82.05 - Análisis Predictivo

05/07/2023

CASO DE NEGOCIO

PREDICCIÓN DE DROPOUT DE ESTUDIANTES

MANUEL SAÚL HANONO

AGENDA

01 Introducción
Desafíos y objetivos del trabajo, hipótesis planteada y composición de la base.

02 Análisis Exploratorio
EDA, tratamiento de columnas, imputaciones y codificación de variables. Gráficos exploratorios.

03 Selección de Modelos
Modelos utilizados, comparaciones y proceso de selección.

04 Conclusiones
Conclusiones que se obtienen del desarrollo del modelo y la implementación del caso de negocios.

OBJETIVO

El objetivo de este caso de estudio será entrenar un modelo para poder **predecir si un alumno del Tecnológico de Monterrey continuará o no con sus estudios.**
Este modelo servirá luego para asesorar la institución a realizar un *correcto seguimiento de potenciales* *alumnos a perder*, con el fin de **disminuir** al máximo posible el dropout.
Asimismo, de tener una implementación exitosa, se podría expandir su uso hacía otras instituciones.

HIPÓTESIS

- *¿Hay relación significativa entre las características (académicas y personales) de los estudiantes y su tendencia al abandono universitario?*

- *¿Existe una relación significativa entre el pasado de los estudiantes y el dropout?*

Tecnológico de Monterrey

La base contiene información sobre estudiantes del Tecnológico de Monterrey. Contiene, en total, **143.326** registros con **50** variables.

student.id	Integer	Encoded enrollment number of the student	1-121584
generation	String	Unique indicator that denotes the Generation	AD14, AD15, AD16, AD17, AD18, AD19,
educational.model	Binary	Educational model to which the student	1: TEC21 Model, 0: No TEC21 Model
level	String	Level of studies to which the student belongs	High School, Undergraduate
gender	String	Student's gender	Male, Female
age	Integer	Student's age	Range from 13 to 55 years
max.degree.parents	String	Highest level of studies obtained by the	No information, No degree, Undergraduate
father.education.complete	String	Description of the last level of studies	Attended university, but did not graduate;
father.education.summary	String	Last level of studies completed by the father	No information, No degree, Undergraduate
mother.education.complete	String	Description of the last level of studies	Attended university, but did not graduate;
mother.education.summary	String	Last level of studies completed by the	No information, No degree, Undergraduate
parents.exatec	String	Indicator that denotes if either one of the two	Yes, No, No information
father.exatec	String	Indicator that denotes if the student's father	Yes, No, No information
mother.exatec	String	Indicator that denotes if the student's mother	Yes, No, No information
tec.no.tec	String	Indicator that denotes if the student comes	TEC, NO TEC
foreign	String	Indicator to identify if the student is a	Local, Yes: National, Yes: Foreigner
zone.type	String	Description of the type of zone to which the	Rural, Semiurban, Urban, No information
first.generation	String	It indicates that the student is the first person	Yes, No, No information, Does not apply
school	String	Acronyms of the school to which the	High school, EN = Business school, EMCS =
program	String	Acronyms of the academic program to which	The meaning of the acronyms can be found
region	String	Code of the region to which the campus	RM = Monterrey Region, RO = West Region,
PNA	Float	Previous level score (average)	Range from 0 to 100
admission.test	Integer and String	Admission test score	Range from 0 to 1600, Does not apply
online.test	Binary	If the student took the admission test online	1: YES, 0: NO
english.evaluation	Integer		Level 0: No information, Level 1: Beginner,

admission.rubric	Integer	Score generated from the student profile	Range from 0 to 50
general.math.eval	Float and String	Mathematics grade from the admission test	Range from 0 to 100, Does not apply, No
retention	Binary	Value that indicates if the student continues	1: Retention, 0: Dropout
FTE	Float	Indicates if the student is a full-time student	Range from 0.04 to 1.44
scholarship.perc	Float	Scholarship percentage	Range from 0 to 1
scholarship.type	String	Scholarship type	Academic talent, Army/Navy scholarship,
loan.perc	Float	Percentage of scholarship loan	Range from 0 to .50
total.scholarship.loan	Float	Total percentage of scholarship	Range from 0 to 1
school.cost	String	Cost level of the student's tuition from the	Public, Low cost, Medium cost, Medium high
id.school.origin	String	Encoded identifier of the school where the	
socioeconomic.level	String	Socioeconomic level	Level 1, Level 2, Level 3, Level 4, Level 5,
social.lag	String	Social Gap Index	Low, Medium, High, No information
average.first.period	Float	Average obtained in the first period	Range from 0 to 100
failed.subject.first.period	Integer	Number of subjects failed in the first period	Range from 0 to 8
dropped.subject.first.period	Integer	Number of subjects dropped out in the first	Range from 0 to 9
dropout.semester	Integer	Value that indicates the semester when the	0,1,2,3,4
physical.education	Binary and String	Value that indicates if the student enrolled in	0, 1, Does not apply, No information
cultural.diffusion	Binary and String	Value that indicates if the student enrolled in	0, 1, Does not apply, No information
student.society	Binary and String	Value that indicates if the student enrolled in	0, 1, Does not apply, No information
total.life.activities	Integer and String	Number of LIFE (student leadership and	0, 1, 2, 3, 4, 5, Does not apply, No
athletic.sports	Binary and String	Value that indicates if the student enrolled in	0, 1, Does not apply, No information
art.culture	Binary and String	Value that indicates if the student enrolled in	0, 1, Does not apply, No information
student.society.leadership	Binary and String	Value that indicates if the student enrolled in	0, 1, Does not apply, No information
life.work.mentoring	Binary and String	Value that indicates if the student received	0, 1, Does not apply, No information
wellness.activities	Binary and String	Value that indicates if the student enrolled in	0, 1, Does not apply, No information

PREPARACIÓN DE LA BASE

➔ Se ha verificado que las variables no contienen nulos. Están categorizados por defecto.

```
df['first.generation'].value_counts()

Does not apply     65809
No information     37372
No                 34752
Yes                 5393
```

```
df['mother.exatec'].value_counts()

No                 94020
No information     24904
Yes                24402
```

```
Undergraduate degree     49888
No information           49351
Master degree            22860
No degree                17667
PhD                       3560
Name: father.education.summary, dtype: int64
```

```
Undergraduate degree     53453
No information           50458
No degree                24741
Master degree            12892
PhD                       1782
Name: mother.education.summary, dtype: int64
```

PREPARACIÓN DE LA BASE

➔ Se ha verificado que la base no contiene duplicados. Si hay alumnos.

```
df['student.id'].duplicated().sum()
```

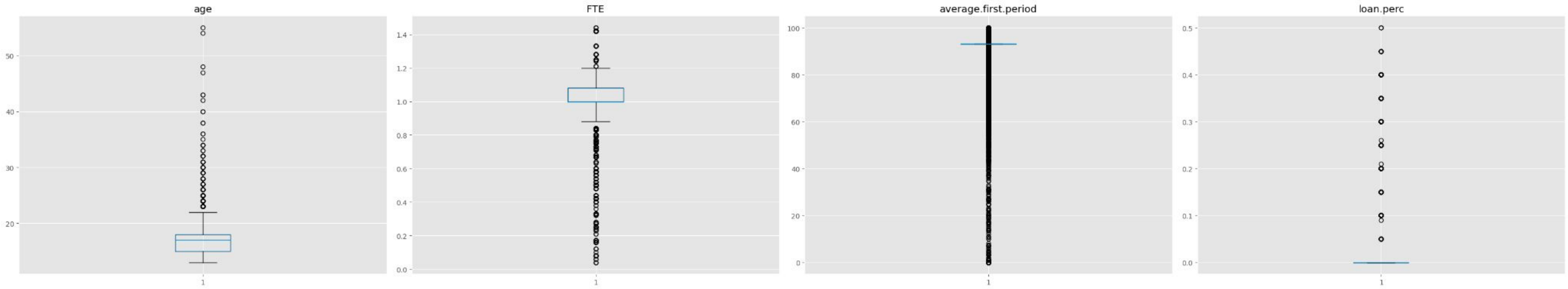
```
21742
```

```
df['student.id'].value_counts().value_counts()
```

```
1     99877
2     21672
3        35
```

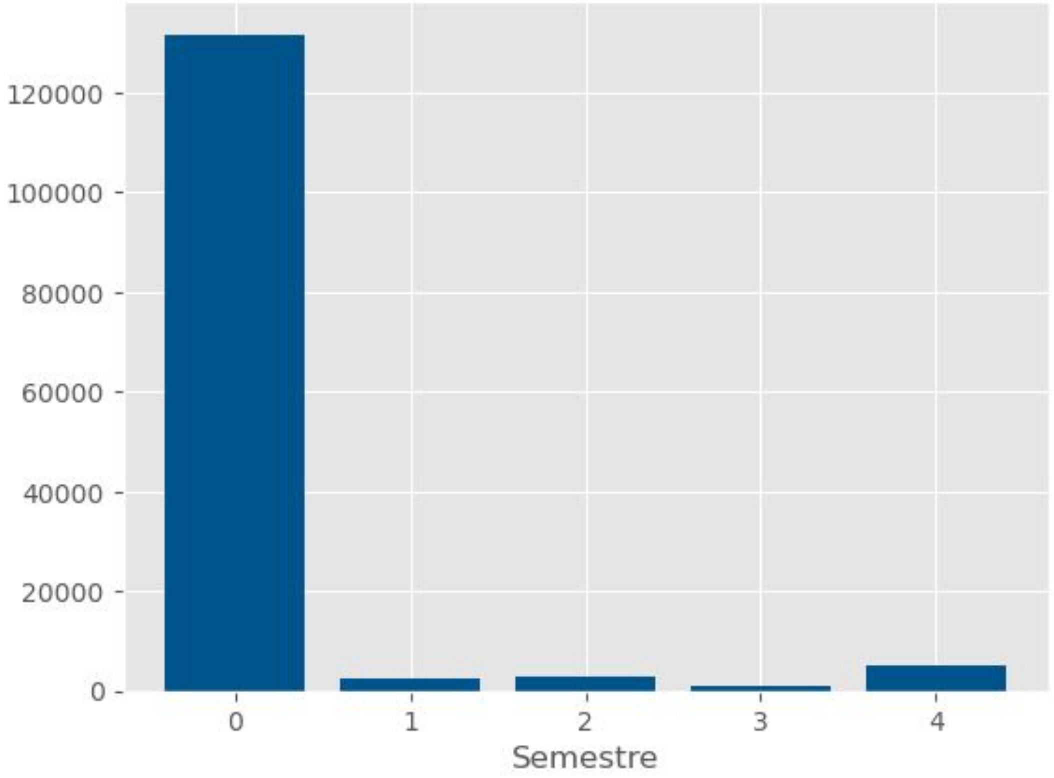
PREPARACIÓN DE LA BASE

➔ Tratamiento de Outliers ⟶ **No** se excluyen.



PREPARACIÓN DE LA BASE

➔ Se han ~~eliminado~~ columnas que afectan el estudio (*dropout.semester*).

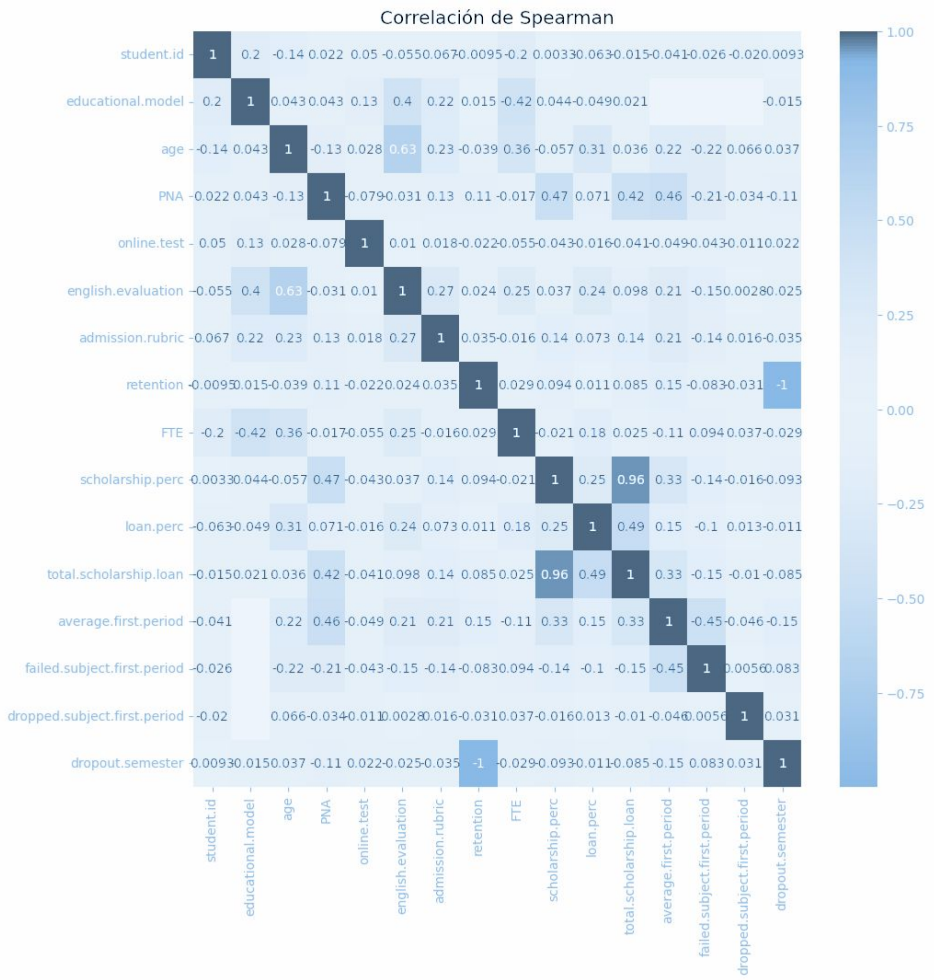


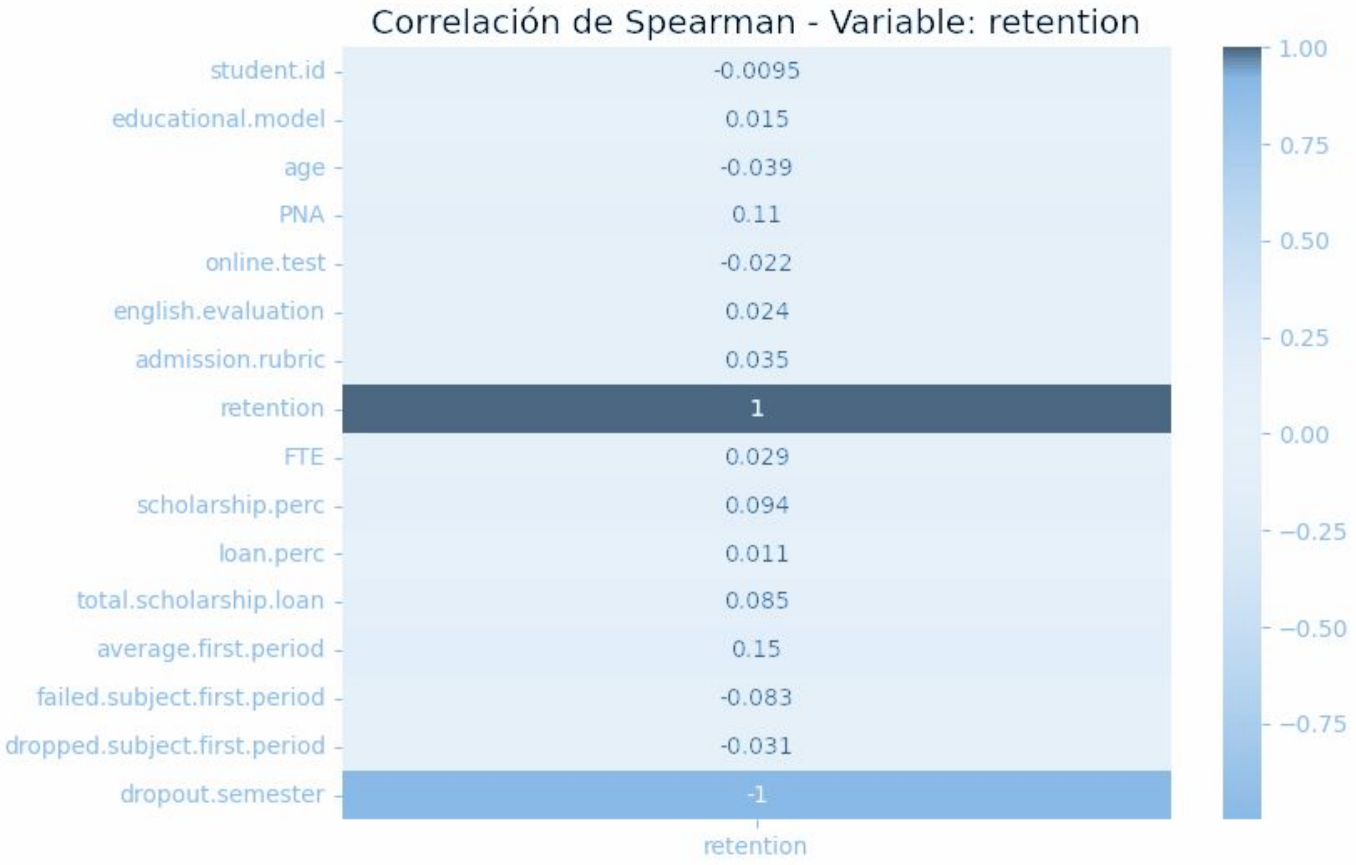
PREPARACIÓN DE LA BASE

→ Pipeline para **transformar** variables categóricas a númericas.

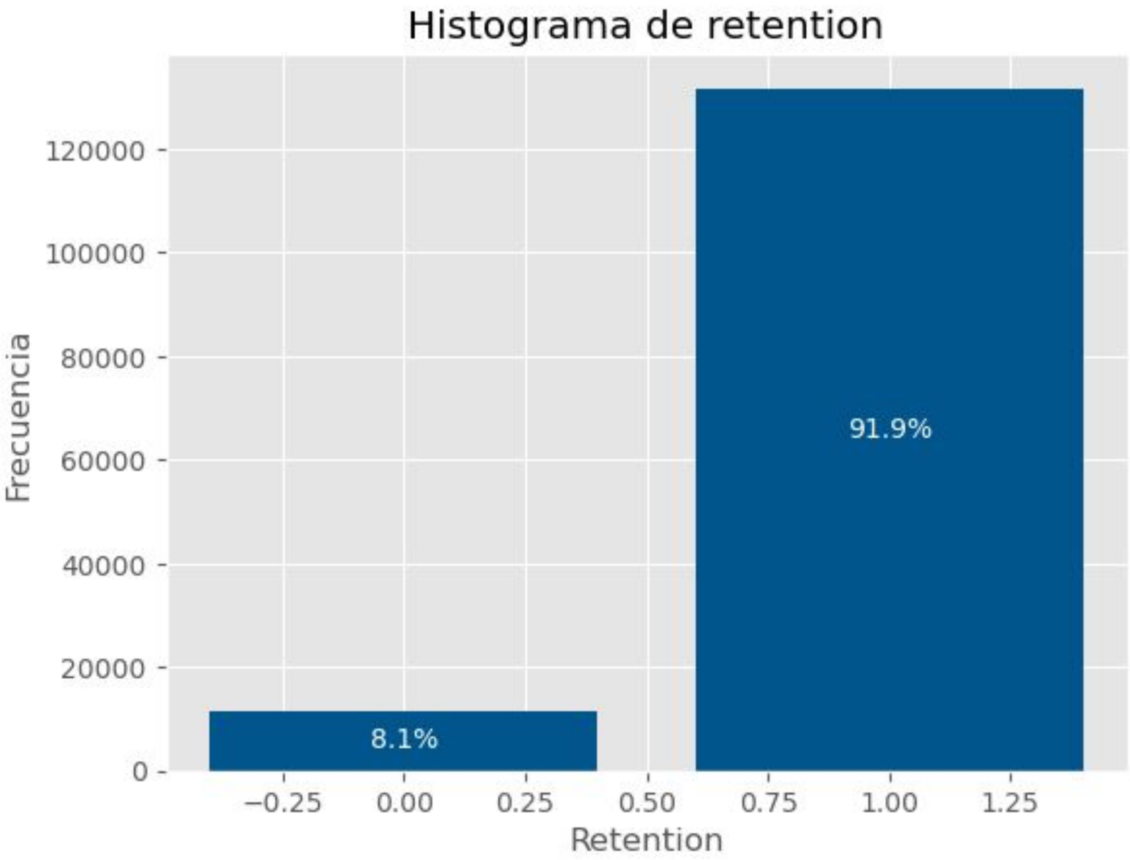
```
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OrdinalEncoder
from sklearn.compose import make_column_transformer

categorical_columns = [c for c in df.columns if df[c].dtype == 'object']

categorical_pipeline = Pipeline([
    ('encoder', OrdinalEncoder())
])


df2 = df

df2[categorical_columns] = categorical_pipeline.fit_transform(df2[categorical_columns])
```

ANÁLISIS EXPLORATORIO: VARIABLES NUMÉRICAS



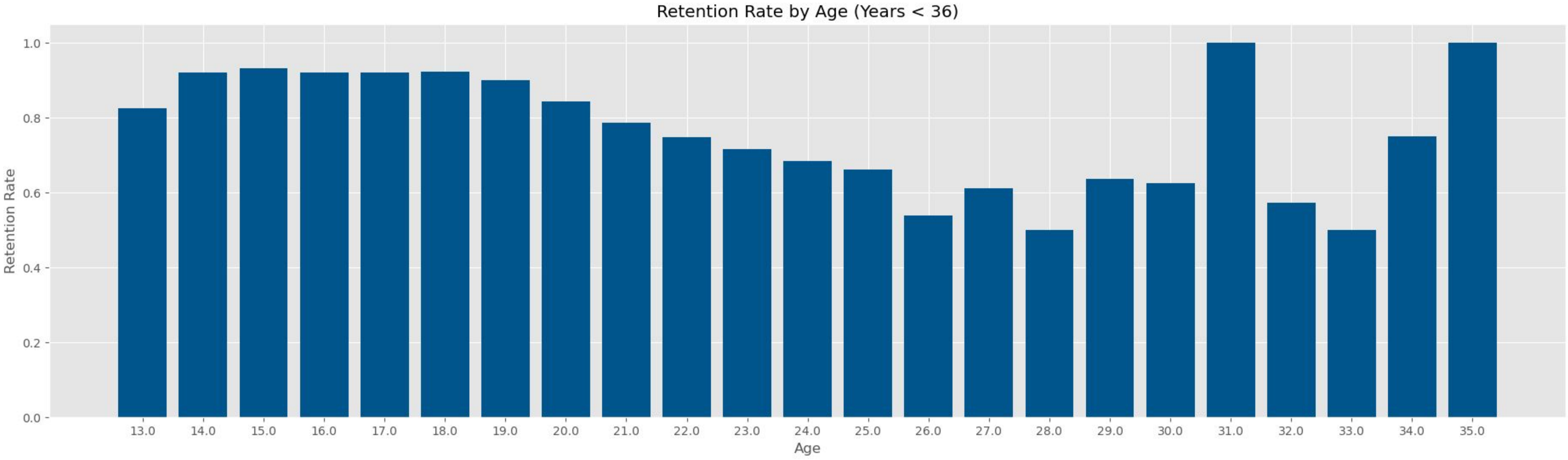


ANÁLISIS EXPLORATORIO: VARIABLES CATEGÓRICAS

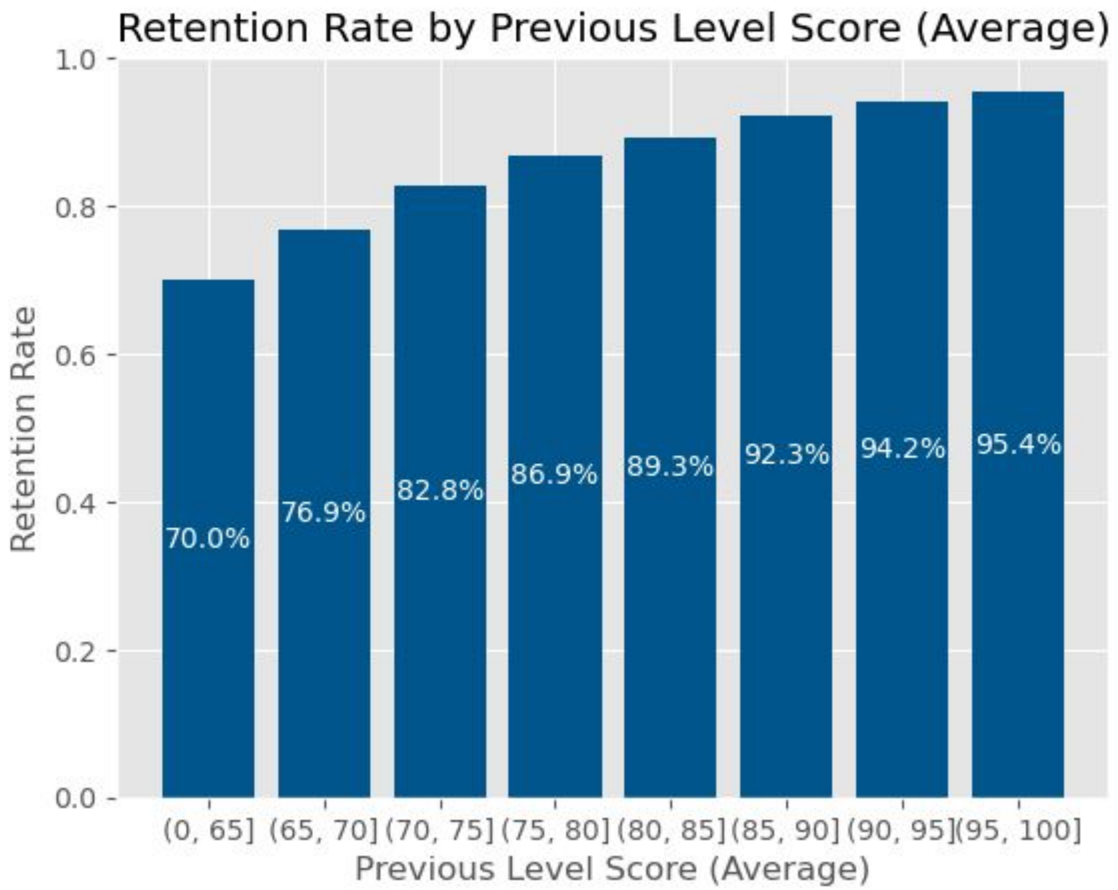


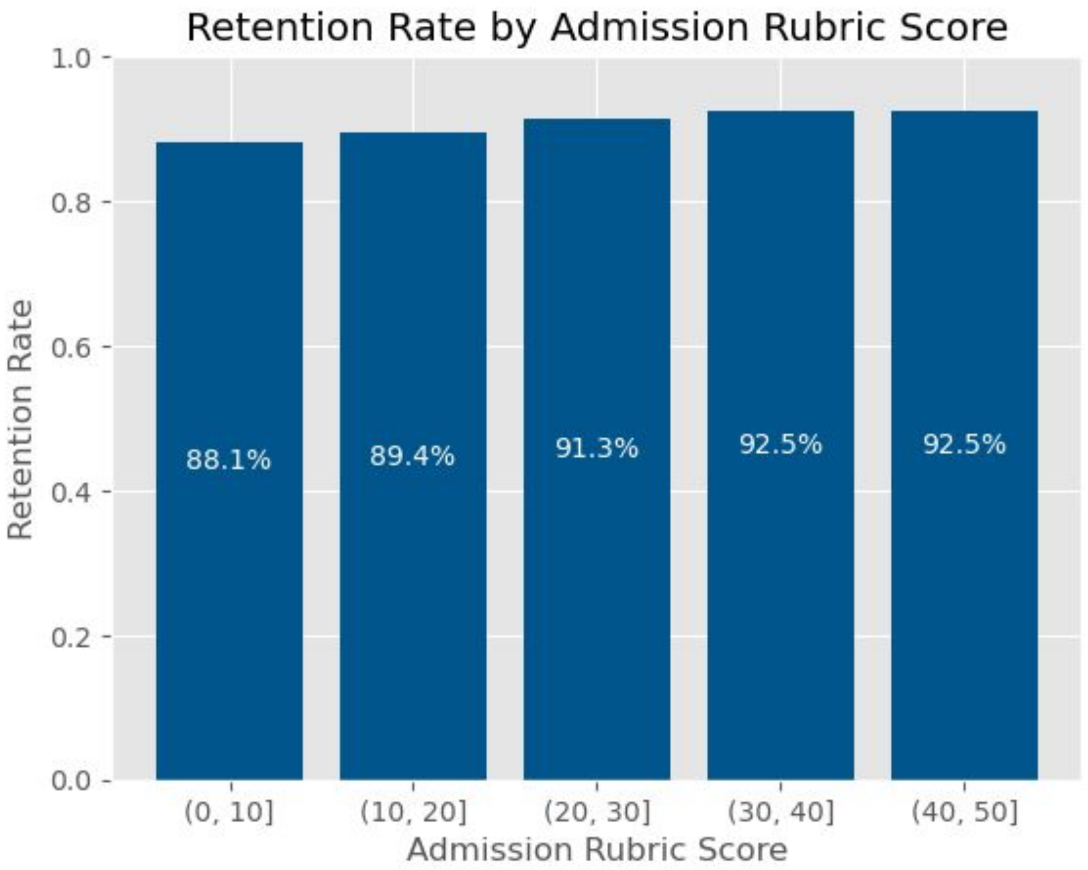

```
Cramer's V correlation between generation and retention: 0.025938607398461 31
Cramer's V correlation between level and retention: 0.02639986988238043
Cramer's V correlation between gender and retention: 0.012296088153849912
Cramer's V correlation between max.degree.parents and retention: 0.0794213829509962
Cramer's V correlation between father.education.complete and retention: 0.07649117325604864
Cramer's V correlation between father.education.summary and retention: 0.07601948372223484
Cramer's V correlation between mother.education.complete and retention: 0.07581956584941359
Cramer's V correlation between mother.education.summary and retention: 0.07426729142879385
Cramer's V correlation between parents.exatec and retention: 0.07864540379499366
Cramer's V correlation between father.exatec and retention: 0.07379621037124894
Cramer's V correlation between mother.exatec and retention: 0.07353034527637337
Cramer's V correlation between tec.no.tec and retention: 0.014494978884522518
Cramer's V correlation between foreign and retention: 0.03853038380984946
Cramer's V correlation between zone.type and retention: 0.017348171015218437
Cramer's V correlation between first.generation and retention: 0.0371958009012185
Cramer's V correlation between school and retention: 0.04004027357662622
Cramer's V correlation between program and retention: 0.06783794329688755
Cramer's V correlation between region and retention: 0.05279574167331518
Cramer's V correlation between admission.test and retention: 0.08100340957366693
Cramer's V correlation between general.math.eval and retention: 0.09301594838688339
Cramer's V correlation between scholarship.type and retention: 0.10351324982762634
Cramer's V correlation between school.cost and retention: 0.07105467118644741
Cramer's V correlation between id.school.origin and retention: 0.25060008451322846
Cramer's V correlation between socioeconomic.level and retention: 0.012560278773612 49
Cramer's V correlation between social.lag and retention: 0.015346450397523814
Cramer's V correlation between physical.education and retention: 0.2254809273253195
Cramer's V correlation between cultural.diffusion and retention: 0.22426607131596074
Cramer's V correlation between student.society and retention: 0.22512449832137893
Cramer's V correlation between total.life.activities and retention: 0.11409358526993703
Cramer's V correlation between athletic.sports and retention: 0.10695832739378586
Cramer's V correlation between art.culture and retention: 0.10346600459862106
Cramer's V correlation between student.society.leadership and retention: 0.10327171934621902
Cramer's V correlation between life.work.mentoring and retention: 0.10109344314961677
Cramer's V correlation between wellness.activities and retention: 0.10372456880100855
```

RETENTION & AGE

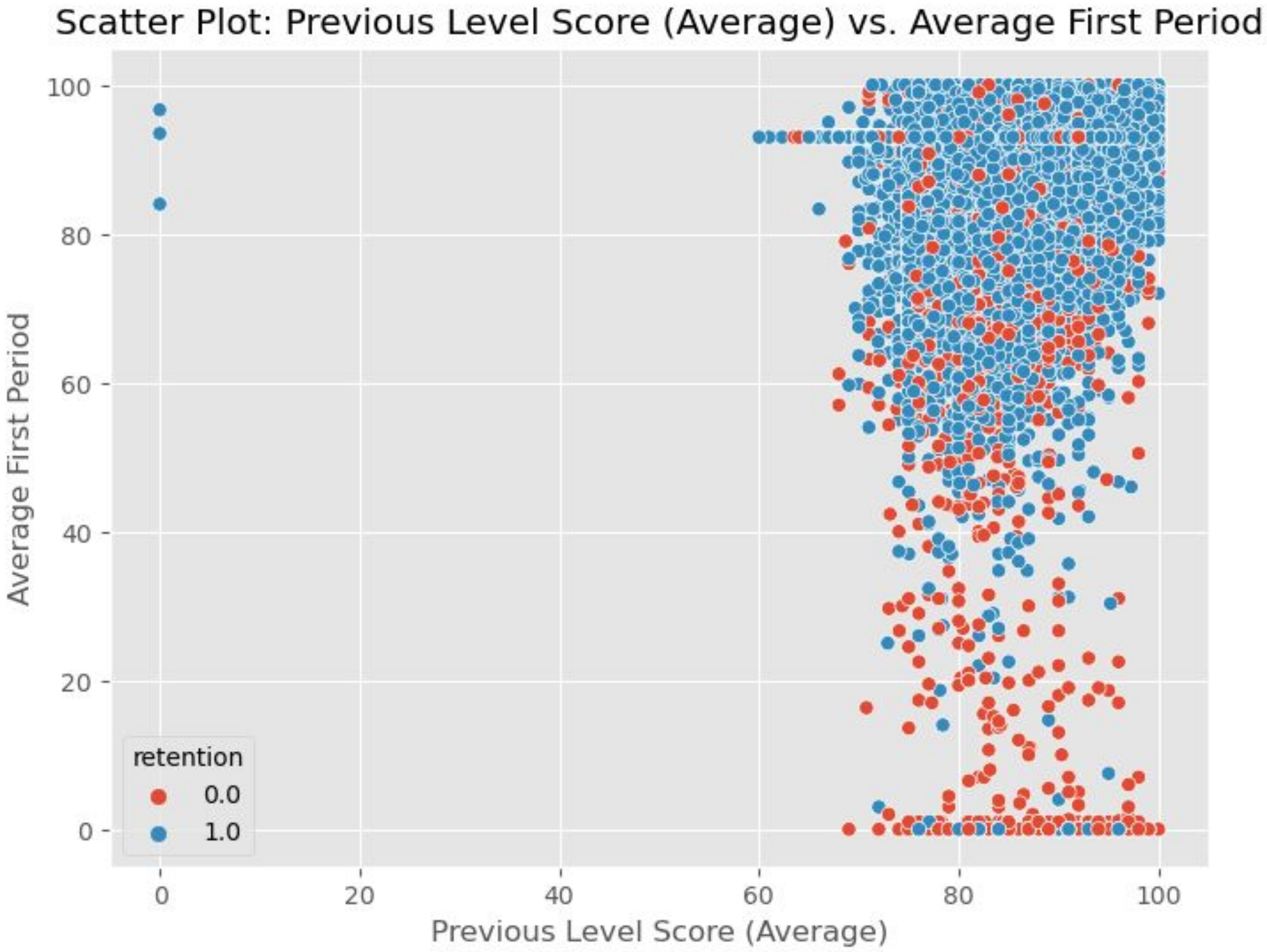


RETENTION & PREVIOUS PERFORMANCE

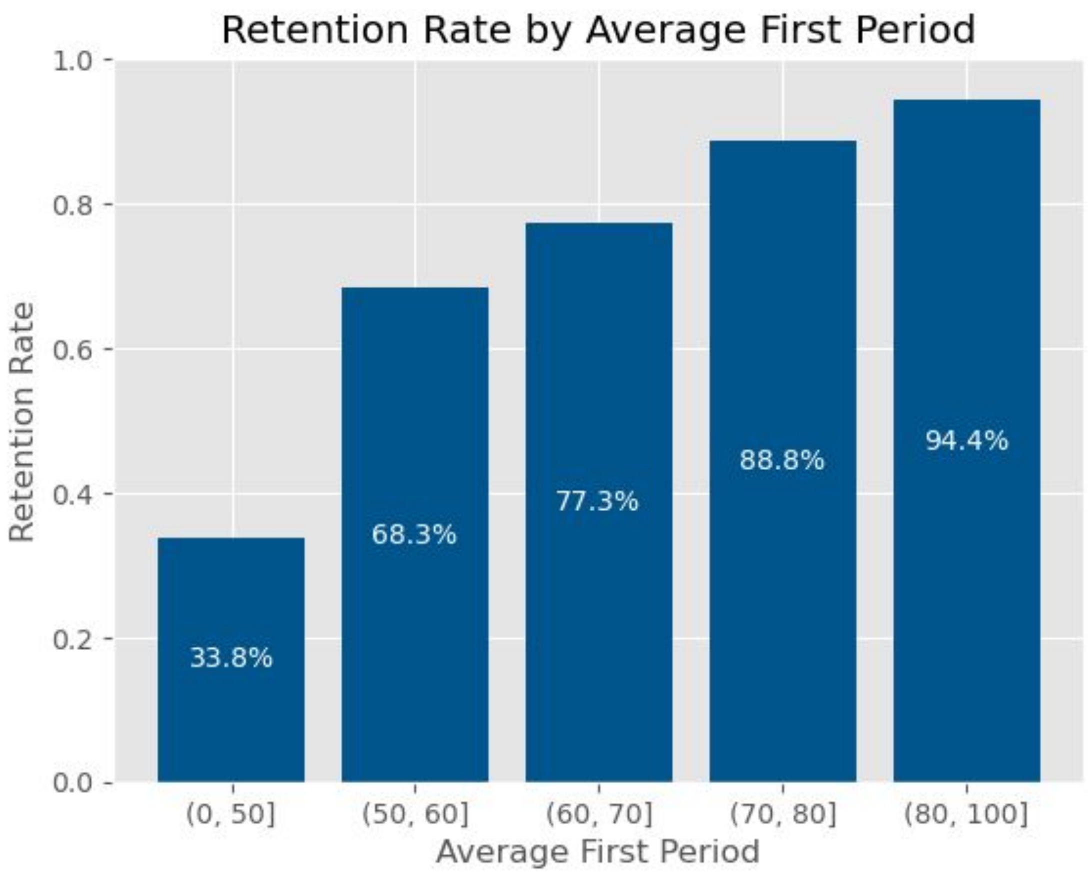


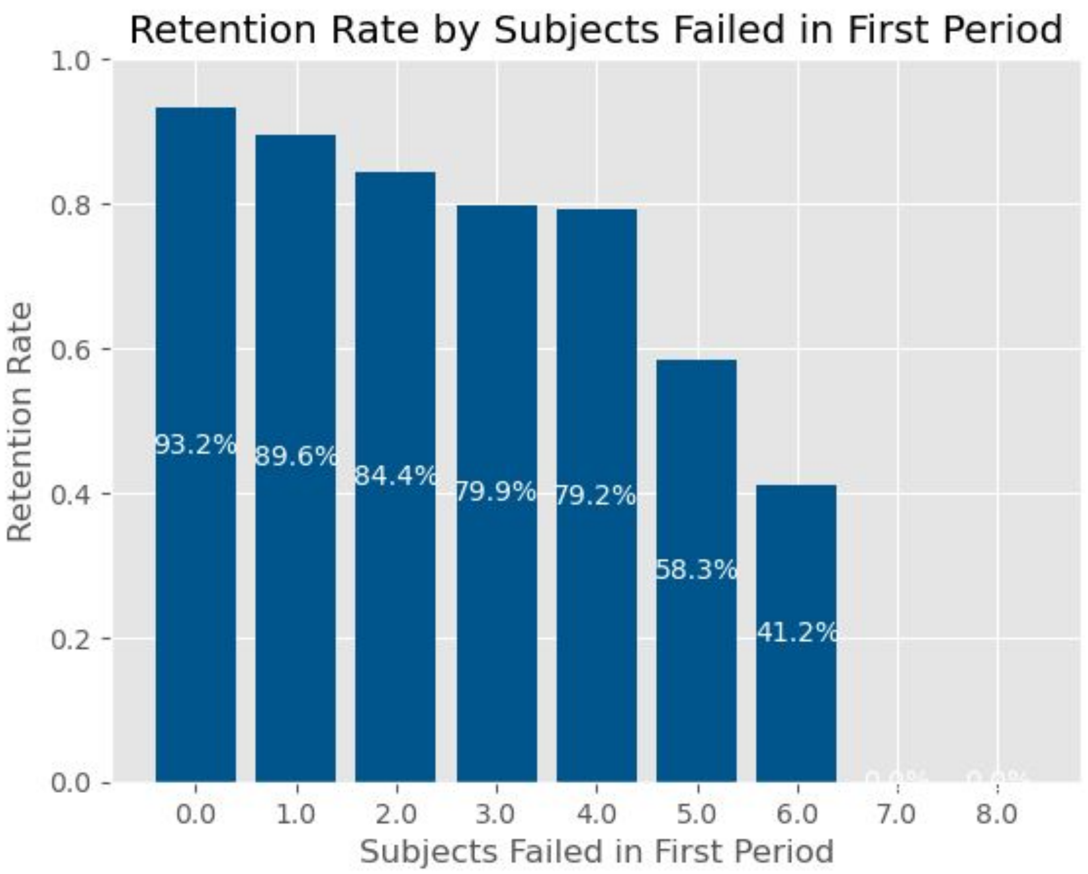


FIRST PERIOD & PREVIOUS PERFORMANCE

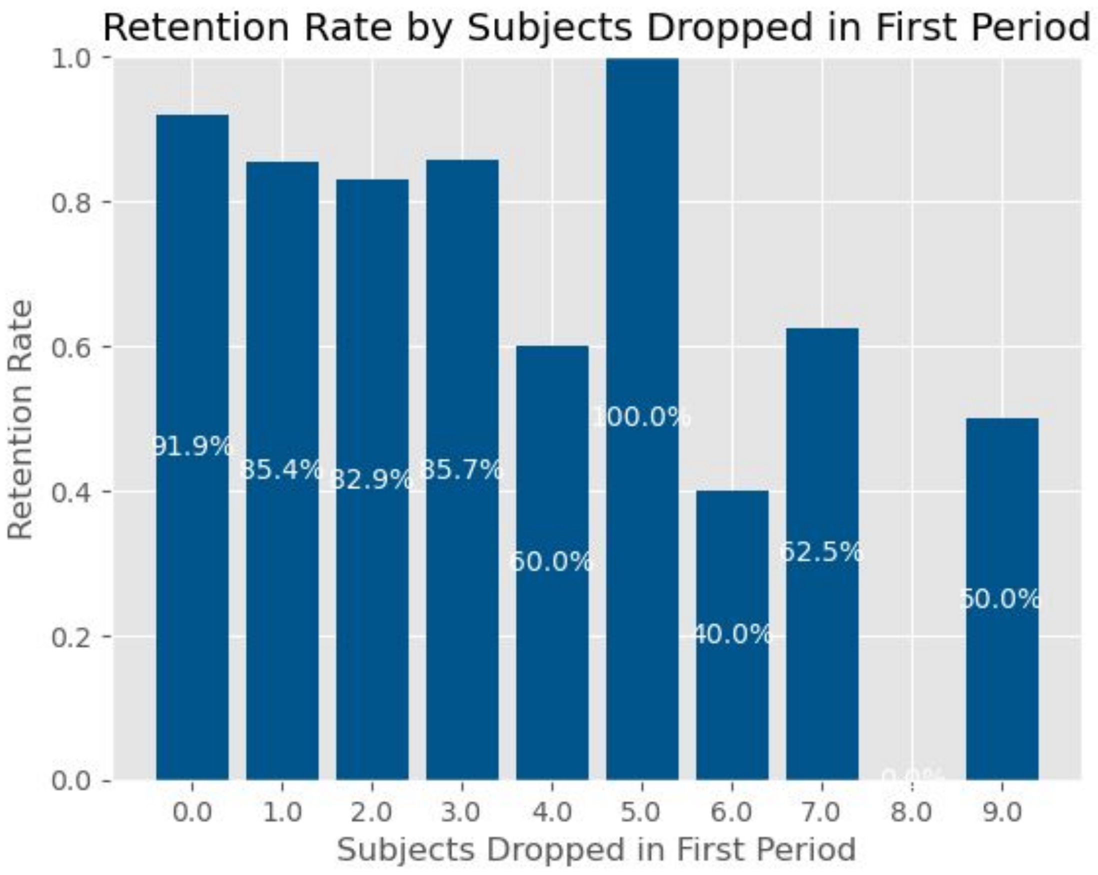


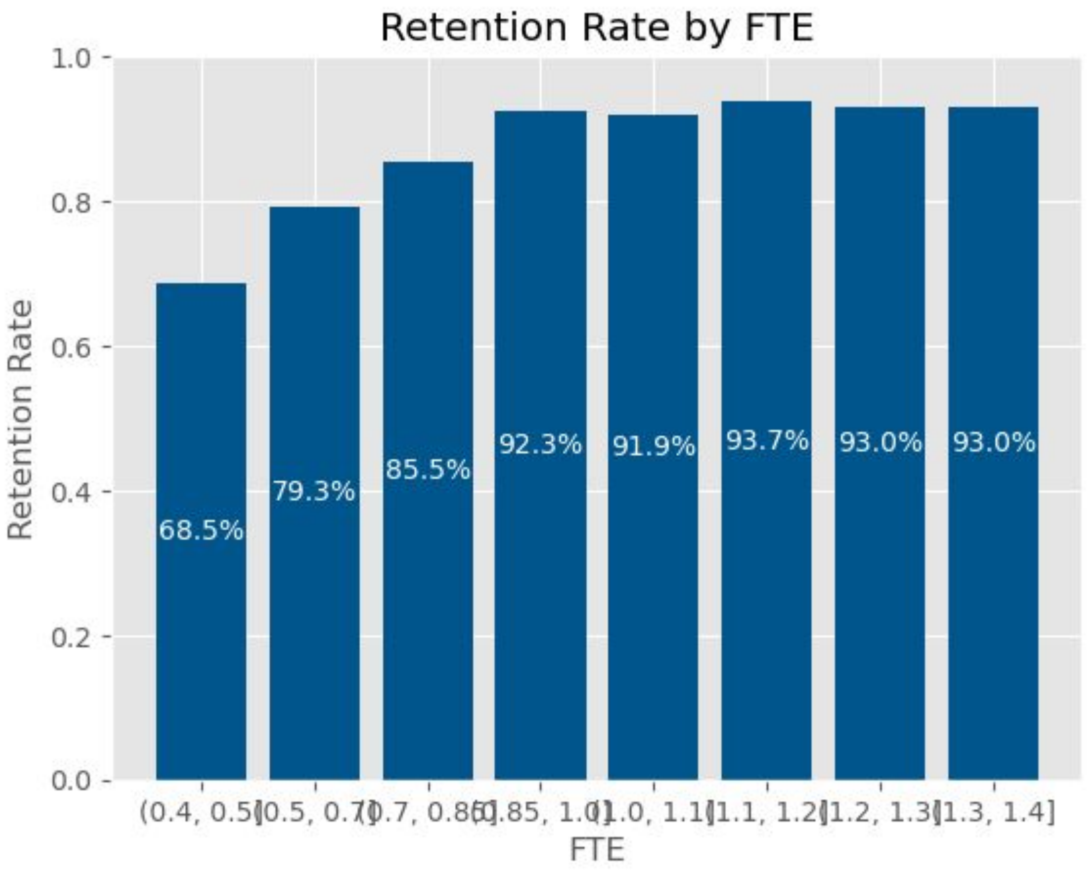
RETENTION & FIRST PERIOD PERFORMANCE



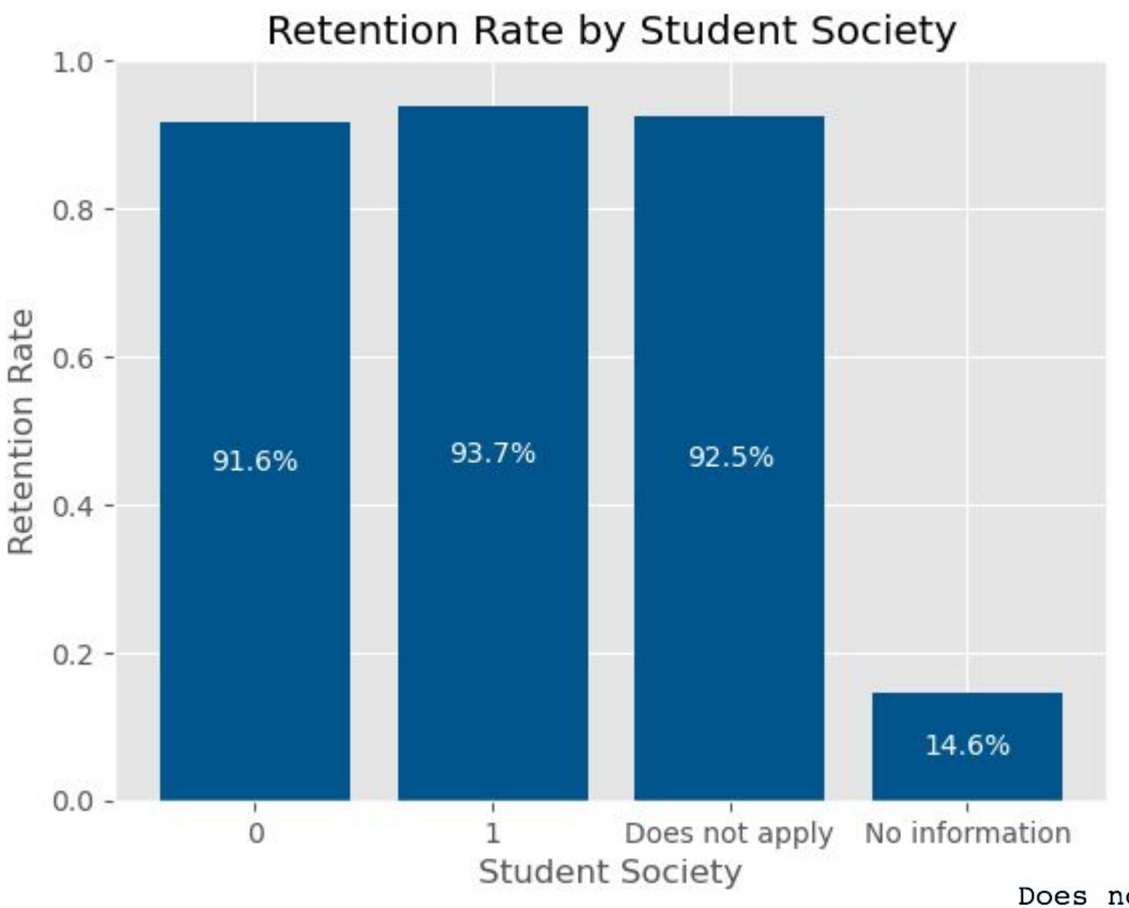


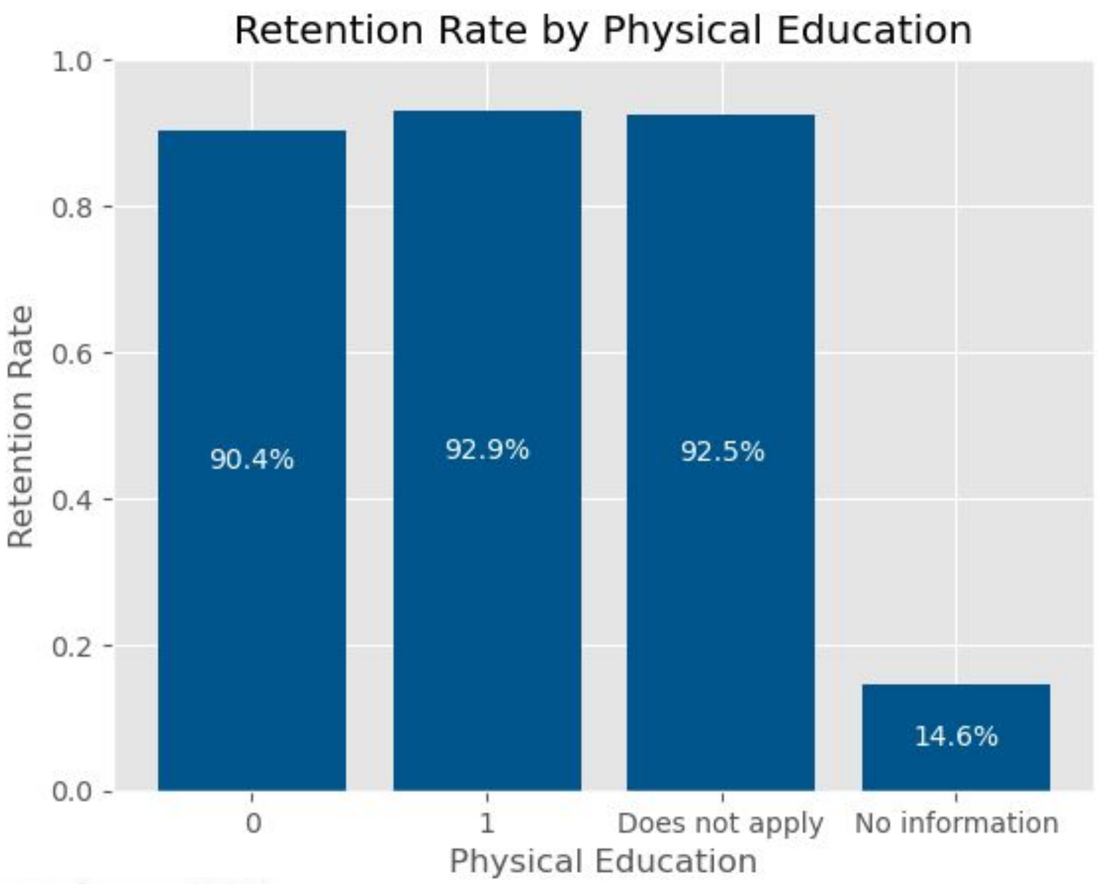
RETENTION & FIRST PERIOD PERFORMANCE





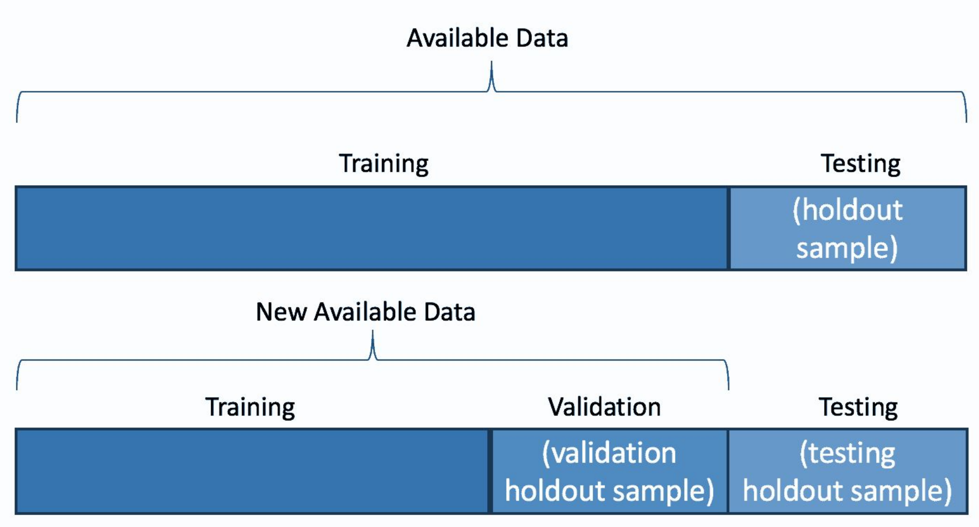
RETENTION & STUDENT LIFE





Does not apply	64611
0	52710
1	25115
No information	890

SPLIT DE LA BASE



1. Se divide la base en 80-20 de manera **estratificada**, generando una base *train* y una base *val*.
2. Se divide en 80-20 la base *train*, obteniendo una de *train* y otra de *test* para evaluar a los modelos que se van entrenando.
3. Se utiliza *val* para verificar los resultados de los modelos entrenados en una *base desconocida*.

```
X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2, stratify = y, random_state=1411)
val = pd.concat([X_val, y_val], axis=1)

train = pd.concat([X_train, y_train], axis=1)

X = train.drop('retention', axis=1)
y = train['retention']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, stratify=y, random_state=141102)
```

MODELOS

Se realizaron pruebas con los siguientes modelos, siempre teniendo en cuenta que es un problema de clasificación:

1. Regresión Logística
2. Random Forest
3. XGBoost
4. ExtraTrees
5. RUSBoost

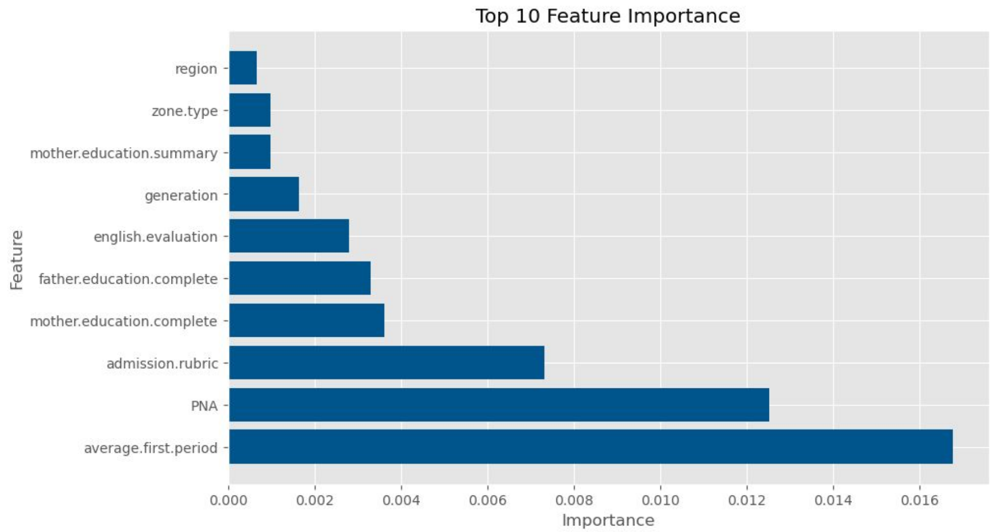
Para todos ellos, se realizó una búsqueda de hiperparametros **Bayesiana** con **CV** (*para evitar overfitting*).

MODELO 1: REGRESIÓN LOGÍSTICA

Modelo de aprendizaje supervisado que se utiliza para predecir una variable categórica binaria en función de un conjunto de variables predictoras. Se basa en el concepto de la regresión lineal, pero utiliza una función logística para modelar la relación entre las variables predictoras y la probabilidad de pertenecer a una clase en particular.

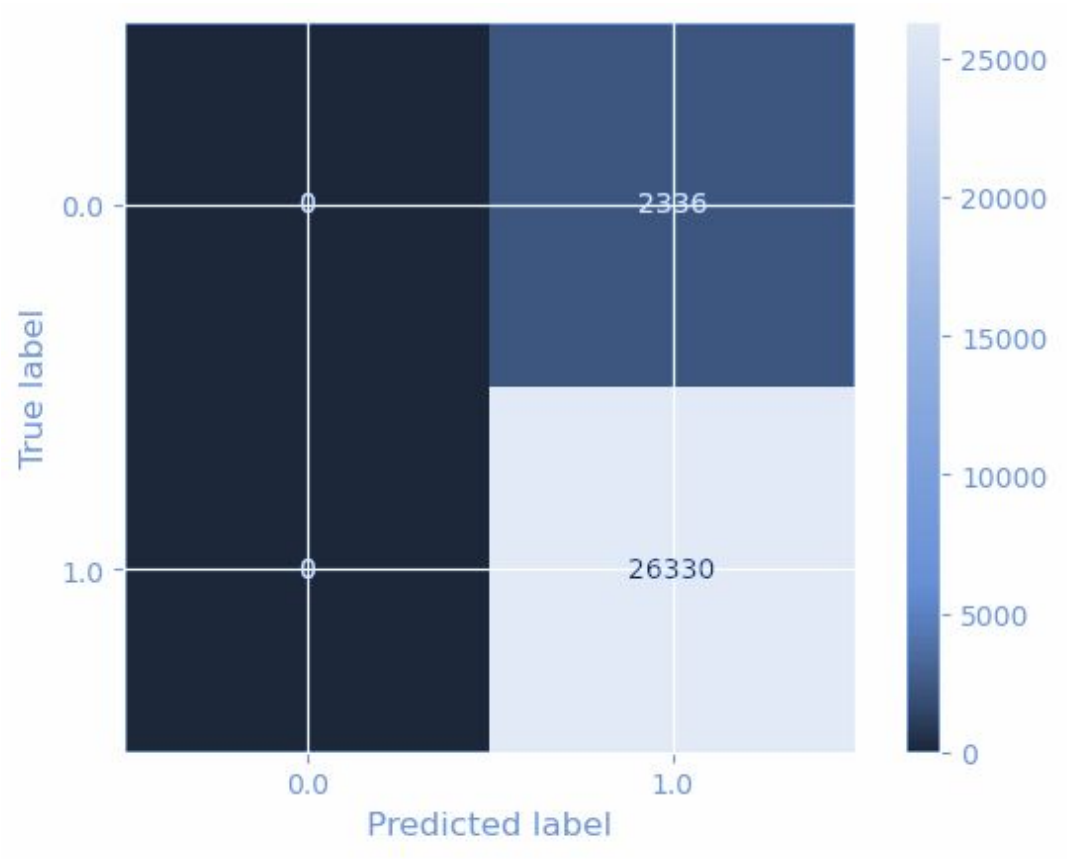
Ventajas del modelo para esta problemática

- Simplicidad y velocidad.




```
Accuracy: 0.9176260247688819
F1-Score: 0.9570437748720864
Precision: 0.9176260247688819
Recall: 1.0
AUC-ROC: 0.606857440084556
```

MODELO 1: REGRESIÓN LOGÍSTICA




```
Accuracy: 0.9185097327844833
Precision: 1.0
Recall: 0.9185097327844833
F1-Score: 0.9575241835769874
```

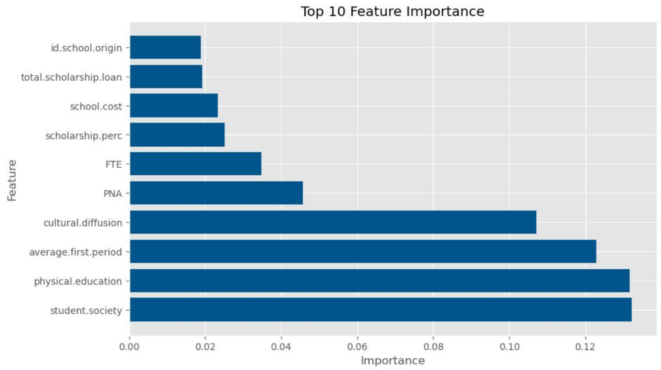
➔ Este modelo **NO** es bueno para la problemática, ya que no captura bien a quienes efectivamente realizan un abandono de la universidad.

MODELO 2: RANDOM FOREST

Algoritmo que toma múltiples muestras bootstrap del conjunto de datos original y entrena árboles de decisión en cada una de estas muestras. Una vez que se han construido todos los árboles, las predicciones finales se obtienen promediando las predicciones de cada árbol (en el caso de la clasificación).

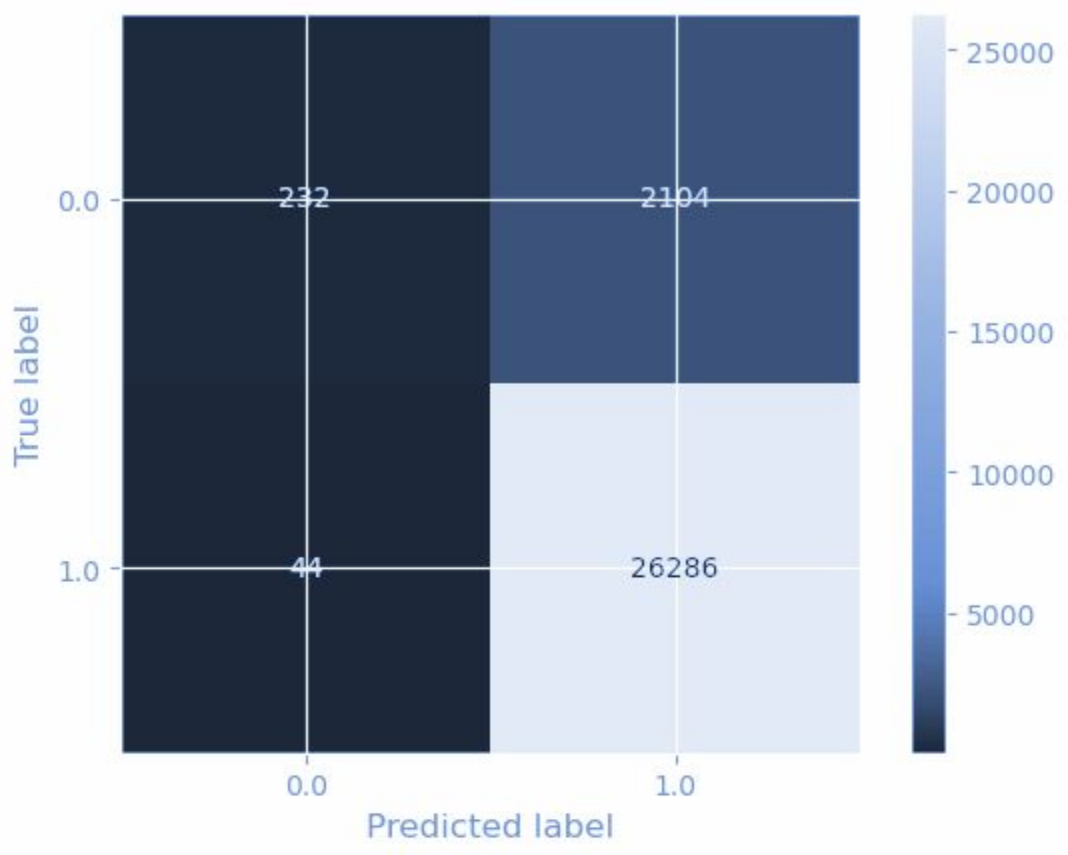
Ventajas del modelo para esta problemática

- Robustez ante outliers
- Buena capacidad de generalización
- Reducción del sobreajuste (múltiples árboles de decisión)



Accuracy: 0.9229897087039944
F1-Score: 0.9596324403401298
Precision: 0.924510019819423
Recall: 0.997528869457777
AUC-ROC: 0.736789779842012

MODELO 2: RANDOM FOREST



Accuracy: 0.9250680248377869
Precision: 0.9983289023927079
Recall: 0.925889397675237 8
F1-Score: 0.9607456140350876
AUC-ROC: 0.8832345539100827

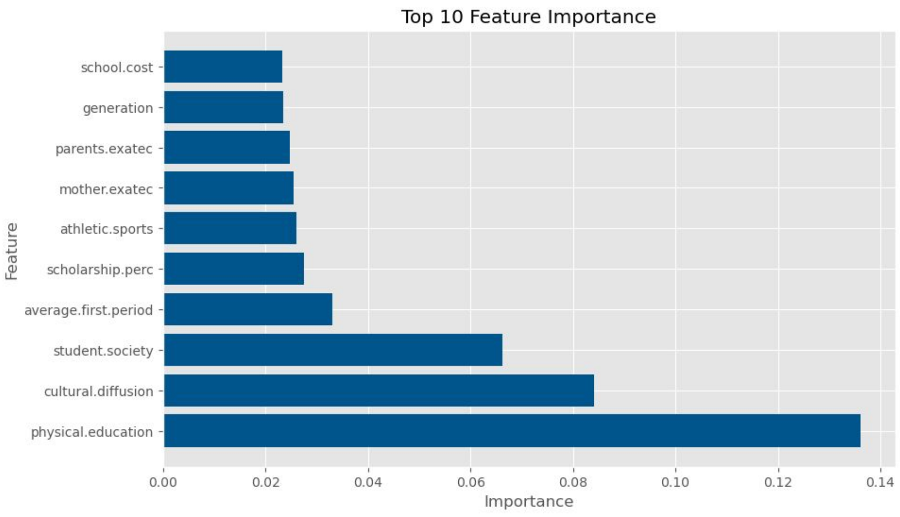
➔ Este modelo **mejora** al anterior, ya que comienza a capturar mejor los valores. A su vez, lleva **menor tiempo** de cómputo.

MODELO 3: XGBoost

XGBoost (Extreme Gradient Boosting) es una biblioteca optimizada y escalable para realizar tareas de aprendizaje automático basadas en árboles de decisión y algoritmos de boosting.

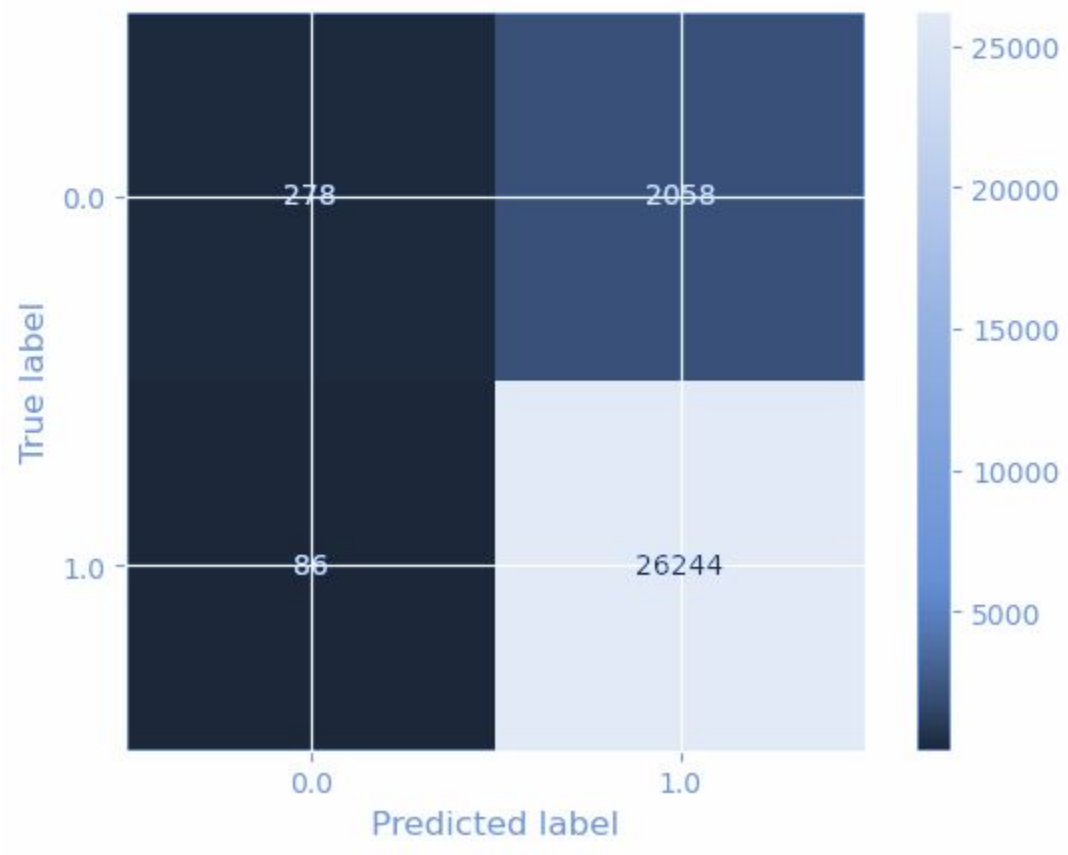
Ventajas del modelo para esta problemática

- Rendimiento y eficiencia
- Regularización avanzada




```
Accuracy: 0.9236438165009594
F1-Score: 0.9599212616448076
Precision: 0.925947187141217
Recall: 0.9964833911514518
AUC-ROC: 0.7660735622968895
```

MODELO 3: XGBoost




```
Accuracy: 0.9252075629665806
Precision: 0.9967337637675655
Recall: 0.9272842908628366
F1-Score: 0.9607556011129008
AUC-ROC: 0.8455102772995501
```

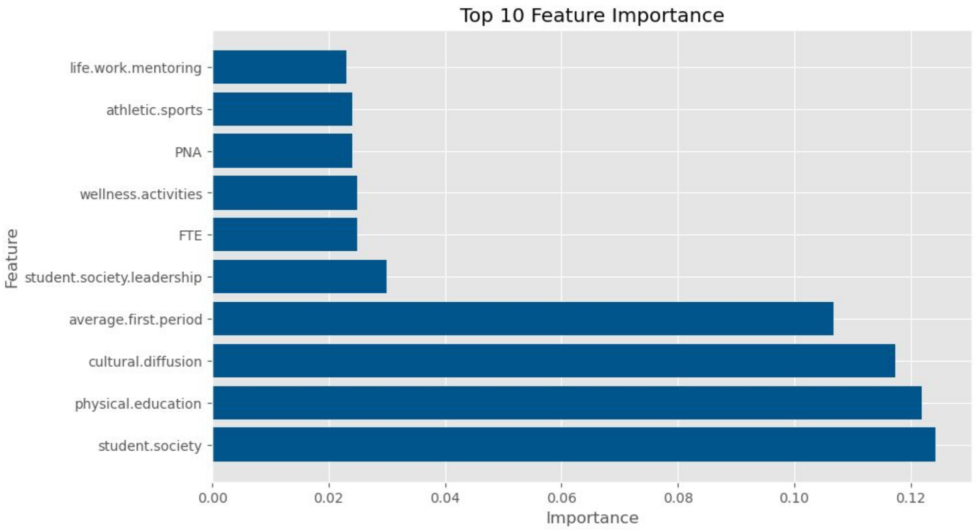
➔ Este modelo **no mejora** al anterior, ya que todas las métricas dan similares pero lleva **mayor tiempo** de cómputo. Además, el AUC-ROC es menor.

MODELO 4: ExtraTrees

Crea muchos árboles de decisión, pero el muestreo de cada árbol es aleatorio, sin reemplazo. Esto crea un conjunto de datos para cada árbol con muestras únicas.

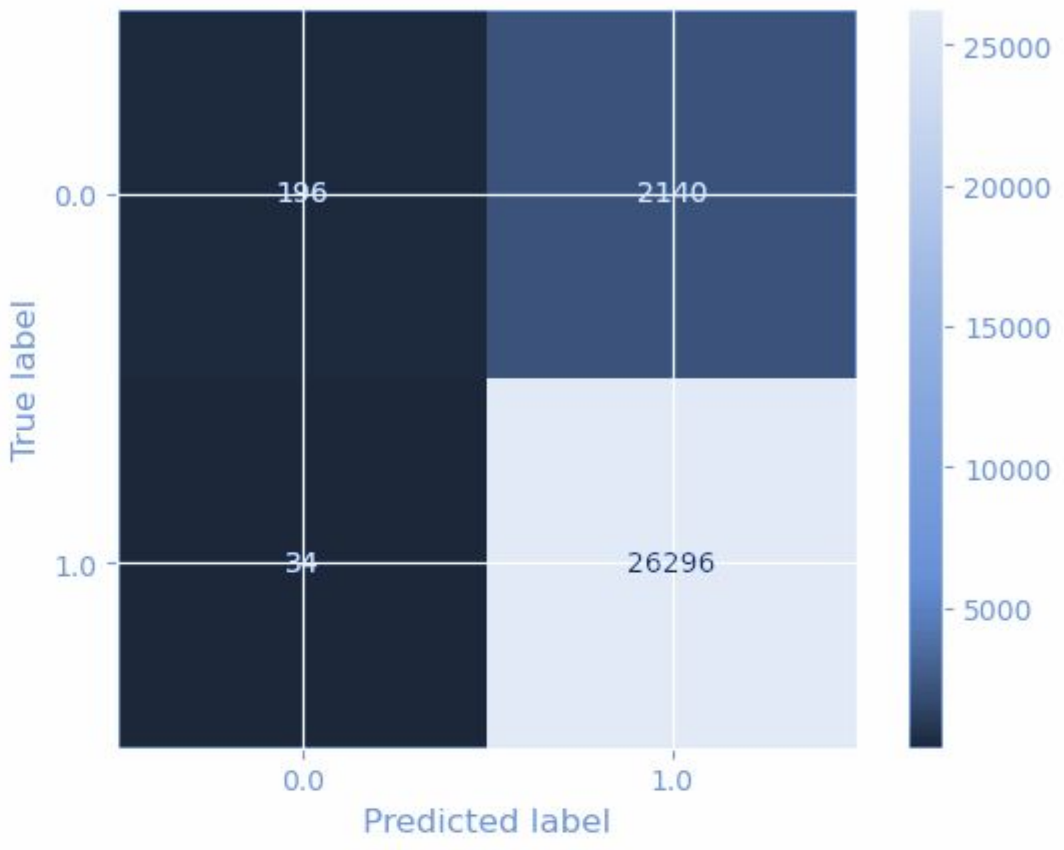
Ventajas del modelo para esta problemática

- Más veloz que el Random Forest. Se debe a que, en lugar de buscar la división óptima en cada nodo, lo hace aleatoriamente.
- Usa toda la muestra



Accuracy: 0.9248212105354963
F1-Score: 0.9606662103582021
Precision: 0.9250406432620062
Recall: 0.9991457453371932
AUC-ROC: 0.7362173655554308

MODELO 4: ExtraTrees




```
Accuracy: 0.924161027000628
Precision: 0.998708697303456l
Recall: 0.9247432831621888
F1-Score: 0.960303838147756
AUC-ROC: 0.888458598102833 5
```

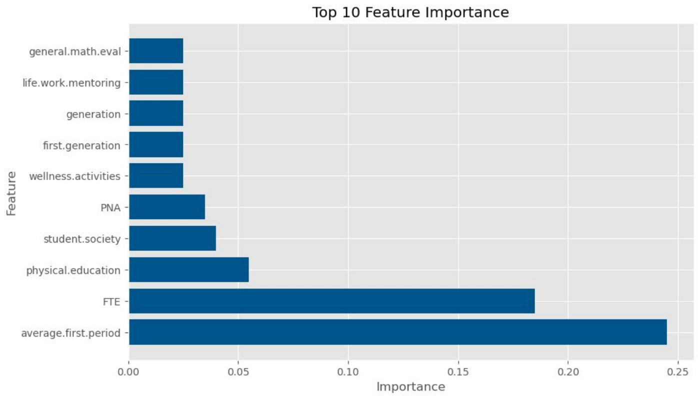
➔ Este modelo **mejora** al anterior, ya que todas las métricas dan similares al de RandomForest.

MODELO 5: RUSBoost

RUSBoost es un algoritmo diseñado para manejar el desequilibrio de clases en conjuntos de datos. La sigla RUS significa Random Under-Sampling, que se refiere a la técnica de submuestreo aleatorio utilizado en este algoritmo. Este combina la técnica de submuestreo aleatorio con el algoritmo de boosting. El submuestreo aleatorio se utiliza para reducir la proporción de la clase mayoritaria (clase dominante) en el conjunto de datos, mientras que el boosting se utiliza para construir un modelo fuerte a partir de clasificadores débiles.

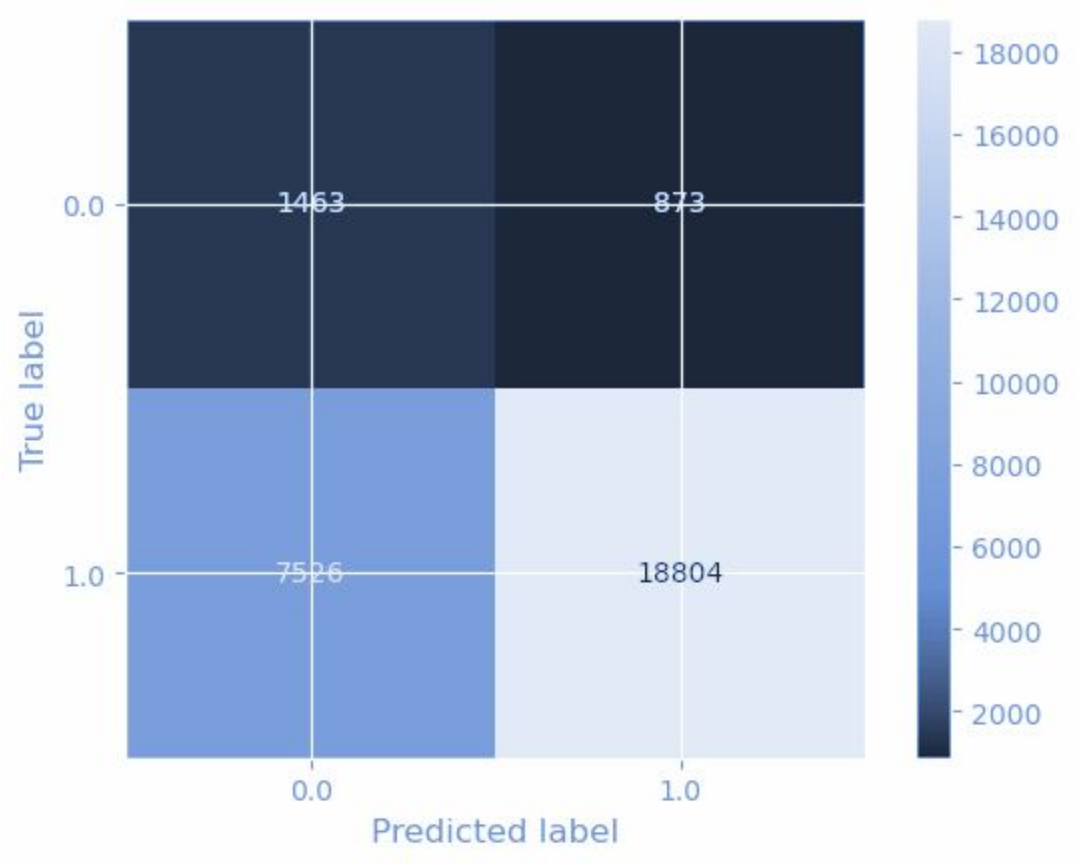
Ventajas del modelo para esta problemática

- Mitiga el sesgo de clases, pero conserva la información de las clases.



Accuracy: 0.7091400662829235
Precision: 0.9547205557309757
Recall: 0.7174789995728726
F1-Score: 0.8192705793095975
AUC-ROC: 0.6661011333167963

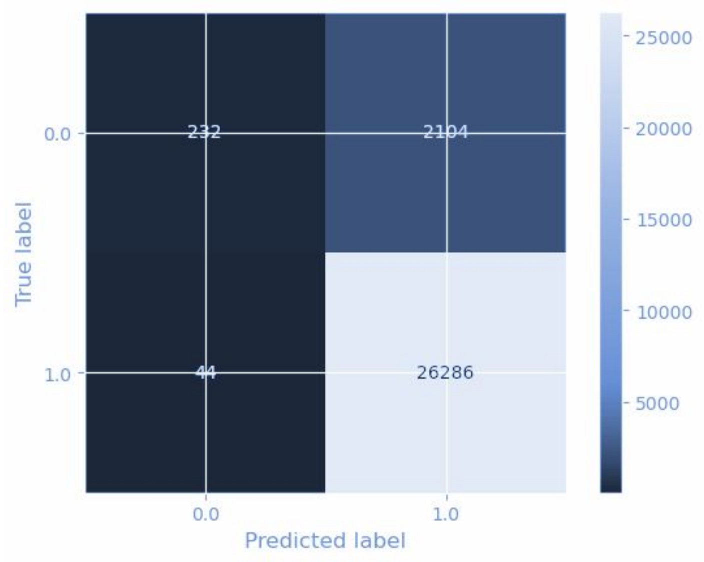
MODELO 5: RUSBoost




```
Accuracy: 0.7070048140654434
Precision: 0.7141663501709077
Recall: 0.9556334807135234
F1-Score: 0.8174408242224009
AUC-ROC: 0.5591939792042 42
```

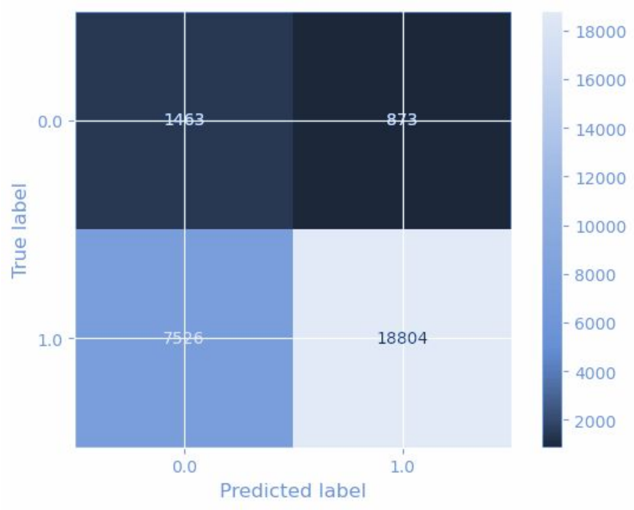
➔ Este modelo **es el peor en cuanto a las métricas.** Esto se debe a que le da más peso a la clase minoritaria, pero termina empeorando la performance.

MODELOS: COMPARACIÓN DE MÉTRICAS




```
Accuracy: 0.9250680248377869
Precision: 0.9983289023927079
Recall: 0.9258893976752378
F1-Score: 0.9607456140350876
AUC-ROC: 0.8832345539100827
```

RANDOM FOREST




```
Accuracy: 0.7070048140654434
Precision: 0.7141663501709077
Recall: 0.9556334807135234
F1-Score: 0.8174408242224009
AUC-ROC: 0.559193979204242
```

RUSBOOST

CONCLUSIONES

- ➔ El modelo ganador es Random Forest, el cual tiene el mayor nivel de acierto.
- ➔ Los modelos basados en árboles resultaron muy eficientes para esta problemática.

HIPÓTESIS

- Se puede predecir la variable dropout, a través de:
  - ✓ Características de los estudiantes. Las features más importantes en todos los modelos tenían relación al estudiante, su presente, sus capacidades, etc.
  - X El pasado. El impacto de los estudios de los padres, rendimiento previo y proveniencia no tuvieron la importancia esperada en el desarrollo de los modelos.

RECOMENDACIONES

En cuanto a los **modelos**:

1. Seleccionar el modelo que tenga mayor precisión para la clase positiva pero no perder de vista la clase minoritaria.
2. Es importante darle importancia a quienes no se retienen, ya que en ellos debe estar el foco. Propuesta modelos mixtos / inclusión del RUSBoost para el análisis.

En cuanto al **negocio**:

1. Crear un programa de seguimiento con tutorías, enfocado en alumnos con peores rendimientos promedio (*first.period.average*) en el primer semestre.
2. Fomentar los grupos de estudio y el trabajo en equipo, principalmente juntando alumnos con mayor tendencia al dropout.
3. Brindar apoyo emocional y psicológico a la generalidad de los alumnos, ya que muchas veces el estrés en la facultad puede llevar a dejarla.

Instituto Tecnológico de Buenos Aires

Tecnológico de Monterrey

MUCHAS GRACIAS