UEA: Datos a gran escala

Impartida por Pedro Pablo González Pérez

Caso de estudio

Predict students' dropout and academic success

DataSet

UC Irvine Machine Learning

Metodología

CRISP-DM (Cross Industry Standard Process for Data Mining)

Desarrollado por:

2193076785 - Eduardo Isaac Davila Bernal

Índice del contenido:

Marco Teórico
Fase 1
- 1.1.- Determinación de los objetivos del proyecto.
- 1.2.- Valoración de la situación actual del objetivo del proyecto.
- 1.3.- Datos disponibles
- 1.4.- Disponibilidad de recursos humanos y tecnológicos
- 1.5.- Identificación de riesgos.
- 1.6.- Determinación de los objetivos de minería de datos.
- 1.7.- Propuesta del enfoque metodológico.
- 1.8.- Exploración de los datos

Fase 2
- 2.1.- Recopilación de los datos
- 2.2.- Descripción de los datos
- 2.3 Descripción de los atributos
- 2.4 Exploración de los datos (exhaustiva)
- 2.5.- Verificación de la calidad de los datos
- 2.6.- 5V en la Minería de Datos

Fase 3
- 3.1.- Selección de datos
- 3.2.- Limpieza de los datos
- 3.3.- Construcción de nuevos datos
- 3.4.- Integración de datos
- 3.5.- Formato de datos

Fase 4
- 4.1.- Objetivo principal
  - 4.1.1.- Utilizando el Software IBM SPSS Modeler
    - 4.1.1.1.- Selección de características con PCA
      - 4.1.1.1.1.- Balanceo del DataSet
      - 4.1.1.1.2.- Aplicación de los modelos sin balanceo ni selección de características
      - 4.1.1.1.3.- Resultados de los modelos sin balanceo ni selección de características:
      - 4.1.1.1.4.- Aplicación de los modelos con balanceo sobremuestreo y con selección de características.
      - 4.1.1.1.5.- Resultados de los modelos con balanceo sobremuestreo y con selección de características
      - 4.1.1.1.6.- Aplicación de los modelos con balanceo submuestreo y con selección de características
      - 4.1.1.1.7.- Resultados de los modelos con balanceo submuestreo y con selección de características
    - 4.1.1.2.- Selección de características con Neural Net
      - 4.1.1.2.1.- Balanceo del DataSet
      - 4.1.1.2.2.- Aplicación de los modelos sin balanceo ni selección de características
      - 4.1.1.2.3.- Resultados de los modelos sin balanceo ni selección de características:

4.1.1.2.4.- Aplicación de los modelos con balanceo sobremuestreo y con selección de características
4.1.1.2.5.- Resultados de los modelos con balanceo sobremuestreo y con selección de características
4.1.1.2.6.- Aplicación de los modelos con balanceo submuestreo y con selección de características
4.1.1.2.7.- Resultados de los modelos con balanceo submuestreo y con selección de características
4.1.2.- Utilizando Python
4.1.2.1.- Librerías utilizadas
4.1.2.2.- Correlación con la variable objetivo (Target)
4.1.2.3.- Selección de las técnicas de modelado
4.1.2.4.- Error Medio Absoluto (Mean Absolute Error)
4.1.2.5.- Generación de los modelos
4.1.2.5.1.- Random Forest Regressor (Todos los atributos)
4.1.2.5.2.- Random Forest Regressor (Solo los mejores atributos)
4.1.2.5.3.- Logistic Regression (Todos los atributos)
4.1.2.5.4.- Logistic Regression (Solo los mejores atributos)
Fase 5
5.1.- Evaluaciones con IBM SPSS Modeler
5.1.1 - Evaluación aplicando PCA
5.1.1.1 - Evaluación de los modelos sin balanceo ni selección de características
5.1.1.1.1 - Neural Net
5.1.1.1.2 - Random Trees
5.1.1.1.3 - Tree-AS
5.1.1.1.4 - C&R Tree
5.1.1.1.5 - C5.0
5.1.1.1.6 - Logistic
5.1.1.1.7 - GLE
5.1.1.1.8 - SVM
5.1.1.1.9 - LSVM
5.1.1.1.10 - Bayes Net
5.1.1.1.11 - Mejor modelo
5.1.1.2 - Evaluación de los modelos con balanceo y selección de características.
5.1.1.2.1 - Neural Net
5.1.1.2.2 - Random Trees
5.1.1.2.3 - Trees-AS
5.1.1.2.4 - C&R Tree
5.1.1.2.5 - C5.0
5.1.1.2.6 - Logistic
5.1.1.2.7 - GLE
5.1.1.2.8 - SVM
5.1.1.2.9 - LSVM
5.1.1.2.10 - Bayes Net
5.1.1.2.11 - Mejor modelo

5.1.1.3 - Evaluación de los modelos con balanceo submuestreo y selección de características
5.1.1.3.1 - Neural Net
5.1.1.3.2 - Random Trees
5.1.1.3.3 - Trees-AS
5.1.1.3.4 - C&R Tree
5.1.1.3.5 - C5.0
5.1.1.3.6 - Logistic
5.1.1.3.7 - GLE
5.1.1.3.8 - SVM
5.1.1.3.9 - LSVM
5.1.1.3.10 - Bayes Net
5.1.1.3.11 - Mejor Modelo
5.1.2 - Evaluación aplicando Neural Net
5.1.2.1 - Evaluación de los modelos sin balanceo ni selección de características
5.1.2.1.1 - Neural Net
5.1.2.1.2 - Random Trees
5.1.2.1.3 - Tree-AS
5.1.2.1.4 - C&R Tree
5.1.2.1.5 - C5.0
5.1.2.1.6 - Logistic
5.1.2.1.7 - GLE
5.1.2.1.8 - SVM
5.1.2.1.9 - LSVM
5.1.2.1.10 - Bayes Net
5.1.2.1.11 - Mejor modelo
5.1.2.2 - Evaluación de los modelos con balanceo y selección de características
5.1.2.2.1 - Neural Net
5.1.2.2.2 - Random Trees
5.1.2.2.3 - Tree-AS
5.1.2.2.4 - C&R Tree
5.1.2.2.5 - C5.0
5.1.2.2.6 - Logistic
5.1.2.2.7 - GLE
5.1.2.2.8 - SVM
5.1.2.2.9 - LSVM
5.1.2.2.10 - Bayes Net
5.1.2.2.11 - Mejor modelo
5.1.2.3 - Evaluación de los modelos con balanceo submuestreo y selección de características
5.1.2.3.1 - Neural Net
5.1.2.3.2 - Random Trees
5.1.2.3.3 - Tree-AS
5.1.2.3.4 - C&R Tree
5.1.2.3.5 - C5.0

5.1.2.3.6 - Logistic
5.1.2.3.7 - GLE
5.1.2.3.8 - SVM
5.1.2.3.9 - LSVM
5.1.2.3.10 - Bayes Net
5.1.2.3.11 - Mejor modelo
5.2 - Mejores modelos de IBM SPSS Modeler
5.3 - Mejores modelos en Python
5.4 - Conclusiones del análisis de datos
Referencias

Marco Teórico

Deserción escolar

Se puede definir como el abandono de una persona de una institución educativa antes de recibir su diploma. El Departamento Nacional de Estadística (DANE) señala que la deserción se produce cuando el estudiante se desvincula definitiva o temporalmente del sistema educativo formal.

Causas de deserción en México

Las principales causas de la deserción se vinculan con:

- Deficiencias en cuanto a la preparación académica previa de los estudiantes, lo cual dificulta el proceso de integración social en el nuevo nivel educativo.
- Factores económicos que impiden a los estudiantes de bajos recursos solventar los costos asociados a los estudios universitarios.
- Factores inter sistémicos que se refieren a la oferta educativa, la desigualdad en la calidad de los servicios educativos y los mecanismos de acceso, asociados a la asignación de plantel, modalidad y turno.
- Factores intra sistémicos que se vinculan con las prácticas pedagógicas inadecuadas, formación docente limitada y condiciones laborales precarias, infraestructura y equipamiento insuficiente, incompatibilidad entre la cultura juvenil y escolar, currículo poco pertinente, gestión escolar deficiente, y participación limitada de padres y estudiantes en la escuela.
- Embarazo adolescente que afecta el desarrollo normal de las actividades académicas.
- Entorno familiar.
- Actitud o la desmotivación del estudiante.

Posgrado

Los posgrados son los estudios universitarios que se realizan una vez que el estudiante obtuvo su título de Licenciatura. Por lo general, estos fomentan actividades de investigación y actualización de contenidos y se especializan en ciertas áreas del conocimiento.

Fase 1
Comprensión del dominio del problema

1.1.- Determinación de los objetivos del proyecto.

El objetivo del proyecto es desarrollar y aplicar modelos de predicción trabajando con la metodología ***CRISP-DM*** para:

- **Objetivo Principal:** Predecir la deserción estudiantil y el éxito académico en las instituciones de educación superior.
- **Objetivo Secundario:** Predecir si el alumno es apto para un posgrado una vez terminados los estudios superiores.

El problema clave que se busca abordar es tanto la deserción escolar como el fracaso educativo en la educación superior. Esto representa un obstáculo para el crecimiento económico, el empleo, la competitividad y la productividad, impactando directamente en la vida de los estudiantes, sus familias, las instituciones de educación superior y la sociedad en su conjunto. Lo que se requiere es construir modelos para predecir el rendimiento académico y la deserción para hacer una estimación del riesgo de deserción y fracaso escolar.

La pregunta central que este proyecto busca responder es:

"¿Cómo podemos prever y abordar de manera efectiva la deserción estudiantil en la educación superior?"

Se busca entender los factores que llevan a los estudiantes a abandonar sus estudios y utilizar estos conocimientos para desarrollar modelos de predicción que permitan anticipar la deserción. También se requiere abordar qué factores están a favor del éxito académico.

1.2.- Valoración de la situación actual del objetivo del proyecto.

La información fue adquirida mediante diferentes bases de datos (no relacionadas entre sí) de estudiantes inscritos en diferentes carreras universitarias, como agronomía, diseño, educación, enfermería, periodismo, administración, trabajo social y tecnologías.

Para resolver el problema que se plantea necesitamos saber cuales son los principales factores que influyen en la deserción y el fracaso académico en una institución de educación superior. Existen diferentes causas que ocasionan este problema, por ejemplo:

- Estado civil
- Condiciones socioeconómicas
- Falta de preparación académica
- Compromisos laborales
- Entre otros factores

Para comprender de mejor manera el problema que se quiere resolver es necesario entender algunos conceptos importantes que están presentes en el DataSet que se usará. A continuación presentamos las definiciones de estos conceptos:

Término	Definición
Deserción Estudiantil	Abandono prematuro de un programa educativo antes de su finalización.
Éxito Académico	Se refiere al logro positivo de los objetivos académicos que implican completar un programa educativo de manera satisfactoria.
Fracaso Académico	No alcanzar los estándares académicos o no completar con éxito un programa educativo.
Datos Demográficos	Información estadística sobre características de la población, como edad, género, nacionalidad, etc.
Datos Socioeconomicos	Información que combina aspectos sociales y económicos, como ingresos, nivel educativo de los padres y condiciones sociales.
Datos Macroeconomicos	Información económica a nivel nacional o global, como tasas de empleo, PIB y otros indicadores económicos amplios.
Posgrado	Tienen la finalidad de actualizar y/o especializar a los profesionales con una alta capacidad crítica y creativa, a través de investigaciones originales.
Curricular units 1st sem (grade)	Promedio final del primer semestre del alumno.
Curricular units 2nd sem (grade)	Promedio final del segundo semestre del alumno.
Target	Estado actual del estudiante, si se encuentra estudiando, fue dado de baja o se ha graduado.

1.3.- Datos disponibles

La información disponible con la que se cuenta es un volumen de datos que contiene **4,424 registros** y **37 atributos** donde cada registro representa un estudiante.

El dataset incluye los datos del momento de la inscripción de los estudiantes y datos al final del primer y segundo semestre. Las fuentes de datos utilizadas consisten en datos internos y externos de la institución, así como de ***Academic Management System*** **(AMS)**, la institución de ***Support System for the Teaching Activity*** **(denominada internamente como PAE)**, la ***General Directorate of Higher Education*** **(DGES)** y por la ***Contemporary Portugal Database*** **(PORDATA)**

En la siguiente tabla se agruparon por clase los atributos que contiene el dataset, se dividieron en datos demográficos, socioeconómicos, macroeconómicos y académicos al momento de la inscripción y del primer y segundo semestre:

Clase de atributo	Atributo
Datos demográficos	Marital status Nationality Displaced Gender Age at enrollment Internationality
Datos socioeconómicos	Mother's qualification Father's qualification Mother's occupation Father's occupation Educational special needs Debtor Tuition fees up to date Scholarship holder
Datos macroeconómicos	Unemployment rate Inflation rate GDP
Datos académicos al momento de la inscripción	Application mode Application order Course Daytime/evening attendance Previous qualification
Datos académicos al final del 1er semestre	Curricular units 1st sem (credited) Curricular units 1st sem (enrolled) Curricular units 1st sem (evaluations) Curricular units 1st sem (approved) Curricular units 1st sem (grade) Curricular units 1st sem (without evaluations)
Datos académicos al final del 2do semestre	Curricular units 2nd sem (credited) Curricular units 2nd sem (enrolled) Curricular units 2nd sem (evaluations) Curricular units 2nd sem (approved) Curricular units 2nd sem (grade) Curricular units 2nd sem (without evaluations)

Marital status	Application mode	Application order	Course	Daytime/evening attendance	Previous qualification	Previous qualification (grade)	Nacionality	Mothers qualification	Fathers qualification	Mothers occupation	Fathers occupation	Admission grade
1	17	5	171	1	1	122.0	1	19	12	5	9	127.3
1	15	1	9254	1	1	160.0	1	1	3	3	3	142.5
1	1	5	9070	1	1	122.0	1	37	37	9	9	124.8
1	17	2	9773	1	1	122.0	1	38	37	5	3	119.6
2	39	1	8014	0	1	100.0	1	37	38	9	9	141.5
2	39	1	9991	0	19	133.1	1	37	37	9	7	114.8
1	1	1	9500	1	1	142.0	1	19	38	7	10	128.4
1	18	4	9254	1	1	119.0	1	37	37	9	9	113.1
1	1	3	9238	1	1	137.0	62	1	1	9	9	129.3
1	1	1	9238	1	1	138.0	1	1	19	4	7	123.0
1	1	1	9670	1	1	139.0	1	38	19	5	7	130.6
1	1	1	9500	1	1	136.0	1	19	38	9	9	119.3
1	1	2	9853	1	1	133.0	1	19	37	4	9	130.2
1	53	1	9254	1	42	110.0	1	1	1	4	7	111.8
1	1	1	9085	1	1	149.0	1	38	37	5	5	137.1
1	1	1	9773	1	1	127.0	1	19	37	9	3	120.7
1	18	1	9238	1	1	137.0	1	19	38	5	8	137.4
1	17	2	9500	1	1	135.0	1	19	1	5	4	127.3
1	1	1	9130	1	1	137.0	1	3	19	3	5	136.3
1	1	1	9853	1	1	140.0	1	19	19	7	7	124.6
1	1	3	171	1	1	122.0	1	1	1	9	8	120.3
1	18	4	9556	1	1	127.0	1	1	38	4	7	121.8
1	1	4	9500	1	1	142.0	1	19	19	1	1	125.5
1	1	4	9670	1	1	125.0	1	1	38	4	7	114.9
1	1	1	9500	1	1	126.0	1	19	19	3	7	123.9
1	1	1	9238	1	1	151.0	1	19	38	9	9	157.0
1	17	2	9238	1	1	115.0	1	19	38	7	8	116.4
1	1	1	9085	1	1	138.0	1	19	19	3	3	131.0
1	1	1	9147	1	1	135.0	1	19	1	9	9	122.1
1	17	2	9500	1	1	127.0	1	3	3	1	10	118.8
1	44	1	9003	1	39	150.0	1	4	5	4	2	150.0
1	51	1	9070	1	1	125.0	1	42	3	125	124	130.0
1	1	1	9147	1	1	143.0	1	1	1	4	6	138.8
1	18	1	8014	0	1	138.0	1	19	1	9	5	134.5
1	1	1	9991	0	1	130.0	1	2	37	3	9	131.4
1	39	1	33	1	1	130.0	1	38	37	9	6	102.5
1	39	1	9119	1	10	133.1	1	34	37	0	0	128.8
1	43	1	9147	1	1	140.0	1	37	37	6	6	122.9
4	39	1	9991	0	1	120.0	1	19	19	9	3	113.9
1	44	1	9130	1	39	120.0	6	4	4	3	6	120.0
1	1	5	9773	1	1	126.0	1	3	3	2	2	121.1
1	1	2	9147	1	1	133.0	1	3	19	2	7	120.4
1	1	6	9500	1	1	103.0	1	1	3	2	2	100.6
1	17	1	9773	1	1	120.0	1	1	38	5	5	121.4
1	39	1	9991	0	1	133.1	1	37	37	9	9	109.7
1	17	1	9991	0	1	154.0	1	38	38	4	5	134.1
1	17	2	9238	1	1	130.0	1	19	19	9	9	127.6
1	17	6	9500	1	1	132.0	1	37	19	7	7	132.4
1	44	1	8014	0	39	130.0	1	1	19	9	9	133.4
1	1	1	9238	1	1	139.0	1	37	38	9	9	126.1

Ilustración 1.- Sección del DataSet "Students.csv"

1.4.- Disponibilidad de recursos humanos y tecnológicos

No contamos con expertos/especialistas en la materia para asesorarnos en la elaboración de este proyecto.

Los recursos tecnológicos con los que contamos son:

- IBM SPSS Modeler
- Conocimiento del lenguaje R y manejo de RStudio
- Microsoft Excel
- Manejo de librerías para análisis de datos en Python (numpy, pandas, matplotlib)
- Chat GPT Plus como segunda perspectiva para interpretación y análisis de datos.

1.5.- Identificación de riesgos.

- Los riesgos que se pueden presentar es la equivocación de no dar una interpretación adecuada de los datos, eso puede afectar a la hora de poner en marcha el enfoque de la metodología de la minería de datos **(CRISP-DM)**.
- Debido a la gran cantidad de atributos es posible que se sobre analicen los datos o no se analicen lo suficiente por la inexperiencia del equipo en la materia.
- Al avanzar en el proyecto la cantidad de datos con los que inicialmente se cuenta no serán los suficientes para trabajar de manera adecuada en los objetivos del proyecto.

1.6.- Determinación de los objetivos de minería de datos.

Problema de predicción: Lo que se requiere es construir modelos para predecir el rendimiento académico y la deserción para hacer una estimación del riesgo de deserción y fracaso escolar, así como la de conocer qué factores son los que lo ocasionan la deserción, el fracaso y el éxito académico.
Problema de clasificación: Indicar al usuario si es apto para entrar a un posgrado a partir de su historial académico.

1.7.- Propuesta del enfoque metodológico.

Etapa	Tiempo asignado en semanas	Recursos utilizados	Riesgos	Actividades
Comprensión de dominio del problema	**1-2**	- DataSet. - Equipo humano de trabajo.	Entendimiento errónea de algún proceso en el enfoque de minería de datos.	- Asignar objetivo - Valoración del proyecto - Objetivo a la minería de datos - Propuesta de enfoque
Comprensión de los datos	**2-4**	- Dataset. - Software para visualización de datos.	Interpretación de los datos equivocada al contar con una gran cantidad de ellos	- Recopilación - Descripción - Exploración - Verificación - Calidad
Preparación de los datos	**5-7**	- Dataset - Software para manejo de datos	Inexperiencia en el manejo de grandes cantidades de datos	- Selección - Limpieza - Creación - Integración - Formato
Modelado	**6-8**	- Dataset - Software para modelado de datos	Falta de experiencia para analizar y relacionar el conjunto de datos	- Técnicas de modelado - Comprobación - Generación - Evaluación
Evaluación	**8-10**	- Dataset - Software para presentación de resultados	Falta de criterio al intentar llegar al objetivo del proyecto	- Asignar respuestas al problema
Presentación de resultados	**11-12**	- Software para edición de texto e imagen	Llegar a conclusiones erróneas o sesgadas	- Exposición de resultados obtenidos

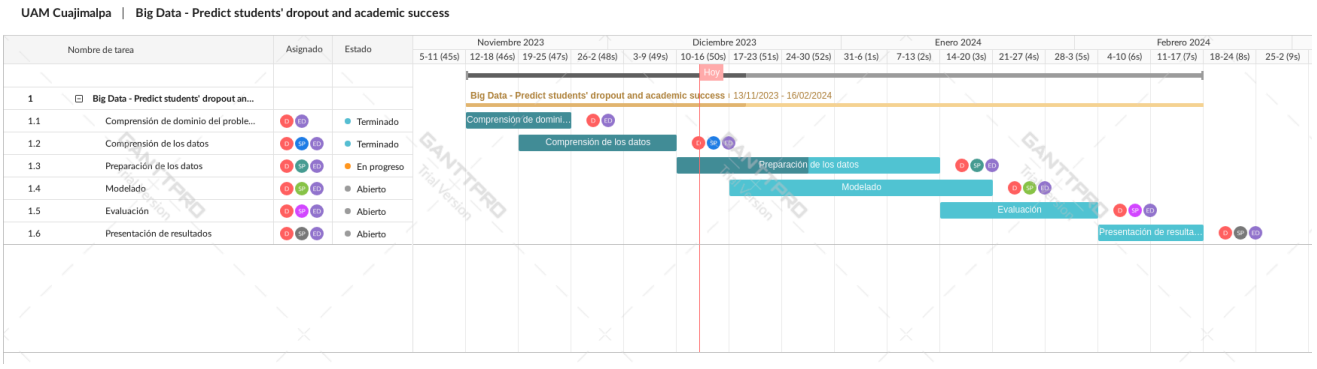


Ilustración 2.- Aquí se presenta el gráfico de Gantt, el cual posibilita la visualización de las actividades y los recursos a lo largo del tiempo en el proyecto.

1.8.- Exploración de los datos

Target: Estado de los estudiantes en el DataSet

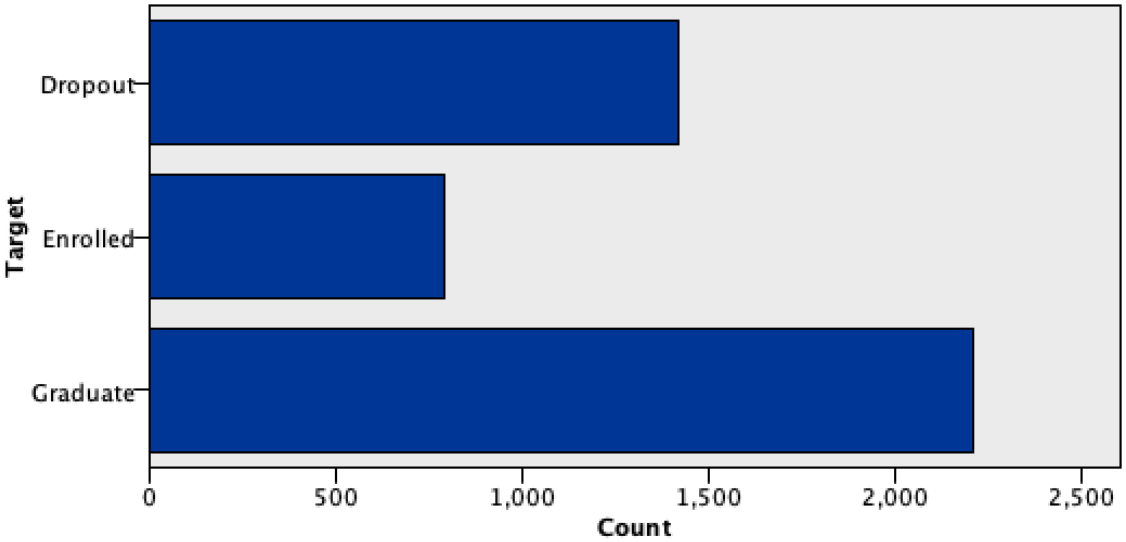


Ilustración 3.- La primera gráfica muestra la distribución del target en el conjunto de datos, el cual indica el estado actual del estudiante. Encontramos tres estados principales: "Graduado", "Desertado" y "Enrolado".

- **Graduado:** representa aquellos estudiantes que han completado exitosamente sus estudios.
- **Desertado:** corresponde a los estudiantes que han abandonado sus estudios antes de completarlos.
- **Enrolado:** indica aquellos estudiantes que están actualmente matriculados en la institución.

Al observar la distribución de estos estados, podemos notar que hay un desbalance en los datos. Hay aproximadamente 2200 registros de estudiantes graduados, alrededor de 1300 registros de estudiantes desertores y cerca de 700 registros de estudiantes enrolados.

Aunque hay una diferencia en la cantidad de registros entre los diferentes estados, este desbalance no parece ser significativo. Esto significa que no hay una discrepancia excesiva entre las clases objetivo en términos de la cantidad de datos disponibles para el análisis. Por lo tanto, este desbalance no debería representar un problema importante para el análisis de datos y la construcción de modelos predictivos.

Marital status: Estado civil de los estudiantes del DataSet

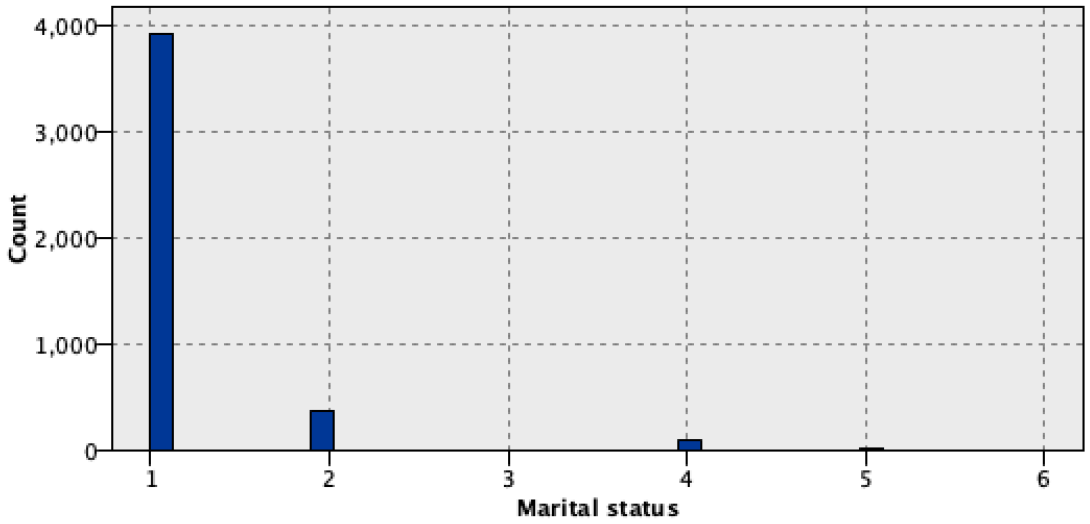


Ilustración 4.- La segunda gráfica muestra la distribución del estado civil de los estudiantes en el conjunto de datos. Los estados civiles representados incluyen "Soltero", "Casado", "Divorciado", "Unión de hecho", "Separado legalmente" y "Viudo".

Observamos un notable desbalance en los datos en función del estado civil de los estudiantes. La mayoría de los registros, aproximadamente 3,800, corresponden a estudiantes solteros. Por otro lado, hay alrededor de 200 registros de estudiantes casados, menos de 100 registros para los estados civiles de divorciado, unión de hecho, separado legalmente y viudo.

Este desbalance en la distribución de los datos según el estado civil podría tener implicaciones en el análisis de datos y la construcción de modelos predictivos. Es importante tener en cuenta este desbalance al interpretar los resultados y al realizar el modelado, ya que puede afectar la precisión de los modelos, especialmente si alguna categoría minoritaria es relevante para el objetivo del análisis. Será necesario considerar técnicas de manejo de desbalance de clases durante el proceso de modelado para abordar este problema y garantizar que los modelos sean representativos y generalizables.

Admission Grade: Notas de admisión de los estudiantes del DataSet

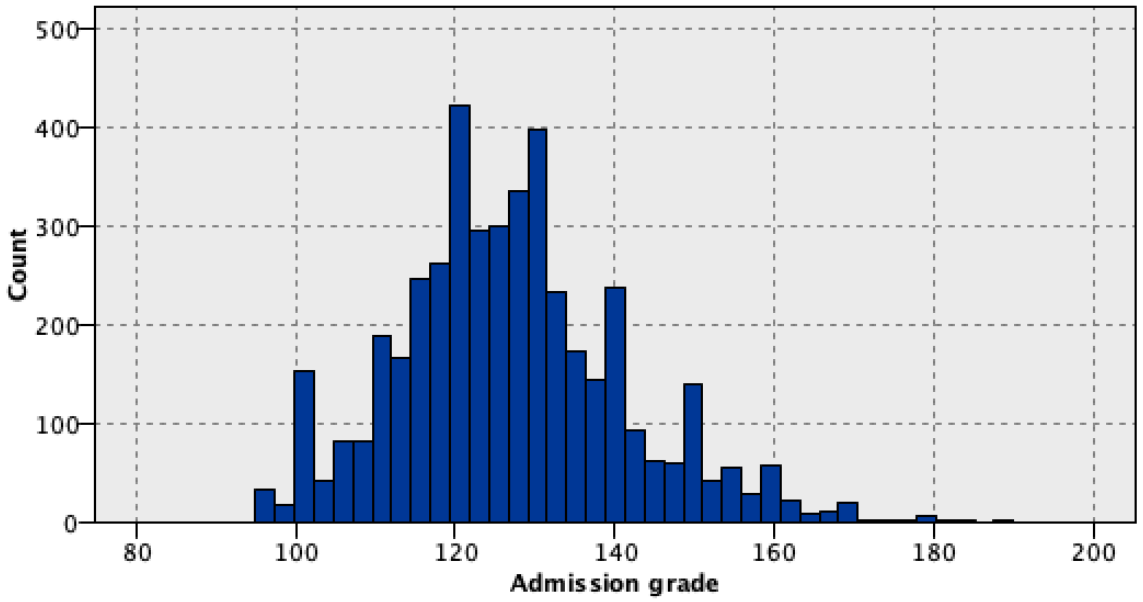


Ilustración 5.- La tercera gráfica muestra la distribución de las notas de admisión de los estudiantes en el conjunto de datos. Las notas de admisión están representadas en una escala de 0 a 200.

Observamos que la distribución de las notas de admisión sigue una forma de campana de Gauss, también conocida como distribución normal. Esto implica que la mayoría de los estudiantes tienen notas de admisión cercanas a un valor central, mientras que las notas extremadamente altas o bajas son menos comunes.

En particular, notamos que hay una mayor concentración de casos en el rango de notas entre 80 y 140. Esto sugiere que la mayoría de los estudiantes obtuvieron notas de admisión dentro de este rango. A medida que nos alejamos de este rango central, la cantidad de casos disminuye gradualmente, lo que indica que las notas extremadamente altas o bajas son menos frecuentes.

Esta distribución de notas de admisión proporciona información importante sobre el perfil académico de los estudiantes en el conjunto de datos. Además, esta distribución puede ser útil para identificar posibles patrones o relaciones entre las notas de admisión y otras variables en el análisis de datos y la construcción de modelos predictivos.

Previous qualification(grade): Calificaciones previas a la toma de datos de los estudiantes del DataSet

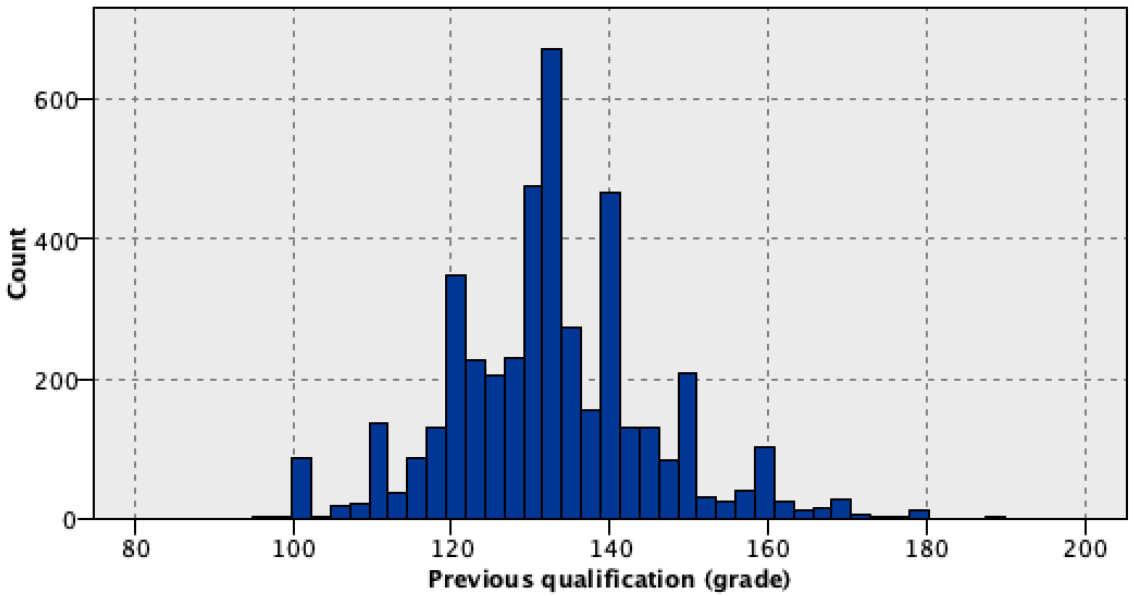


Ilustración 6.- La cuarta gráfica representa la distribución de las calificaciones obtenidas en la calificación previa a la toma de datos por parte de los estudiantes en el conjunto de datos. Estas calificaciones anteriores también están en una escala de 0 a 200.

Al igual que en la gráfica anterior que mostraba las notas de admisión, observamos que la distribución de las calificaciones de cualificaciones anteriores sigue una forma de campana de Gauss. Sin embargo, en este caso, la campana se centra alrededor de un rango ligeramente diferente de valores, aproximadamente entre 100 y 140.

Esto sugiere que la mayoría de los estudiantes tienen calificaciones anteriores dentro de este rango específico. A medida que nos alejamos de este rango central, la cantidad de casos disminuye gradualmente, lo que indica que las calificaciones extremadamente altas o bajas en cualificaciones anteriores son menos comunes.

Esta distribución de calificaciones de cualificaciones anteriores proporciona información valiosa sobre el rendimiento académico previo de los estudiantes en el conjunto de datos. Al igual que con la distribución de notas de admisión, esta distribución puede ser útil para identificar tendencias y relaciones en el análisis de datos y en la construcción de modelos predictivos.

Fase 2
Comprensión de los datos

2.1.- Recopilación de los datos

Datos existentes:

- No se cuenta con datos previos al inicio del proyecto.

Datos adquiridos:

- El dataset fue obtenido del siguiente sitio: https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

  Las fuentes de datos utilizadas consisten en datos internos y externos de la institución, así como de *Academic Management System* (AMS), la institución de *Support System for the Teaching Activity* (denominada internamente como PAE), la *General Directorate of Higher Education* (DGES) y por la *Contemporary Portugal Database* (PORDATA).

Datos adicionales:

De momento la cantidad de datos y atributos parecen ser los suficientes para el desarrollo del proyecto, sin embargo se buscaron datasets con estructuras similares para aumentar el volumen de datos sin resultados positivos.

2.2.- Descripción de los datos

Como ya se ha mencionado anteriormente el dataset cuenta con 4,424 registros y 35 atributos, de los cuales 34 son datos de tipo entero o continuo y el restante es un dato de tipo categórico.

	A	B	C	D	E	F	G	H	I	J	
1	Marital status	Application mode	Application order	Course	Daytime/evening attendance	Previous qualification	Previous qualification (grade)	Nacionality	Mother's qualification	Father's qualification	Mother'
2	1	17	5	171	1	1	122.0	1	19	12	5
3	1	15	1	9254	1	1	160.0	1	1	3	3
4	1	1	5	9070	1	1	122.0	1	37	37	9
5	1	17	2	9773	1	1	122.0	1	38	37	5
6	2	39	1	8014	0	1	100.0	1	37	38	9
7	2	39	1	9991	0	19	133.1	1	37	37	9
8	1	1	1	9500	1	1	142.0	1	19	38	7
9	1	18	4	9254	1	1	119.0	1	37	37	9
10	1	1	3	9238	1	1	137.0	62	1	1	9
11	1	1	1	9238	1	1	138.0	1	1	19	4
12	1	1	1	9670	1	1	139.0	1	38	19	5
13	1	1	1	9500	1	1	136.0	1	19	38	9
14	1	1	2	9853	1	1	133.0	1	19	37	4
15	1	53	1	9254	1	42	110.0	1	1	1	4
16	1	1	1	9085	1	1	149.0	1	38	37	5
17	1	1	1	9773	1	1	127.0	1	19	37	9
18	1	18	1	9238	1	1	137.0	1	19	38	5
19	1	17	2	9500	1	1	135.0	1	19	1	5
20	1	1	1	9130	1	1	137.0	1	3	19	3
21	1	1	1	9853	1	1	140.0	1	19	19	7
22	1	1	3	171	1	1	122.0	1	1	1	9
23	1	18	4	9556	1	1	127.0	1	1	38	4
24	1	1	4	9500	1	1	142.0	1	19	19	1
25	1	1	4	9670	1	1	125.0	1	1	38	4
26	1	1	1	9500	1	1	126.0	1	19	19	3
27	1	1	1	9238	1	1	151.0	1	19	38	9
28	1	17	2	9238	1	1	115.0	1	19	38	7
29	1	1	1	9085	1	1	138.0	1	19	19	3
30	1	1	1	9147	1	1	135.0	1	19	1	9
31	1	17	2	9500	1	1	127.0	1	3	3	1

Ilustración 7.- Representación gráfica de la base de datos mediante el uso de Excel.

2.3 Descripción de los atributos

A continuación, se presenta una tabla que describe las variables presentes en el conjunto de datos, junto con sus posibles valores:

Atributo	Tipo	Descripción
Marital Status	Categórico	1 – single 2 – married 3 – widower 4 – divorced 5 – facto union 6 – legally separated
Application mode	Categórico	1 - 1st phase - general contingent 2 - Ordinance No. 612/93 5 - 1st phase - special contingent (Azores Island) 7 - Holders of other higher courses 10 - Ordinance No. 854-B/99 15 - International student (bachelor) 16 - 1st phase - special contingent (Madeira Island) 17 - 2nd phase - general contingent 18 - 3rd phase - general contingent 26 - Ordinance No. 533-A/99, item b2) (Different Plan) 27 - Ordinance No. 533-A/99, item b3 (Other Institution) 39 - Over 23 years old 42 - Transfer 43 - Change of course 44 - Technological specialization diploma holders 51 - Change of institution/course 53 - Short cycle diploma holders 57 - Change of institution/course (International)
Application order	Nominal	Application order (between 0 - first choice; and 9 last choice)
Course	Categórico	33 - Biofuel Production Technologies 171 - Animation and Multimedia Design 8014 - Social Service (evening attendance) 9003 - Agronomy 9070 - Communication Design 9085 - Veterinary Nursing 9119 - Informatics Engineering 9130 - Equinculture 9147 - Management 9238 - Social Service 9254 - Tourism 9500 - Nursing 9556 - Oral Hygiene 9670 - Advertising and Marketing Management 9773 - Journalism and Communication 9853 - Basic Education 9991 - Management (evening attendance)
Daytime/evening attendance	Categórico	1 – daytime 0 - evening
Previous qualification	Categórico	1 - Secondary education 2 - Higher education - bachelor's degree 3 - Higher education - degree … 42 - Professional higher technical course 43 - Higher education - master (2nd cycle)
Previous qualification (grade)	Ordinal	Grade of previous qualification (between 0 and 200)
Nacionality	Categórico	1 - Portuguese; 2 - German; 6 - Spanish; 11 - Italian; 13 - Dutch; 14 - English; 17 - Lithuanian; 21 - Angolan; 22 - Cape Verdean; 24 - Guinean; 25 - Mozambican; 26 - Santomean; 32 - Turkish; 41 - Brazilian; 62 - Romanian; 100 - Moldova (Republic of); 101 - Mexican; 103 - Ukrainian; 105 - Russian; 108 - Cuban; 109 - Colombian
Mother's qualification	Ordinal	1 - Secondary Education - 12th Year of Schooling or Eq. 2 - Higher Education - Bachelor's Degree … 43 - Higher Education - Master (2nd cycle) 44 - Higher Education - Doctorate (3rd cycle)
Father's qualification	Ordinal	1 - Secondary Education - 12th Year of Schooling or Eq. 2 - Higher Education - Bachelor's Degree ... 43 - Higher Education - Master (2nd cycle) 44 - Higher Education - Doctorate (3rd cycle)
Mother's occupation	Categórico	0 - Student 1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers 2 - Specialists in Intellectual and Scientific Activities … 193 - Unskilled workers in extractive industry, construction, manufacturing and transport 194 - Meal preparation assistants
Father's occupation	Categórico	0 - Student 1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers 2 - Specialists in Intellectual and Scientific Activities … 193 - Unskilled workers in extractive industry, construction, manufacturing and transport 194 - Meal preparation assistants 195 - Street vendors (except food) and street service providers
Admission grade	Ordinal	Admission grade (between 0 and 200)
Displaced	Categórico	1 – yes 0 – no
Educational special needs	Categórico	1 – yes 0 – no
Debtor	Categórico	1 – yes 0 – no
Tuition fees up to date	Categórico	1 – yes 0 – no

Gender	Categórico	1 – male 0 – female
Scholarship holder	Categórico	1 – yes 0 – no
Age at enrollment	Continuo	Age of studend at enrollment
International	Categórico	1 – yes 0 – no
Curricular units 1st sem (credited)	Continuo	Number of curricular units credited in the 1st semester
Curricular units 1st sem (enrolled)	Continuo	Number of curricular units enrolled in the 1st semester
Curricular units 1st sem (evaluations)	Continuo	Number of evaluations to curricular units in the 1st semester
Curricular units 1st sem (approved)	Continuo	Number of curricular units approved in the 1st semester
Curricular units 1st sem (grade)	Continuo	Grade average in the 1st semester (between 0 and 20)
Curricular units 1st sem (without evaluations)	Continuo	Number of curricular units without evalutions in the 1st semester
Curricular units 2nd sem (credited)	Continuo	Number of curricular units credited in the 2nd semester
Curricular units 2nd sem (enrolled)	Continuo	Number of curricular units enrolled in the 2nd semester
Curricular units 2nd sem (evaluations)	Continuo	Number of evaluations to curricular units in the 2nd semester
Curricular units 2nd sem (approved)	Continuo	Number of curricular units approved in the 2nd semester
Curricular units 2nd sem (grade)	Continuo	Grade average in the 2nd semester (between 0 and 20)
Curricular units 2nd sem (without evaluations)	Continuo	Number of curricular units without evalutions in the 1st semester
Unemployment rate	Continuo	Unemployment rate (%)
Inflation rate	Continuo	Inflation rate (%)
GDP	Continuo	GDP
Target	Categórico	Target. The problem is formulated as a three category classification task (dropout, enrolled, and graduate) at the end of the normal duration of the course.

2.4 Exploración de los datos (exhaustiva)

Con la ayuda del lenguaje de programación R nos enfocamos en un atributo en particular para está exhaustiva exploración de los datos, en este caso clasificamos en categorías las cuales son "*Dropout*", "*Enrolled*" y "*Graduate*".

La clase mayoritaria, "*Graduate*", representa el 50% de los registros (2209 de 4424) la clase "*Dropout*" representa el 32% del total de registros (1421 de 4424), mientras que la clase minoritaria, "*Enrolled*", representa el 18% del total de registros (794 de 4424). Esto podría dar como resultado una alta precisión de predicción impulsada por la clase mayoritaria a expensas de un pobre desempeño de la clase minoritaria.

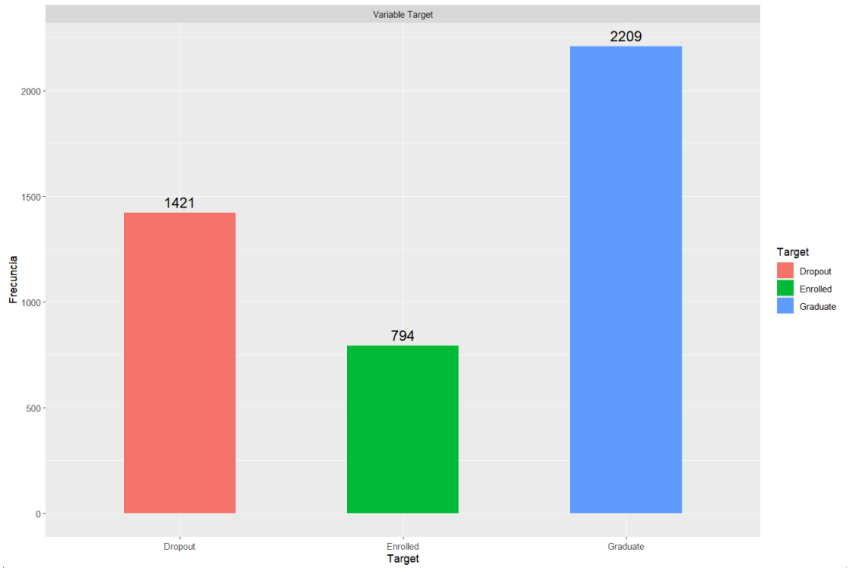


Ilustración 7.- Distribución de expedientes estudiantiles entre las tres categorías consideradas para el éxito académico.

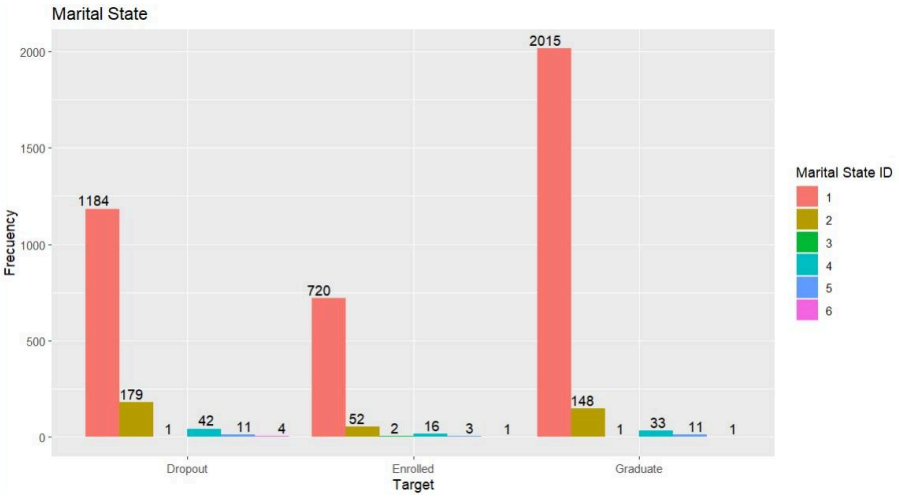


Ilustración 8.- Dispersión de individuos registrados en la base de datos según su estado civil. Distribución de expedientes estudiantiles entre las tres categorías consideradas para el éxito académico.

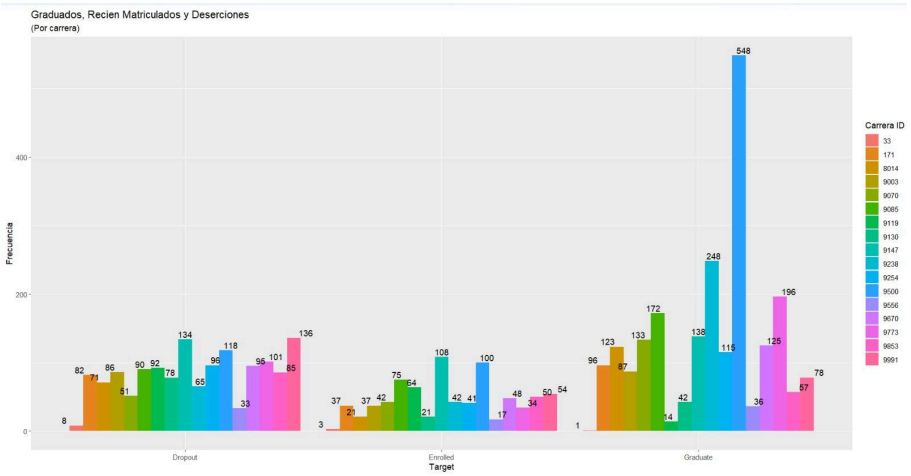


Ilustración 9.- Dispersión del curso inscrito de los individuos registrados en el DataSet.

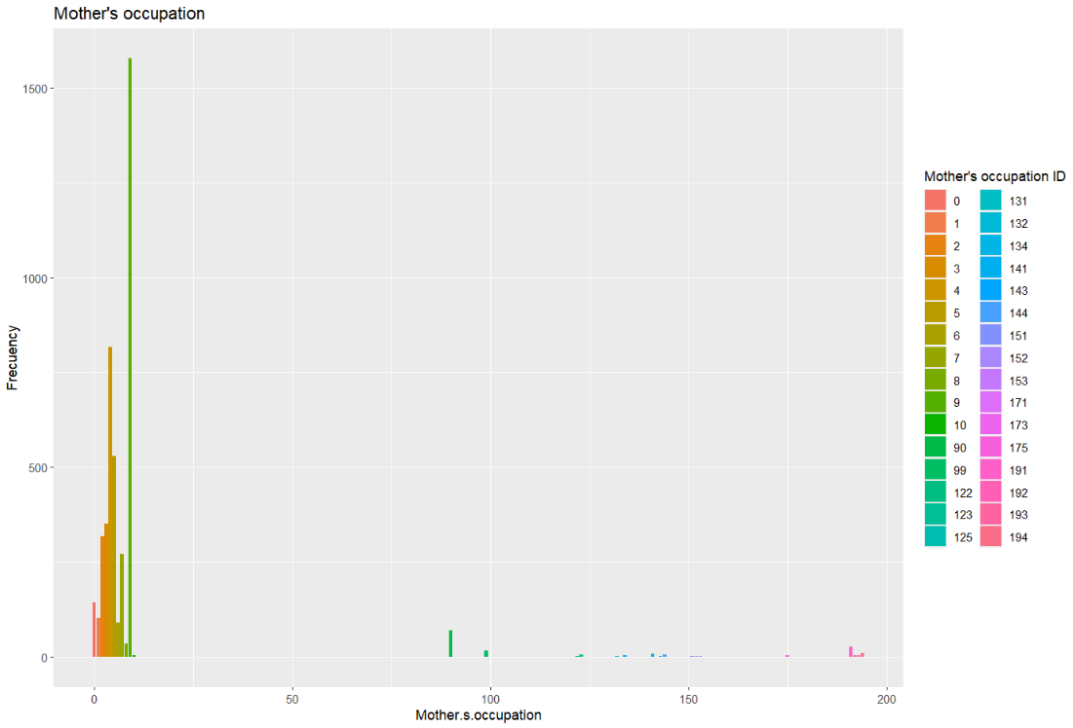


Ilustración 10.- Distribución de la ocupación de las Madres registradas en el DataSet.

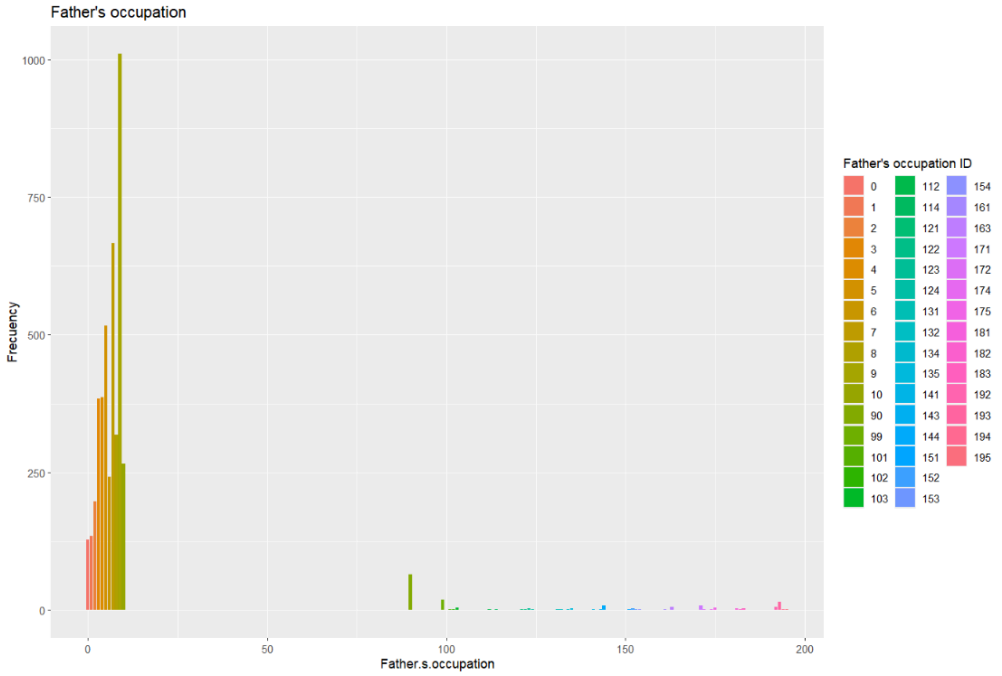


Ilustración 11.- Distribución de la ocupación de los Padres registradas en el DataSet.

Al realizar la primera exploración de los datos se observa que el campo “Target” es el de mayor relevancia por la naturaleza del proyecto, por lo que al analizar el valor del campo con respecto a los demás se ha descubierto un error en los datos, ya que si el campo se encuentra con el valor “Graduate” los campos “‘Curricular units 1st sem (grade)” y “Curricular units 2nd sem (grade)” no deberían tener valores iguales a cero.

Curricular units 1st sem (grade)	Curricular units 2nd sem (grade)	Target
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate
0.000	12.286	Graduate
0.000	0.000	Graduate
0.000	0.000	Graduate

Ilustración 12.- Visualización de los errores encontrados en el DataSet.

La cantidad de datos erróneos son 77 que representan un 1.7 % del total de los datos entonces se llegó al acuerdo que deben ser eliminados de nuestro dataset para evitar fallos en los objetivos del proyecto, una vez detectada la anomalía se eliminarán en una fase posterior del modelo de trabajo (Fase 3. Limpieza de los datos).

2.5.- Verificación de la calidad de los datos

Características	Notas
Datos perdidos o vacíos	El DataSet no se encuentran valores nulos o vacíos, se validó con un Script de Python para evaluar cada uno de los registros.
Errores en los datos	Algunos atributos cambiaron del tipo de dato de entero a real, por lo cual se tuvo que hacer la conversión de acorde a la especificación del dataset. Se eliminará un conjunto de datos que presentan errores y sus valores no son adecuados para el proyecto.
Errores de medición	No se han encontrado errores de medición en los datos.
Errores de codificación	La manipulación del tipo de archivo fuente (.csv) presentaba algunos errores de lectura que imposibilitaba la interpretación de los datos. Esto ya se encuentra corregido con una modificación directa.
Atributos redundantes o de escasa utilidad	Se detectaron algunas columnas que no son relevantes para la resolución de la problemática.

2.6.- 5V en la Minería de Datos

- **1.- Volumen:**

El proyecto maneja un conjunto de datos sustancial, compuesto por 4424 registros de estudiantes con 35 atributos cada uno. Este volumen significativo proporciona una base robusta para la construcción y evaluación de modelos predictivos.

- **2.- Velocidad:**

La velocidad de adquisición y procesamiento de datos es esencial para ofrecer predicciones en tiempo real. La implementación de modelos en una herramienta de Análisis de Aprendizaje permitirá una respuesta rápida al identificar el riesgo de abandono y proporcionar información al equipo de tutoría de manera oportuna.

- **3.- Variedad:**

La diversidad de datos es evidente en la inclusión de información demográfica, macroeconómica, socioeconómica y académica. La variedad de atributos en el conjunto de datos enriquece el análisis, permitiendo una comprensión más completa de los factores que contribuyen al abandono estudiantil.

- **4. Veracidad:**

La precisión y confiabilidad de los datos son cruciales. La metodología empleada en la preparación de datos abordó problemas potenciales de integridad y aseguró que la información utilizada en la construcción de modelos sea precisa y veraz.

- **5. Valor:**

La esencia misma de la minería de datos es extraer valor significativo de los datos. En este proyecto, el valor se traduce en la capacidad de predecir el abandono estudiantil y mejorar el éxito académico. La implementación de modelos en una herramienta práctica agrega valor al proporcionar al equipo de tutoría información útil y accesible.

Fase 3

Preparación de los datos

3.1.- Selección de datos

Se considera que los 37 atributos de la tabla son importantes para el análisis por lo que se mantendrán. Utilizaremos los 4424 registros ya que se consideran suficientes para los objetivos.

3.2.- Limpieza de los datos

Realizando un análisis de datos haciendo uso de la herramienta auditoría de datos podemos observar de manera preliminar que no se cuenta con valores nulos o perdidos.

Campos completos (%): 100 % Registros completos (%): 100 %

Campo	Medida	Valores atípicos	Extremos	Acción	Imputar perdidos	Método	% Completo	Registros válidos	Valor nulo	Cadena vacía	Espacio en blanco	Valor en blanco
Marital status	Nominal	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Application ...	Nominal	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Application o...	Nominal	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Course	Nominal	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Daytime/eve...	Nominal	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Previous qual...	Ordinal	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Previous qual...	Ordinal	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Nacionality	Nominal	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Mother's qua...	Ordinal	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Father's quali...	Ordinal	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Mother's occ...	Nominal	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Father's occu...	Nominal	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Admission gr...	Nominal	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Displaced	Categórico	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Educational s...	Categórico	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Debtor	Categórico	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Tuition fees ...	Categórico	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Gender	Categórico	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Scholarship h...	Categórico	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Age at enroll...	Nominal	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
International	Nominal	--	--	--	Nunca	Fijo	100	4424	0	0	0	0
Curricular uni...	Continuo	102	47	Ninguno	Nunca	Fijo	100	4424	0	0	0	0
Curricular uni...	Continuo	95	11	Ninguno	Nunca	Fijo	100	4424	0	0	0	0
Curricular uni...	Continuo	54	6	Ninguno	Nunca	Fijo	100	4424	0	0	0	0
Curricular uni...	Continuo	56	5	Ninguno	Nunca	Fijo	100	4424	0	0	0	0
Curricular uni...	Continuo	0	0	Ninguno	Nunca	Fijo	100	4424	0	0	0	0
Curricular uni...	Continuo	23	39	Ninguno	Nunca	Fijo	100	4424	0	0	0	0
Curricular uni...	Continuo	56	54	Ninguno	Nunca	Fijo	100	4424	0	0	0	0
Curricular uni...	Continuo	74	8	Ninguno	Nunca	Fijo	100	4424	0	0	0	0
Curricular uni...	Continuo	41	2	Ninguno	Nunca	Fijo	100	4424	0	0	0	0
Curricular uni...	Continuo	21	2	Ninguno	Nunca	Fijo	100	4424	0	0	0	0
Curricular uni...	Continuo	0	0	Ninguno	Nunca	Fijo	100	4424	0	0	0	0
Curricular uni...	Continuo	35	59	Ninguno	Nunca	Fijo	100	4424	0	0	0	0
Unemployme...	Continuo	0	0	Ninguno	Nunca	Fijo	100	4424	0	0	0	0
Inflation rate	Continuo	0	0	Ninguno	Nunca	Fijo	100	4424	0	0	0	0
GDP	Continuo	0	0	Ninguno	Nunca	Fijo	100	4424	0	0	0	0
Target	Categórico	--	--	--	Nunca	Fijo	100	4424	0	0	0	0

Ilustración 13.- Validación de los datos nulos del DataSet.

Sin embargo el hecho de no tener valores nulos o perdidos no excluye a errores en los datos ya que en la fase 2 de exploración de los datos se evidenciaba un conjunto de datos que era necesario eliminar, con lo cual utilizamos la herramienta de filtrado de datos para quitarlos de nuestro dataset.

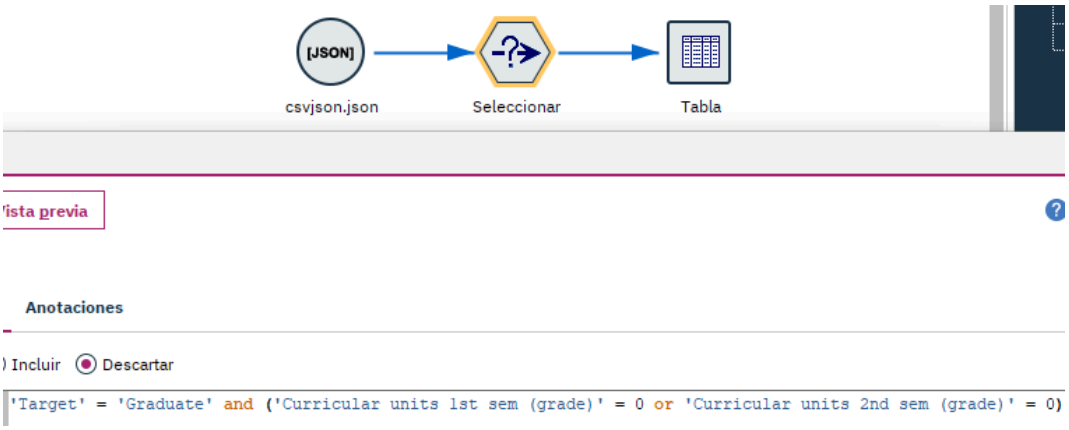


Ilustración 14.- Proceso para eliminar los registros que causan defectos en el DataSet.

Target
Dropout
Graduate
Dropout
Graduate
Graduate
Graduate
Graduate
Dropout
Graduate
Dropout
Graduate

Se realizó un cambio en el tipo dato del atributo Target ya que era un tipo de dato categórico en cadenas de caracteres.

La ruta para realizar el cambio fue rellenar los datos “Graduated”, “Enrolled” y “Dropout” de la siguiente manera:

Condición: Target = "Dropout” **Reemplazar por:** “0”
Condición: Target = " Graduate” **Reemplazar por:** “1”
Condición: Target = "Enrolled” **Reemplazar por:** “2”

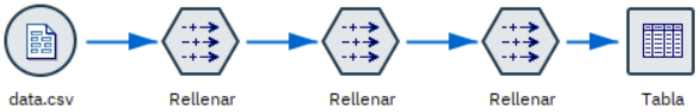


Ilustración 15.- Proceso para convertir Target en un valor Numérico

Target
0
1
0
1
1
1
1
0
1
0
1
1
0
1
1
0
2

En nuestro campo objetivo donde se indica el estado de cada individuo del dataset podemos observar un ligero desbalance entre las tres categorías. Ya que existe una ligera dominancia por parte de los datos “Dropout” y “Graduate” respecto al campo restante “Enrolled”.

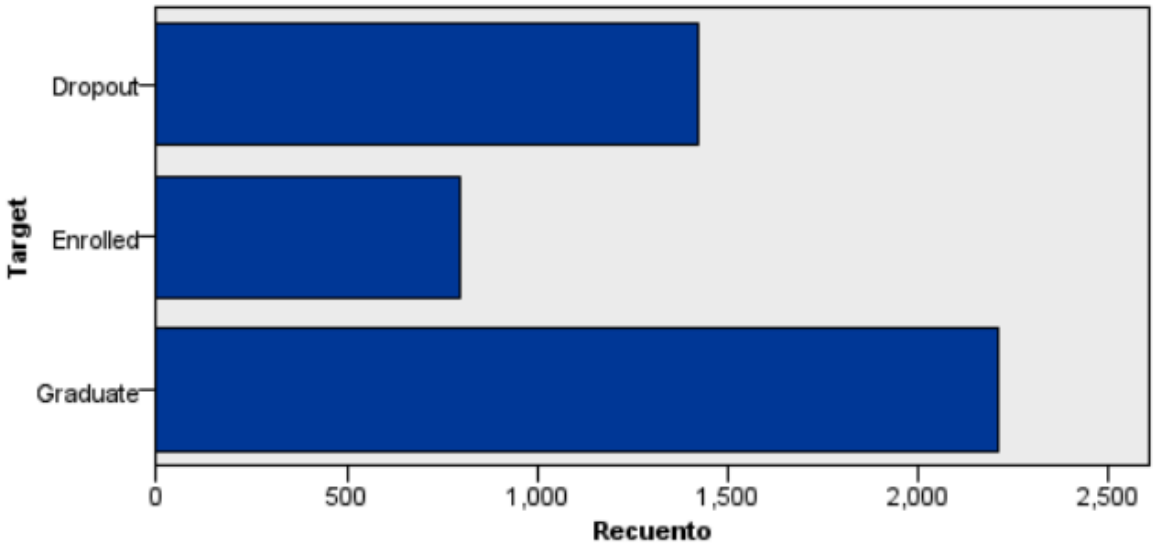


Ilustración 16.- Distribución de el Target en el DataSet.

Se hizo uso de la herramienta de balanceo que SPSS Modeler nos ofrece, la elección que se tomó fue hacer una reducción manual de los datos con mayor cantidad de registros para llevar a cabo el balanceo.

Ilustración 17.- Instrucciones de Nodos para equilibrar los registros del DataSet.

Una vez que se ejecuta la herramienta podemos observar una reducción de los datos acompañado de una gráfica aunque permite observar el correcto balanceo.

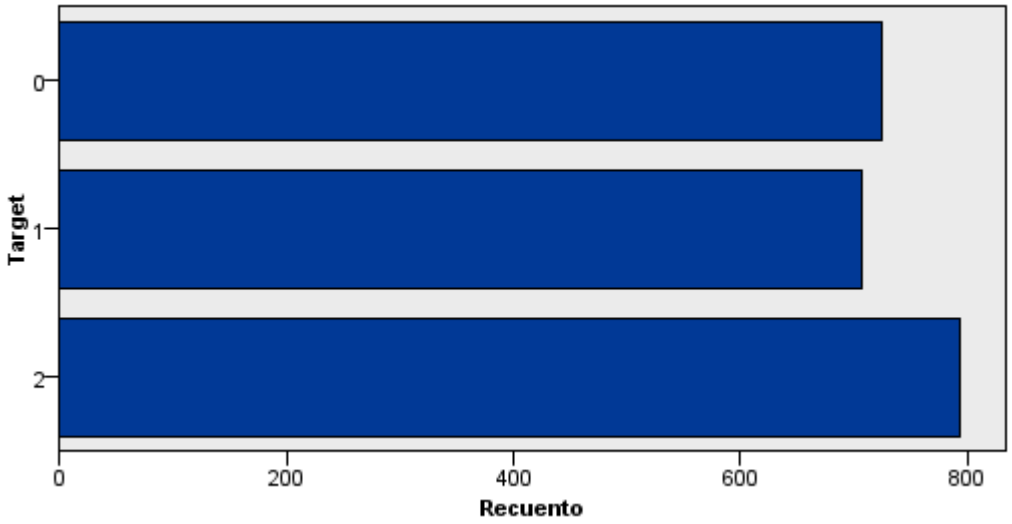


Ilustración 18.- Nuevos datos equilibrados para mayor precisión en el análisis de los datos.

Comunalidades

	Inicial	Extracción
Marital status	1,000	,351
Application mode	1,000	,515
Application order	1,000	,245
Course	1,000	,204
Daytime/evening attendance	1,000	,382
Previous qualification	1,000	,150
Previous qualification (grade)	1,000	,054
Nacionality	1,000	,701
Mother's qualification	1,000	,361
Father's qualification	1,000	,331
Mother's occupation	1,000	,836
Father's occupation	1,000	,833
Displaced	1,000	,320
Educational special needs	1,000	,004
Debtor	1,000	,171
Tuition fees up to date	1,000	,236
Gender	1,000	,151
Scholarship holder	1,000	,176
Age at enrollment	1,000	,675
International	1,000	,702
Curricular units 1st sem (credited)	1,000	,726
Curricular units 1st sem (enrolled)	1,000	,859
Curricular units 1st sem (evaluations)	1,000	,635
Curricular units 1st sem (approved)	1,000	,864
Curricular units 1st sem (grade)	1,000	,701
Curricular units 1st sem (without evaluations)	1,000	,356
Curricular units 2nd sem (credited)	1,000	,718
Curricular units 2nd sem (enrolled)	1,000	,791
Curricular units 2nd sem (evaluations)	1,000	,561
Curricular units 2nd sem (approved)	1,000	,827

Una vez que realizamos el balance de los datos llevamos a cabo un análisis de componentes principales (PCA/Factorial).

Donde podemos observar los valores que tienen mayor variabilidad y son aquellos que hay que tener en cuenta, mientras el valor sea más cercano a uno más importante es.

3.3.- Construcción de nuevos datos

Se decidió no crear ningún dato nuevo ya que al ser un dataset amplio consideramos que tenemos la información suficiente.

3.4.- Integración de datos

No es necesario una integración de datos ya que no contamos con otros conjuntos de datos o datos nuevos.

3.5.- Formato de datos

Todos los datos ya se encuentran en el formato requerido. Solo fue necesario verificar que los tipos de datos fueran los adecuados para el análisis, los cuales fueron previamente definidos en la fase de comprensión de datos.

Campo	Medida	Valores	No se encuentra	Comprobar	
Marital status	Nominal			Ninguno	Entrada
Application mode	Nominal			Ninguno	Entrada
Application order	Nominal			Ninguno	Entrada
Course	Nominal			Ninguno	Entrada
Daytime/evening attendance	Nominal			Ninguno	Entrada
Previous qualification	Ordinal			Ninguno	Entrada
Previous qualification (grade)	Ordinal			Ninguno	Entrada
Nacionality	Nominal			Ninguno	Entrada
Mother's qualification	Ordinal			Ninguno	Entrada
Father's qualification	Ordinal			Ninguno	Entrada
Mother's occupation	Nominal			Ninguno	Entrada
Father's occupation	Nominal			Ninguno	Entrada
Admission grade	Ordinal			Ninguno	Entrada
Displaced	Categórico			Ninguno	Entrada
Educational special needs	Categórico			Ninguno	Entrada
Debtor	Categórico			Ninguno	Entrada
Tuition fees up to date	Categórico			Ninguno	Entrada
Gender	Categórico			Ninguno	Entrada
Scholarship holder	Categórico			Ninguno	Entrada
Age at enrollment	Nominal			Ninguno	Entrada
International	Nominal			Ninguno	Entrada
Curricular units 1st sem (credited)	Continuo			Ninguno	Entrada
Curricular units 1st sem (enrolled)	Continuo			Ninguno	Entrada
Curricular units 1st sem (evaluations)	Continuo			Ninguno	Entrada
Curricular units 1st sem (approved)	Continuo			Ninguno	Entrada
Curricular units 1st sem (grade)	Continuo			Ninguno	Entrada
Curricular units 1st sem (without evaluations)	Continuo			Ninguno	Entrada
Curricular units 2nd sem (credited)	Continuo			Ninguno	Entrada
Curricular units 2nd sem (enrolled)	Continuo			Ninguno	Entrada
Curricular units 2nd sem (evaluations)	Continuo			Ninguno	Entrada
Curricular units 2nd sem (approved)	Continuo			Ninguno	Entrada
Curricular units 2nd sem (grade)	Continuo			Ninguno	Entrada
Curricular units 2nd sem (without evaluations)	Continuo			Ninguno	Entrada
Unemployment rate	Continuo			Ninguno	Entrada
Inflation rate	Continuo			Ninguno	Entrada
GDP	Continuo			Ninguno	Entrada
Target	Categórico			Ninguno	Entrada

Ilustración 19.- Validación de los tipos de datos en el DataSet Update.

Fase 4
Modelado

4.1.- Objetivo principal

4.1.1.- Utilizando el Software IBM SPSS Modeler

4.1.1.1.- Selección de características con PCA

Comunalidades		
	Inicial	Extracción
Marital status	1,000	,337
Application mode	1,000	,519
Application order	1,000	,294
Course	1,000	,123
Daytime/evening attendance	1,000	,365
Previous qualification	1,000	,239
Previous qualification (grade)	1,000	,348
Nacionality	1,000	,472
Mothers qualification	1,000	,378
Fathers qualification	1,000	,343
Mothers occupation	1,000	,778
Fathers occupation	1,000	,771
Admission grade	1,000	,383
Displaced	1,000	,322
Educational special needs	1,000	,004
Debtor	1,000	,176
Tuition fees up to date	1,000	,261
Gender	1,000	,146
Scholarship holder	1,000	,185
Age at enrollment	1,000	,657
International	1,000	,480
Curricular units 1st sem (credited)	1,000	,762
Curricular units 1st sem (enrolled)	1,000	,859
Curricular units 1st sem (evaluations)	1,000	,613
Curricular units 1st sem (approved)	1,000	,873
Curricular units 1st sem (grade)	1,000	,680
Curricular units 1st sem (without evaluations)	1,000	,274
Curricular units 2nd sem (credited)	1,000	,754
Curricular units 2nd sem (enrolled)	1,000	,785
Curricular units 2nd sem (evaluations)	1,000	,527
Curricular units 2nd sem (approved)	1,000	,834
Curricular units 2nd sem (grade)	1,000	,716
Curricular units 2nd sem (without evaluations)	1,000	,247
Unemployment rate	1,000	,148
Inflation rate	1,000	,012
GDP	1,000	,154

Método de extracción: análisis de componentes principales.

En esta sección, se realizó un análisis de selección de características utilizando la técnica de **Análisis de Componentes Principales (PCA) en IBM SPSS Modeler**. El **PCA** es una técnica de reducción de dimensionalidad que se utiliza para identificar las características más importantes en un conjunto de datos al transformar las variables originales en un conjunto de variables no correlacionadas llamadas componentes principales.

Pasos del Proceso:

Preparación de los Datos:
Se cargó el conjunto de datos en **IBM SPSS Modeler** y se realizó un preprocesamiento inicial para tratar valores faltantes, valores atípicos y otras transformaciones necesarias.

Configuración del Nodo PCA:
Se agregó y configuró un nodo **PCA** en el flujo de trabajo de **IBM SPSS Modeler**. En este nodo, se especificaron las opciones de configuración, como el número de componentes principales a retener y cualquier otra configuración relacionada con el análisis.

Ejecución del Análisis:
Se ejecutó el nodo **PCA** para realizar el análisis de selección de características en el conjunto de datos. Durante este proceso, las variables originales fueron transformadas en componentes principales, y se identificaron aquellos componentes que explicaban la mayor variabilidad en los datos.

Selección de Características:
Se examinaron los resultados del análisis de PCA para identificar los componentes principales que contribuyen significativamente a la variabilidad en los datos. Se seleccionaron los componentes más importantes para su uso en análisis posteriores, con el objetivo de reducir la dimensionalidad del conjunto de datos mientras se retiene la información más relevante.

Identificación de Comunalidades Altas y Bajas:

En el análisis de componentes principales **(PCA)**, las comunalidades proporcionan información sobre la relación entre las variables observadas y los factores latentes en el modelo. Las comunalidades representan la proporción de la varianza de una variable que es explicada por los factores en el modelo. Identificar comunalidades altas y bajas es fundamental para comprender qué variables están más estrechamente relacionadas con la estructura latente subyacente de los datos y cuáles pueden no ser relevantes para el análisis.

Comunalidades Altas:
Las variables con comunalidades más altas están más estrechamente relacionadas con los factores en el modelo y tienen una mayor importancia en la estructura latente de los datos. Estas variables explican una proporción significativa de la varianza total en los datos y son importantes para el análisis. Identificar y priorizar estas variables puede ayudar a enfocar el análisis en las características más relevantes.

Descarte de Variables con Comunalidades Bajas:
Las variables con comunalidades bajas, especialmente aquellas cercanas a 0, indican que no están bien representadas por los factores en el modelo. Esto sugiere que estas variables pueden no ser relevantes o contribuir significativamente a la estructura subyacente de los datos. En consecuencia, es posible considerar descartar estas variables si no son importantes para los objetivos de investigación o si afectan negativamente al rendimiento del modelo.

Selección de las características:
Estas son las variables que tienen comunalidades mayores a 0.200, lo que sugiere que están más estrechamente relacionadas con los factores en el modelo y tienen una mayor importancia en la estructura latente de los datos. Las demás variables pueden considerarse menos relevantes en este contexto y podrían ser descartadas del análisis.

Variable	Inicial	Extracción
Marital status	1.000	0.337
Application mode	1.000	0.519
Application order	1.000	0.294
Daytime/evening attendance	1.000	0.365
Previous qualification	1.000	0.239
Previous qualification (grade)	1.000	0.348
Nacionality	1.000	0.472
Mothers qualification	1.000	0.378
Fathers qualification	1.000	0.343
Mothers occupation	1.000	0.778
Fathers occupation	1.000	0.771
Admission grade	1.000	0.383
Displaced	1.000	0.322
Tuition fees up to date	1.000	0.261
Age at enrollment	1.000	0.657
International	1.000	0.480
Curricular units 1st sem (credited)	1.000	0.762
Curricular units 1st sem (enrolled)	1.000	0.859
Curricular units 1st sem (evaluations)	1.000	0.613
Curricular units 1st sem (approved)	1.000	0.873
Curricular units 1st sem (grade)	1.000	0.680
Curricular units 1st sem (without evaluations)	1.000	0.274
Curricular units 2nd sem (credited)	1.000	0.754
Curricular units 2nd sem (enrolled)	1.000	0.785
Curricular units 2nd sem (evaluations)	1.000	0.527
Curricular units 2nd sem (approved)	1.000	0.834
Curricular units 2nd sem (grade)	1.000	0.716
Curricular units 2nd sem (without evaluations)	1.000	0.247

4.1.1.1.1.- Balanceo del DataSet

El balanceo del conjunto de datos es un paso crucial en el análisis de datos cuando se enfrenta a conjuntos de datos desequilibrados, es decir, conjuntos de datos en los que las clases objetivo están representadas de manera desproporcionada. El desequilibrio puede tener un impacto significativo en el rendimiento de los modelos de aprendizaje automático, ya que pueden sesgar el modelo hacia las clases mayoritarias y hacer que sea menos preciso para predecir las clases minoritarias.
El nodo de equilibrado en IBM SPSS Modeler proporciona una solución eficaz para abordar este problema al aplicar técnicas de submuestreo o sobremuestreo para igualar la distribución de clases en el conjunto de datos.

Pasos para Balancear el DataSet con el Nodo de Equilibrar:

Configuración del Nodo de Equilibrar:
Configura el nodo de equilibrado en IBM SPSS Modeler, especificando la variable objetivo y seleccionando el método de equilibrado deseado (submuestreo o sobremuestreo).

Ejecución del Nodo:
Ejecuta el nodo para aplicar la técnica de balanceo seleccionada al conjunto de datos. Durante este paso, el nodo ajustará la distribución de clases para mejorar la representación de las clases minoritarias.

Evaluación del Conjunto de Datos Equilibrado:
Después de balancear el conjunto de datos, es fundamental evaluar el impacto del equilibrado en el rendimiento del modelo. Esto puede implicar la comparación de métricas de rendimiento antes y después del equilibrado, como precisión, sensibilidad, especificidad, etc.

Beneficios del Balanceo del DataSet:

- Mejora la precisión del modelo al reducir el sesgo hacia las clases mayoritarias.
- Ayuda a prevenir la infra-representación de clases minoritarias, lo que podría resultar en decisiones erróneas.
- Proporciona una representación más equitativa de todas las clases en el conjunto de datos, lo que puede mejorar la robustez y generalización del modelo.

Resultados del balanceo:

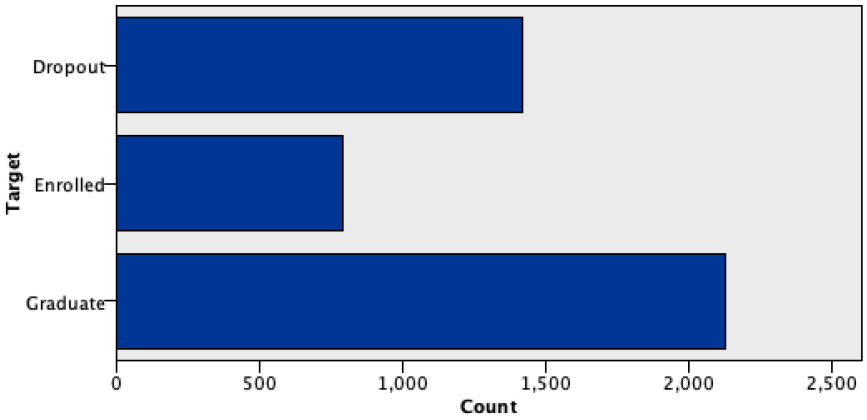


Ilustración 19.- DataSet actual sin balanceo.

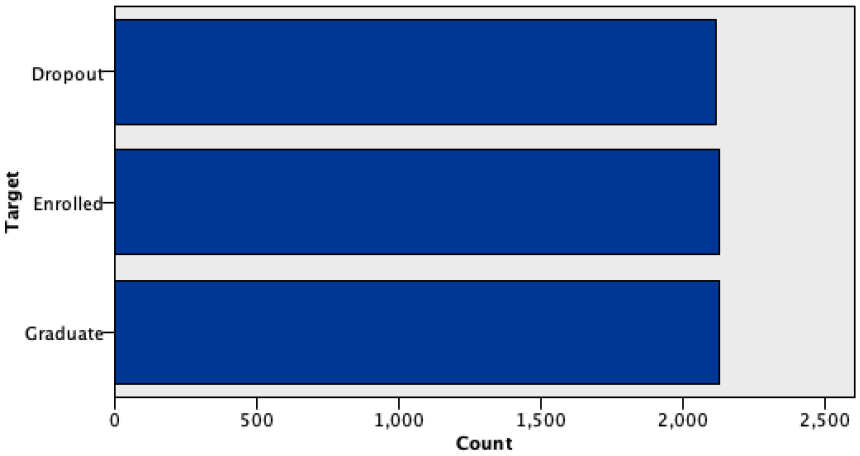


Ilustración 19.- DataSet con balanceo sobremuestreo.

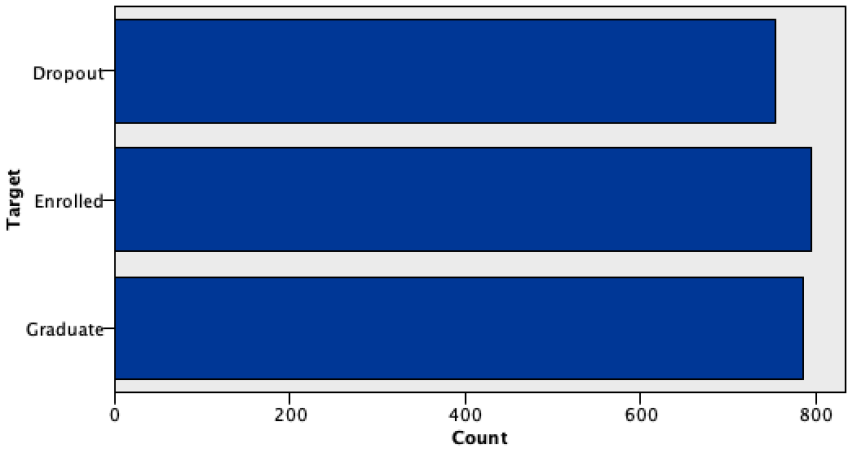
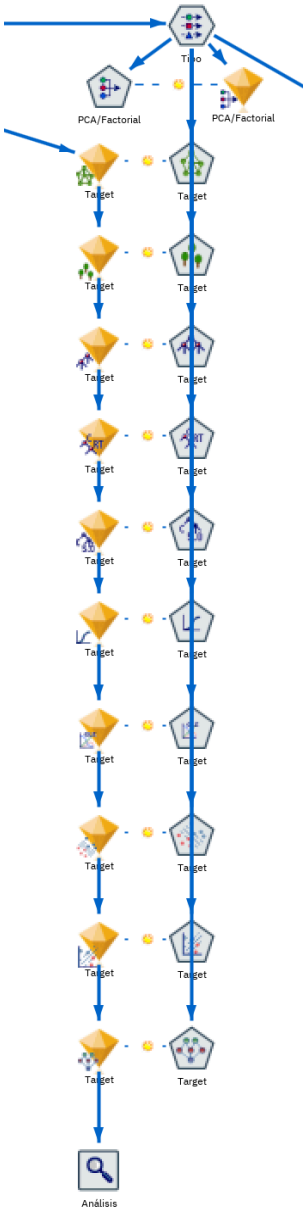


Ilustración 19.- DataSet con balanceo submuestreo.

4.1.1.1.2.- Aplicación de los modelos sin balanceo ni selección de características



Flujo de trabajo IBM SPSS Modeler

Se llevó a cabo la evaluación inicial de los modelos de aprendizaje automático utilizando el conjunto de datos sin aplicar técnicas de balanceo de clases o selección de características.

Este enfoque nos permite comprender el rendimiento de los modelos en su estado original y proporciona una línea base para comparar con los resultados obtenidos después de aplicar técnicas adicionales de preprocesamiento de datos.

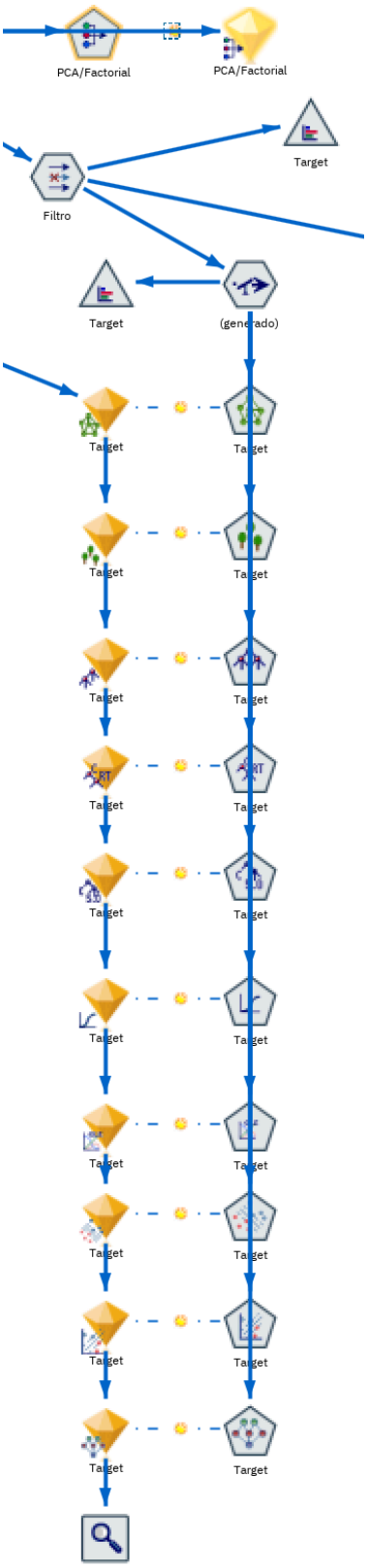
Durante esta etapa, se aplicaron varios algoritmos de aprendizaje automático, incluyendo redes neuronales, árboles de decisión, regresión logística, máquinas de vectores de soporte (SVM), y otros modelos, directamente al conjunto de datos sin modificaciones.

Esta evaluación inicial nos permitió identificar los modelos prometedores y descartar aquellos que mostraron un rendimiento insatisfactorio. Además, nos proporcionó una comprensión de los desafíos inherentes en los datos, como el desequilibrio de clases o la presencia de características irrelevantes o redundantes, que podrían abordarse mediante técnicas adicionales de preprocesamiento en las etapas posteriores del análisis.

4.1.1.1.3.- Resultados de los modelos sin balanceo ni selección de características:

Evaluación de los modelos	Resultados	Atributos importantes
Neural Net	77.43%	1.- Curricular units 2nd sem (approved) 2.- Curricular units 1st sem (approved) 3.- Curricular units 1st sem (enrolled) 4.- Curricular units 2nd sem (enrolled) 5.- Curricular units 2nd sem (evaluations)
Random Trees	84.61%	1.- Curricular units 2nd sem (evaluations) 2.- Fathers Occupation 3.- Age at enrollment 4.- Application Mode 5.- Previous qualification (grade)
Tree-AS	76.4%	La importancia de predictor no está disponible porque todos los predictores son constantes o aparecen como no relacionados con el destino.
C&R Tree	75.64%	1.- Curricular units 2nd sem (approved) 2.- Tuition frees up to date 3.- Curricular units 2nd sem (evaluations) 4.- Age at enrollment 5.- Curricular units 2nd sem (enrolled)
C5.0	84.59%	1.- Curricular units 2nd sem (approved) 2.- Tuition frees up to date 3.- Curricular units 1st sem (approved) 4.- Curricular units 2nd sem (enrolled) 5.- Curricular units 1st sem (without evaluations)
Logistic	77.87%	N/A
GLE	77.87%	1.- Curricular units 2nd sem (approved) 2.- Curricular units 1st sem (approved) 3.- Curricular units 2nd sem (grade) 4.- Tuition frees up to date 5.- Curricular units 2nd sem (evaluations)
SVM	81.6%	N/A
LSVM	77.8%	N/A
Bayes Net	76.79%	N/A

4.1.1.1.4.- Aplicación de los modelos con balanceo sobremuestreo y con selección de características.



Flujo de trabajo IBM SPSS Modeler

Se procedió a realizar una serie de técnicas de preprocesamiento adicionales con el objetivo de mejorar el rendimiento de los modelos de aprendizaje automático. En esta fase, se abordaron dos aspectos importantes: el desbalance de clases y la selección de características relevantes.

En primer lugar, se aplicó la técnica de balanceo sobremuestreo para abordar el desequilibrio de clases en el conjunto de datos. Esta técnica consiste en aumentar la cantidad de muestras de la clase minoritaria mediante la replicación de observaciones o la generación de datos sintéticos. Esto permite que el modelo se entrene de manera más equilibrada y reduzca el sesgo hacia la clase mayoritaria.

Además, se llevó a cabo un proceso de selección de características para identificar y eliminar aquellas que no contribuyen significativamente a la capacidad predictiva del modelo. Esto ayuda a reducir la dimensionalidad del conjunto de datos y a mejorar la eficiencia del modelo al eliminar el ruido y la redundancia.

Posteriormente, se aplicaron los mismos algoritmos de aprendizaje automático utilizados en la sección anterior, pero esta vez sobre el conjunto de datos preprocesado. Se registraron nuevamente las métricas de rendimiento para evaluar la mejora en la capacidad predictiva de los modelos después de aplicar las técnicas de balanceo y selección de características.

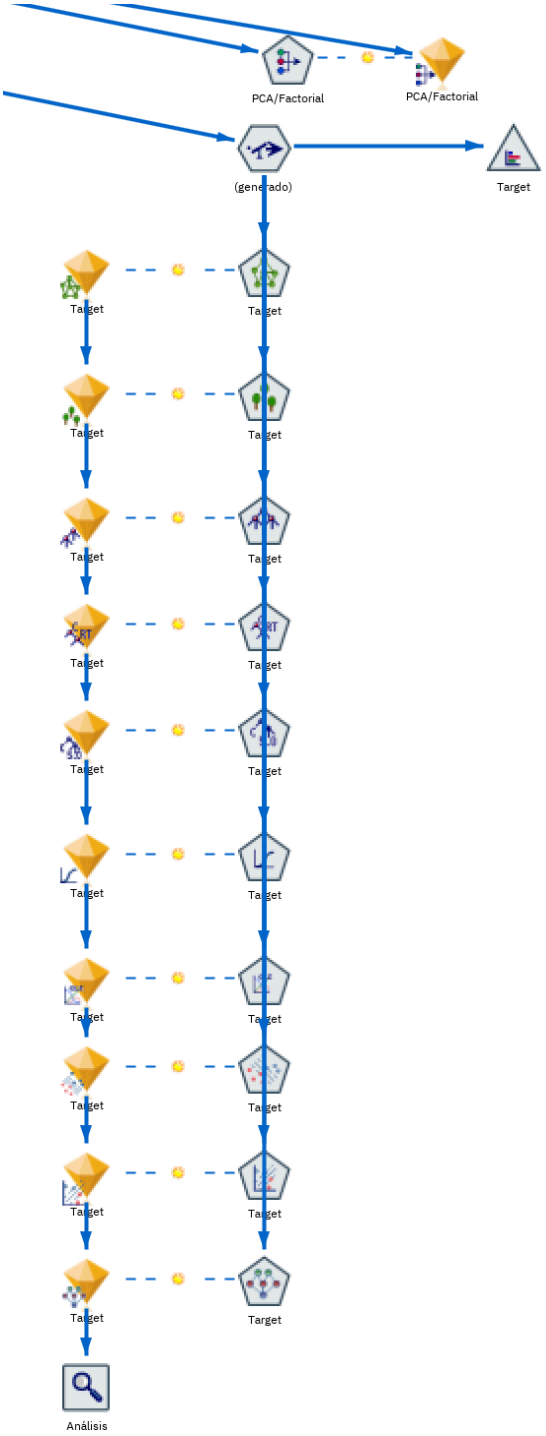
Esta fase nos permitió comparar el rendimiento de los modelos antes y después de aplicar las técnicas de preprocesamiento, identificando cualquier mejora significativa en términos de precisión.

Además, nos proporcionó información sobre la importancia relativa de las características seleccionadas en la predicción del resultado deseado.

4.1.1.1.5.- Resultados de los modelos con balanceo sobremuestreo y con selección de características

Evaluación de los modelos	Resultados	Atributos importantes
Neural Net	73.94%	1.- Curricular units 2nd sem (approved) 2.- Curricular units 1st sem (approved) 3.- Curricular units 1st sem (enrolled) 4.- Curricular units 2nd sem (enrolled) 5.- Curricular units 2nd sem (grade)
Random Trees	83.57%	1.- Previous qualification (grade) 2.- Fathers Occupation 3.- Admission grade 4.- Mothers occupation 5.- Age at enrollment
Tree-AS	72.07%	La importancia de predictor no está disponible porque todos los predictores son constantes o aparecen como no relacionados con el destino.
C&R Tree	72.51%	1.- Curricular units 2nd sem (approved) 2.- Curricular units 1st sem (evaluations) 3.- Age at enrollment 4.- Curricular units 1st sem (approved) 5.- Mothers occupation
C5.0	92.02%	1.- Curricular units 2nd sem (approved) 2.- Age at enrollment 3.- Fathers occupation 4.- Curricular units 1st sem (evaluations) 5.- Curricular units 2nd sem (grade)
Logistic	73.64%	N/A
GLE	73.59%	1.- Curricular units 2nd sem (approved) 2.- Curricular units 1st sem (approved) 3.- Curricular units 2nd sem (grade) 4.- Curricular units 2nd sem (evaluations) 5.- Age at enrollment
SVM	74.4%	N/A
LSVM	72.99%	N/A
Bayes Net	72.23%	N/A

4.1.1.1.6.- Aplicación de los modelos con balanceo submuestreo y con selección de características



Flujo de trabajo IBM SPSS Modeler

Se llevaron a cabo una serie de técnicas de preprocesamiento adicionales con el fin de mejorar la capacidad predictiva de los modelos de aprendizaje automático. En esta fase, se abordaron dos aspectos fundamentales: el desbalance de clases y la reducción de la dimensionalidad mediante la selección de características relevantes.

En primer lugar, se aplicó la técnica de balanceo submuestreo para abordar el desequilibrio de clases en el conjunto de datos. Esta técnica consiste en reducir la cantidad de muestras de la clase mayoritaria, seleccionando aleatoriamente una cantidad igual de muestras de la clase minoritaria. Esto ayuda a evitar el sesgo hacia la clase mayoritaria y a mejorar la capacidad del modelo para reconocer y predecir la clase minoritaria con precisión.

Además, se llevó a cabo un proceso de selección de características para identificar y retener únicamente aquellas que contribuyen de manera significativa a la capacidad predictiva del modelo. Esto se logró mediante técnicas de análisis de importancia de características o métodos de filtrado y envoltura, que permiten identificar las características más relevantes para la tarea de clasificación.

Posteriormente, se aplicaron los mismos algoritmos de aprendizaje automático utilizados en las secciones anteriores, pero esta vez sobre el conjunto de datos preprocesado. Se registraron nuevamente las métricas de rendimiento para evaluar la mejora en la capacidad predictiva de los modelos después de aplicar las técnicas de balanceo y selección de características.

Esta fase de preprocesamiento nos permitió comparar el rendimiento de los modelos antes y después de aplicar las técnicas de balanceo y selección de características, identificando cualquier mejora significativa en términos de precisión. Además, nos proporcionó información sobre la importancia relativa de las características seleccionadas en la predicción del resultado deseado.

4.1.1.1.7.- Resultados de los modelos con balanceo submuestreo y con selección de características

Evaluación de los modelos	Resultados	Atributos importantes
Neural Net	71.15%	1.- Curricular units 2nd sem (approved) 2.- Curricular units 2nd sem (enrolled) 3.- Curricular units 1st sem (approved) 4.- Curricular units 1st sem (enrolled) 5.- Curricular units 2nd sem (grade)
Random Trees	79.14%	1.- Age at enrollment 2.- Fathers qualification 3.- Admission grade 4.- Mothers qualification 5.- Previous qualification (grade)
Tree-AS	67.7%	La importancia de predictor no está disponible porque todos los predictores son constantes o aparecen como no relacionados con el destino.
C&R Tree	67.38%	1.- Curricular units 2nd sem (approved) 2.- Curricular units 1st sem (approved) 3.- Curricular units 1st sem (evaluations) 4.- Curricular units 2nd sem (enrolled) 5.- Curricular units 2nd sem (grade)
C5.0	77%	1.- Curricular units 2nd sem (approved) 2.- Curricular units 1st sem (approved) 3.- Curricular units 2nd sem (grade) 4.- Curricular units 1st sem (evaluations) 5.- Curricular units 2nd sem (enrolled)
Logistic	73.2%	N/A
GLE	73.2%	1.- Curricular units 2nd sem (approved) 2.- Curricular units 1st sem (approved) 3.- Curricular units 2nd sem (grade) 4.- Age at enrollment 5.- Curricular units 2nd sem (evaluations)
SVM	72.53%	N/A
LSVM	72.81%	N/A
Bayes Net	71.24%	N/A

4.1.1.2.- Selección de características con Neural Net

Se empleó el algoritmo de redes neuronales para llevar a cabo un proceso de selección de características en el conjunto de datos. Este proceso tiene como objetivo identificar las variables más relevantes o informativas que contribuyen significativamente a la capacidad predictiva del modelo.

Para realizar la selección de características, se utilizó el enfoque de las redes neuronales, que tiene la capacidad de aprender automáticamente las representaciones más importantes de los datos durante el proceso de entrenamiento. Durante este proceso, las neuronas en las capas ocultas de la red neuronal aprenden a ponderar las características de entrada de manera que las más relevantes tengan un mayor impacto en la salida del modelo.

El algoritmo de redes neuronales empleado se entrenó utilizando el conjunto de datos completo, y durante este proceso, se observó cómo cada característica contribuía a la capacidad predictiva del modelo. Las características que tenían un mayor impacto en la salida del modelo fueron consideradas como las más relevantes y fueron seleccionadas para su inclusión en el modelo final.

Este enfoque de selección de características con redes neuronales ofrece varias ventajas, como la capacidad de manejar relaciones no lineales entre las características y la capacidad de detectar interacciones complejas entre ellas. Además, al dejar que el algoritmo de redes neuronales identifique automáticamente las características más relevantes, se puede evitar el sesgo humano en la selección de características.

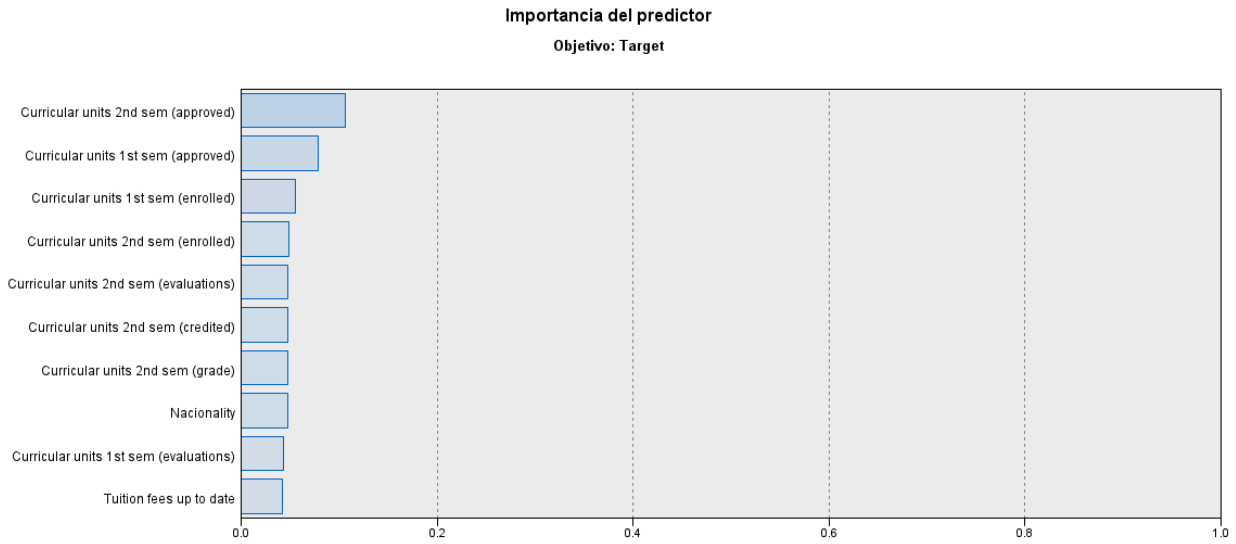


Ilustración 19.- Se visualizan en un gráfica las características más importantes

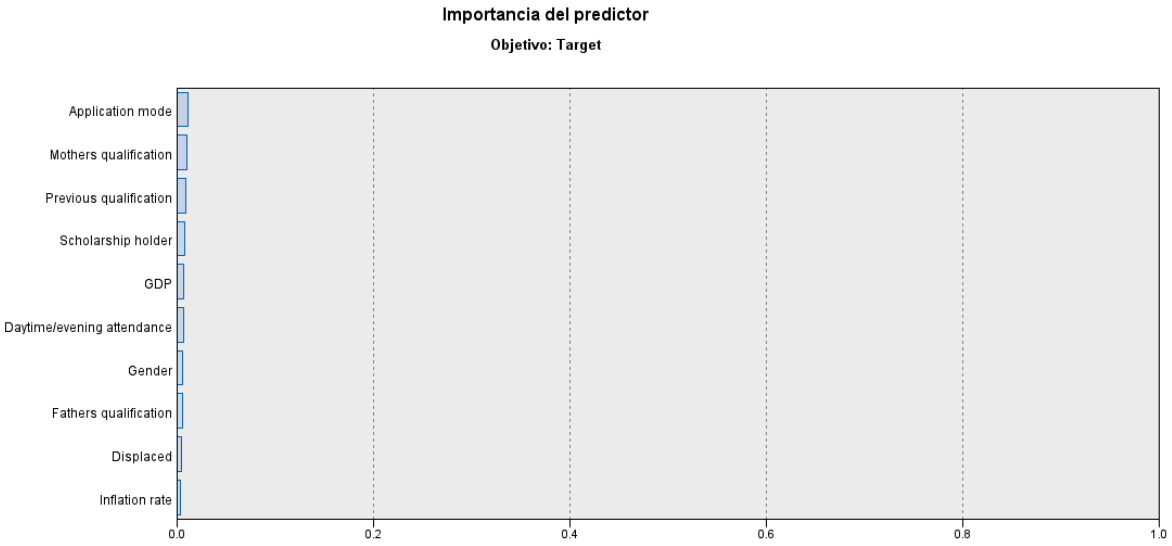


Ilustración 19.- Se visualizan en un gráfica las características menos importantes

Atributos menos importantes	Importancia del predictor
Application Mode	0.01
Mothers qualification	0.01
Previous qualification	0.01
Scholarship holder	0.01
GDP	0.01
Daytime/evening attendance	0.01
Gender	0.01
Fathers qualification	0.01
Displaced	0.01
Inflation rate	0.00

Atributos más importantes	Importancia del predictor
Curricular units 2nd sem (approved)	0.11
Curricular units 1st sem (approved)	0.08
Curricular units 1st sem (enrolled)	0.06
Curricular units 2nd sem (enrolled)	0.05
Curricular units 2nd sem (evaluations)	0.05
Curricular units 2nd sem (credited)	0.05
Curricular units 2nd sem (grade)	0.05
Nacionality	0.05
Curricular units 1st sem (evaluations)	0.04
Tuition fees up to date	0.04

4.1.1.2.1.- Balanceo del DataSet

El balanceo del conjunto de datos es un paso crucial en el análisis de datos cuando se enfrenta a conjuntos de datos desequilibrados, es decir, conjuntos de datos en los que las clases objetivo están representadas de manera desproporcionada. El desequilibrio puede tener un impacto significativo en el rendimiento de los modelos de aprendizaje automático, ya que pueden sesgar el modelo hacia las clases mayoritarias y hacer que sea menos preciso para predecir las clases minoritarias.

El nodo de equilibrado en IBM SPSS Modeler proporciona una solución eficaz para abordar este problema al aplicar técnicas de submuestreo o sobremuestreo para igualar la distribución de clases en el conjunto de datos.

Pasos para Balancear el DataSet con el Nodo de Equilibrar:

Configuración del Nodo de Equilibrar:
Configura el nodo de equilibrado en IBM SPSS Modeler, especificando la variable objetivo y seleccionando el método de equilibrado deseado (submuestreo o sobremuestreo).

Ejecución del Nodo:
Ejecuta el nodo para aplicar la técnica de balanceo seleccionada al conjunto de datos. Durante este paso, el nodo ajustará la distribución de clases para mejorar la representación de las clases minoritarias.

Evaluación del Conjunto de Datos Equilibrado:
Después de balancear el conjunto de datos, es fundamental evaluar el impacto del equilibrado en el rendimiento del modelo. Esto puede implicar la comparación de métricas de rendimiento antes y después del equilibrado, como precisión, sensibilidad, especificidad, etc.

Beneficios del Balanceo del DataSet:

- Mejora la precisión del modelo al reducir el sesgo hacia las clases mayoritarias.
- Ayuda a prevenir la infra-representación de clases minoritarias, lo que podría resultar en decisiones erróneas.
- Proporciona una representación más equitativa de todas las clases en el conjunto de datos, lo que puede mejorar la robustez y generalización del modelo.

Resultados del balanceo:

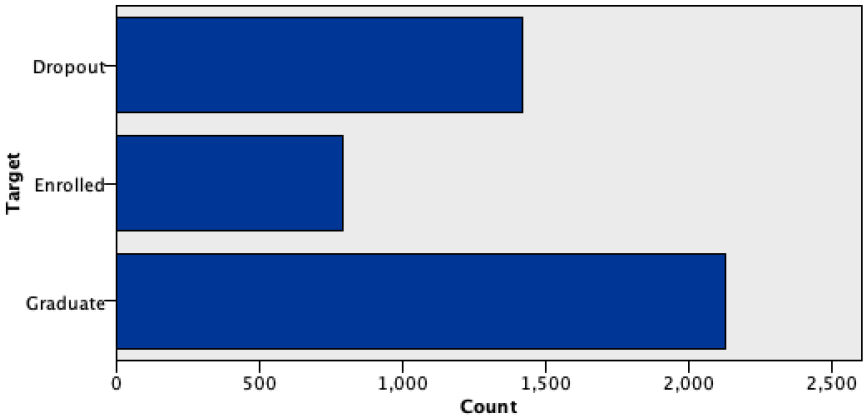


Ilustración 19.- Validación de los tipos de datos en el DataSet Update.

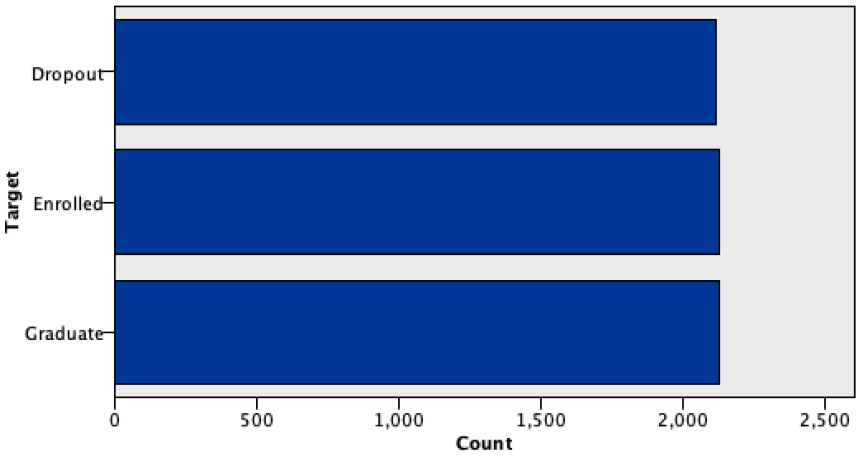


Ilustración 19.- Validación de los tipos de datos en el DataSet Update.

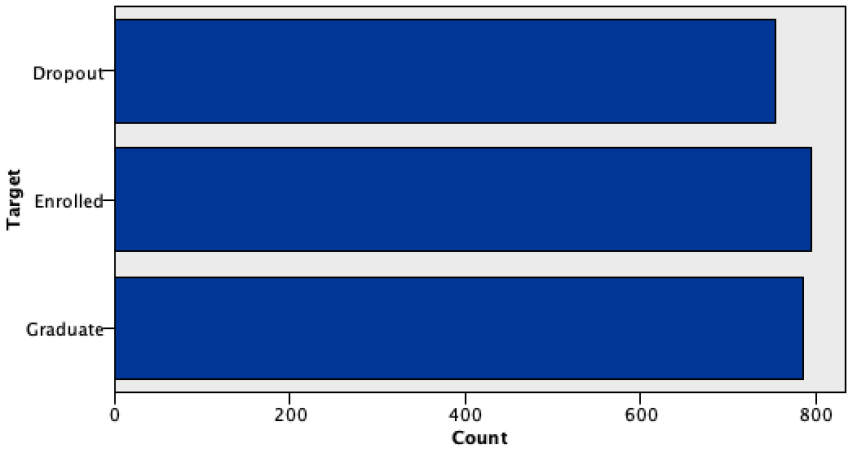
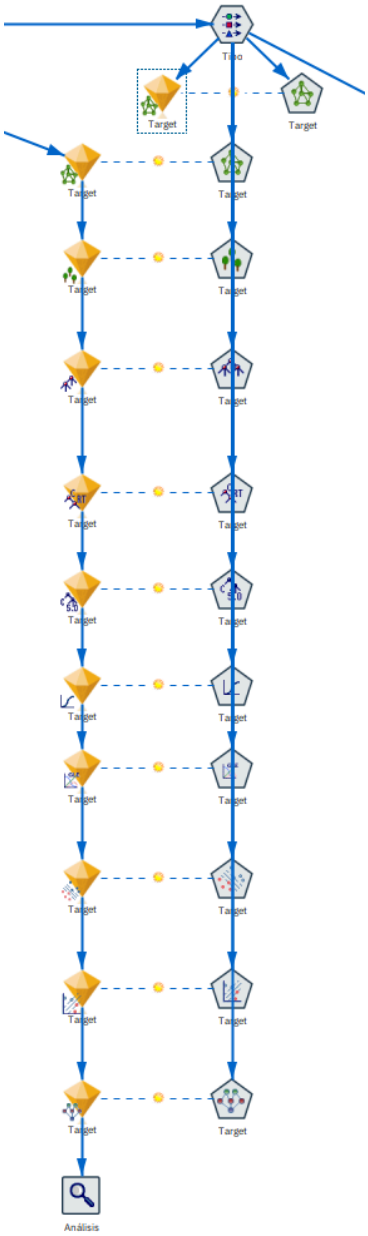


Ilustración 19.- Validación de los tipos de datos en el DataSet Update.

4.1.1.2.2.- Aplicación de los modelos sin balanceo ni selección de características



Flujo de trabajo IBM SPSS Modeler

Se llevó a cabo la evaluación inicial de los modelos de aprendizaje automático utilizando el conjunto de datos sin aplicar técnicas de balanceo de clases o selección de características.

Este enfoque nos permite comprender el rendimiento de los modelos en su estado original y proporciona una línea base para comparar con los resultados obtenidos después de aplicar técnicas adicionales de preprocesamiento de datos.

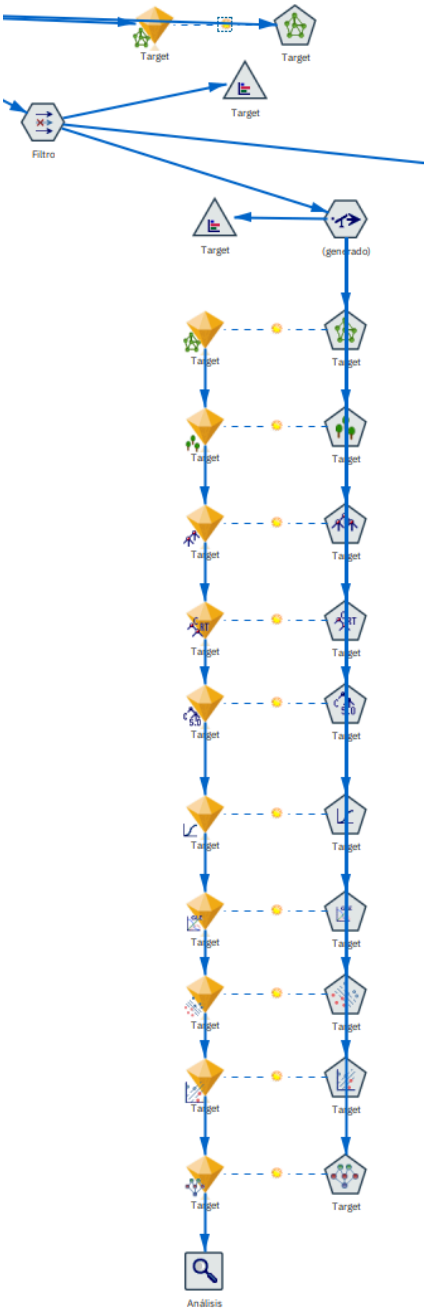
Durante esta etapa, se aplicaron varios algoritmos de aprendizaje automático, incluyendo redes neuronales, árboles de decisión, regresión logística, máquinas de vectores de soporte (SVM), y otros modelos, directamente al conjunto de datos sin modificaciones.

Esta evaluación inicial nos permitió identificar los modelos prometedores y descartar aquellos que mostraron un rendimiento insatisfactorio. Además, nos proporcionó una comprensión de los desafíos inherentes en los datos, como el desequilibrio de clases o la presencia de características irrelevantes o redundantes, que podrían abordarse mediante técnicas adicionales de preprocesamiento en las etapas posteriores del análisis.

4.1.1.2.3.- Resultados de los modelos sin balanceo ni selección de características:

Evaluación de los modelos	Resultados	Atributos importantes
Neural Net	78.28%	1.- Curricular units 2nd sem (approved) 2.- Curricular units 1st sem (approved) 3.- Curricular units 2nd sem (enrolled) 4.- Curricular units 2nd sem (credited) 5.- Curricular units 1st sem (enrolled)
Random Trees	84.54%	1.- Previous qualification (grade) 2.- Course 3.- Curricular units 2nd sem (approved) 4.- Admission grade 5.- Fathers qualification
Tree-AS	76.37%	La importancia de predictor no está disponible porque todos los predictores son constantes o aparecen como no relacionados con el destino.
C&R Tree	75.62%	1.- Curricular units 2nd sem (approved) 2.- Tuition frees up to date 3.- Curricular units 2nd sem (evaluations) 4.- Age at enrollment 5.- Curricular units 2nd sem (enrolled)
C5.0	84.56%	1.- Curricular units 2nd sem (approved) 2.- Tuition frees up to date 3.- Curricular units 1st sem (approved) 4.- Curricular units 2nd sem (enrolled) 5.- Curricular units 1st sem (without evaluations)
Logistic	77.85%	N/A
GLE	77.85%	1.- Curricular units 2nd sem (approved) 2.- Curricular units 1st sem (approved) 3.- Curricular units 2nd sem (grade) 4.- Tuition frees up to date 5.- Curricular units 2nd sem (evaluations)
SVM	81.57%	N/A
LSVM	77.8%	N/A
Bayes Net	76.77%	N/A

4.1.1.2.4.- Aplicación de los modelos con balanceo sobremuestreo y con selección de características



Flujo de trabajo IBM SPSS Modeler

Se procedió a realizar una serie de técnicas de preprocesamiento adicionales con el objetivo de mejorar el rendimiento de los modelos de aprendizaje automático. En esta fase, se abordaron dos aspectos importantes: el desbalance de clases y la selección de características relevantes.

En primer lugar, se aplicó la técnica de balanceo sobremuestreo para abordar el desequilibrio de clases en el conjunto de datos. Esta técnica consiste en aumentar la cantidad de muestras de la clase minoritaria mediante la replicación de observaciones o la generación de datos sintéticos. Esto permite que el modelo se entrene de manera más equilibrada y reduzca el sesgo hacia la clase mayoritaria.

Además, se llevó a cabo un proceso de selección de características para identificar y eliminar aquellas que no contribuyen significativamente a la capacidad predictiva del modelo. Esto ayuda a reducir la dimensionalidad del conjunto de datos y a mejorar la eficiencia del modelo al eliminar el ruido y la redundancia.

Posteriormente, se aplicaron los mismos algoritmos de aprendizaje automático utilizados en la sección anterior, pero esta vez sobre el conjunto de datos preprocesado. Se registraron nuevamente las métricas de rendimiento para evaluar la mejora en la capacidad predictiva de los modelos después de aplicar las técnicas de balanceo y selección de características.

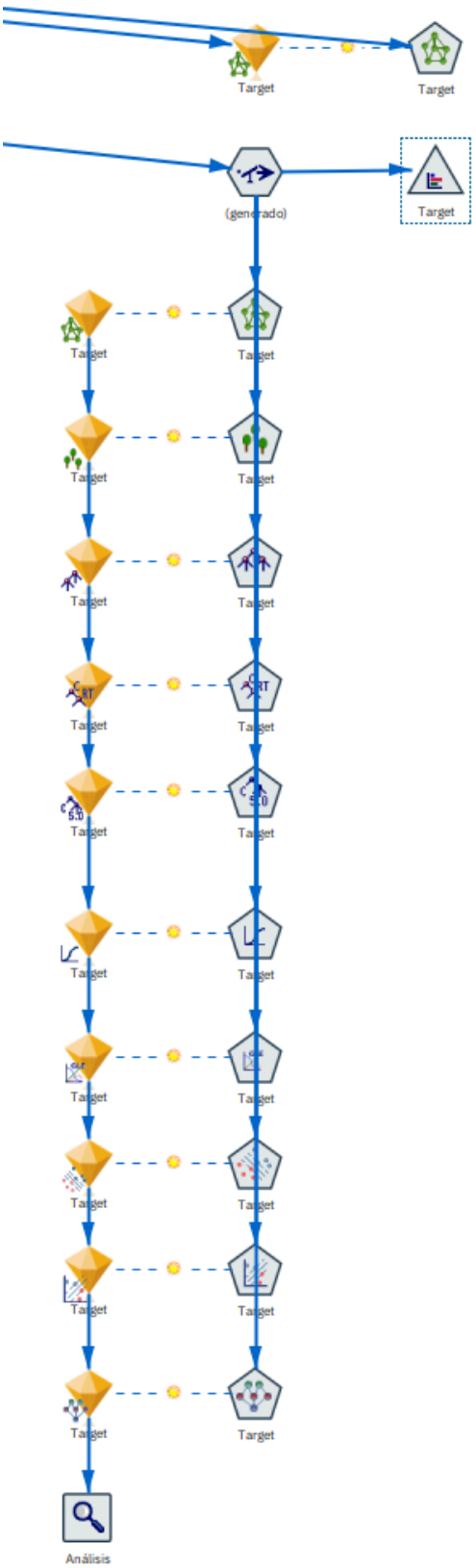
Esta fase nos permitió comparar el rendimiento de los modelos antes y después de aplicar las técnicas de preprocesamiento, identificando cualquier mejora significativa en términos de precisión.

Además, nos proporcionó información sobre la importancia relativa de las características seleccionadas en la predicción del resultado deseado.

4.1.1.2.5.- Resultados de los modelos con balanceo sobremuestreo y con selección de características

Evaluación de los modelos	Resultados	Atributos importantes
Neural Net	75.98%	1.- Curricular units 2nd sem (approved) 2.- Curricular units 2nd sem (enrolled) 3.- Curricular units 1st sem (approved) 4.- Nacionality 5.- Curricular units 2nd sem (grade)
Random Trees	85.67%	1.- Admission grade 2.- Age at enrollment 3.- Previous qualification (grade) 4.- Fathers Occupation 5.- Curricular units 2nd sem (evaluations)
Tree-AS	75.66%	La importancia de predictor no está disponible porque todos los predictores son constantes o aparecen como no relacionados con el destino.
C&R Tree	73.2%	1.- Curricular units 2nd sem (approved) 2.- Curricular units 1st sem (evaluations) 3.- Tuition fees up to date 4.- Curricular units 1st sem (approved) 5.- Curricular units 2nd sem (enrolled)
C5.0	92.34%	1.- Curricular units 2nd sem (approved) 2.- Curricular units 2nd sem (evaluations) 3.- Application order 4.- Previous qualification (grade) 5.- Tuition fees up to date
Logistic	75.8%	N/A
GLE	75.8%	1.- Curricular units 2nd sem (approved) 2.- Curricular units 1st sem (approved) 3.- Curricular units 2nd sem (grade) 4.- Tuition fees up to date 5.- Curricular units 2nd sem (evaluations)
SVM	75.89%	N/A
LSVM	75.36%	N/A
Bayes Net	73.38%	N/A

4.1.1.2.6.- Aplicación de los modelos con balanceo submuestreo y con selección de características



Flujo de trabajo IBM SPSS Modeler

Se llevaron a cabo una serie de técnicas de preprocesamiento adicionales con el fin de mejorar la capacidad predictiva de los modelos de aprendizaje automático. En esta fase, se abordaron dos aspectos fundamentales: el desbalance de clases y la reducción de la dimensionalidad mediante la selección de características relevantes.

En primer lugar, se aplicó la técnica de balanceo submuestreo para abordar el desequilibrio de clases en el conjunto de datos. Esta técnica consiste en reducir la cantidad de muestras de la clase mayoritaria, seleccionando aleatoriamente una cantidad igual de muestras de la clase minoritaria. Esto ayuda a evitar el sesgo hacia la clase mayoritaria y a mejorar la capacidad del modelo para reconocer y predecir la clase minoritaria con precisión.

Además, se llevó a cabo un proceso de selección de características para identificar y retener únicamente aquellas que contribuyen de manera significativa a la capacidad predictiva del modelo. Esto se logró mediante técnicas de análisis de importancia de características o métodos de filtrado y envoltura, que permiten identificar las características más relevantes para la tarea de clasificación.

Posteriormente, se aplicaron los mismos algoritmos de aprendizaje automático utilizados en las secciones anteriores, pero esta vez sobre el conjunto de datos preprocesado. Se registraron nuevamente las métricas de rendimiento para evaluar la mejora en la capacidad predictiva de los modelos después de aplicar las técnicas de balanceo y selección de características.

Esta fase de preprocesamiento nos permitió comparar el rendimiento de los modelos antes y después de aplicar las técnicas de balanceo y selección de características, identificando cualquier mejora significativa en términos de precisión. Además, nos proporcionó información sobre la importancia relativa de las características seleccionadas en la predicción del resultado deseado.

4.1.1.2.7.- Resultados de los modelos con balanceo submuestreo y con selección de características

Evaluación de los modelos	Resultados	Atributos importantes
Neural Net	75.66%	1.- Curricular units 2nd sem (approved) 2.- Curricular units 1st sem (approved) 3.- Curricular units 2nd sem (enrolled) 4.- Curricular units 1st sem (enrolled) 5.- Curricular units 2nd sem (grade)
Random Trees	79.55%	1.- Previous qualification (grade) 2.- Mothers occupation 3.- Fathers occupation 4.- Admission grade 5.- Curricular units 1st sem (grade)
Tree-AS	73.08%	La importancia de predictor no está disponible porque todos los predictores son constantes o aparecen como no relacionados con el destino.
C&R Tree	75.43%	1.- Curricular units 2nd sem (approved) 2.- Tuition fees up to date 3.- Curricular units 1st sem (approved) 4.- Curricular units 1st sem (evaluations) 5.- Course
C5.0	79.55%	1.- Curricular units 2nd sem (approved) 2.- Curricular units 2nd sem (grade) 3.- Tuition fees up to date 4.- Curricular units 1st sem (approved) 5.- Application order
Logistic	75.96%	N/A
GLE	75.96%	1.- Curricular units 2nd sem (approved) 2.- Curricular units 1st sem (approved) 3.- Curricular units 2nd sem (grade) 4.- Tuition fees up to date 5.- Curricular units 2nd sem (evaluations)
SVM	74.63%	N/A
LSVM	75.55%	N/A
Bayes Net	72.51%	N/A

4.1.2.- Utilizando Python

4.1.2.1.- Librerías utilizadas

En este proyecto, se han empleado varias herramientas y bibliotecas de Python para llevar a cabo el análisis y la predicción:

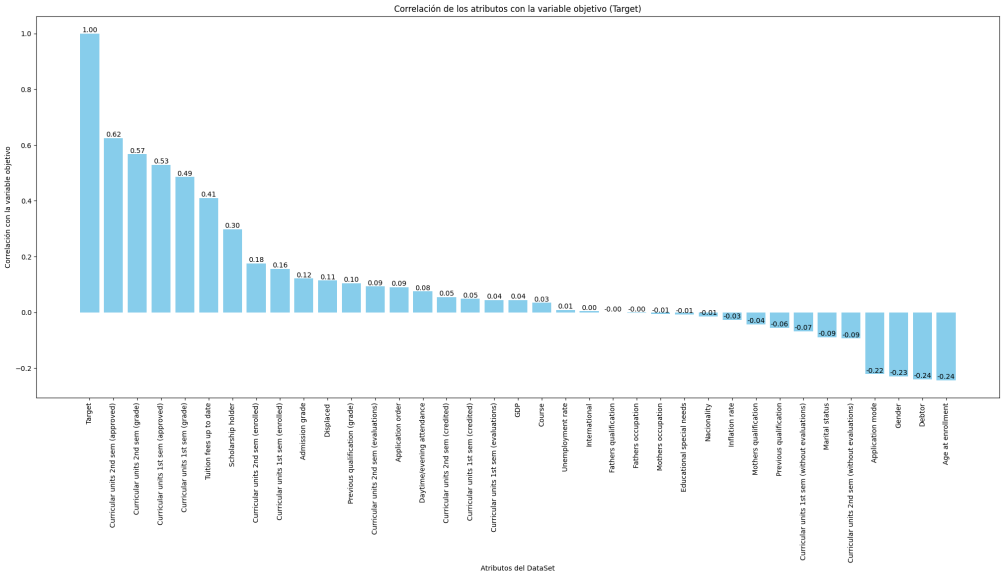
- **NumPy:** Una biblioteca fundamental para realizar operaciones numéricas en Python, utilizada para el manejo eficiente de matrices y operaciones matemáticas.

- **Pandas:** Una biblioteca que proporciona estructuras de datos y herramientas de análisis de datos que facilitan la manipulación y el análisis de conjuntos de datos tabulares.

- **Matplotlib:** Una biblioteca de visualización en 2D de Python que permite crear gráficos estáticos, histogramas, gráficos de barras, dispersión, etc., con solo unas pocas líneas de código.

- **Scikit-learn (sklearn):** Una biblioteca de aprendizaje automático en Python que ofrece herramientas simples y eficientes para el análisis predictivo de datos. Se utiliza para implementar modelos de aprendizaje automático y realizar evaluaciones de su rendimiento.

4.1.2.2.- Correlación con la variable objetivo (Target)



4.1.2.3.- Selección de las técnicas de modelado

Para abordar la tarea de predecir el estado de los alumnos (graduados, enrolados o desertores) en nuestro proyecto de análisis de datos, hemos optado por emplear dos algoritmos de modelado principales en Python: Random Forest Regressor y Logistic Regression.

- **Random Forest Regressor**

El Random Forest Regressor se seleccionó debido a su capacidad para manejar conjuntos de datos complejos y su habilidad para trabajar tanto con variables numéricas como categóricas. Este algoritmo es especialmente útil en nuestro caso, donde tenemos una variedad de características que podrían influir en la predicción del estado de los alumnos.

- **Logistic Regression**

Por otro lado, elegimos la Logistic Regression por su capacidad para modelar relaciones entre variables independientes y la probabilidad de que ocurra un evento binario. Aunque su nombre sugiere una aplicación binaria, puede ser extendida para problemas de clasificación multiclase, lo cual se adapta perfectamente a nuestra situación de predecir entre tres estados distintos.

Al utilizar estos dos algoritmos de modelado en conjunto, buscamos maximizar nuestra capacidad para prever con precisión el estado futuro de los alumnos en nuestro estudio. Este enfoque nos permite aprovechar las fortalezas individuales de cada algoritmo y obtener una comprensión más completa y precisa de los factores que influyen en la graduación, enrolamiento o deserción de los estudiantes.

4.1.2.4.- Error Medio Absoluto (Mean Absolute Error)

El error medio absoluto (MAE por sus siglas en inglés, Mean Absolute Error) es una medida de la diferencia entre los valores predichos por un modelo y los valores reales observados. Proporciona una indicación de qué tan cerca están las predicciones del modelo de los valores reales en promedio.

El cálculo del MAE es bastante simple: para cada punto de datos en el conjunto de prueba, se calcula la diferencia absoluta entre la predicción del modelo y el valor real, y luego se toma el promedio de todas estas diferencias absolutas.

Matemáticamente, el MAE se define como:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

El MAE tiene la ventaja de ser fácil de interpretar, ya que representa la magnitud promedio de los errores en la misma escala que los datos originales. Por ejemplo, si estamos prediciendo la edad de una persona y el MAE es 2 años, significa que, en promedio, nuestras predicciones están desviadas en ±2 años de las edades reales observadas.

4.1.2.5.- Generación de los modelos

4.1.2.5.1.- Random Forest Regressor (Todos los atributos)

En este proceso, se utilizó el algoritmo Random Forest Regressor para crear un modelo predictivo con el objetivo de predecir el estado de los alumnos (graduados, enrolados o desertores) en función de diversas características. A continuación se detalla paso a paso:

- **Preprocesamiento de datos:**

Se eliminó la columna "Target" del conjunto de datos original (new_data), ya que esta columna representa la variable objetivo que deseamos predecir. Las características restantes se almacenaron en la variable attributes.

- **Definición de la variable objetivo:**

La columna "Target" se asignó a la variable objetive, la cual representa la variable objetivo del problema de predicción.

- **División del conjunto de datos:**

El conjunto de datos se dividió en conjuntos de entrenamiento y validación utilizando la función train_test_split. El 80% de los datos se asignaron al conjunto de entrenamiento (train_X, train_y), mientras que el 20% se reservó para el conjunto de validación (val_X, val_y). Esta división nos permite evaluar el rendimiento del modelo en datos no vistos.

- **Creación y entrenamiento del modelo:**

Se creó una instancia del modelo Random Forest Regressor (rf_model) con una semilla aleatoria fija para garantizar la reproducibilidad de los resultados. Luego, el modelo se entrenó utilizando los datos de entrenamiento (train_X, train_y) mediante el método fit.

- **Predicciones y evaluación del modelo:**

Se realizaron predicciones sobre los datos de validación (val_X) utilizando el modelo entrenado (rf_model.predict). Posteriormente, se calculó el error medio absoluto (MAE) entre las predicciones y los valores reales (val_y) utilizando la función mean_absolute_error. Este valor de error nos proporciona una medida de qué tan cerca están las predicciones del modelo de los valores reales en el conjunto de validación.

```python
attributes = new_data.drop('Target', axis=1) # Quitamos la columna Target

objetive = new_data.Target # Definimos la columna Target como la variable objetivo

train_X, val_X, train_y, val_y = train_test_split(attributes, objetive, test_size=0.2, random_state=1) # Dividimos el dataset en entrenamiento y validación

# Crear un modelo de regresión con Random Forest
rf_model = RandomForestRegressor(random_state=1) # Definimos el modelo con una semilla aleatoria para reproducibilidad de resultados

rf_model.fit(train_X, train_y) # Entrenamos el modelo con los datos de entrenamiento

rf_val_predictions = rf_model.predict(val_X) # Realizamos predicciones con los datos de validación

rf_val_mae = mean_absolute_error(rf_val_predictions, val_y) # Calculamos el error medio absoluto
```

Ilustración 19.- Código en Python del modelo.

- **Resultados del modelo:**

Se utilizó un modelo de regresión con Random Forest para predecir el estado de los alumnos (graduados, enrolados o desertores) utilizando un conjunto de características seleccionadas. Después de entrenar el modelo y realizar predicciones sobre el conjunto de validación, se obtuvo un error medio absoluto (MAE) de 0.355839.

El MAE representa la magnitud promedio de los errores entre las predicciones del modelo y los valores reales en el conjunto de validación. En este contexto, un MAE de 0.355839 indica que, en promedio, las predicciones del modelo están desviadas en ±0.355839 unidades del valor real del estado de los alumnos.

Resultados del modelo	
MAE	0.355839

Dado que el MAE es una métrica de evaluación de la precisión del modelo, un valor más bajo indica un mejor rendimiento del modelo. En este caso, un MAE de 0.355839 sugiere que el modelo de regresión con Random Forest tiene una precisión aceptable en la predicción del estado de los alumnos.

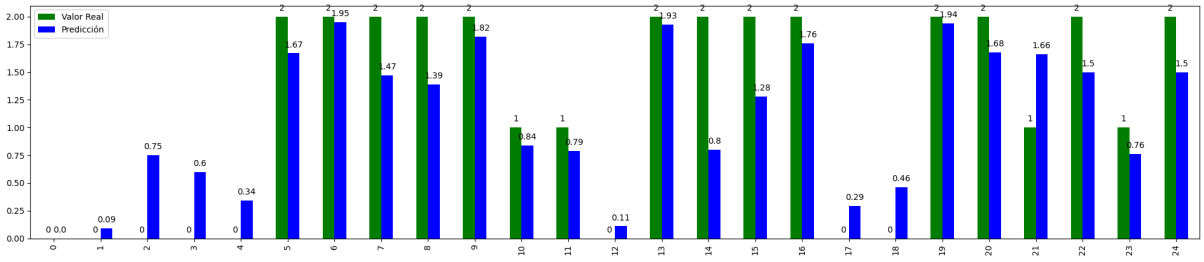


Ilustración 19.- Diferencias entre el valor real y su predicción.

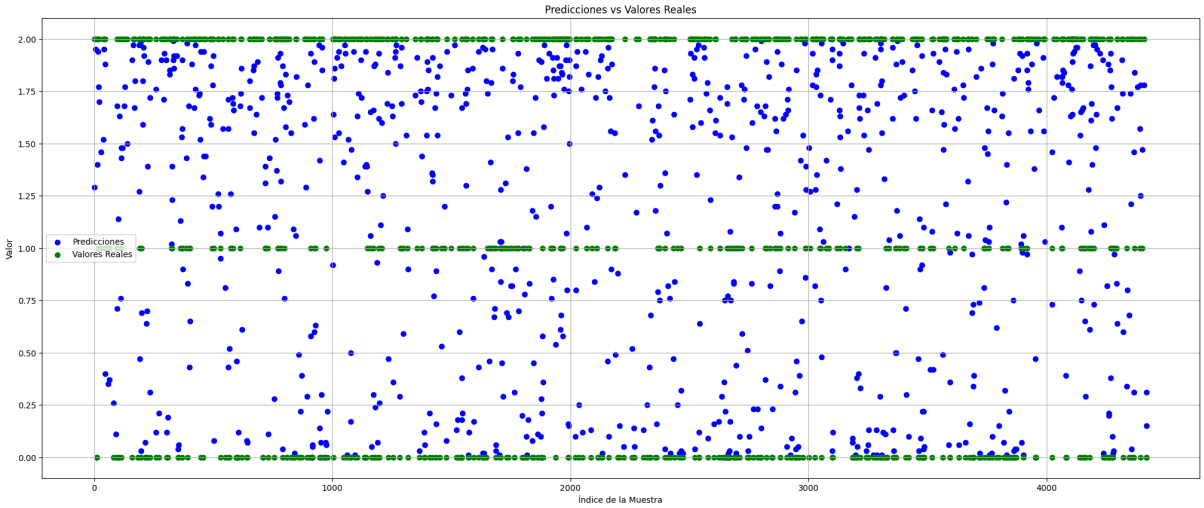


Ilustración 19.- Distribución de los valores de evaluación de predicción vs los valores reales.

4.1.2.5.2.- Random Forest Regressor (Solo los mejores atributos)

En este proceso, se aplicó el algoritmo Random Forest Regressor para desarrollar un modelo predictivo con el fin de predecir el estado de los alumnos (graduados, enrolados o desertores), utilizando un conjunto seleccionado de características. A continuación se desglosa el proceso paso a paso:

- **Selección de características relevantes:**

Se especificaron las características consideradas relevantes para la predicción del estado de los alumnos.

- **Preprocesamiento de datos:**

Se seleccionaron las características especificadas del conjunto de datos original (new_data), formando así el conjunto de características (attributes). La columna "Target" se mantuvo como la variable objetivo (objetive).

- **División del conjunto de datos:**

Se dividió el conjunto de datos en conjuntos de entrenamiento y validación utilizando la función train_test_split. El 80% de los datos se asignaron al conjunto de entrenamiento (train_X, train_y), mientras que el 20% restante se reservó para el conjunto de validación (val_X, val_y).

- **Creación y entrenamiento del modelo:**

Se creó una instancia del modelo Random Forest Regressor (rf_model) con una semilla aleatoria fija para garantizar la reproducibilidad de los resultados. Luego, el modelo se entrenó utilizando los datos de entrenamiento (train_X, train_y) mediante el método fit.

- **Predicciones y evaluación del modelo:**

Se realizaron predicciones sobre los datos de validación (val_X) utilizando el modelo entrenado (rf_model.predict). Posteriormente, se calculó el error medio absoluto (MAE) entre las predicciones y los valores reales (val_y) utilizando la función mean_absolute_error. Este valor de error proporciona una medida de qué tan cerca están las predicciones del modelo de los valores reales en el conjunto de validación.

```python
features = ['Curricular units 2nd sem (approved)', 'Curricular units 2nd sem (grade)', 'Curricular units 1st sem (approved)',
            'Curricular units 1st sem (grade)', 'Tuition fees up to date','Scholarship holder','Curricular units 2nd sem (enrolled)',
            'Curricular units 1st sem (enrolled)']

attributes = new_data[features] # Quitamos la columna Target

objetive = new_data.Target # Definimos la columna Target como la variable objetivo

train_X, val_X, train_y, val_y = train_test_split(attributes, objetive, test_size=0.2, random_state=1) # Dividimos el dataset en entrenamiento y validación

# Crear un modelo de regresión con Random Forest
rf_model = RandomForestRegressor(random_state=1) # Definimos el modelo con una semilla aleatoria para reproducibilidad de resultados

rf_model.fit(train_X, train_y) # Entrenamos el modelo con los datos de entrenamiento

rf_val_predictions = rf_model.predict(val_X) # Realizamos predicciones con los datos de validación

rf_val_mae = mean_absolute_error(rf_val_predictions, val_y) # Calculamos el error medio absoluto
```

Ilustración 19.- Código en Python del modelo.

- **Resultados del modelo:**

Se utilizó un modelo de regresión con Random Forest para predecir el estado de los alumnos (graduados, enrolados o desertados) utilizando un conjunto específico de características seleccionadas. Después de entrenar el modelo y realizar predicciones sobre el conjunto de validación, se obtuvo un error medio absoluto (MAE) de 0.354149.

El MAE, en este contexto, representa la magnitud promedio de los errores entre las predicciones del modelo y los valores reales en el conjunto de validación. Con un MAE de 0.354149, se puede interpretar que, en promedio, las predicciones del modelo están desviadas en ±0.354149 unidades del valor real del estado de los alumnos.

Resultados del modelo	
MAE	0.354149

Como con el primer modelo, un MAE más bajo indica un mejor rendimiento del modelo en términos de precisión. En este caso, un MAE de 0.354149 sugiere que el modelo de regresión con Random Forest y las características seleccionadas tiene una precisión similar al primer modelo, con una ligera mejora en la precisión.

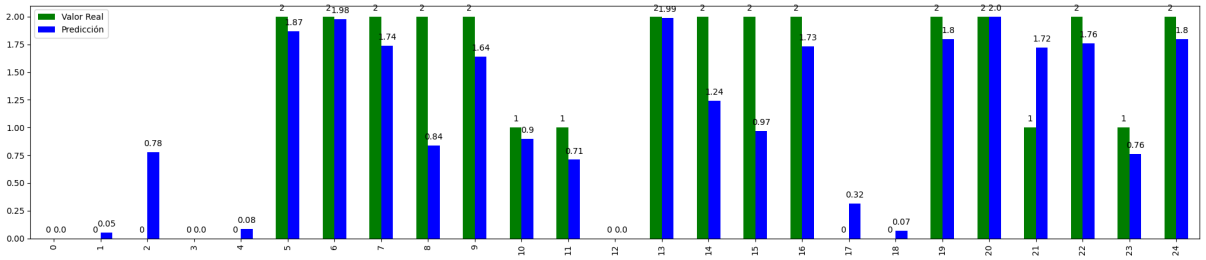


Ilustración 19.- Diferencias entre el valor real y su predicción.

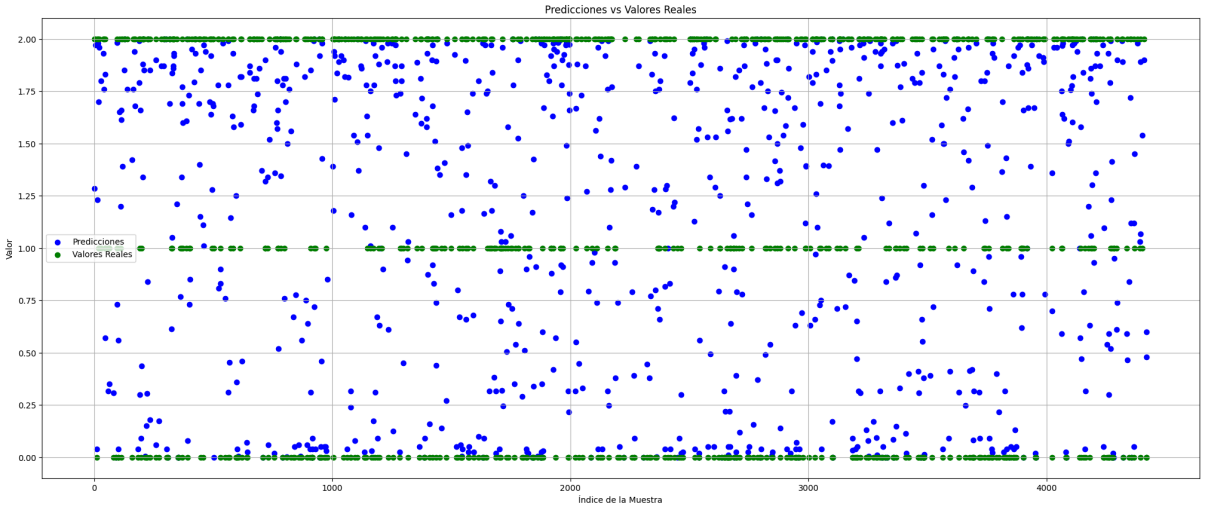


Ilustración 19.- Distribución de los valores de evaluación de predicción vs los valores reales.

4.1.2.5.3.- Logistic Regression (Todos los atributos)

En este proceso, se utilizó el algoritmo de Regresión Logística para crear un modelo predictivo con el objetivo de predecir el estado de los alumnos (graduados, enrolados o desertados). A continuación se desglosa paso a paso:

- **División del conjunto de datos:**

Se dividió el conjunto de datos en dos partes: los atributos (características) y la variable objetivo (la columna "Target").

- **División del conjunto de datos en entrenamiento y validación:**

El conjunto de datos se dividió en conjuntos de entrenamiento y validación utilizando la función train_test_split de scikit-learn. El 80% de los datos se asignaron al conjunto de entrenamiento (train_X, train_y), mientras que el 20% restante se reservó para el conjunto de validación (val_X, val_y). Esta división nos permite evaluar el rendimiento del modelo en datos no vistos.

- **Creación del modelo de regresión logística:**

Se creó una instancia del modelo de Regresión Logística (logistic_model) utilizando la clase LogisticRegression de scikit-learn. Se especificó una semilla aleatoria (random_state=1) para garantizar la reproducibilidad de los resultados. Además, se eligió el solver 'liblinear' para el cálculo de la optimización.

- **Entrenamiento del modelo:**

El modelo de Regresión Logística se entrenó utilizando los datos de entrenamiento (train_X, train_y) mediante el método fit.

- **Predicciones y evaluación del modelo:**

Se realizaron predicciones sobre los datos de validación (val_X) utilizando el modelo entrenado (logistic_model.predict). Posteriormente, se calculó el error medio absoluto (MAE) entre las predicciones y los valores reales (val_y) utilizando la función mean_absolute_error. Este valor de error proporciona una medida de qué tan cerca están las predicciones del modelo de los valores reales en el conjunto de validación.

```python
# Dividir el dataset en atributos y variable objetivo
attributes = new_data.drop('Target', axis=1) # Quitamos la columna Target
objetive = new_data.Target # Definimos la columna Target como la variable objetivo

# Dividir el dataset en entrenamiento y validación
train_X, val_X, train_y, val_y = train_test_split(attributes, objetive, test_size=0.2, random_state=1)

# Crear un modelo de regresión logística
logistic_model = LogisticRegression(random_state=1, solver='liblinear')

#logistic_model = LogisticRegression(random_state=1, max_iter=10000) # Definimos el modelo con una semilla aleatoria para reproducibilidad de resultados

# Entrenar el modelo con los datos de entrenamiento
logistic_model.fit(train_X, train_y)

# Realizar predicciones con los datos de validación
logistic_val_predictions = logistic_model.predict(val_X)

# Calcular el error medio absoluto
logistic_val_mae = mean_absolute_error(logistic_val_predictions, val_y)
```

Ilustración 19.- Código en Python del modelo.

- **Resultados del modelo:**

Se utilizó un modelo de regresión logística para predecir el estado de los alumnos (graduados, enrolados o desertados) utilizando todas las características disponibles en el conjunto de datos. Después de entrenar el modelo y realizar predicciones sobre el conjunto de validación, se obtuvo un error medio absoluto (MAE) de 0.282759.

El MAE representa la magnitud promedio de los errores entre las predicciones del modelo y los valores reales en el conjunto de validación. Con un MAE de 0.282759, se puede interpretar que, en promedio, las predicciones del modelo están desviadas en ±0.282759 unidades del valor real del estado de los alumnos.

Resultados del modelo	
MAE	0.282759

Un MAE más bajo indica un mejor rendimiento del modelo en términos de precisión. En este caso, un MAE de 0.282759 sugiere que el modelo de regresión logística tiene una precisión superior a los modelos de regresión con Random Forest evaluados anteriormente.

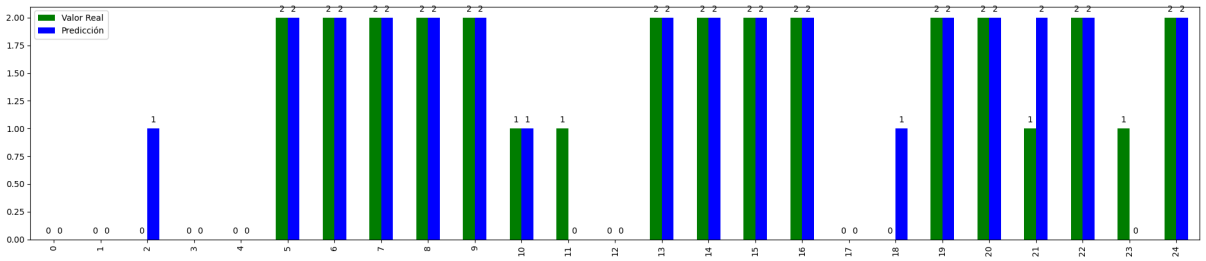


Ilustración 19.- Diferencias entre el valor real y su predicción.

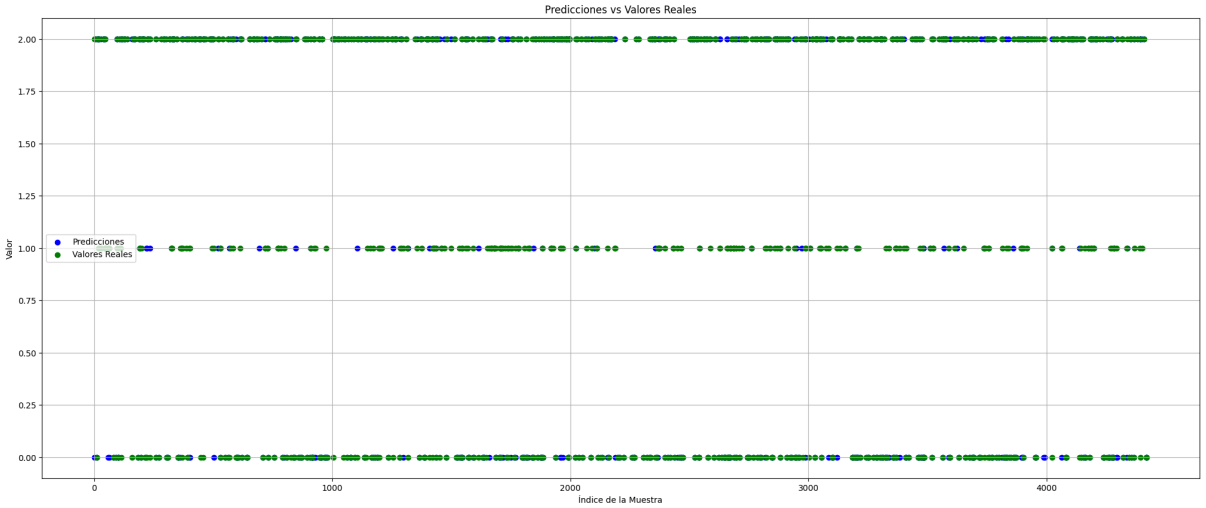


Ilustración 19.- Distribución de los valores de evaluación de predicción vs los valores reales.

4.1.2.5.4.- Logistic Regression (Solo los mejores atributos)

En este proceso, se aplicó el algoritmo de Regresión Logística para crear un modelo predictivo con el objetivo de predecir el estado de los alumnos (graduados, enrolados o desertores), utilizando un conjunto seleccionado de características. A continuación se desglosa el proceso paso a paso:

- **Selección de características relevantes:**

Se especificaron las características consideradas relevantes para la predicción del estado de los alumnos.

- **Preprocesamiento de datos:**

Se seleccionaron las características especificadas del conjunto de datos original (new_data), formando así el conjunto de características (attributes). La columna "Target" como variable objetivo (objetive).

- **División del conjunto de datos en entrenamiento y validación:**

Se dividió el conjunto de datos en conjuntos de entrenamiento y validación utilizando la función train_test_split de scikit-learn. El 80% de los datos se asignaron al conjunto de entrenamiento (train_X, train_y), mientras que el 20% restante se reservó para el conjunto de validación (val_X, val_y).

- **Creación del modelo de regresión logística:**

Se creó una instancia del modelo de Regresión Logística (logistic_model) utilizando la clase LogisticRegression de scikit-learn. Se especificó una semilla aleatoria (random_state=1) para garantizar la reproducibilidad de los resultados. Además, se eligió el solver 'liblinear' para el cálculo de la optimización.

- **Entrenamiento del modelo:**

El modelo de Regresión Logística se entrenó utilizando los datos de entrenamiento (train_X, train_y) mediante el método fit.

- **Predicciones y evaluación del modelo:**

Se realizaron predicciones sobre los datos de validación (val_X) utilizando el modelo entrenado (logistic_model.predict). Posteriormente, se calculó el error medio absoluto (MAE) entre las predicciones y los valores reales (val_y) utilizando la función mean_absolute_error.

```python
features = ['Curricular units 2nd sem (approved)', 'Curricular units 2nd sem (grade)', 'Curricular units 1st sem (approved)',
            'Curricular units 1st sem (grade)', 'Tuition fees up to date','Scholarship holder','Curricular units 2nd sem (enrolled)',
            'Curricular units 1st sem (enrolled)']

attributes = new_data[features] # Quitamos la columna Target

objetive = new_data.Target # Definimos la columna Target como la variable objetivo

# Dividir el dataset en entrenamiento y validación
train_X, val_X, train_y, val_y = train_test_split(attributes, objetive, test_size=0.2, random_state=1)

# Crear un modelo de regresión logística
logistic_model = LogisticRegression(random_state=1, solver='liblinear')

#logistic_model = LogisticRegression(random_state=1, max_iter=10000) # Definimos el modelo con una semilla aleatoria para reproducibilidad de resultados

# Entrenar el modelo con los datos de entrenamiento
logistic_model.fit(train_X, train_y)

# Realizar predicciones con los datos de validación
logistic_val_predictions = logistic_model.predict(val_X)

# Calcular el error medio absoluto
logistic_val_mae = mean_absolute_error(logistic_val_predictions, val_y)
```

Ilustración 19.- Código en Python del modelo.

- **Resultados del modelo:**

Se utilizó un modelo de regresión logística para predecir el estado de los alumnos (graduados, enrolados o desertados) utilizando un conjunto específico de características seleccionadas. Después de entrenar el modelo y realizar predicciones sobre el conjunto de validación, se obtuvo un error medio absoluto (MAE) de 0.291954.

El MAE representa la magnitud promedio de los errores entre las predicciones del modelo y los valores reales en el conjunto de validación. Con un MAE de 0.291954, se puede interpretar que, en promedio, las predicciones del modelo están desviadas en ±0.291954 unidades del valor real del estado de los alumnos.

Resultados del modelo	
MAE	0.291954

Un MAE más bajo indica un mejor rendimiento del modelo en términos de precisión. En este caso, un MAE de 0.291954 sugiere que el modelo de regresión logística con las características seleccionadas tiene una precisión comparable al modelo anterior, aunque ligeramente inferior.

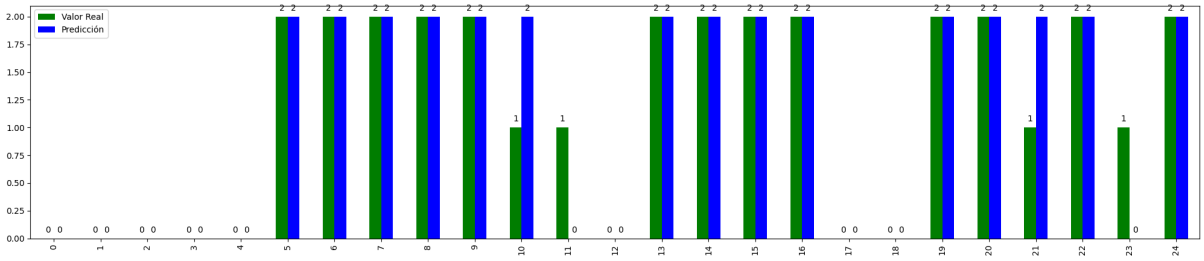


Ilustración 19.- Diferencias entre el valor real y su predicción.

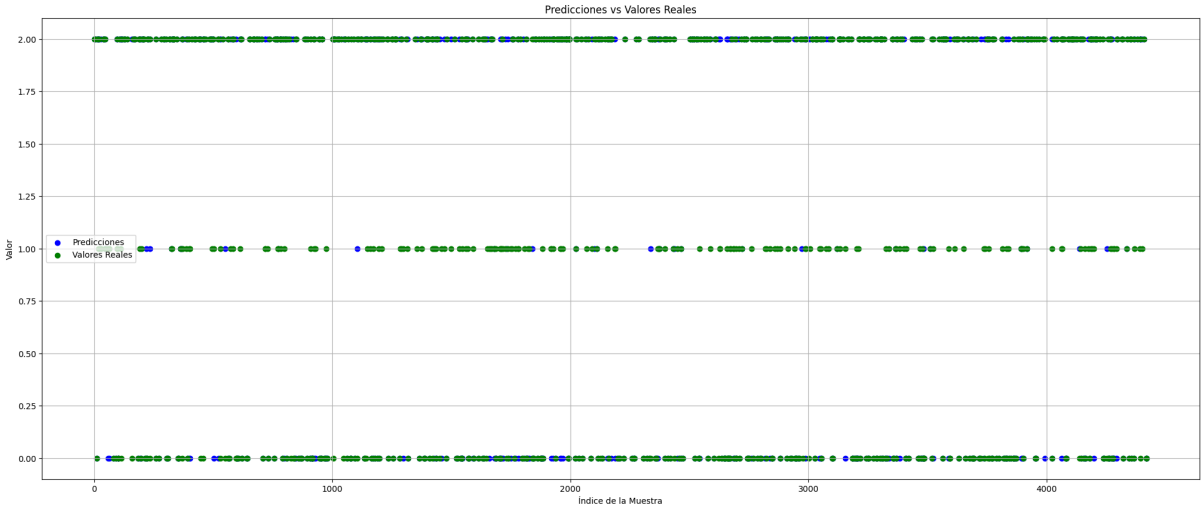


Ilustración 19.- Distribución de los valores de evaluación de predicción vs los valores reales.

Fase 5
Evaluación

5.1.- Evaluaciones con IBM SPSS Modeler

5.1.1 - Evaluación aplicando PCA

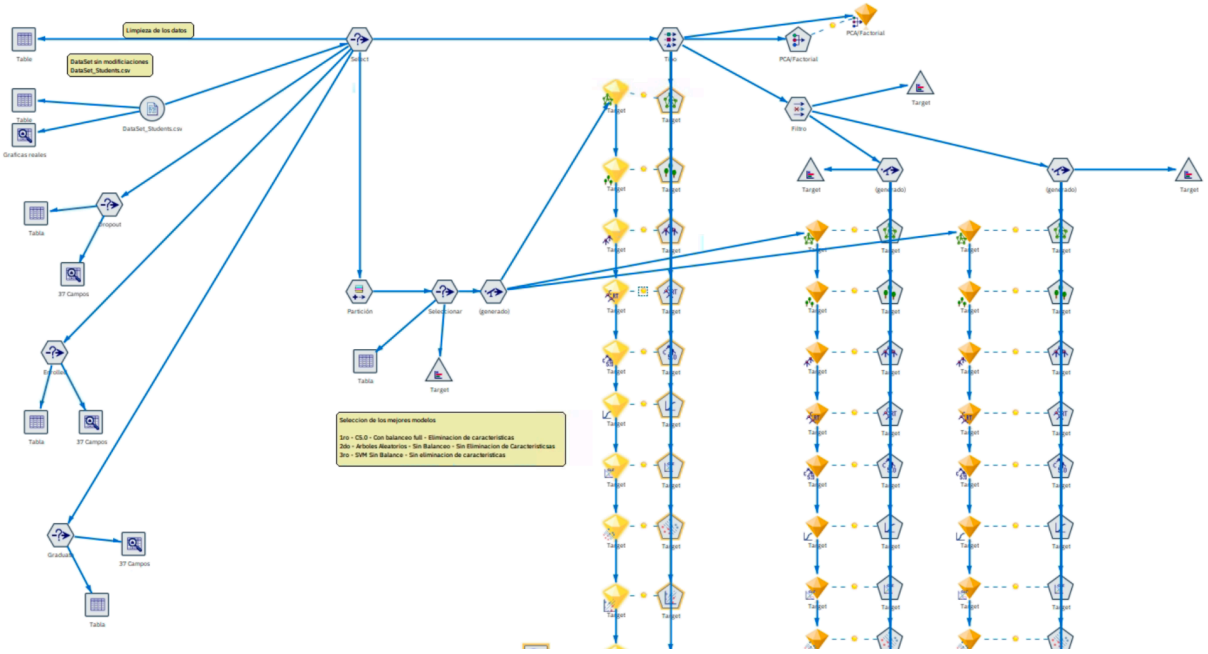


Ilustración 19.- Flujo de trabajo de IBM SPSS Modeler para generar nuevos datos.

5.1.1.1 - Evaluación de los modelos sin balanceo ni selección de características

5.1.1.1.1 - Neural Net

Comparando $N-Target con Target

'Partición'	2_Comprobación	
Correctos	1,000	78.37 %
Erróneos	276	21.63 %
Total	1,276	

Matriz de coincidencias para $N-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	347	35	52
Enrolled	57	76	87
Graduate	14	31	577

5.1.1.1.2 - Random Trees

Comparando $R-Target con Target

'Partición'	2_Comprobación	
Correctos	1,087	85.19 %
Erróneos	189	14.81 %
Total	1,276	

Matriz de coincidencias para $R-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	342	60	32
Enrolled	6	191	23
Graduate	8	60	554

5.1.1.1.3 - Tree-AS

Comparando $R1-Target con Target

'Partición'	2_Comprobación	
Correctos	993	77.82 %
Erróneos	283	22.18 %
Total	1,276	

Matriz de coincidencias para $R1-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
1.79	1	0	0
Dropout	346	29	59
Enrolled	60	68	92
Graduate	18	24	579

5.1.1.1.4 - C&R Tree

Comparando $R2-Target con Target

'Partición'	2_Comprobación	
Correctos	980	76.8 %
Erróneos	296	23.2 %
Total	1,276	

Matriz de coincidencias para $R2-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	310	49	75
Enrolled	36	59	125
Graduate	3	8	611

5.1.1.1.5 - C5.0

Comparando $C-Target con Target

'Partición'	2_Comprobación	
Correctos	1,083	84.87 %
Erróneos	193	15.13 %
Total	1,276	

Matriz de coincidencias para $C-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	379	25	30
Enrolled	45	112	63
Graduate	6	24	592

5.1.1.1.6 - Logistic

Comparando $L-Target con Target

'Partición'	2_Comprobación	
Correctos	1,012	79.31 %
Erróneos	264	20.69 %
Total	1,276	

Matriz de coincidencias para $L-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	344	43	47
Enrolled	47	85	88
Graduate	12	27	583

5.1.1.1.7 - GLE

Comparando $L1-Target con Target

'Partición'	2_Comprobación	
Correctos	1,012	79.31 %
Erróneos	264	20.69 %
Total	1,276	

Matriz de coincidencias para $L1-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	344	43	47
Enrolled	47	85	88
Graduate	12	27	583

5.1.1.1.8 - SVM

Comparando $S-Target con Target

'Partición'	2_Comprobación	
Correctos	1,067	83.62 %
Erróneos	209	16.38 %
Total	1,276	

Matriz de coincidencias para $S-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	362	34	38
Enrolled	40	117	63
Graduate	8	26	588

5.1.1.1.9 - LSVM

Comparando $L2-Target con Target

'Partición'	2_Comprobación	
Correctos	1,012	79.31 %
Erróneos	264	20.69 %
Total	1,276	

Matriz de coincidencias para $L2-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	340	47	47
Enrolled	45	87	88
Graduate	13	24	585

5.1.1.1.10 - Bayes Net

Comparando $B-Target con Target

'Partición'	2_Comprobación	
Correctos	995	77.98 %
Erróneos	281	22.02 %
Total	1,276	

Matriz de coincidencias para $B-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate	$null$
Dropout	336	55	43	0
Enrolled	34	122	64	0
Graduate	32	52	537	1

5.1.1.1.11 - Mejor modelo

	1	2	3	4	5	6	7	8	9	10
Resultado	78.37%	85.19%	77.82%	76.8%	84.87%	79.31%	79.31%	83.62%	79.31%	77.98%

5.1.1.2 - Evaluación de los modelos con balanceo y selección de características.

5.1.1.2.1 - Neural Net

Comparando $N-Target con Target

'Partición'	2_Comprobación	
Correctos	951	74.53 %
Erróneos	325	25.47 %
Total	1,276	

Matriz de coincidencias para $N-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	295	94	45
Enrolled	42	134	44
Graduate	7	93	522

5.1.1.2.2 - Random Trees

Comparando $R-Target con Target

'Partición'	2_Comprobación	
Correctos	1,064	83.39 %
Erróneos	212	16.61 %
Total	1,276	

Matriz de coincidencias para $R-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	326	70	38
Enrolled	9	192	19
Graduate	7	69	546

5.1.1.2.3 - Trees-AS

Comparando $R1-Target con Target

'Partición'	2_Comprobación	
Correctos	914	71.63 %
Erróneos	362	28.37 %
Total	1,276	

Matriz de coincidencias para $R1-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	279	113	42
Enrolled	29	151	40
Graduate	6	132	484

5.1.1.2.4 - C&R Tree

Comparando $R2-Target con Target

'Partición'	2_Comprobación	
Correctos	931	72.96 %
Erróneos	345	27.04 %
Total	1,276	

Matriz de coincidencias para $R2-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	280	107	47
Enrolled	35	138	47
Graduate	11	98	513

5.1.1.2.5 - C5.0

Comparando $C-Target con Target

'Partición'	2_Comprobación	
Correctos	1,174	92.01 %
Erróneos	102	7.99 %
Total	1,276	

Matriz de coincidencias para $C-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	398	18	18
Enrolled	8	207	5
Graduate	16	37	569

5.1.1.2.6 - Logistic

Comparando $L-Target con Target

'Partición'	2_Comprobación	
Correctos	960	75.24 %
Erróneos	316	24.76 %
Total	1,276	

Matriz de coincidencias para $L-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	290	102	42
Enrolled	31	145	44
Graduate	7	90	525

5.1.1.2.7 - GLE

Comparando $L1-Target con Target

'Partición'	2_Comprobación	
Correctos	958	75.08 %
Erróneos	318	24.92 %
Total	1,276	

Matriz de coincidencias para $L1-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	290	103	41
Enrolled	32	143	45
Graduate	9	88	525

5.1.1.2.8 - SVM

Comparando $S-Target con Target

'Partición'	2_Comprobación	
Correctos	950	74.45 %
Erróneos	326	25.55 %
Total	1,276	

Matriz de coincidencias para $S-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	283	114	37
Enrolled	27	151	42
Graduate	6	100	516

5.1.1.2.9 - LSVM

Comparando $L2-Target con Target

'Partición'	2_Comprobación	
Correctos	942	73.82 %
Erróneos	334	26.18 %
Total	1,276	

Matriz de coincidencias para $L2-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	275	116	43
Enrolled	29	147	44
Graduate	7	95	520

5.1.1.2.10 - Bayes Net

Comparando $B-Target con Target

'Partición'	2_Comprobación	
Correctos	915	71.71 %
Erróneos	361	28.29 %
Total	1,276	

Matriz de coincidencias para $B-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	289	111	34
Enrolled	29	137	54
Graduate	37	96	489

5.1.1.2.11 - Mejor modelo

	1	2	3	4	5	6	7	8	9	10
Resultado	74.53%	83.39%	71.63%	72.96%	92.01%	75.24%	75.08%	74.45%	73.82%	71.71%

5.1.1.3 - Evaluación de los modelos con balanceo submuestreo y selección de características

5.1.1.3.1 - Neural Net

Comparando $N-Target con Target

'Partición'	2_Comprobación	
Correctos	919	72.02 %
Erróneos	357	27.98 %
Total	1,276	

Matriz de coincidencias para $N-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	282	114	38
Enrolled	38	143	39
Graduate	8	120	494

5.1.1.3.2 - Random Trees

Comparando $R-Target con Target

'Partición'	2_Comprobación	
Correctos	1,016	79.62 %
Erróneos	260	20.38 %
Total	1,276	

Matriz de coincidencias para $R-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	324	72	38
Enrolled	20	176	24
Graduate	24	82	516

5.1.1.3.3 - Trees-AS

Comparando $R1-Target con Target

'Partición'	2_Comprobación	
Correctos	856	67.08 %
Erróneos	420	32.92 %
Total	1,276	

Matriz de coincidencias para $R1-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	264	134	36
Enrolled	34	157	29
Graduate	16	171	435

5.1.1.3.4 - C&R Tree

Comparando $R2-Target con Target

'Partición'	2_Comprobación	
Correctos	860	67.4 %
Erróneos	416	32.6 %
Total	1,276	

Matriz de coincidencias para $R2-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	216	170	48
Enrolled	17	171	32
Graduate	0	149	473

5.1.1.3.5 - C5.0

Comparando $C-Target con Target

'Partición'	2_Comprobación	
Correctos	973	76.25 %
Erróneos	303	23.75 %
Total	1,276	

Matriz de coincidencias para $C-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	303	87	44
Enrolled	15	159	46
Graduate	19	92	511

5.1.1.3.6 - Logistic

Comparando $L-Target con Target

'Partición'	2_Comprobación	
Correctos	953	74.69 %
Erróneos	323	25.31 %
Total	1,276	

Matriz de coincidencias para $L-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	286	109	39
Enrolled	29	150	41
Graduate	8	97	517

5.1.1.3.7 - GLE

Comparando $L1-Target con Target

'Partición'	2_Comprobación	
Correctos	953	74.69 %
Erróneos	323	25.31 %
Total	1,276	

Matriz de coincidencias para $L1-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	286	109	39
Enrolled	29	150	41
Graduate	8	97	517

5.1.1.3.8 - SVM

Comparando $S-Target con Target

'Partición'	2_Comprobación	
Correctos	936	73.35 %
Erróneos	340	26.65 %
Total	1,276	

Matriz de coincidencias para $S-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	275	121	38
Enrolled	25	152	43
Graduate	8	105	509

5.1.1.3.9 - LSVM

Comparando $L2-Target con Target

'Partición'	2_Comprobación	
Correctos	943	73.9 %
Erróneos	333	26.1 %
Total	1,276	

Matriz de coincidencias para $L2-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	276	116	42
Enrolled	29	149	42
Graduate	8	96	518

5.1.1.3.10 - Bayes Net

Comparando $B-Target con Target

'Partición'	2_Comprobación	
Correctos	912	71.47 %
Erróneos	364	28.53 %
Total	1,276	

Matriz de coincidencias para $B-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate	$null$
Dropout	286	108	38	2
Enrolled	29	142	49	0
Graduate	43	94	484	1

5.1.1.3.11 - Mejor Modelo

	1	2	3	4	5	6	7	8	9	10
Resultado	72.02%	79.62%	67.08%	67.4%	76.25%	74.69%	74.69%	73.35%	73.9%	71.47%

5.1.2 - Evaluación aplicando Neural Net

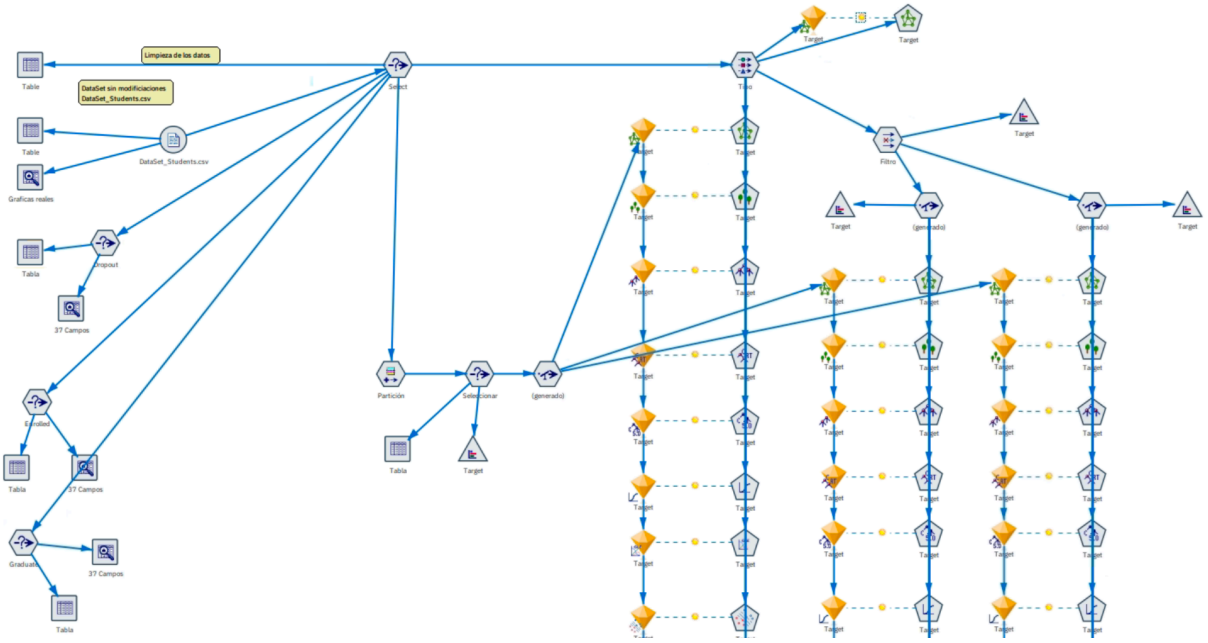


Ilustración 19.- Flujo de trabajo de IBM SPSS Modeler para generar nuevos datos.

5.1.2.1 - Evaluación de los modelos sin balanceo ni selección de características

5.1.2.1.1 - Neural Net

Comparando $N-Target con Target

'Partición'	2_Comprobación	
Correctos	1,046	78.65 %
Erróneos	284	21.35 %
Total	1,330	

Matriz de coincidencias para $N-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	367	36	42
Enrolled	62	81	105
Graduate	10	29	598

5.1.2.1.2 - Random Trees

Comparando $R-Target con Target

'Partición'	2_Comprobación	
Correctos	1,111	83.53 %
Erróneos	219	16.47 %
Total	1,330	

Matriz de coincidencias para $R-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	338	72	35
Enrolled	8	211	29
Graduate	5	70	562

5.1.2.1.3 - Tree-AS

Comparando $R1-Target con Target

'Partición'	2_Comprobación	
Correctos	1,010	75.94 %
Erróneos	320	24.06 %
Total	1,330	

Matriz de coincidencias para $R1-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	360	32	53
Enrolled	74	64	110
Graduate	17	34	586

5.1.2.1.4 - C&R Tree

Comparando $R2-Target con Target

'Partición'	2_Comprobación	
Correctos	1,000	75.19 %
Erróneos	330	24.81 %
Total	1,330	

Matriz de coincidencias para $R2-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	318	47	80
Enrolled	46	68	134
Graduate	11	12	614

5.1.2.1.5 - C5.0

Comparando $C-Target con Target

'Partición'	2_Comprobación	
Correctos	1,125	84.59 %
Erróneos	205	15.41 %
Total	1,330	

Matriz de coincidencias para $C-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	391	24	30
Enrolled	54	127	67
Graduate	10	20	607

5.1.2.1.6 - Logistic

Comparando $L-Target con Target

'Partición'	2_Comprobación	
Correctos	1,039	78.12 %
Erróneos	291	21.88 %
Total	1,330	

Matriz de coincidencias para $L-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	353	48	44
Enrolled	49	98	101
Graduate	16	33	588

5.1.2.1.7 - GLE

Comparando $L1-Target con Target

'Partición'	2_Comprobación	
Correctos	1,039	78.12 %
Erróneos	291	21.88 %
Total	1,330	

Matriz de coincidencias para $L1-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	353	48	44
Enrolled	49	98	101
Graduate	16	33	588

5.1.2.1.8 - SVM

Comparando $S-Target con Target

'Partición'	2_Comprobación	
Correctos	1,087	81.73 %
Erróneos	243	18.27 %
Total	1,330	

Matriz de coincidencias para $S-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	366	41	38
Enrolled	51	125	72
Graduate	12	29	596

5.1.2.1.9 - LSVM

Comparando $L2-Target con Target

'Partición'	2_Comprobación	
Correctos	1,028	77.29 %
Erróneos	302	22.71 %
Total	1,330	

Matriz de coincidencias para $L2-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	343	59	43
Enrolled	50	93	105
Graduate	15	30	592

5.1.2.1.10 - Bayes Net

Comparando $B-Target con Target

'Partición'	2_Comprobación	
Correctos	1,030	77.44 %
Erróneos	300	22.56 %
Total	1,330	

Matriz de coincidencias para $B-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	341	68	36
Enrolled	38	128	82
Graduate	27	49	561

5.1.2.1.11 - Mejor modelo

	1	2	3	4	5	6	7	8	9	10
Resultado	78.65%	83.53%	75.94%	75.19%	84.59%	78.12%	78.12%	81.73%	77.29%	77.44%

5.1.2.2 - Evaluación de los modelos con balanceo y selección de características

5.1.2.2.1 - Neural Net

Comparando $N-Target con Target

'Partición'	2_Comprobación	
Correctos	1,023	76.92 %
Erróneos	307	23.08 %
Total	1,330	

Matriz de coincidencias para $N-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	329	89	27
Enrolled	34	163	51
Graduate	7	99	531

5.1.2.2.2 - Random Trees

Comparando $R-Target con Target

'Partición'	2_Comprobación	
Correctos	1,151	86.54 %
Erróneos	179	13.46 %
Total	1,330	

Matriz de coincidencias para $R-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	362	57	26
Enrolled	8	208	32
Graduate	3	53	581

5.1.2.2.3 - Tree-AS

Comparando $R1-Target con Target

'Partición'	2_Comprobación	
Correctos	1,015	76.32 %
Erróneos	315	23.68 %
Total	1,330	

Matriz de coincidencias para $R1-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	304	103	38
Enrolled	26	164	58
Graduate	6	84	547

5.1.2.2.4 - C&R Tree

Comparando $R2-Target con Target

'Partición'	2_Comprobación	
Correctos	979	73.61 %
Erróneos	351	26.39 %
Total	1,330	

Matriz de coincidencias para $R2-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	288	106	51
Enrolled	33	152	63
Graduate	1	97	539

5.1.2.2.5 - C5.0

Comparando $C-Target con Target

'Partición'	2_Comprobación	
Correctos	1,235	92.86 %
Erróneos	95	7.14 %
Total	1,330	

Matriz de coincidencias para $C-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	410	21	14
Enrolled	1	244	3
Graduate	9	47	581

5.1.2.2.6 - Logistic

Comparando $L-Target con Target

'Partición'	2_Comprobación	
Correctos	1,008	75.79 %
Erróneos	322	24.21 %
Total	1,330	

Matriz de coincidencias para $L-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	315	103	27
Enrolled	33	167	48
Graduate	10	101	526

5.1.2.2.7 - GLE

Comparando $L1-Target con Target

'Partición'	2_Comprobación	
Correctos	1,008	75.79 %
Erróneos	322	24.21 %
Total	1,330	

Matriz de coincidencias para $L1-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	315	103	27
Enrolled	33	167	48
Graduate	10	101	526

5.1.2.2.8 - SVM

Comparando $S-Target con Target

'Partición'	2_Comprobación	
Correctos	1,012	76.09 %
Erróneos	318	23.91 %
Total	1,330	

Matriz de coincidencias para $S-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	314	103	28
Enrolled	30	170	48
Graduate	8	101	528

5.1.2.2.9 - LSVM

Comparando $L2-Target con Target

'Partición'	2_Comprobación	
Correctos	1,002	75.34 %
Erróneos	328	24.66 %
Total	1,330	

Matriz de coincidencias para $L2-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	309	107	29
Enrolled	30	168	50
Graduate	8	104	525

5.1.2.2.10 - Bayes Net

Comparando $B-Target con Target

'Partición'	2_Comprobación	
Correctos	969	72.86 %
Erróneos	361	27.14 %
Total	1,330	

Matriz de coincidencias para $B-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	310	101	34
Enrolled	36	148	64
Graduate	20	106	511

5.1.2.2.11 - Mejor modelo

	1	2	3	4	5	6	7	8	9	10
Resultado	76.92%	86.54%	76.32%	73.61%	92.86%	75.79%	75.79%	76.09%	75.34%	72.86%

5.1.2.3 - Evaluación de los modelos con balanceo submuestreo y selección de características

5.1.2.3.1 - Neural Net

Comparando $N-Target con Target

'Partición'	2_Comprobación	
Correctos	1,009	75.86 %
Erróneos	321	24.14 %
Total	1,330	

Matriz de coincidencias para $N-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	317	96	32
Enrolled	40	161	47
Graduate	9	97	531

5.1.2.3.2 - Random Trees

Comparando $R-Target con Target

'Partición'	2_Comprobación	
Correctos	1,063	79.92 %
Erróneos	267	20.08 %
Total	1,330	

Matriz de coincidencias para $R-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	334	84	27
Enrolled	12	213	23
Graduate	6	115	516

5.1.2.3.3 - Tree-AS

Comparando $R1-Target con Target

'Partición'	2_Comprobación	
Correctos	979	73.61 %
Erróneos	351	26.39 %
Total	1,330	

Matriz de coincidencias para $R1-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	284	115	46
Enrolled	25	163	60
Graduate	3	102	532

5.1.2.3.4 - C&R Tree

Comparando $R2-Target con Target

'Partición'	2_Comprobación	
Correctos	1,006	75.64 %
Erróneos	324	24.36 %
Total	1,330	

Matriz de coincidencias para $R2-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	300	89	56
Enrolled	36	136	76
Graduate	2	65	570

5.1.2.3.5 - C5.0

Comparando $C-Target con Target

'Partición'	2_Comprobación	
Correctos	1,081	81.28 %
Erróneos	249	18.72 %
Total	1,330	

Matriz de coincidencias para $C-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	372	53	20
Enrolled	25	189	34
Graduate	17	100	520

5.1.2.3.6 - Logistic

Comparando $L-Target con Target

'Partición'	2_Comprobación	
Correctos	1,014	76.24 %
Erróneos	316	23.76 %
Total	1,330	

Matriz de coincidencias para $L-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	319	100	26
Enrolled	35	167	46
Graduate	10	99	528

5.1.2.3.7 - GLE

Comparando $L1-Target con Target

'Partición'	2_Comprobación	
Correctos	1,014	76.24 %
Erróneos	316	23.76 %
Total	1,330	

Matriz de coincidencias para $L1-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	319	100	26
Enrolled	35	167	46
Graduate	10	99	528

5.1.2.3.8 - SVM

Comparando $S-Target con Target

'Partición'	2_Comprobación	
Correctos	997	74.96 %
Erróneos	333	25.04 %
Total	1,330	

Matriz de coincidencias para $S-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	307	108	30
Enrolled	31	169	48
Graduate	8	108	521

5.1.2.3.9 - LSVM

Comparando $L2-Target con Target

'Partición'	2_Comprobación	
Correctos	1,010	75.94 %
Erróneos	320	24.06 %
Total	1,330	

Matriz de coincidencias para $L2-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate
Dropout	315	102	28
Enrolled	32	167	49
Graduate	8	101	528

5.1.2.3.10 - Bayes Net

Comparando $B-Target con Target

'Partición'	2_Comprobación	
Correctos	957	71.95 %
Erróneos	373	28.05 %
Total	1,330	

Matriz de coincidencias para $B-Target (las filas muestran las reales)

'Partición' = 2_Comprobación	Dropout	Enrolled	Graduate	$null$
Dropout	306	101	32	6
Enrolled	39	156	53	0
Graduate	25	110	495	7

5.1.2.3.11 - Mejor modelo

	1	2	3	4	5	6	7	8	9	10
Resultado	75.86%	79.92%	73.61%	75.64%	81.28%	76.24%	76.24%	74.96%	75.94%	71.95%

5.2 - Mejores modelos de IBM SPSS Modeler

Top 3 - Modelos IBM SPSS Modeler			
Selección de Características	**Neural Net**	**PCA**	**PCA**
Configuración	Con balanceo sobremuestreo y selección de características	Con balanceo sobremuestreo y selección de características.	Sin balanceo ni selección de características
Modelo	C5.0	C5.0	Random Trees
Porcentaje de Precisión	92.86%	92.01%	85.19%

5.3 - Mejores modelos en Python

Top 3 - Modelos en Python			
Selección de Características	**N/A**	**Matriz de correlación de Pearson**	**Matriz de correlación de Pearson**
Configuración	Sin selección de características	Con selección de características.	Con selección de características
Modelo	**Logistic Regression**	**Logistic Regression**	**Random Forest Regressor**
MAE	**0.282759**	**0.291954**	**0.354149**
Porcentaje de Precisión	76.43%	75.51%	N/A

5.4 - Conclusiones del análisis de datos

Algunos puntos importantes que destacar son los siguientes:

- **Mejor rendimiento con IBM SPSS Modeler:** Tras analizar y comparar los resultados obtenidos con las herramientas de IBM SPSS Modeler y Python, se evidencia un mejor rendimiento utilizando IBM SPSS Modeler. Este hallazgo destaca la eficacia y la capacidad de la herramienta de IBM para abordar los desafíos específicos del análisis de datos en el contexto del problema estudiado. La mayor eficiencia puede atribuirse a la amplia gama de funcionalidades integradas, la facilidad de uso y la capacidad de optimización de procesos que ofrece IBM SPSS Modeler.

- **El mejor modelo y configuración encontrada:** Después de explorar diversos modelos y configuraciones, se identificó que el modelo C5.0, junto con la configuración de balanceo sobremuestreo y la eliminación de características con Neural Net, proporciona el mejor rendimiento en términos de precisión y capacidad predictiva. Este resultado resalta la importancia de realizar un análisis exhaustivo de las diferentes técnicas y configuraciones disponibles para encontrar la combinación óptima que maximice el rendimiento del modelo. Además, subraya la utilidad de las técnicas de balanceo y selección de características para mejorar la eficacia de los modelos predictivos en situaciones de desequilibrio de clases y alta dimensionalidad de los datos.

- La mayoría de los modelos tienden a seleccionar características más relacionadas con aspectos académicos y escolares, como las unidades curriculares aprobadas y las calificaciones de los semestres anteriores. Esto sugiere que el rendimiento académico previo es un factor importante en la predicción del resultado estudiantil. Sin embargo, algunos modelos también muestran una inclinación hacia características relacionadas con aspectos sociales, como el estado de becario, la ocupación de los padres y la situación de las cuotas escolares. Esto indica que factores socioeconómicos y familiares también pueden influir en la probabilidad de abandono o graduación de un estudiante. Esta combinación de características académicas y sociales resalta la complejidad y multidimensionalidad de los factores que influyen en el éxito estudiantil. Un enfoque integral que considere tanto el desempeño académico como el entorno socioeconómico podría ser crucial para desarrollar intervenciones efectivas destinadas a mejorar las tasas de retención estudiantil y el éxito académico.

- **La modalidad de aplicación como factor significativo:** La modalidad de aplicación emerge como un factor de gran influencia en la tasa de deserción estudiantil. Es importante destacar que el tipo 39, referente a los mayores de 23 años, exhibe la tasa más elevada de abandono en comparación con otros tipos de aplicación. Esta observación sugiere que los estudiantes que ingresan a la institución bajo esta modalidad enfrentan desafíos únicos o presentan características distintas que los hacen más propensos a abandonar sus estudios. Este hallazgo resalta la importancia de comprender las necesidades y dificultades específicas de los estudiantes que acceden a la educación superior bajo esta modalidad, con el fin de implementar estrategias efectivas de retención.

- **Impacto positivo de las becas en la retención de estudiantes:** Se ha identificado que los estudiantes beneficiarios de becas presentan tasas de abandono significativamente inferiores en comparación con aquellos que no reciben apoyo financiero. Este resultado sugiere que las becas tienen un efecto positivo en la retención de estudiantes, posiblemente al reducir las barreras económicas y proporcionar un estímulo adicional para la permanencia en los estudios. Es fundamental reconocer el papel crucial que desempeñan las políticas de becas en la promoción de la equidad y el acceso a la educación superior, así como en la reducción de las disparidades en términos de finalización de estudios.

- **Disparidades de género en la finalización de estudios:** A pesar de no observarse diferencias significativas en la tasa de deserción entre hombres y mujeres, se destaca una disparidad en el caso de los graduados. Aquí, se observa una proporción más alta de graduadas en comparación con graduados, lo que indica una tendencia favorable para el género femenino en términos de completar sus estudios. Este hallazgo sugiere que las mujeres muestran una mayor resiliencia o compromiso con la finalización de sus estudios universitarios en comparación con los hombres. Es esencial investigar más a fondo los factores que contribuyen a esta disparidad de género y desarrollar estrategias que fomenten la igualdad de oportunidades y la retención de estudiantes de todos los géneros en la educación superior.

Puntos clave para una posible solución:

Modalidad de Aplicación:

- Realizar un análisis exhaustivo de las características y necesidades específicas de los estudiantes que optan por la modalidad de aplicación tipo 39 (Mayores de 23 años) para comprender mejor los factores que contribuyen a su alta tasa de deserción.

- Implementar programas de orientación y apoyo específicamente diseñados para este grupo de estudiantes, que aborden sus necesidades académicas, sociales y emocionales.

- Ofrecer oportunidades de aprendizaje flexibles y adaptadas a sus horarios y responsabilidades laborales o familiares, lo que podría mejorar su compromiso y retención.

Becas y Ayudas Financieras:

- Ampliar y mejorar los programas de becas y ayudas financieras disponibles para los estudiantes, especialmente aquellos en riesgo de abandonar sus estudios debido a dificultades económicas.

- Proporcionar asesoramiento financiero y apoyo adicional para ayudar a los estudiantes a administrar mejor sus recursos financieros y afrontar los costos asociados con la educación universitaria.

- Implementar incentivos adicionales o beneficios para los becarios que mantienen un buen desempeño académico y una asistencia regular, fomentando así su compromiso y motivación para completar sus estudios.

Equidad de Género:

- Continuar promoviendo la equidad de género en el acceso a la educación superior y en las oportunidades académicas y profesionales.

- Implementar programas y políticas que aborden los posibles sesgos de género en el aula y en el entorno universitario, promoviendo un ambiente inclusivo y respetuoso para todos los estudiantes.

- Ofrecer programas de mentoría y apoyo específicamente diseñados para mujeres en áreas académicas subrepresentadas, lo que podría ayudar a fomentar su éxito académico y profesional y reducir las disparidades de género en la graduación.

Referencias

1. Rochin Berumen, F. L. (2021). Deserción escolar en la educación superior en México: revisión de literatura. RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo, 11(22).

2. UCI machine learning repository. (s/f). Uci.edu. Recuperado el 18 de febrero de 2024, de https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

3. (S/f). Ibm.com. Recuperado el 18 de febrero de 2024, de https://www.ibm.com/docs/es/SS3RA7_18.3.0/pdf/ModelerApplications.pdf

4. scikit-learn user guide. (s/f). Scikit-learn.org. Recuperado el 18 de febrero de 2024, de https://scikit-learn.org/0.22//_downloads/scikit-learn-docs.pdf

5. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer New York.

6. The Devastator. (2023). Predict students' dropout and academic success [Data set].

7. Infórmese, C. A.-B. [@CALAAnalytics]. (2022, mayo 12). Machine Learning con IBM SPSS Modeler. Youtube. https://www.youtube.com/watch?v=VatQZrkvDds

8. Fürnkranz, J., Chan, P. K., Craw, S., Sammut, C., Uther, W. T. B., Ratnaparkhi, A., Jin, X., Han, J., Yang, Y., Morik, K., Dorigo, M., Birattari, M., Stützle, T., Brazdil, P., Vilalta, R., Giraud-Carrier, C., Soares, C., Rissanen, J., Baxter, R., … De Raedt, L. (2010). Mean Absolute Error. 652. https://doi.org/10.1007/978-1-4899-7687-1_953

9. Thomas, T. (2023, septiembre 8). Mean_absolute_error() or mean_squared_error() or explained_variance_score(). Bitstrapped. https://medium.com/bitstrapped/mean-absolute-error-or-mean-squared-error-or-explained-variance-score-f2872b1800ba