Data for Education 2025

Universidad Tecnológica de Pereira

Desarrollo y comparación de modelos predictivos supervisados de Machine Learning para la identificación temprana de la deserción estudiantil. Caso UTP

Angelica López Gómez, Leonardo Evelio Gaviria Grisales

Universidad Tecnológica de Pereira

Conocer más

Inteligencia Artificial para la Transformación Educativa

Deserción en Colombia

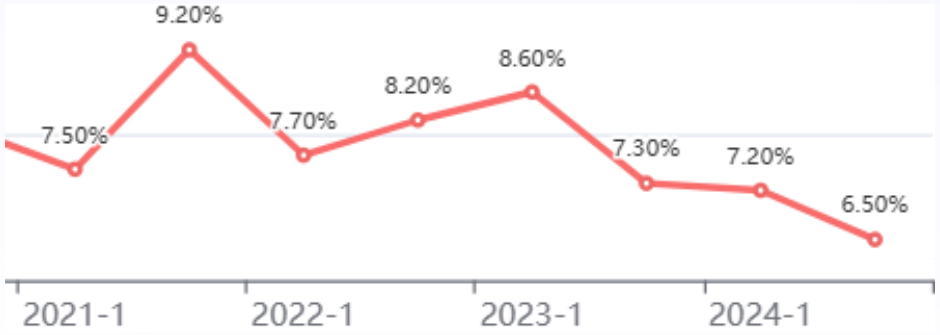
- La deserción estudiantil en Colombia es un problema en crecimiento. **Más del 50% de los casos de deserción se concentran en los primeros semestres**.
- Las cifras nacionales muestran que la matrícula universitaria disminuyó un 5,1 % entre 2021 y 2022.
- Además, la desaparición de Instituciones de Educación Superior (IES) es una realidad preocupante. **Según el SNIES, de más de 300 IES activas, sólo 276 registraron matrícula en 2022 y 272 en 2023.**

Este panorama evidencia que el sistema educativo enfrenta un desafío multidimensional, donde no solo se compromete el acceso a la educación superior, sino también la permanencia de las IES en el país.

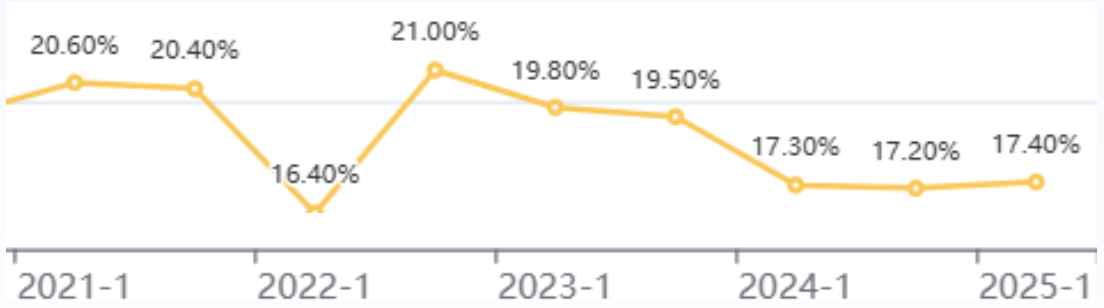
Deserción en la UTP

De acuerdo con los datos institucionales proporcionados por el Área de Administración de la Información Estratégica de la Oficina de Planeación de la UTP

- Tendencia de la **tasa de deserción institucional** interanual:

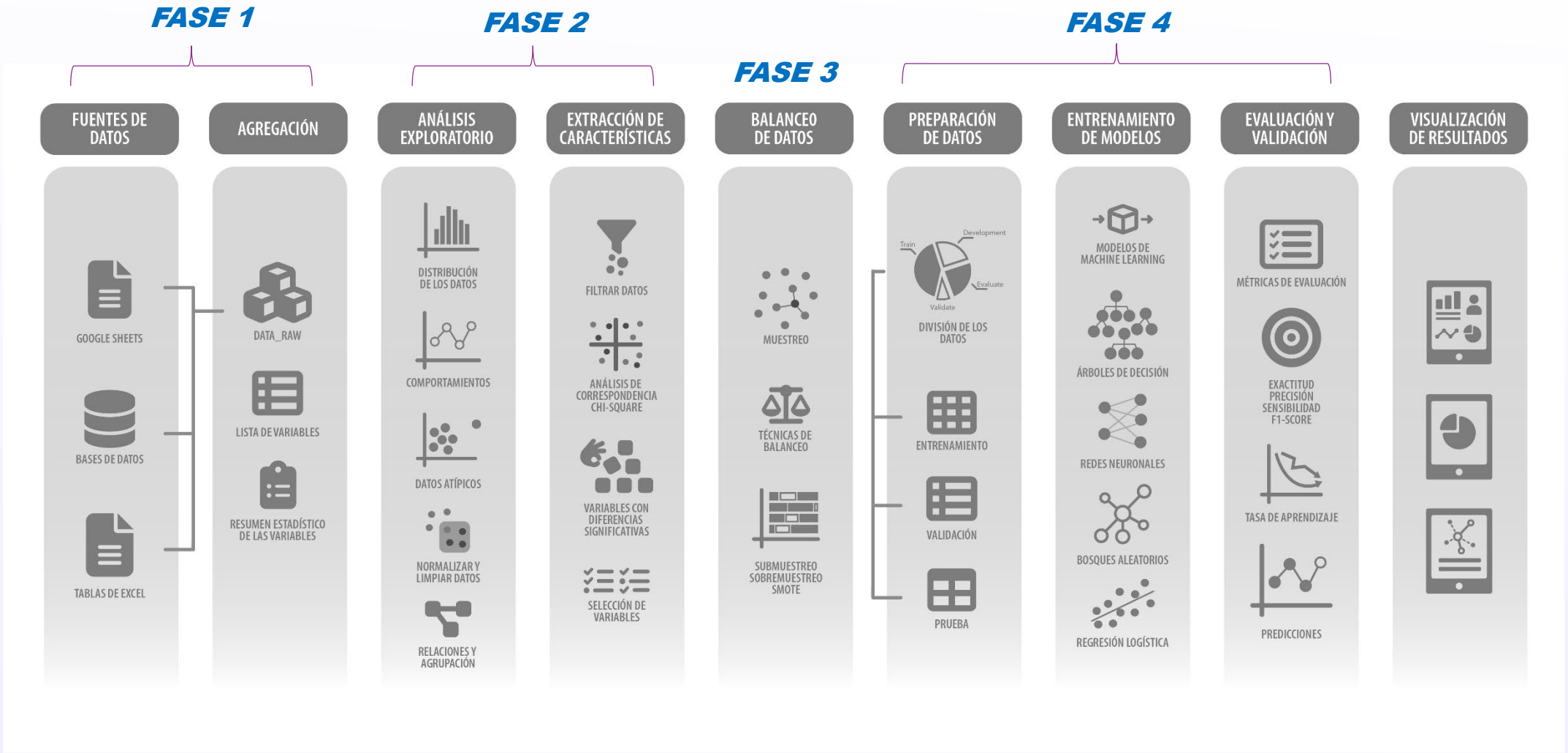


- Tendencia de la **tasa de deserción del primer curso** intersemestral:



Universidad Tecnológica de Pereira

Metodología



Data for Education 2025

Inteligencia Artificial para la Transformación Educativa

Fase 1: Recolección y Preprocesamiento

Estudiantes de primer curso en los periodos 2021-1 al 2024-2:

- **Sociodemográficas:** edad, estrato, sexo, naturaleza del colegio, municipio y departamento de residencia. *(Registro y control académico).*
- **Académicas:** Nivel, subnivel, la facultad a la que pertenece y programa académico. (Vicerrectoría académica).
- **Estado de ingreso:** Periodo, estado de ingreso, puntaje en las pruebas Saber 11. *(Registro y control académico)*
- **Apoyos socioeconómicos:** Becas, alimentación, transporte, monitorias.
- **Salud mental:** Resultados de las Pruebas de ansiedad, depresión, consumo de sustancias *(ASSIST)*
- **Sistema de Alertas Tempranas:** Considerando si el estudiante presentó o no el cuestionario.
- **Variable objetivo:** Estado de matrícula al siguiente semestre (0 = desertor, 1 = permanece).

Consolidación y preparación del conjunto de datos para desarrollar un modelo predictivo

1. Cargar datos: **DISPC_2021_2024, PRUEBAS_2021_2024**
2. Identificación y eliminación de duplicados (**PRUEBAS**)
3. Unificación de Datos (Merge para tener única tabla)
4. Revisión de columnas e eliminación de variables redundantes o calculadas por otras. **APOYADO Y NUMERO_APOYOS**.
5. Análisis de Valores Faltantes (Imputar datos pruebas)
6. Análisis de la variable target (**ESTADO**)

COD_PERIODO	dispc_df	merged_df	diferencia	Porcentaje
2021-1	2231	2040	-191	8.56
2021-2	1876	1823	-53	2.83
2022-1	2416	2364	-52	2.15
2022-2	1911	1869	-42	2.20
2023-1	2221	2188	-33	1.49
2023-2	1895	1852	-43	2.27
2024-1	2590	2464	-126	4.86
2024-2	2094	2072	-22	1.05

Fase 2: Análisis Exploratorio y Selección de Características

Variable	Tipo	Prueba estadística recomendada
COD_PERIODO	Categórica	Chi-cuadrado
IDTERCERO	Identificador	❌ Ninguna
ESTADO (target)	Categórica	-
COD_AIE	Categórica	Chi-cuadrado
NIVEL	Categórica	Chi-cuadrado
SUBNIVEL	Categórica	Chi-cuadrado
FACULTAD	Categórica	Chi-cuadrado
NOMBRE_PROGRAMA	Categórica	Chi-cuadrado
DURACION_SEMESTRES	Numérica	ANOVA
TIPO_VINCULACION	Categórica	Chi-cuadrado
PUNTAJE_ICFES	Numérica	ANOVA
EDAD	Numérica	ANOVA
ESTRATO	Ordinal	Chi-cuadrado
SEXO	Categórica	Chi-cuadrado
TIPO_COLEGIO	Categórica	Chi-cuadrado
MUNICIPIO_RESIDENCIA	Categórica	Chi-cuadrado
DEPTO_RESIDENCIA	Categórica	Chi-cuadrado
APOYADO	Categórica	Chi-cuadrado
APOYO_UTP	Categórica	Chi-cuadrado

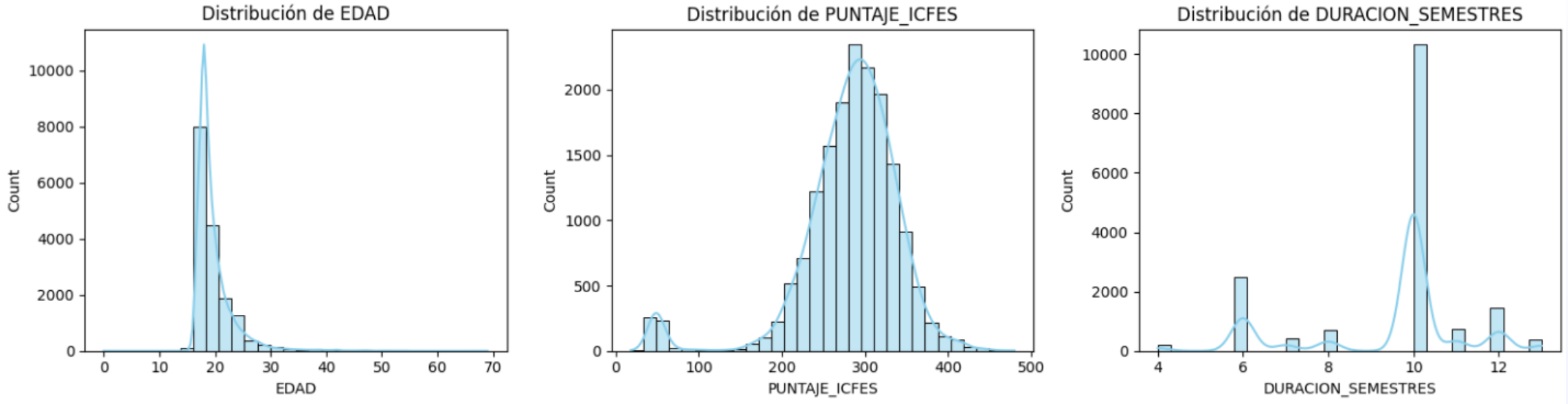
Variable	Tipo	Prueba estadística recomendada
APOYO_CONVENIO	Categórica	Chi-cuadrado
NUMERO_APOYOS	Numérica	ANOVA
PRUEBA_SAT	Categórica	Chi-cuadrado
PRUEBA_ASSIST	Categórica	Chi-cuadrado
PRUEBA_ANSIEDAD_DEPRESION	Categórica	Chi-cuadrado
DEPRESION	Categórica	Chi-cuadrado
ANSIEDAD	Categórica	Chi-cuadrado
NIVEL_TABACO	Categórica	Chi-cuadrado
NIVEL_ALCOHOL	Categórica	Chi-cuadrado
NIVEL_CANNABIS	Categórica	Chi-cuadrado
NIVEL_COCAINA	Categórica	Chi-cuadrado
NIVEL_ANFETAMINA	Categórica	Chi-cuadrado
NIVEL_INHALANTE	Categórica	Chi-cuadrado
NIVEL_SEDANTE	Categórica	Chi-cuadrado
NIVEL_ALUCINOGENO	Categórica	Chi-cuadrado
NIVEL_OPIACEO	Categórica	Chi-cuadrado
NIVEL_OTRADROGA	Categórica	Chi-cuadrado
INGLES_APROBADO	Categórica	Chi-cuadrado

Se definió una tabla de pruebas estadísticas para cada tipo de variable:

- **Categóricas: prueba Chi-cuadrado**
- **Numéricas: prueba ANOVA**

Se aplicaron estas pruebas a **37 variables** para identificar su asociación con la deserción **`ESTADO`**

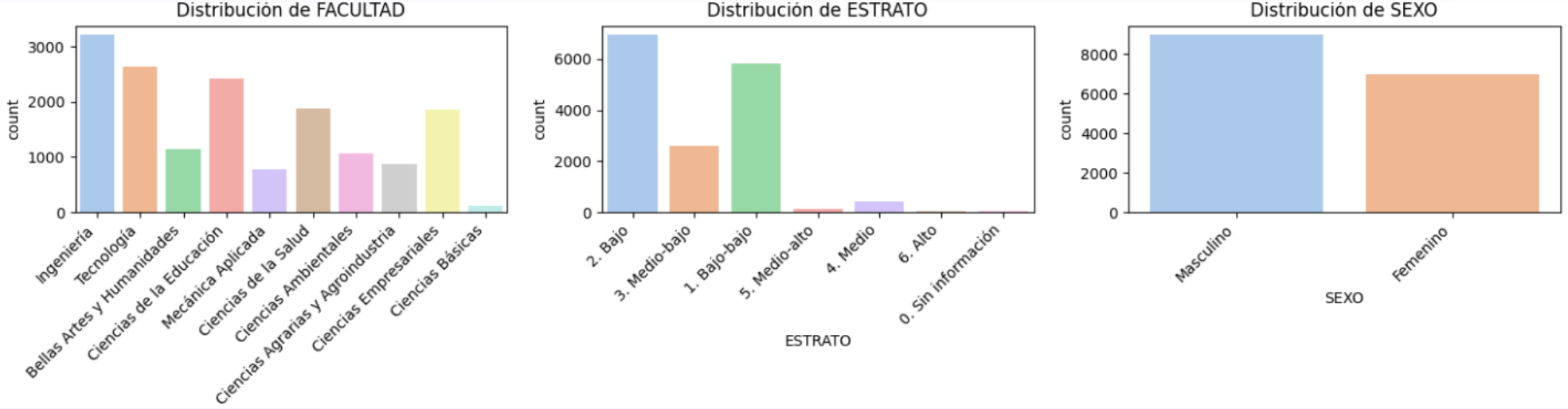
Distribución variables numéricas

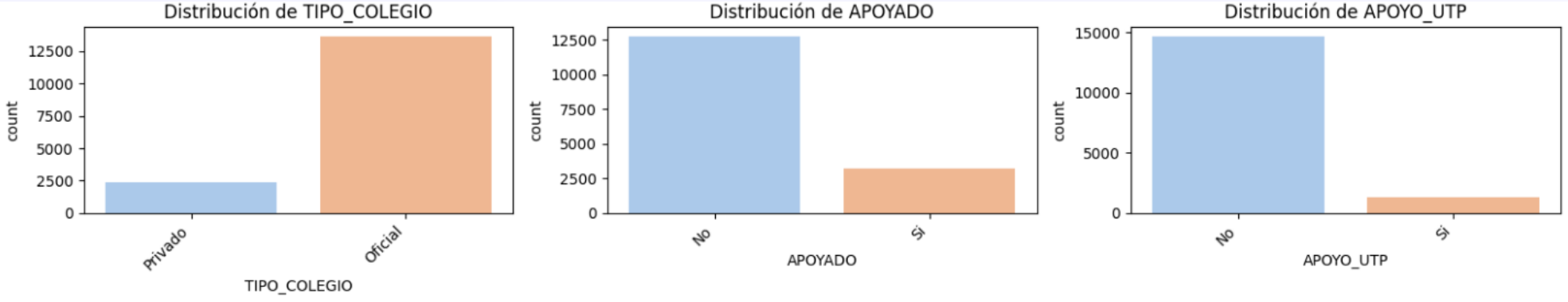


Como se detectaron datos en **PUNTAJE_ICFES** menores que **100**, y **EDAD** superior a **40**.
Se eligió el **z-score** como criterio para la detección y eliminación de valores atípicos.

Universidad Tecnológica de Pereira

Distribución variables sociodemográficas y apoyos

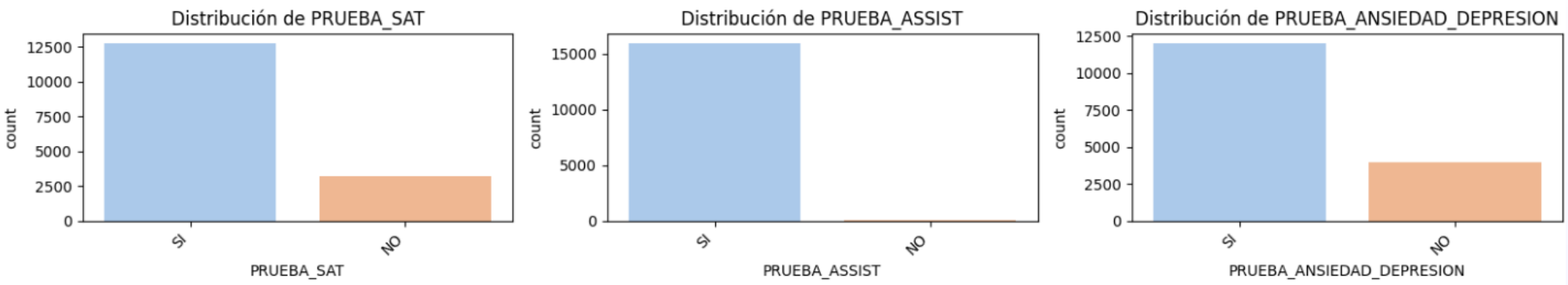




Data for Education 2025

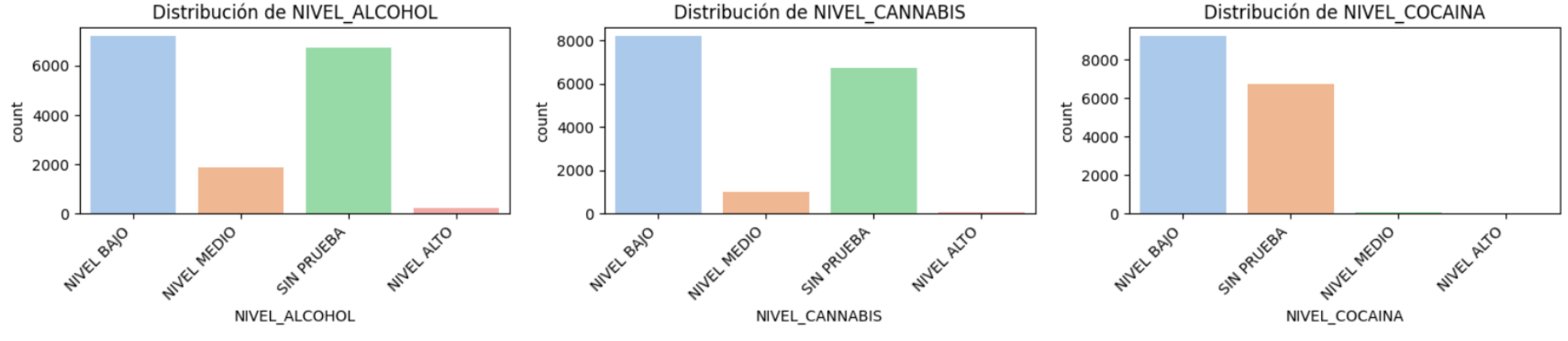
Inteligencia Artificial para la Transformación Educativa

Distribución de pruebas SAT y *ASSIST*



Universidad Tecnológica de Pereira

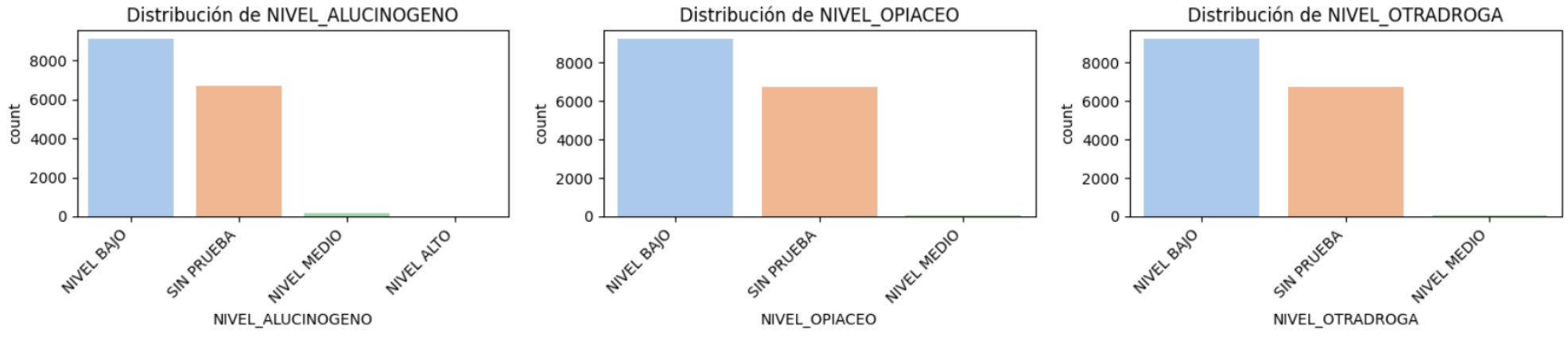
Distribución de consumo de sustancias

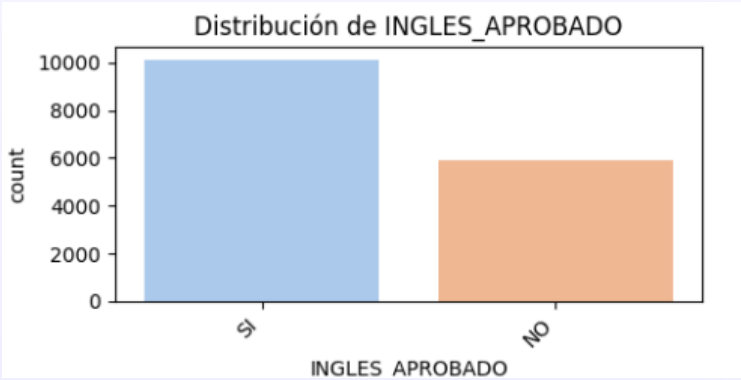


Data for Education 2025

Inteligencia Artificial para la Transformación Educativa

Distribución de otros consumos y aprobación inglés





PRUEBA_ASSIST	Categórica	Chi-cuadrado	7.62	0.0058	✓ Asociación significativa
TIPO_COLEGIO	Categórica	Chi-cuadrado	4.52	0.0335	✓ Asociación significativa
APOYO_CONVENIO	Categórica	Chi-cuadrado	0.00	1.0000	✗ Sin asociación significativa
NIVEL	Categórica	Chi-cuadrado	0.00	1.0000	✗ Sin asociación significativa
TIPO_VINCULACION	Categórica	Chi-cuadrado	0.00	1.0000	✗ Sin asociación significativa

Fase 3: Balanceo de Clases - Preparación de datos

```
ESTADO
2. Permanece programa    12952
1. No matriculado         3146
3. Cambio de programa      574
```

ESTADO	Matriculado	Conteo	Porcentaje (%)
0	SI	13526	81.13
1	NO	3146	18.87

- Random Undersampling
- Random Oversampling
- SMOTE + Random Undersampling

```python
target = 'ESTADO'
id = 'IDTERCERO'

numeric_vars = ['EDAD', 'PUNTAJE_ICFES', 'DURACION_SEMESTRES']

categorical_vars = [
    'COD_PERIODO', 'SUBNIVEL', 'FACULTAD', 'ESTRATO', 'SEXO', 'COD_AIE',
    'TIPO_COLEGIO', 'APOYADO', 'APOYO_UTP', 'NUMERO_APOYOS', 'PRUEBA_SAT',
    'PRUEBA_ASSIST', 'PRUEBA_ANSIEDAD_DEPRESION', 'DEPRESION', 'ANSIEDAD',
    'NIVEL_TABACO', 'NIVEL_ALCOHOL', 'NIVEL_CANNABIS', 'NIVEL_COCAINA',
    'NIVEL_ANFETAMINA', 'NIVEL_INHALANTE', 'NIVEL_SEDANTE', 'NIVEL_ALUCINOGENO',
    'NIVEL_OPIACEO', 'NIVEL_OTRADROGA', 'INGLES_APROBADO'
]

# Eliminar columna de identificación
df = merged_df.drop(columns=[id], errors='ignore')

# Definir variable de target
y = df[target]
X = df.drop(columns=[target])
X = pd.get_dummies(X, columns=categorical_vars, drop_first=True)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)

print("Train shape:", X_train.shape)
print("Test shape:", X_test.shape)
```

Universidad Tecnológica de Pereira

Fase 4: Definición de modelos y entrenamiento

Modelos:

- LogisticRegression
- DecisionTreeClassifier
- RandomForestClassifier
- SVC
- VotingClassifier - Soft
- VotingClassifier - Hard

GRIDSEARCH CV

```python
# Pipeline
pipe = Pipeline([
    ('resample', SMOTE()),  # se sobreescribirá en el GridSearch
    ('scaler', StandardScaler()),
    ('modelo', LogisticRegression(max_iter=1000))
])

# Parámetros: combinación válida de penalty + solver + resampling
parametros = [
    {
        'resample': [SMOTE(), RandomOverSampler(), RandomUnderSampler()],
        'modelo__penalty': ['l1'],
        'modelo__solver': ['liblinear', 'saga'],
        'modelo__C': [0.001, 0.01, 0.1, 1]
    },
    {
        'resample': [SMOTE(), RandomOverSampler(), RandomUnderSampler()],
        'modelo__penalty': ['l2'],
        'modelo__solver': ['liblinear', 'saga', 'lbfgs'],
        'modelo__C': [0.001, 0.01, 0.1, 1]
    }
]
```

```python
pipe_dt = Pipeline([
    ('resample', SMOTE()),
    ('scaler', StandardScaler()),
    ('modelo', DecisionTreeClassifier(random_state=42))
])

parametros_dt = {
    'resample': [SMOTE(), RandomOverSampler(), RandomUnderSampler()],
    'modelo__criterion': ['gini', 'entropy'],
    'modelo__max_depth': [None, 5, 10, 20],
    'modelo__min_samples_split': [2, 5, 10]
}
```

```python
pipe_rf = Pipeline([
    ('resample', SMOTE()),
    ('scaler', StandardScaler()),
    ('modelo', RandomForestClassifier(random_state=42))
])

parametros_rf = {
    'resample': [SMOTE(), RandomOverSampler(), RandomUnderSampler()],
    'modelo__n_estimators': [100, 200],
    'modelo__max_depth': [None, 10, 20],
    'modelo__min_samples_split': [2, 5],
    'modelo__criterion': ['gini', 'entropy']
}
```

Evaluación de los modelos

- El **recall** es especialmente importante en la evaluación de modelos de ***machine learning*** para predecir la deserción estudiantil porque mide la capacidad del modelo para identificar correctamente a los estudiantes que realmente van a desertar.

- En este contexto, un **falso negativo** (no detectar a un estudiante en riesgo) puede ser mucho más costoso que un **falso positivo** (clasificar a alguien como en riesgo cuando no lo está), ya que perder esa alerta significa dejar de intervenir a tiempo y aumentar la probabilidad de abandono.

```
Mejores Hiperparametros:
LogisticRegression: {'modelo__C': 0.1, 'modelo__penalty': 'l2', 'modelo__solver': 'liblinear', 'resample': RandomUnderSampler()}
DecisionTreeClassifier: {'modelo__criterion': 'entropy', 'modelo__max_depth': 5, 'modelo__min_samples_split': 2, 'resample': RandomUnderSampler()}
RandomForestClassifier: {'modelo__criterion': 'gini', 'modelo__max_depth': 10, 'modelo__min_samples_split': 5, 'modelo__n_estimators': 200, 'resample': RandomOverSampler()}
SVC {'kernel': 'rbf', 'gamma': '0.1, 'resample':SMOTE()'}
```

Comparación de modelos

	Modelo	Accuracy	Precision	Recall	F1	AUC
0	LogisticRegression	0.727131	0.376623	0.738540	0.498853	0.731550
1	DecisionTreeClassifier	0.728067	0.376748	0.731749	0.497403	0.729493
2	RandomForestClassifier	0.750859	0.401135	0.719864	0.515188	0.738853
3	SVC	0.796753	0.395973	0.200340	0.266065	0.565740
4	VotingClassifier Soft	0.717452	0.368552	0.752122	0.494696	0.730881
5	VotingClassifier Hard	0.725882	0.377021	0.752122	0.502268	0.736046

PREDECIR: 8 de cada 10 desertores

Universidad Tecnológica de Pereira

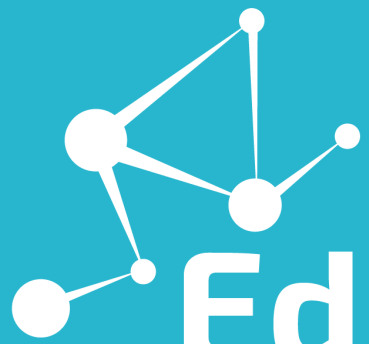
RESULTADOS

- Se logró **optimizar el *recall***, aunque con una ligera disminución en la *precisión*, lo que representa un balance razonable cuando el objetivo principal es **maximizar la detección de estudiantes en riesgo de deserción**.

- El uso de **validación cruzada** combinada con **GridSearchCV** permitió seleccionar de manera más confiable los **hiperparámetros óptimos** en los modelos de *Logistic Regression, Decision Tree, Random Forest y SVM*, incrementando la robustez de los resultados y brindando modelos listos para un entorno de producción.

LECCIONES APRENDIDAS

- Una vez entrenado y validado, el modelo debe ser puesto en producción para que trascienda la fase experimental y se convierta en una **herramienta de articulación dentro del plan de fomento académico institucional**. Su implementación permitirá a directores de programa y decanos contar con evidencia objetiva para **diseñar, ejecutar y evaluar estrategias de prevención de la deserción**.



Data for Education 2025

Universidad Tecnológica de Pereira