividad académica de los estudiantes, la segunda fase se llama comprensión de los datos que se encarga de recopilar, describir, explorar y verificar datos mostrando sus características con ayuda de tablas para que estén mejor organizados y de esta forma se pueda tener una buena calidad de los datos, la tercera fase es donde se procede a preparar los datos y se toman en cuenta las siguientes tareas, selección del subconjunto de datos considerando los de mayor importancia, limpieza de datos que consiste en eliminar, corregir o reemplazar datos, agregación de registros es donde se asignan nuevos registros a un atributo, derivación de atributos de un atributo existente, la cuarta fase es aquella donde se emplearán técnicas de minería para obtener los modelos de predicción que serán valorados a través de sus métricas de evaluación, en la quinta fase se realiza el proceso de comparación de modelos obtenidos el resultado del rendimiento, por último en la sexta fase se debe elegir el mejor modelo después de valorar su rendimiento y posteriormente mostrar los resultados de predicción a los interesados.

Para presentación de los resultados se empleará un dashboard interactivo desarrollado en el programa POWER BI, se mostrará información clave que permitirá a los usuarios identificar diversas variables o características del estudiante asociadas al rendimiento, comparando las cifras de los datos originales con los valores que resultaron de la predicción, dichos valores corresponden a una estimación que permitirá predecir el rendimiento académico del estudiante y tomar decisiones de manera oportuna para mejorar los resultados de aprendizaje. La información expuesta servirá de ayuda para la búsqueda de nuevas estrategias de aprendizaje y poder aplicarlas a grupos de estudiantes en riesgo de reprobación.

### **Situación de la problemática.**

En la Unidad Educativa John F. Kennedy, ubicada en la Parroquia José Luis Tamayo del Cantón Salinas se realizan las juntas de curso académicas amparadas en la Ley Orgánica de Educación Intercultural (Dirección Nacional de Normativa Jurídica del Ministerio de Educación, 2017) y que sirven para analizar el rendimiento educativo de los estudiantes de conformidad con el currículo nacional y los estándares de calidad educativa, para este proceso se toma en cuenta dos aspectos importantes:

- ✓ Análisis de aprovechamiento.
- ✓ Análisis de comportamiento.

El presente trabajo se basa en el primer aspecto, donde se analizan los promedios por parciales o quimestre, el objetivo de la junta es determinar quiénes son los estudiantes que tienen bajo rendimiento académico o su promedio es menor a 7 en las diferentes materias según el pénsum en cada nivel, posteriormente que se conoce la cantidad de alumnos reflejado en porcentajes mediante un diagrama estadístico circular o de barras, el docente que imparte la materia debe preparar y realizar el respectivo refuerzo académico que consiste en brindar asistencia pedagógica al grupo.

Lo preocupante es que este proceso es repetitivo y no ha cambiado durante años, para obtener los resultados se hace uso de técnicas estadísticas remotas, para que como resultado se muestre en un libro de Excel los porcentajes de estudiantes aprobados y reprobados, es evidente que no hay un estudio a fondo que indique por qué el grupo de estudiantes tiende a reprobado y mucho menos evitar que el próximo periodo ocurra lo mismo.

Estas falencias están contribuyendo de manera negativa en el proceso de aprendizaje, el análisis tradicional no proporciona datos que evidencien las causas de la deficiencia en el rendimiento del estudiante y lo conlleve a no obtener el promedio que necesita para aprobar una materia. Si al realizar el análisis se pudiera conocer con anticipación estos detalles se podría evitar que los alumnos presenten novedades al finalizar un periodo o reducir el número de estudiantes que se encuentran dentro del grupo.

Se necesita cambiar este esquema tradicional, mucho más que hoy en día se han aumentado las exigencias en el campo educativo, las entidades educativas están destinadas en hacer prevalecer los logros y también identificar ciertas deficiencias en el grupo estudiantil.

### **Objetivo general.**

Aplicar técnicas de minería de datos mediante la evaluación de patrones en el rendimiento académico de estudiantes para establecer mejoras en las estrategias de aprendizaje.

### **Objetivos específicos.**

- Emplear técnicas de minería de datos con la herramienta Rapidminer para construir modelos de predicción.
- Comparar modelos de predicción mediante métricas de evaluación para estimar su rendimiento.
- Presentar los resultados de predicción por medio de un dashboard interactivo para análisis del rendimiento académico.

### **Alcance del proyecto.**

El alcance del presente proyecto se basa en emplear técnicas de minería de datos procesadas en la herramienta Rapidminer que ayudará a la evaluación de modelos de predicción para el posterior análisis de resultados para que sean presentados en el proceso que se establece en las juntas de curso de la Unidad Educativa John F. Kennedy, situada en la Parroquia José Luis Tamayo Cantón Salinas, con ayuda de este trabajo se podrá determinar qué factores inciden en el rendimiento académico y posteriormente se obtener estrategias de aprendizaje basados en los resultados del modelo de predicción seleccionado y mostrarlos en un dashboard que permite realizar a los interesados las siguientes acciones:

- El vicerrector puede consultar la cantidad de estudiantes que se encuentran en cada rango de las escalas de aprendizaje por curso o materia.
- Permite al vicerrector analizar gráficas de trayectoria donde se observan los promedios finales por cada una de las materias.
- Los coordinadores de área pueden agrupar las materias por áreas y conocer el porcentaje de estudiantes aprobados y reprobados por curso o nivel.
- Permite a los docentes comparar promedios de los estudiantes para contrastar los puntajes alcanzados.
- Los docentes pueden observar las materias que tienen un alto porcentaje de promedios inferiores a 7 por curso o nivel.

## **JUSTIFICACIÓN.**

En el plano educativo las instituciones particulares compiten por una razón, es que la demanda de estudiantes crezca y esto abarca muchos aspectos tales como estructura, tecnología, enseñanza y aprendizaje. De dichos aspectos se desprenden factores que pueden afectar al rendimiento de los estudiantes y que si tienden a mejorar se puede obtener buenos resultados. Por ejemplo, del aprendizaje que reciben los estudiantes se derivan sus calificaciones, son datos que año a año representan grandes cantidades, por tal razón no se analizan de manera global sino por curso. Desde este punto al aplicar técnicas de minería de datos se podrán analizar grandes volúmenes de datos y de estas obtener resultados eficientes que ayudarán a reducir la cantidad de estudiantes reprobados o mejorar el rendimiento. La minería de datos proporciona una solución al problema planteado por medio de la construcción de modelos en busca de patrones de predicción que servirán para encontrar factores que influyan en el rendimiento académico.

En este contexto podemos indicar que la minería de datos mediante la aplicación de las técnicas que ofrece la herramienta Rapidminer, es la opción más viable para predecir a partir de los datos las anomalías que afectan el rendimiento académico. De tal forma que los interesados de la institución en mención, tendrán disponible información que les ayudará a fortalecer y mejorar las estrategias de aprendizaje para beneficio de los estudiantes que necesitan atención pedagógica en alguna asignatura.

### **Eje social.**

Educación diversa y de calidad.

Objetivo 7. Potenciar las capacidades de la ciudadanía y promover una educación innovadora, inclusiva y de calidad en todos los niveles.

## **DESARROLLO.**

### **Metodología CRISP – DM.**

En el presente proyecto se utilizará la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), (Carlos et al., 2015), se basa en un modelo estructurado de seis fases, algunas son bidireccionales, es decir que una fase en concreto puede volver a otra para su revisión y mejorarla, además su estructura es consolidada para aplicarla a un proyecto de minería de datos.

Las fases de CRISP-DM son las siguientes:

- ✓ Comprensión del negocio.
- ✓ Comprensión de los datos.
- ✓ Preparación de los datos.
- ✓ Modelado.
- ✓ Evaluación.
- ✓ Despliegue.

### **Fase 1: comprensión del negocio.**

La primera fase tiene la finalidad de determinar los objetivos y requisitos del trabajo de proyecto desde una perspectiva del negocio, para que se puedan materializar y se conviertan en un plan de proyecto.

#### **Determinar los objetivos del negocio.**

Uno de los pasos más importantes es determinar los objetivos, el principal es localizar diversas estrategias de aprendizaje para mejorar el rendimiento de los estudiantes.

#### **Objetivos del negocio.**

- Brindar educación de calidad y calidez.
- Detectar dificultades de aprendizaje en el grupo estudiantil.
- Lograr el bienestar de los integrantes de la comunidad educativa.

### **Criterios del éxito del negocio.**

El éxito del proyecto se basa en determinar patrones de predicción que ayuden a mejorar el rendimiento académico de los estudiantes, buscando estrategias mediante la construcción de modelos, de los cuales se elegirá el que ayude en la búsqueda de nuevas estrategias de aprendizaje para beneficio del departamento de vicerrectorado y que mediante este propósito los estudiantes obtengan un beneficio en mejora de su nivel de aprendizaje y no tengan la necesidad de someterse a reforzamientos académicos o reprobar alguna materia.

### **Evaluación de la Situación.**

Para la realización del proyecto se cuenta con una base de datos en Excel de los períodos lectivos 2020-2021, en la cual se encuentran registradas las calificaciones de los estudiantes por parciales y quimestres en las diferentes materias, además información de quienes están en las instancias de supletorios, remediales y de gracia.

### **Inventario de recursos.**

Los recursos que se consideran para la realización del proyecto en términos de Software son:

- Rapidminer (limpieza, transformación, manejo de algoritmos de predicción).
- Microsoft Power BI (visualización de resultados)

Los recursos que pertenecen al Hardware comprenden el uso de una Laptop con las siguientes características:

- Marca: Lenovo
- Procesador: Intel Core I7 1.9 Ghz
- Memoria RAM: 8 Gb
- Capacidad de almacenamiento: 2 Tb
- Sistema Operativo: Windows 10
- Tarjeta gráfica: Radeon

### **Costos.**

Para los programas con licencia:

- ✓ POWER BI PRO incluido en el paquete de Microsoft 365, se utiliza licencia con costo de \$8,70.
- ✓ Rapidminer Studio con licencia gratuita para estudiantes universitarios.

Debemos indicar que la gestión de información la genera la institución sin costo alguno, porque posee su propia plataforma para el registro de notas y por último se cuenta con una laptop Lenovo como herramienta de trabajo valorada en \$1200.

### **Establecer el plan de Proyecto.**

El proyecto constará de las siguientes etapas en los tiempos establecidos, tal como se aprecia en la tabla 1, a continuación:

**Tabla 1. Plan del Proyecto**

Fases	Detalle	Tiempo
<b>Comprensión del negocio</b>	Establecer los objetivos y criterios del proyecto basados en el negocio.	1 semana
<b>Comprensión de los datos.</b>	Analizar la estructura de los datos.	1 semana
	Realizar la descripción, exploración y verificación de los datos.	2 semanas
<b>Preparación de datos.</b>	Proceder a la limpieza de los datos.	2 semanas
	Integrar datos a partir de los existentes.	1 semana
<b>Modelado.</b>	Emplear técnicas de modelado y generar el plan de prueba.	2 semanas
	Construir el modelo a partir de métricas de evaluación para su descripción.	1 semana
<b>Evaluación.</b>	Comparar los diferentes modelos seleccionados.	1 semana
<b>Implantación.</b>	Elegir el modelo de mejor rendimiento y se mostrar los resultados a los interesados.	2 semanas

Elaborado por: Miguel Gonzabay P, año 2022

### **Evaluar herramientas y técnicas de minería.**

Rapidminer es un software de uso libre de análisis y minería de datos. Permite el desarrollo de procesos de análisis mediante el encadenamiento de operadores a través de un entorno gráfico. (*RapidMiner - Wikipedia, La Enciclopedia Libre*, n.d.)

Además de ser una herramienta de inteligencia de negocios, proporciona un entorno integrado de aprendizaje automático para la minería de datos, muy útil para el ámbito educativo, investigación, creación de prototipos y desarrollo de aplicaciones.

Este software permite:

- ✓ Cargar y transformar datos.
- ✓ Preprocesamiento de datos y su visualización.
- ✓ Análisis predictivos y modelos estadísticos.
- ✓ Evaluación y despliegue. (*RapidMiner / Aplicaciones y Herramientas de Marketing Digital*, n.d.)

Al verificar cada una de sus funciones es recomendable para el desarrollo del proyecto de aplicación de técnicas de minería de datos para el análisis del rendimiento académico.

En cuanto a las técnicas de minería el modelo más utilizado en este tipo de estudios es el árbol de decisión, la ventaja que permiten interpretar resultados a partir de observaciones o construcciones lógicas, pueden ser utilizados para problemas de regresión o clasificación. Este modelo presenta una estructura jerárquica para establecer relaciones entre la variable dependiente y el conjunto de variables predictoras, en el análisis propuesto un buen ejemplo a tomar sería que estudiantes superan o no los aprendizajes requeridos. Otro algoritmo útil para este tipo de predicciones es el de regresión logística, este puede identificar variables importantes que pueden predecir resultados de promedios de estudiantes en el estudio presentado. Los datos se presentan en tablas denominadas matriz de confusión que sirven para mostrar distribuciones de las clases. (Menacho Chiok, 2017)



## Fase 2: Comprensión de los datos.

En la segunda fase de la metodología CRISP-DM se hace énfasis en las características de los datos, además se verifica que los datos estén en un formato apto para su análisis, al final se comprueba que los datos tengan una calidad óptima.

### Datos iniciales.

Los datos recopilados de los estudiantes para el análisis son: código de estudiantes, curso, paralelo, materia, promedios parciales, promedio quimestral y luego del ciclo regular las calificaciones de examen supletorio, remedial y de gracia.

**Figura 1. Base datos en Excel**

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	codest	curso	especiali	paralelo	materia	q1p1	q1p2	q1ex	q1	q2p1	q2p2	q2ex	q2	supl	reme	gracia	final
2	190007	B02	-	A	LENGUA Y LITERATURA	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	0	0		0
3	190007	B02	-	A	MATEMÁTICA	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	0	0		0
4	190007	B02	-	A	CIENCIAS NATURALES	9.88	10.00	10.00	9.95	9.70	10.00	10.00	9.88	0	0		0
5	190007	B02	-	A	ESTUDIOS SOCIALES	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	0	0		0
6	190007	B02	-	A	EDUCACIÓN ESTÉTICA	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	0	0		0
7	190007	B02	-	A	EDUCACIÓN FÍSICA	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	0	0		0
8	190007	B02	-	A	LENGUA EXTRANJERA	10.00	10.00	10.00	10.00	9.90	9.78	10.00	9.87	0	0		0
9	190007	B02	-	A	COMPUTACION	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	0	0		0
10	190007	B02	-	A	PROYECTOS ESCOLARES	EX	EX	EX	EX	EX	EX		MB	0	0		0
11	190007	B02	-	A	DESARROLLO HUMANO INTEGRAL	EX	EX	EX	EX	EX	EX		MB	0	0		0
12	190007	B02	-	A	Comportamiento	B	B			A	B		B	0	0		0
13	190009	B02	-	A	LENGUA Y LITERATURA	9.94	9.88	9.50	9.83	9.78	9.93	10.00	9.89	0	0		0
14	190009	B02	-	A	MATEMÁTICA	9.81	9.63	9.50	9.68	9.94	10.00	10.00	9.98	0	0		0
15	190009	B02	-	A	CIENCIAS NATURALES	9.88	9.44	9.00	9.53	9.53	9.86	10.00	9.76	0	0		0
16	190009	B02	-	A	ESTUDIOS SOCIALES	9.75	9.88	10.00	9.86	10.00	10.00	10.00	10.00	0	0		0
17	190009	B02	-	A	EDUCACIÓN ESTÉTICA	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	0	0		0
18	190009	B02	-	A	EDUCACIÓN FÍSICA	9.91	9.81	10.00	9.89	10.00	9.95	10.00	9.98	0	0		0
19	190009	B02	-	A	LENGUA EXTRANJERA	9.84	10.00	10.00	9.94	8.91	9.70	9.70	9.39	0	0		0
20	190009	B02	-	A	COMPUTACION	9.80	10.00	10.00	9.92	9.80	9.60	10.00	9.76	0	0		0
21	190009	B02	-	A	PROYECTOS ESCOLARES	EX	EX	EX	EX	EX	EX		MB	0	0		0
22	190009	B02	-	A	DESARROLLO HUMANO INTEGRAL	EX	EX	EX	EX	EX	EX		MB	0	0		0
23	190009	B02	-	A	Comportamiento	R	R			R	A		A	0	0		0

Elaborado por: Miguel Gonzabay P, año 2022

Los datos mostrados en la figura 1, comprenden calificaciones de todo el año lectivo, se pueden observar n promedios parciales, promedios quimestrales y notas de examen en cada una de las materias, en cuanto a información de los estudiantes se tienen el código, curso, especialidad, paralelo y materia. Estas variables se convertirán en atributos en el trabajo de minería de datos. A continuación, se describe cada uno de ellos:

- codest: tipo de dato entero, contiene los códigos que se le asigna al estudiante al matricularse.
- curso: tipo de dato texto, al año al que pertenece el estudiante.
- paralelo: tipo de dato texto, subdivisión de los cursos.

- materia: tipo de dato texto, cada una de las asignaturas impartidas en los cursos.
- q1p1: tipo de dato numérico, se refiere al primer parcial del primer quimestre, contiene el promedio de los insumos generados en es parcial.
- q1p2: tipo de dato numérico, corresponde al parcial dos.
- q1ex: tipo de dato numérico, nota proveniente del examen del primer quimestre.
- q1: tipo de dato numérico, promedio del primer quimestre, cabe recalcar que sucede lo mismo con el promedio del segundo quimestre.
- supl: tipo de dato entero, corresponde a nota del examen supletorio.
- rem: tipo de dato entero, corresponde a nota del examen remedial.
- gracia: tipo de dato entero, corresponde a nota del examen de gracia.
- final: tipo de dato numérico, contiene el promedio de los dos quimestres.

### Descripción de los datos.

En la figura 2, se identifican los atributos que se van a seleccionar y para esto se consideran las características importantes y su relación con la propuesta del análisis del rendimiento académico. Luego se define un tipo de estructura para los datos, se puede presentar una ficha técnica, tal como se muestra en la tabla 2.

**Figura 2. Datos estadísticos de los atributos**

Attribute	Type	Count	Least	Most	Values
q1p1	Polynomial	0	9.99 (1)	10.00 (1657)	10.00 (1657), 10 (431), ... [266 more]
q1p2	Polynomial	0	9.99 (1)	10.00 (1202)	10.00 (1202), 10 (424), ... [321 more]
q1ex	Polynomial	0	9.97 (1)	10.00 (2868)	10.00 (2868), 9.00 (901), ... [123 more]
q2p1	Polynomial	0	8.51 (1)	10.00 (1260)	10.00 (1260), 10 (474), ... [383 more]
q2p2	Polynomial	0	9.92 (1)	10.00 (1349)	10.00 (1349), 10 (425), ... [398 more]
codest	Integer	0	80009	200065	Average 163380.008

Elaborado por: Miguel Gonzabay P, año 2022

**Tabla 2. Ficha técnica de un atributo**

Identificador	Codest
<b>NOMBRE</b>	Código del estudiante
<b>DETALLE</b>	Contiene el código del estudiante al matricularse
<b>TIPO</b>	Integer
<b>FUENTE</b>	Base de datos del sistema académico

Elaborado por: Miguel Gonzabay P, año 2022

En la tabla 3, se muestran los datos útiles para la siguiente fase y se coloca el tipo de datos que corresponden ya sea de tipo entero, numérico etc.

**Tabla 3. Tipos de datos**

Datos	Tipo
<b>codest</b>	entero
<b>q1p1</b>	numérico
<b>q1p2</b>	numérico
<b>q1ex</b>	numérico
<b>q1</b>	numérico
<b>supl</b>	entero
<b>rem</b>	entero
<b>gracia</b>	entero
<b>final</b>	numérico

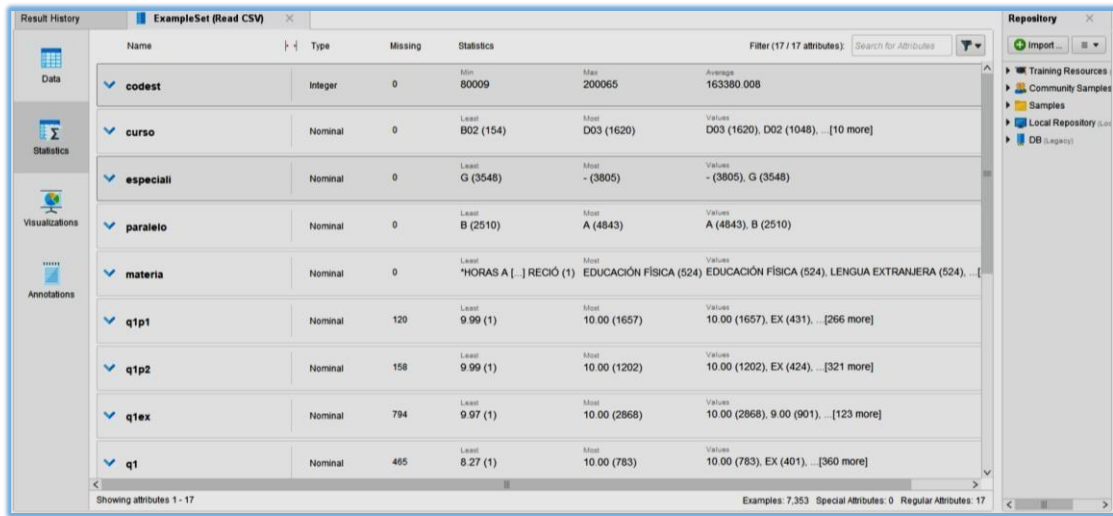
Elaborado por: Miguel Gonzabay P, año 2022

### **Exploración de los datos.**

Al terminar la tarea de descripción de los datos se debe proceder a la exploración de los mismos con el fin de encontrar una estructura general para los datos, se puede aplicar pruebas estadísticas para observar valores inferenciales, que serían útiles en esta fase.

La figura 3, muestra información que despliega cifras estadísticas de la base de datos que los procesa la herramienta Rapidminer.

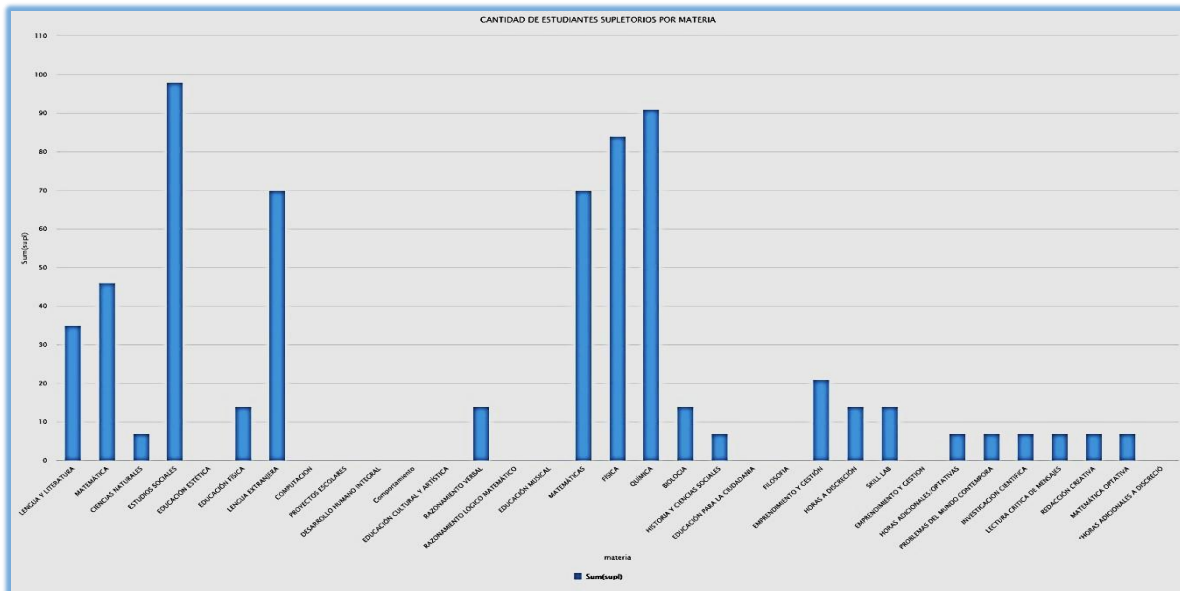
**Figura 3. Cifras estadísticas**



Elaborado por: Miguel Gonzabay P, año 2022

La figura 4 muestra la cantidad de estudiantes en instancias supletorios por materia, información que sirve de mucho para analizar qué cantidad de estudiantes por nivel han aprobado los exámenes supletorios.

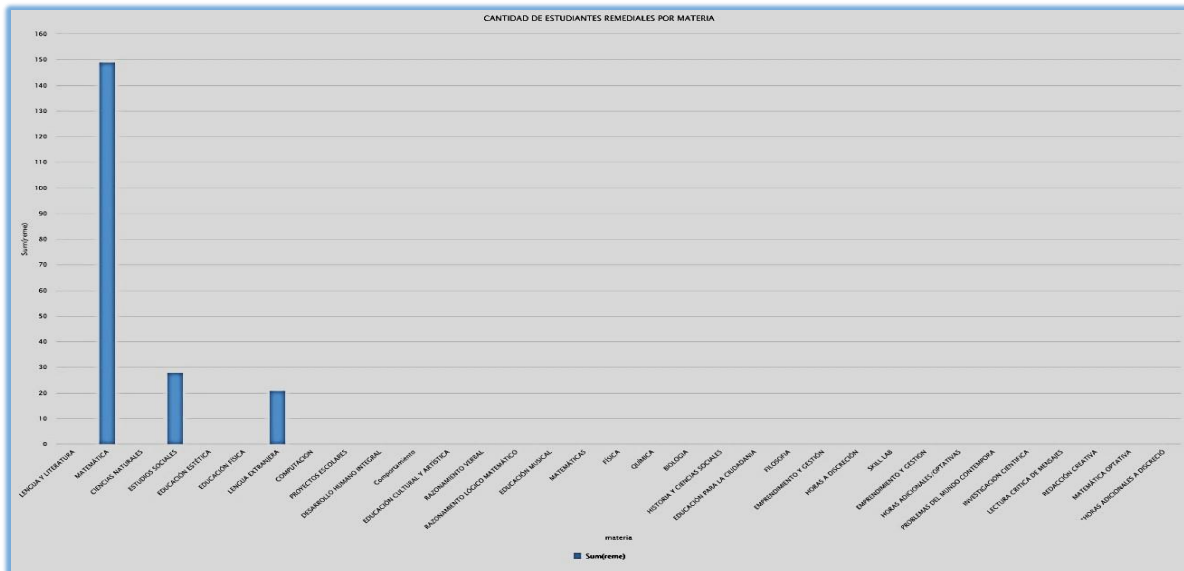
**Figura 4. Cantidad de estudiantes en etapa supletorio**



Elaborado por: Miguel Gonzabay P, año 2022

En la figura 5, se muestra la cantidad de estudiantes que rindieron exámenes remediales por materia.

**Figura 5. Cantidad de estudiantes en etapa remedial**



Elaborado por: Miguel Gonzabay P, año 2022

### Verificar la calidad de los datos.

En este paso se procede a ver que anomalías poseen los datos que pasaron por la fase de exploración, como, por ejemplo, la consistencia de los valores, la cantidad de valores nulos y valores que estén fuera de rango, este tipo de datos pueden convertirse en ruido durante el proceso de análisis, luego de esto se garantiza que los datos estén aptos para la siguiente etapa.

En los ejemplos de la figura 6 y 7 se observan valores no consistentes y nulos respectivamente.

**Figura 6. Valores con inconsistencias**

#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	190007	B02	-	A	LENGUA Y LITERATURA	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	0	0	0	10.00		
2	190007	B02	-	A	MATEMATICA	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	0	0	0	10.00		
3	190007	B02	-	A	CIENCIAS NATURALES	9.88	10.00	10.00	9.95	9.70	10.00	10.00	9.88	0	0	0	9.91		
4	190007	B02	-	A	ESTUDIOS SOCIALES	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	0	0	0	10.00		
5	190007	B02	-	A	EDUCACIÓN ESTÉTICA	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	0	0	0	10.00		
6	190007	B02	-	A	EDUCACIÓN FÍSICA	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	0	0	0	10.00		
7	190007	B02	-	A	LENGUA EXTRANJERA	10.00	10.00	10.00	10.00	9.90	9.78	10.00	9.87	0	0	0	9.93		
8	190007	B02	-	A	COMPUTACION	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	0	0	0	10.00		
9	190007	B02	-	A	PROYECTOS ESCOLARES	EX	EX	EX	EX	EX	EX	MB	0	0	0	EX			
10	190007	B02	-	A	DESARROLLO HUMANO INTEGRAL	EX	EX	EX	EX	EX	EX	MB	0	0	0	EX			
11	190007	B02	-	A	Comportamiento	B	B	B	A	B	B	0	0	0	0	B			
12	190009	B02	-	A	LENGUA Y LITERATURA	9.94	9.88	9.50	9.83	9.78	9.93	10.00	9.89	0	0	0	9.88		
13	190009	B02	-	A	MATEMATICA	9.81	9.63	9.50	9.68	9.94	10.00	9.98	0	0	0	9.83			
14	190009	B02	-	A	CIENCIAS NATURALES	9.88	9.44	9.00	9.53	9.53	9.86	10.00	9.76	0	0	0	9.64		
15	190009	B02	-	A	ESTUDIOS SOCIALES	9.75	9.88	10.00	9.86	10.00	10.00	10.00	10.00	0	0	0	9.93		
16	190009	B02	-	A	EDUCACIÓN ESTÉTICA	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	0	0	0	10.00		
17	190009	B02	-	A	EDUCACIÓN FÍSICA	9.91	9.81	10.00	9.88	10.00	9.95	10.00	9.88	0	0	0	9.93		
18	190009	B02	-	A	LENGUA EXTRANJERA	9.84	10.00	10.00	9.94	8.91	9.70	9.70	9.39	0	0	0	9.66		
19	190009	B02	-	A	COMPUTACION	9.80	10.00	10.00	9.92	9.80	9.60	10.00	9.76	0	0	0	9.84		
20	190009	B02	-	A	PROYECTOS ESCOLARES	EX	EX	EX	EX	EX	EX	MB	0	0	0	EX			
21	190009	B02	-	A	DESARROLLO HUMANO INTEGRAL	EX	EX	EX	EX	EX	EX	MB	0	0	0	EX			
22	190009	B02	-	A	Comportamiento	B	B	B	A	A	0	0	0	0	0	A			
23	190022	B02	-	A	LENGUA Y LITERATURA	8.75	8.06	8.00	8.33	8.50	9.00	9.00	8.80	0	0	0	8.56		
24	190022	B02	-	A	MATEMATICA	8.88	8.63	9.00	8.81	9.38	9.99	9.00	9.55	0	0	0	9.18		
25	190022	B02	-	A	CIENCIAS NATURALES	9.25	8.50	8.50	8.80	8.99	9.38	9.00	9.15	0	0	0	8.97		
26	190022	B02	-	A	ESTUDIOS SOCIALES	9.25	9.00	9.00	9.10	9.69	9.44	9.00	9.46	0	0	0	9.28		
27	190022	B02	-	A	EDUCACIÓN ESTÉTICA	10.00	10.00	10.00	10.00	9.80	10.00	10.00	9.92	0	0	0	9.96		
28	190022	B02	-	A	EDUCACIÓN FÍSICA	9.83	9.85	10.00	9.87	9.83	9.95	10.00	9.95	0	0	0	9.91		
29	190022	B02	-	A	LENGUA EXTRANJERA	9.38	9.47	10.00	9.54	8.35	8.69	8.00	8.42	0	0	0	8.98		

Elaborado por: Miguel Gonzabay P, año 2022

La calidad de los datos representa la precisión de dichos valores, por tal razón es necesario verificar con qué frecuencia presentan errores o existen valores perdidos, que son muy comunes en grandes volúmenes de datos.

**Figura 7. Valores nulos**

#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
5221	160118	D02	G	B	FILOSOFIA	9.30	9.80	10.00	9.64	10.00	9.00	9.00	9.40	0	0	0	9.52		
5222	160118	D02	G	B	LENGUA Y LITERATURA	10.00	9.40	9.00	9.56	10.00	10.00	10.00	10.00	0	0	0	9.78		
5223	160118	D02	G	B	LENGUA EXTRANJERA	10.00	10.00	10.00	10.00	9.80	9.80	10.00	9.76	0	0	0	9.88		
5224	160118	D02	G	B	EDUCACIÓN CULTURAL Y ARTÍSTICA	10.00	10.00	10.00	10.00	10.00	9.00	10.00	9.80	0	0	0	9.80		
5225	160118	D02	G	B	EDUCACIÓN FÍSICA	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	0	0	0	10.00		
5226	160118	D02	G	B	EMPRESARIADO Y GESTION	10.00	8.60	10.00	9.44	8.80	8.80	10.00	9.04	0	0	0	9.24		
5227	160118	D02	G	B	HORAS A DISCRECIÓN	9.70	9.60		9.62	9.50	9.40		9.56	0	0	0	9.59		
5228	160118	D02	G	B	RAZONAMIENTO LÓGICO MATEMÁTICO	9.40	9.20	9.00	9.24	9.00	9.20	10.00	9.28	0	0	0	9.26		
5229	160118	D02	G	A	SKILL LAB	10.00	10.00		10.00	10.00	9.80	10.00	9.94	0	0	0	9.92		
5230	160118	D02	G	B	Comportamiento	A	B		A	A		A	0	0	0	A			
5231	160121	D02	G	A	MATEMÁTICAS	7.50	8.80	8.00	8.12	8.00	7.60	9.00	8.04	0	0	0	8.08		
5232	160121	D02	G	A	FÍSICA	7.60	6.40	6.00	6.80	6.00	5.60	7.00	6.04	7	0	0	7.00		
5233	160121	D02	G	A	QUÍMICA	7.40	8.00	9.00	7.96	7.60	6.20	8.00	7.12	0	0	0	7.54		
5234	160121	D02	G	A	BIOLOGIA	7.90	6.80	8.00	7.48	8.00	7.70	8.00	7.88	0	0	0	7.68		
5235	160121	D02	G	A	HISTORIA Y CIENCIAS SOCIALES	9.60	9.60	10.00	9.68	9.80	8.60	9.00	9.16	0	0	0	9.42		
5236	160121	D02	G	A	EDUCACIÓN PARA LA CIUDADANÍA	7.20	7.90	8.00	7.64	7.20	7.50	8.00	7.48	0	0	0	7.56		
5237	160121	D02	G	A	FILOSOFIA	8.10	8.70	9.00	8.52	8.80	8.00	8.00	8.32	0	0	0	8.42		
5238	160121	D02	G	A	LENGUA Y LITERATURA	10.00	9.00	8.00	9.20	7.30	7.40	7.00	7.26	0	0	0	8.24		
5239	160121	D02	G	A	LENGUA EXTRANJERA	9.20	9.32	10.00	9.41	8.96	8.00	2.00	7.18	0	0	0	8.29		
5240	160121	D02	G	A	EDUCACIÓN CULTURAL Y ARTÍSTICA	10.00	10.00	10.00	10.00	10.00	8.80	10.00	9.52	0	0	0	9.76		
5241	160121	D02	G	A	EDUCACIÓN FÍSICA	10.00	10.00	10.00	10.00	10.00	8.20	8.00	8.88	0	0	0	9.44		
5242	160121	D02	G	A	EMPRESARIADO Y GESTION	10.00	8.00	8.00	8.80	8.00	7.60	8.00	7.84	0	0	0	8.32		
5243	160121	D02	G	A	HORAS A DISCRECIÓN	8.50	8.20		8.53	9.20	9.00		8.48	0	0	0	8.50		
5244	160121	D02	G	A	RAZONAMIENTO LÓGICO MATEMÁTICO	7.60	7.20	8.50	7.62	8.40	8.00	8.00	8.16	0	0	0	7.89		
5245	160121	D02	G	A	SKILL LAB	9.40	9.20	10.00	9.44	10.00	10.00	4.00	8.80	0	0	0	9.12		
5246	160121	D02	G	A	Comportamiento	B	B		B	B		B	0	0	0	B			
5247	170082	D02	G	B	MATEMÁTICAS	8.00	8.40	9.00	8.36	8.40	8.20	10.00	8.64	0	0	0	8.50		
5248	170082	D02	G	B	FÍSICA	8.00	7.80	9.00	8.12	7.20	7.20	9.00	7.56	0	0	0	7.84		
5249	170082	D02	G	B	QUÍMICA	8.80	8.50	9.00	8.72	8.20	9.40	10.00	9.04	0	0	0	8.88		
5250	170082	D02	G	B	BIOLOGIA	6.90	8.00	9.00	7.76	8.40	7.90	8.00	8.12	0	0	0	7.94		

Elaborado por: Miguel Gonzabay P, año 2022

**Fase 3. Preparación de los datos.**

En esta fase se seleccionan los atributos que se consideren de importancia según los siguientes criterios:

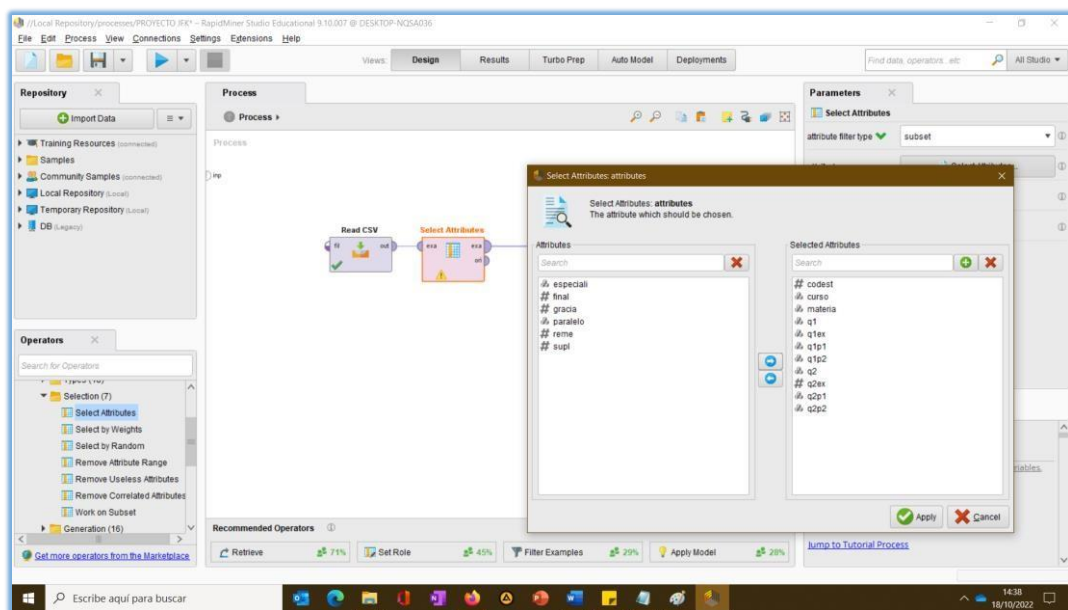
- ✓ Revisar que contengan datos completos y sin errores.
- ✓ Verificar que los datos sean precisos, de eso depende la calidad.
- ✓ Comprobar que los datos sean íntegros, esto garantiza la fiabilidad.

Para cumplir con estas tareas, la herramienta Rapidminer ofrece operadores que facilitará su procesamiento y se describen a continuación:

### Seleccionar los datos.

En la figura 8, se muestra el operador **select atributos** que servirá para seleccionar los atributos que se consideran importantes y se descartan los que no cumple con los criterios para seguir con el proceso.

**Figura 8. Selección de atributos**



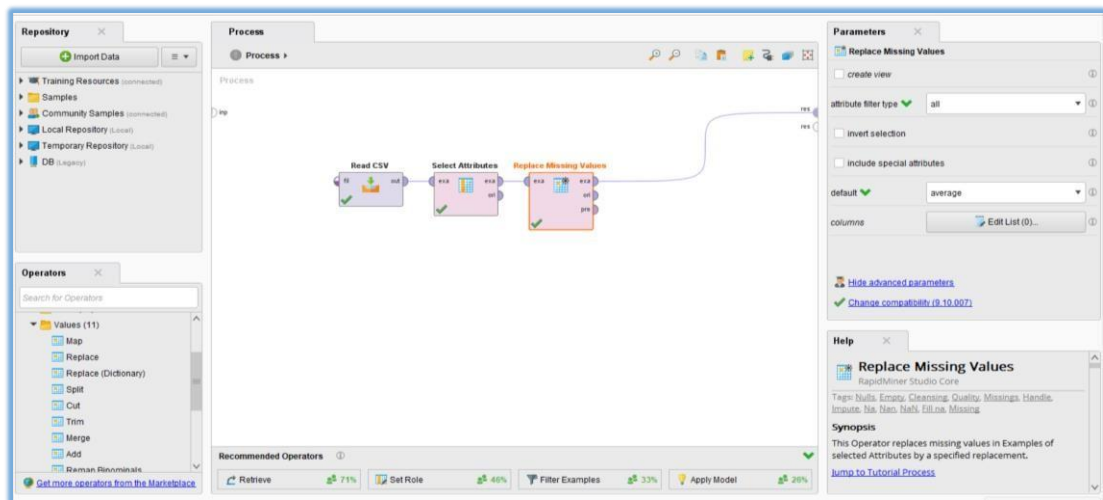
Elaborado por: Miguel Gonzabay P, año 2022

### Limpiar los datos.

Luego se debe reemplazar los valores nulos, para lo cual se utiliza el operador **replace missing values**, una de las opciones es buscar y reemplaza los valores por el promedio que se genera con los valores de dicho atributo para tener una nueva estimación con los valores reemplazados, también se puede ubicar el valor máximo o mínimo, el proceso se muestra en la figura 9.



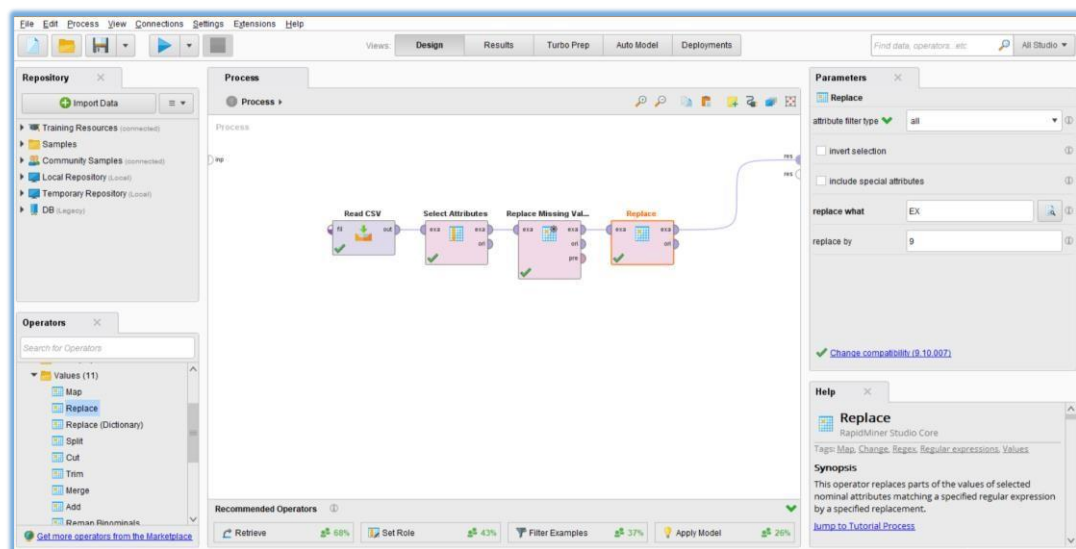
**Figura 9. Asignación de valores a datos nulos**



Elaborado por: Miguel Gonzabay P, año 2022

Para reemplazar datos que tienen valores cualitativos, en este caso las materias optativas utilizamos el operador **replace**, el cual ubicará la calificación de EX (10), MB (9), B (8) y R (7), la razón en que en estas materias al momento de calificar se otorgan valores entre 7 y 10 pero al establecer la nota se da un valor cualitativo, proceso que se muestra en la figura 10.

**Figura 10. Reemplazar valores cualitativos en materias optativas**



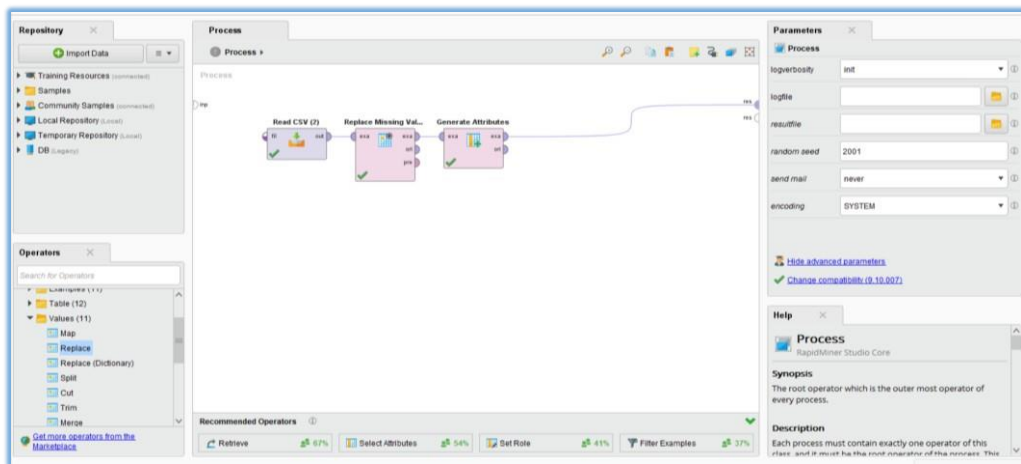
Elaborado por: Miguel Gonzabay P, año 2022



## Construir los datos.

Esta tarea consiste en generar nuevos atributos a partir de otros, para esto se puede utilizar el operador **generate attribute**, en este caso los siguientes atributos; “Promedio” que va a contener la media del primer y segundo quimestre, luego de esto se genera el atributo “Escala de aprendizaje”, me permitirá ver en que escala de aprendizaje se encuentra ubicado el estudiante dependiendo de su promedio, los rangos de cada escala se muestran en la tabla 4, por último se deriva el atributo “Estado” que mostrará a los estudiantes aprobados si el promedio es mayor a 7, caso contrario será reprobado, el proceso se muestra en la figura 11.

**Figura 11. Proceso para generar nuevos atributos**



Elaborado por: Miguel Gonzabay P, año 2022

**Tabla 4. Escalas de aprendizaje**

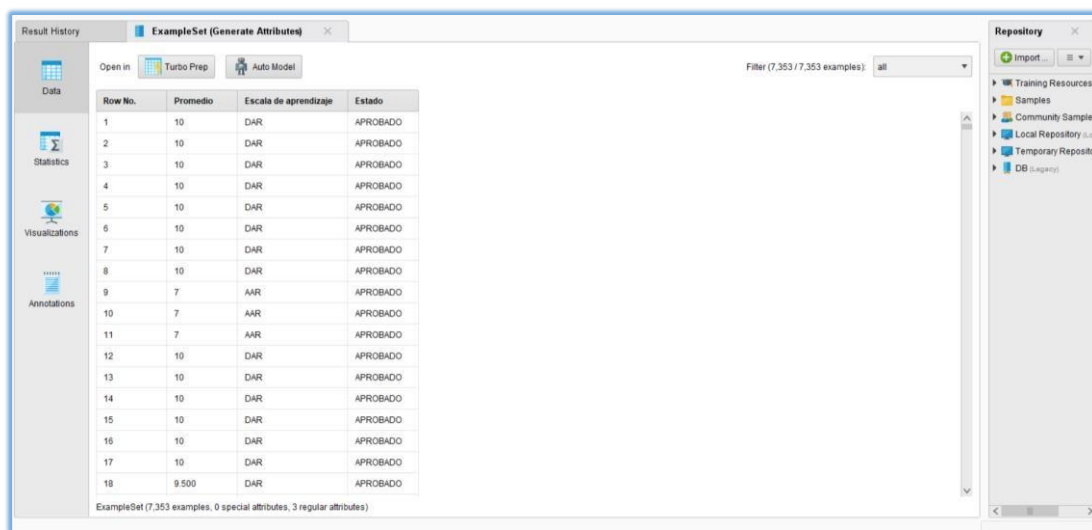
Escala cualitativa	Significado	Escala cuantitativa
DAR	Domina los aprendizajes requeridos	9-10
AAR	Alcanza los aprendizajes requeridos	7-8,99
PAR	Próximo a alcanzar los aprendizajes requeridos	4,01-6,99
NAR	No alcanza los aprendizajes requeridos	<4

Elaborado por: Miguel Gonzabay P, año 2022

## Integrar los datos.

En esta parte se crean nuevas estructuras, como por ejemplo creación de nuevos campos, registros o nuevas tablas, en la figura 12 se observa un ejemplo en los cuales se han ingresado nuevos registros a las tablas a partir de los existentes.

**Figura 12. Mostar nuevos registros**



Row No.	Promedio	Escala de aprendizaje	Estado
1	10	DAR	APROBADO
2	10	DAR	APROBADO
3	10	DAR	APROBADO
4	10	DAR	APROBADO
5	10	DAR	APROBADO
6	10	DAR	APROBADO
7	10	DAR	APROBADO
8	10	DAR	APROBADO
9	7	AAR	APROBADO
10	7	AAR	APROBADO
11	7	AAR	APROBADO
12	10	DAR	APROBADO
13	10	DAR	APROBADO
14	10	DAR	APROBADO
15	10	DAR	APROBADO
16	10	DAR	APROBADO
17	10	DAR	APROBADO
18	9.500	DAR	APROBADO

Elaborado por: Miguel Gonzabay P, año 2022

## Fase 4. Modelado.

En esta etapa se seleccionan las técnicas de modelados que sean apropiadas para el proyecto de minería. Se valora el rendimiento del modelo y se toman en cuenta los siguientes criterios:

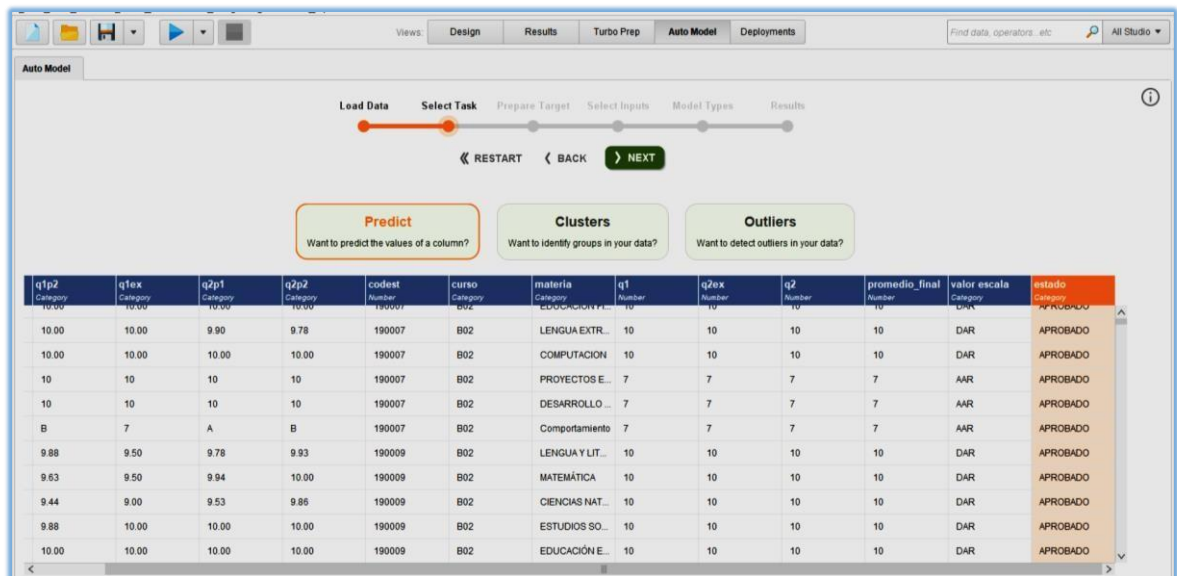
- ✓ Precisión del modelo.
- ✓ Mejor tiempo de ejecución del modelo.
- ✓ Mejor tiempo de respuesta.

### Selección de la técnica de modelado.

Se debe escoger la técnica de minería de datos más apropiada para el proyecto, la herramienta Rapidminer permite escoger tres que son: predicción (PREDICT), clasificación (CLUSTERS) y la de valores atípicos (OUTLIERS), para el caso se escogerá la de predicción, conociendo que uno de los objetivos del proyecto es emplear técnicas

de predicción y sería la técnica más apropiada para aplicar al proceso, en esta tarea se selecciona el atributo “estado” como variable de predicción y genera dos posibles resultados “APROBADO” y “REPROBADO”, considerando esto tenemos un caso de clasificación binaria, se observa detalle en la figura 13.

**Figura 13. Técnicas de predicción propuestas por Rapidminer**



Elaborado por: Miguel Gonzabay P, año 2022

### Generar el plan de prueba.

Una vez que termina de ejecutar el proceso anterior, la herramienta Rapidminer nos brinda diferentes modelos, cada uno desglosa características que poseen cifras métricas las cuales permitirán valorar su rendimiento, además se indica que el programa por defecto escogió para los modelos seleccionados 4412 de 7353 números de datos que corresponde un 60% para entrenamiento, el restante 40% para datos de prueba. A continuación, se detallan las métricas que el programa elige de acuerdo a los resultados obtenidos mediante sus métricas de evaluación, en primer lugar, se sitúa la precisión (ACCURACY), que muestra la exactitud con que se ejecuta el modelo y la cifra es un valor porcentual, en segundo lugar, tenemos el tiempo total de ejecución, que muestra el tiempo en segundos en que tarda en ejecutarse el modelo (TOTAL\_TIME). Tiempo de

puntuación, (SCORING\_TIME) se refiere al mejor tiempo de respuesta al estimar valores de predicción, tal como se observa en la figura 14.

**Figura 14. Características de los modelos**

Model	Accuracy	Standard Deviation	Gains	Total Time	Training Time (1,000 ...	Sco
Deep Learning	98.3%	± 3.7%	56	21 s	4 s	3 s
Decision Tree	100.0%	± 0.0%	58	9 s	919 ms	905
Random Forest	100.0%	± 0.0%	58	39 s	1 s	18 s
Gradient Boosted Trees	78.3%	± 4.6%	32	56 s	2 s	2 s
Support Vector Machine	100.0%	± 0.0%	58	16 s	805 ms	738

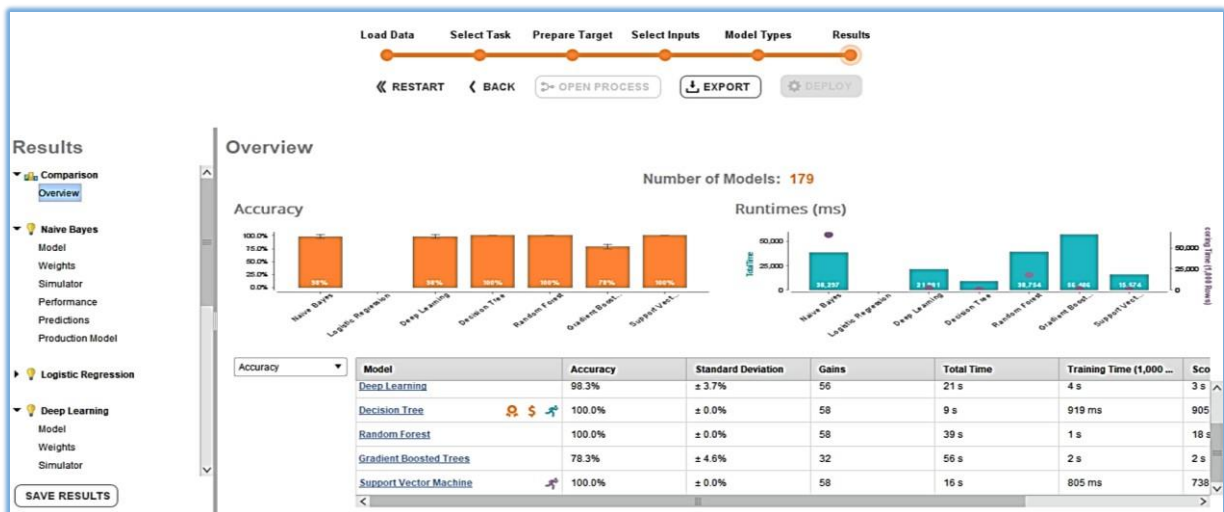
Elaborado por: Miguel Gonzabay P, año 2022

### Construir el modelo.

En una de las etapas anteriores, el plan de prueba, se tomaron como ejemplo métricas de evaluación que se contemplan para cada modelo, basándose en aquellas y orientados al primer objetivo del proyecto que es emplear técnicas de predicción en minería de datos para construir modelos de algoritmos, se detalla lo siguiente:

El programa Rapidminer ejecuta los modelos sobre el conjunto de datos, la herramienta por defecto genera un grupo de modelos tal como se muestra en la figura 15.

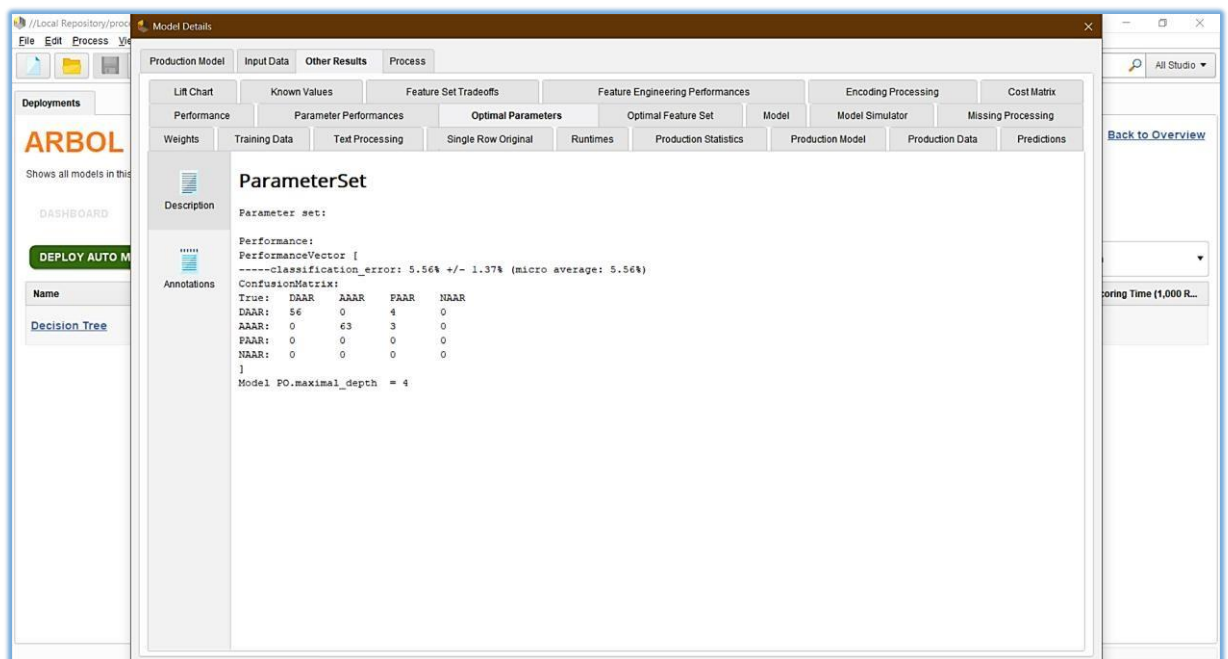
**Figura 15. Detalle del modelo**



Elaborado por: Miguel Gonzabay P, año 2022

Cada uno de los modelos despliega un menú donde se observan todos los parámetros que contienen diferentes métricas y ayudan a entender su funcionamiento. Se procede hacer la revisión del contenido de cada una de las pestañas, para observar el comportamiento del modelo y considerar con que métricas se debe trabajar para la siguiente etapa. En la figura 16, se visualiza el desempeño del modelo de árbol de decisión con los detalles mostrados en la matriz de confusión.

**Figura 16. Matriz de confusión**



Elaborado por: Miguel Gonzabay P, año 2022

### **Descripción del modelo.**

En este apartado se describen los resultados que se obtuvieron al ejecutar los modelos propuestos, en la figura 14 se muestra como ejemplo las métricas del modelo SUPPORT\_VECTOR\_MACHINE, que en español significa máquina de vector de soporte y a continuación se describe cada uno de sus parámetros:

La precisión (ACCURACY), es de 100%, un valor porcentual que representa cuanto se maximiza la precisión del modelo, la desviación estándar (STANDARD\_DEVIATION), representa una medida de dispersión en este caso no hay variación porque tiende a un valor de 0, la ganancia (GAINS) es de 58, es una cantidad que representa el valor de

relevancia de un atributo, el tiempo total en que se ejecuta el modelo (TOTAL\_TIME) que es de 16 segundos, el tiempo de entrenamiento de los datos (TRAINING\_TIME) que es de 805 milisegundos y el tiempo de puntuación o también denominada tiempo de respuesta (SCORING\_TIME) que es de 730 milisegundos.

Las cifras anteriores ayudan a medir el rendimiento del modelo, las métricas de evaluación nos ayudan a verificar la calidad de las predicciones que ofrece, en el ejemplo se muestra que el valor de la exactitud del modelo es de 100%, que indica que ha acertado 10 de 10 predicciones. La ganancia mide la efectividad en comparación de los resultados del análisis descriptivo y del análisis predictivo, marca un valor de 58 que corresponde al 58% de efectividad a favor del modelo. Por último, el tiempo de ejecución y el tiempo de entrenamiento deben tener respuesta inmediata para garantizar un modelo eficiente.

## **Fase 5. Evaluación.**

### **Evaluar el modelo.**

En la figura 14, se muestran las cifras de las métricas, el programa Rapidminer hace una valoración y presenta los mejores modelos basados en su rendimiento, tomando en cuenta la precisión de la predicción, el mejor tiempo de ejecución y el mejor tiempo de puntuación, que fueron descritos en la etapa anterior.

En la tabla 5, se ha ubicado los modelos que por defecto el programa eligió, valorando el rendimiento de cada uno, basados en los parámetros de predicción y que se describieron en el apartado anterior para su fundamentación.

**Tabla 5. Cuadro comparativo entre los modelos**

<b>Modelos</b>	<b>ACCURACY</b>	<b>RUN_TIME</b>	<b>SCORING_TIME</b>
<b>NAIVE_BAYES</b>	98.3%	38 s	38 s
<b>DECISION_TREE</b>	100%	9 s	905 ms
<b>SUPPORT_VECTOR_MACHINE</b>	100%	16 s	730 ms

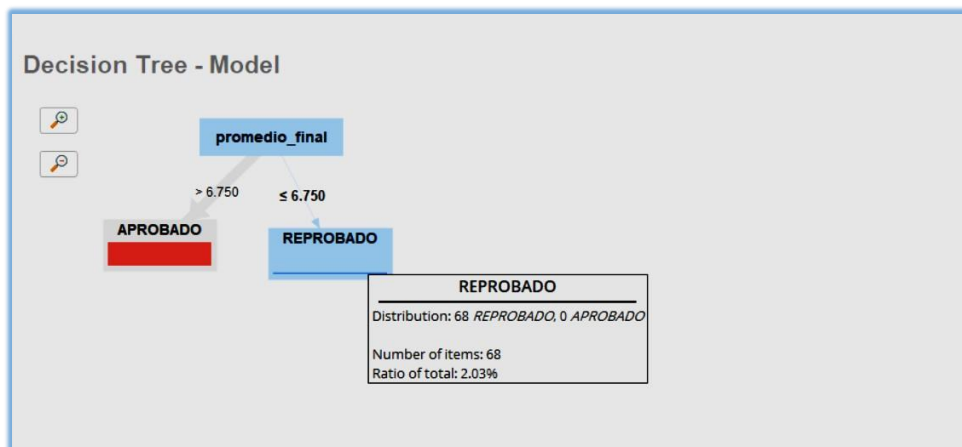
Elaborado por: Miguel Gonzabay P, año 2022

En esta etapa se procede a verificar si se ha cumplido con el objetivo de comparar modelos de predicción, si todo ha sido a favor se puede pasar a la parte de implantación, en el caso de haber algo inconcluso debe darse solución dentro de la misma para proceder.

### **Evaluar resultados.**

Comparando los resultados tenemos en primer lugar el modelo de DECISION TREE con una precisión del 100%, con el mejor tiempo de ejecución realizado en 9 segundos y un tiempo de respuesta de 905 milisegundos. En segundo lugar tenemos el SUPPORT VECTOR MACHINE con una precisión del 100%, un tiempo de ejecución de 16 s y un tiempo de respuesta de 730 milisegundos y en tercer lugar está el NAIVE BAYES con una precisión del 98.3%, un tiempo de ejecución de 38 segundos, con un tiempo de respuesta de 38 segundos, de esta manera se concluye que el modelo prioritario por su rendimiento es DECISION TREE o árbol de decisión, porque posee uno de los mayores porcentajes de precisión, la mejor marca en tiempo de ejecución, que son parámetros relevantes al evaluar un modelo en Rapidminer, véase un esquema del modelo en la figura 17.

**Figura 17. Estructura del modelo**



Elaborado por: Miguel Gonzabay P, año 2022

### **Fase 6. Implantación.**

Una vez valorado y seleccionado el modelo ideal, se somete a la prueba final que servirá para cumplir con el tercer objetivo del proyecto, que es presentar un dashboard o un panel interactivo que permitirá visualizar de forma gráfica los datos de origen interactuando con los resultados del modelo de predicción y que a su vez ayudará a los interesados a realizar acciones como:

- ✓ Monitoreo y análisis de los resultados finales de un período académico.
- ✓ Comparaciones de nivel de porcentajes del rendimiento académico por cursos.
- ✓ Verificar que grupo de estudiantes tiende a un bajo rendimiento según la escala de aprendizaje.
- ✓ Identificar estudiantes por materia o área que posean un índice alto o bajo de promedio sea este por parcial o quimestre.

### Presentar resultados de predicción en el dashboard.

El modelo de predicción elegido para el ejemplo fue DECISION TREE que en este proyecto muestra reglas jerárquicas mediante los nodos que son los elementos que lo constituyen y conducen a una decisión final, tal como se muestra en la figura 17.

Para la presentación se tomó los datos descargados del programa Rapidminer del modelo elegido, tal como se muestra en la figura 18. Luego se procesó los valores de predicción en la herramienta POWER BI que sirvió para el diseño del dashboard interactivo. Para el proceso se tomaron las calificaciones de los estudiantes de todos los niveles de estudio en función del atributo “estado” y son direccionados a las escalas de aprendizaje y los promedios como se observa en la tabla 4.

**Figura 18. Valores de predicción del modelo**

Ro...	codest	estado	prediction(e...	confidence(REPROBADO)	confidence(APROBADO)	materia	valor escala	q1ex	q1	q...	q2	promedio_fl...
302	200010	REPROBADO	REPROBADO	1	0	MATEMÁTICAS	PAR	6.00	5	5	5	5
303	200010	APROBADO	APROBADO	0.167	0.833	QUÍMICA	AAR	8.00	8	8	7	7.500
304	200010	APROBADO	APROBADO	0.167	0.833	BIOLOGIA	AAR	7.00	7	7	8	7.500
305	200010	APROBADO	APROBADO	0.167	0.833	EDUCACIÓN ...	DAR	8.00	9	10	9	9
306	200010	APROBADO	APROBADO	0.167	0.833	EDUCACIÓN ...	AAR	8.00	9	7	7	8
307	200010	APROBADO	APROBADO	0.167	0.833	SKILL LAB	DAR	10.00	10	7	8	9
308	200010	APROBADO	APROBADO	0.167	0.833	RAZONAME...	AAR	7.00	8	7	7	7.500
309	200016	REPROBADO	REPROBADO	1	0	MATEMÁTICAS	PAR	5.00	5	7	6	5.500
310	200016	APROBADO	APROBADO	0.167	0.833	HISTORIA Y ...	AAR	8.00	8	8	9	8.500
311	200016	APROBADO	APROBADO	0.167	0.833	EDUCACIÓN ...	AAR	8.50	8	9	8	8
312	200016	APROBADO	APROBADO	0.167	0.833	FILOSOFIA	AAR	8.00	7	9	8	7.500
313	200016	APROBADO	APROBADO	0.167	0.833	LENGUA Y LL...	AAR	6.00	8	8...	7	7.500
314	200016	APROBADO	APROBADO	0.333	0.667	LENGUA EXT...	AAR	7.00	7	7	7	7
315	200016	APROBADO	APROBADO	0.167	0.833	EDUCACIÓN ...	DAR	10.00	10	8	9	8.500

Elaborado por: Miguel Gonzabay P, año 2022



En la presentación podemos visualizar la primera columna que contiene el código del estudiante, en la siguiente encontramos el atributo a predecir que corresponde a “estado” que asigna el valor de acuerdo a la condicionante que indica si el promedio es mayor igual a 7 “APROBADO” caso contrario “REPROBADO”, en las columnas siguientes se observan valores de predicción entre 0 y 1 que se establecen al obtener la probabilidad condicional y por último las columnas de los atributos que se han seleccionado para relacionar la variable con los datos proporcionados en cada uno de los campos.

Para mejor explicación se toma el caso de dos estudiantes referente a la variable de predicción como se muestra en la figura 18, donde apreciamos diferentes valores de confianza en la columna reprobado, tenemos el valor de uno (1) en calidad de verdadero corresponde al 100% , por ende indica que el estudiante ha reprobado, otro valor que aparece es 0,167 que corresponden el 17% asignado a promedios entre 7,5 y 9, es un valor bajo de porcentaje indica que los promedios que se encuentre dentro del rango tienen una gran probabilidad de aprobar, luego tenemos un valor de 0,333 que corresponde a un 33% en un rango de promedios entre 7 y 7,5 donde debe existir una alerta al tener una probabilidad aproximada al valor de condición que es 7. Así mientras el valor de predicción se acerque más a 1 aumenta su probabilidad de reprobar. La otra columna corresponde a los valores de confianza de aprobados y contienen los valores de predicción en caso de ocurrir el suceso contrario a los reprobados por ejemplo que si el resultado tiene cero (0) en calidad de falso será reprobado entonces mientras el valor de predicción se acerque más a 1 aumenta su probabilidad de aprobar.

A continuación, se muestra el diseño del dashboard en la figura 19, se toma como ejemplo los resultados del primer año de bachillerato, mostrando una comparación entre los datos iniciales y los datos predictivos, con los que se hará el análisis del rendimiento según la variable elegida para la predicción. En el gráfico de barra apiladas se puede apreciar el número de los estudiantes por materia aprobados y reprobados, en el gráfico de líneas permite visualizar variación del número de estudiantes a lo largo del rango de las diferentes escalas generado en cada materia, en el gráfico circular se presenta los valores del resultado de la predicción de promedios en forma de porcentaje y por último en el

gráfico de columnas apiladas se muestran los valores de confianza de la predicción en cuanto a estudiantes reprobados.

**Figura 19. Dashboard con valores predictivos**



Elaborado por: Miguel Gonzabay P, año 2022

## CONCLUSIONES

El presente trabajo se basó en el empleo de técnicas de minería de datos para construir modelos de predicción del rendimiento académico de estudiantes. Luego se evaluaron los modelos tomando en cuenta las siguientes métricas de valoración: la precisión, el mejor tiempo de ejecución y el mejor tiempo de puntuación, además con los resultados obtenidos enfocándose en las métricas se pudo comparar cada uno de los modelos valorando su rendimiento.

El mejor proceso predictivo se alcanzó con el modelo basado en DECISION TREE con una precisión del 100%, un tiempo de ejecución y de respuesta de 9 segundos y 905 milisegundos, respectivamente.

Cabe recalcar que en las entidades educativas al finalizar algún periodo académico se hace un análisis de rendimiento de los estudiantes usando técnicas monótonas que solo dan resultados superficiales y no se profundiza en el problema en el caso de existir un alto porcentaje de estudiantes con bajo rendimiento.

Al emplear técnicas de minería de datos facilitó la búsqueda de patrones de predicción que influyen en el rendimiento académico de los estudiantes, utilizando modelos de predicción que se eligieron con la ayuda de la herramienta Rapidminer basados en la metodología CRISP-DM, con los resultados se obtuvieron ideas para mejorar las estrategias de aprendizaje que se utilizan en la institución educativa John F. Kennedy y así desterrar factores negativos que son derivados del bajo rendimiento académico de los estudiantes. En el caso presentado se observó en todos los niveles un porcentaje alto de aprobados, un porcentaje muy bajo de reprobados y un porcentaje considerable con riesgo de reprobación, lo que da a entender que existe un buen desempeño de parte de los actores del proceso educativo en la institución, pero también necesitan atender al grupo en riesgo. En conclusión, se recomienda que los establecimientos educativos para los diferentes procesos de análisis, utilicen técnicas de minería de datos, porque brindan un gran beneficio en este ámbito.

A futuro se puede agregar al proyecto más tablas con atributos tales como lugar de residencia, condición social, estabilidad económica, relación familiar que son variables

no académicas, pero guardan mucha relación con los estudiantes y también influyen en el estudio como factores externos y con la aplicación de técnicas de minería de datos brinde más información relacionada con el rendimiento académico.

## **RECOMENDACIONES.**

El proyecto de aplicación de técnicas de minería mediante el análisis del rendimiento académico, servirá para orientar trabajos futuros, que permitan buscar nuevas estrategias en beneficio del aprendizaje de los estudiantes de las instituciones educativas.

Para optimizar el proyecto se recomienda incluir o trabajar con otros atributos que tengan relación con el entorno educativo del estudiante de tal manera que se pueda brindar mayor información y esto encamine a obtener más alternativas en un proyecto similar, con esto se podrán tener más modelos predictivos con la ayuda de otras técnicas de minería en función a las variables propuestas, lo que garantiza un trabajo con mejor resultados de predicción.

La metodología CRISP-DM es recomendable para el desarrollo de los procesos de minería de datos, cada una de sus fases se adapta a las necesidades propuestas en el proyecto.

Se recomienda usar la herramienta Rapidminer por la eficiencia al trabajar con operadores en cadena, logrando efectividad en menos tiempo, su entorno grafico también es importante en la interacción con el usuario.

## ANEXOS

Anexo 1: Aval para usar el nombre de la institución.



**UNIDAD EDUCATIVA "JOHN F. KENNEDY"**

José Luis Tamayo – Salinas – Ecuador

Telf. 2775-218

Email: [24h00341johnfkennedy@gmail.com](mailto:24h00341johnfkennedy@gmail.com)



José Luis Tamayo, septiembre 15 de 2022  
Oficio No. 308 - UEJFK-2022

Lcdo.  
**Miguel Gonzabay Panchana**  
Docente Unidad Educativa John F. Kennedy  
Ciudad.

De mis consideraciones:

La Superioridad de la Unidad Educativa "John F. Kennedy", le expresa un cordial saludo. En relación a la solicitud presentada por usted, el día 14 de septiembre del presente año, sobre el uso del nombre de la Institución como propuesta de tesis, tengo a bien comunicarle que ha sido autorizada.

Sin otro particular, y deseando éxitos en la culminación de su Maestría, me suscribo de usted.

Atentamente,

  
Tcnlgo. Luis Montenegro Peña  
Gerente Compañía Colegio Kennedy



c.c. Rectorado

## REFERENCIAS

1. Carlos, U., De Madrid, I., Galán, V., Tutora, C., & Castro Galán, E. (2015). *Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario*.
2. Dirección Nacional de Normativa Jurídica del Ministerio de Educación. (2017). *LEY ORGÁNICA DE EDUCACIÓN INTERCULTURAL*. [www.educacion.gob.ec](http://www.educacion.gob.ec)
3. Menacho Chiok, C. H. (2017). Predicción del rendimiento académico aplicando técnicas de minería de datos. *Anales Científicos*, 78(1), 26.  
<https://doi.org/10.21704/ac.v78i1.811>
4. *RapidMiner - Wikipedia, la enciclopedia libre*. (n.d.). Retrieved September 16, 2022, from <https://es.wikipedia.org/wiki/RapidMiner>
5. *RapidMiner | Aplicaciones y herramientas de Marketing digital*. (n.d.). Retrieved September 16, 2022, from <https://www.martechforum.com/herramienta/rapidminer/>
6. Rico Páez, A., & Gaytán Ramírez, N. D. (2022). Modelos predictivos del rendimiento académico a partir de características de estudiantes de ingeniería. *IE Revista de Investigación Educativa de La REDIECH*, 13, e1426.  
[https://doi.org/10.33010/IE\\_RIE\\_REDIECH.V13I0.1426](https://doi.org/10.33010/IE_RIE_REDIECH.V13I0.1426)
7. *Metodología CRISP-DM - Adictos al trabajo Tutoriales*. (n.d.). Retrieved September 16, 2022, from <https://www.adictosaltrabajo.com/2021/01/14/metodologia-crisp-dm/>
8. Marques, M. P. (2014). *Mineria de Datos. La Metodología Cris-DM de Ibm. El Lenguaje Clem E IBM SPSS Modeler*:
9. Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of Students' Academic Performance Based on Courses' Grades Using Deep Neural Networks. *IEEE Access*, 9, 140731–140746.  
<https://doi.org/10.1109/ACCESS.2021.3119596>
10. Timaran, S. R., Hernandez, I., Caicedo, S. J., Hidalgo, A., & Alvarado, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. *Ingenierías*, 8(26), 37–47.  
<https://ediciones.ucc.edu.co/index.php/ucc/catalog/download/36/40/230-1?inline=1#:~:text=El Descubrimiento de conocimiento en,que el usuario los analice>.
11. Beneyto Sánchez, S. (2015). *Entorno familiar y rendimiento académico*. 96.  
[https://books.google.com/books/about/ENTORNO\\_FAMILIAR\\_Y\\_RENDIMIENTO\\_ACADEMIC.html?hl=es&id=7nVTBwAAQBAJ](https://books.google.com/books/about/ENTORNO_FAMILIAR_Y_RENDIMIENTO_ACADEMIC.html?hl=es&id=7nVTBwAAQBAJ)
12. Carlos, U., De Madrid, I., Galán, V., Tutora, C., & Castro Galán, E. (2015). *Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario*.
13. Contreras, L. E., Fuentes, H. J., Rodríguez, J. I., Contreras, L. E., Fuentes, H. J., & Rodríguez, J. I. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. *Formación Universitaria*, 13(5), 233–246.  
<https://doi.org/10.4067/S0718-50062020000500233>

14. Menacho Chiok, C. H. (2017). Predicción del rendimiento académico aplicando técnicas de minería de datos. *Anales Científicos*, 78(1), 26.  
<https://doi.org/10.21704/ac.v78i1.811>
15. Grasso, P. (2020). Rendimiento académico: un recorrido conceptual que aproxima a una definición unificada para el ámbito superior. *Revista de Educación*, 11(20), 87–102.
16. Carlos, U., De Madrid, I., Galán, V., Tutora, C., & Castro Galán, E. (2015). *Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario*.
17. Gustavo, I., Cordero Páez, E., Hugo, I. V., & Galarza, S. (n.d.). “Proceso para el descubrimiento de conocimiento en las bases de datos de la Universidad de Cuenca mediante técnicas de Data Mining” AUTOR.
18. *Big Data en Educación: El futuro digital del aprendizaje, la política y la ...* - Ben Williamson - Google Libros. (n.d.). Retrieved November 7, 2022, from [https://books.google.com.ec/books?id=upojEAAAQBAJ&pg=PT148&dq=mineria+de+datos+y+el+rendimiento+academico+de+colegio&hl=es-419&sa=X&ved=2ahUKEwiLle2espr7AhXN44UKHcAFC3IQ6AF6BAgIEAI#v=onepage&q=mineria de datos y el rendimiento academico de colegio&f=false](https://books.google.com.ec/books?id=upojEAAAQBAJ&pg=PT148&dq=mineria+de+datos+y+el+rendimiento+academico+de+colegio&hl=es-419&sa=X&ved=2ahUKEwiLle2espr7AhXN44UKHcAFC3IQ6AF6BAgIEAI#v=onepage&q=mineria+de+datos+y+el+rendimiento+academico+de+colegio&f=false)
19. *La investigación del rendimiento estudiantil: problemas y paradigmas* - Eleazar Narváez - Google Libros. (n.d.). Retrieved November 7, 2022, from [https://books.google.com.ec/books?id=ztS1fmBreBAC&printsec=frontcover&dq=dificultades+en+el++rendimiento+academico&hl=es-419&sa=X&redir\\_esc=y#v=onepage&q=dificultades en el rendimiento academico&f=false](https://books.google.com.ec/books?id=ztS1fmBreBAC&printsec=frontcover&dq=dificultades+en+el++rendimiento+academico&hl=es-419&sa=X&redir_esc=y#v=onepage&q=dificultades+en+el+rendimiento+academico&f=false)