Corporación Universitaria Autónoma de Nariño
AUNAR VILLAVICENCIO

PREDICCIÓN DE MATRÍCULAS EN PRIMER SEMESTRE DE EDUCACIÓN SUPERIOR EN COLOMBIA CON MACHINE LEARNING

Shirley Vanessa Daza Riveros
Jhulian Mateo Espinoza Zipaquira

28 de febrero del 2025

#AUNAR Villavicencio

/ @Aunarvillavo
www.aunarvillavicencio.edu.co

Objetivo General

Analizar y modelar las tendencias de matrículas en instituciones de educación superior de primer semestre en Colombia, utilizando técnicas de aprendizaje automático para predecir datos futuros, y apoyar la toma de decisiones en el ámbito educativo.

Objetivos Específicos

- Explorar y limpiar los datos históricos de matrículas, seleccionando las variables relevantes y aplicando estadísticas descriptivas y visualización de las variables.
- Implementar y evaluar modelos de aprendizaje automático para la predicción de matrículas.
- Optimizar las predicciones utilizando un enfoque basado en Random Forest.
- Visualizar y comprar los resultados de predicción con los datos reales.

Resumen

El análisis de las tendencias de matrícula en instituciones de educación superior en Colombia permite comprender la evolución del número de estudiantes inscritos y anticipar su comportamiento en el futuro. Para ello, se emplearon técnicas de aprendizaje automático con el fin de predecir las matrículas en semestres posteriores y facilitar la toma de decisiones en el ámbito educativo.

Se trabajó con un conjunto de datos históricos, los cuales fueron procesados y analizados para identificar patrones clave. Posteriormente, se entrenaron modelos de Regresión Lineal y Random Forest Regressor, evaluando su desempeño mediante métricas como MAE, MSE Y R2. Los resultados fueron visualizados a través de gráficos comparativos y se generaron reportes en Excel con las predicciones obtenidas.

Los hallazgos de este estudio pueden ser de utilidad para instituciones educativas y organismos gubernamentales en la planificación y asignación de recursos, proporcionando una base cuantitativa para la formulación de estrategias en el sector educativo.

Introducción

La educación superior desempaña un papal fundamental en el desarrollo social y económico de Colombia. Comprender las tendencias de matrícula en las instituciones universitarias es esencial para la planificación educativa y la asignación eficiente de recursos. Sin embargo, la variabilidad en el número de inscritos a lo largo del tiempo plantea desafíos para la toma de decisiones basadas en datos históricos.

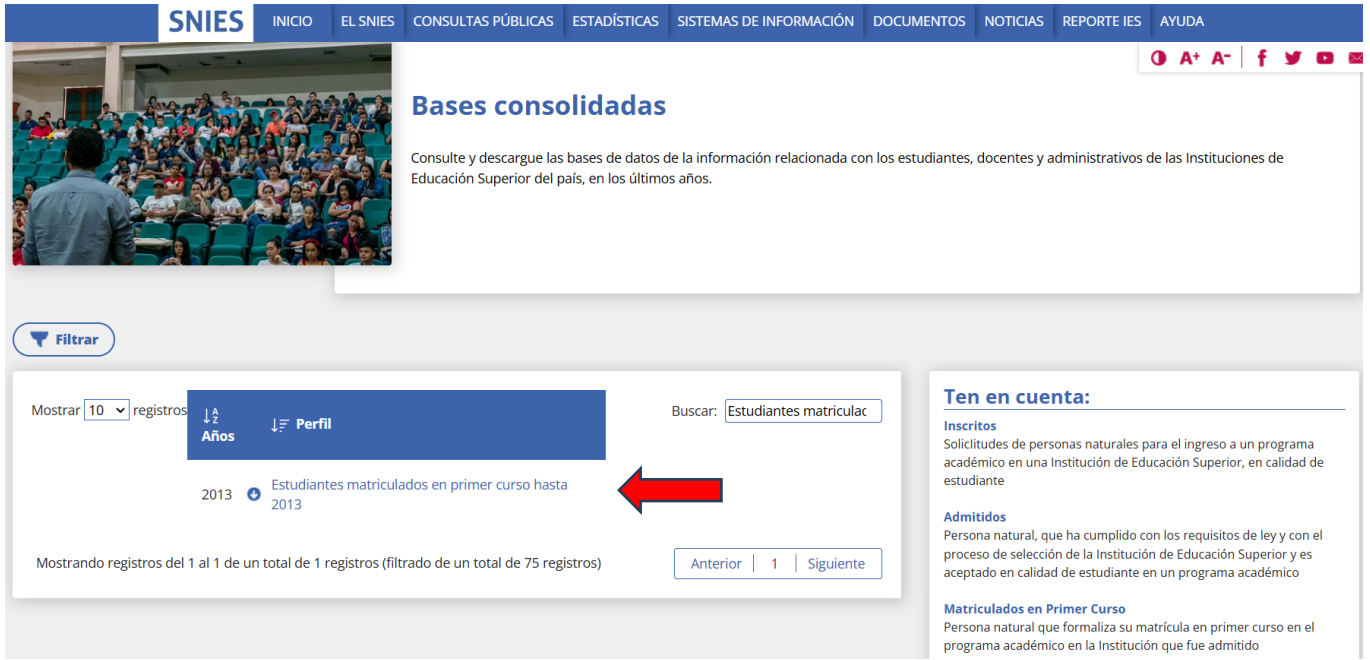
En este contexto, el aprendizaje automático se presenta como una herramienta poderosa para analizar patrones y predecir el comportamiento futuro de las matrículas. Mediante la aplicación de modelos estadísticos y de machine learning, es posible anticipar cambios en la demanda educativa y diseñar estrategias fundamentadas en evidencia.

Este proyecto busca aplicar técnicas de aprendizaje automático para modelar y predecir las matrículas en educación superior de primer semestre en Colombia, utilizando datos históricos y algoritmos de predicción. A través del análisis exploratorio y el uso de modelos como la Regresión Lineal y Random Forest Regressor, se busca estimar la cantidad de estudiantes matriculados en distintos semestres y generar información relevante para el sector educativo.

Métodos (Materiales y Métodos)

Para la proyección de la matrícula en primer semestre en educación superior en Colombia, se empleó un enfoque basado en Machine Learning, Los datos fueron obtenidos del Sistema Nacional de Información de la Educación Superior (SNIES), los cuales incluyen registros de matrícula a nivel nacional desglosado por institución, metodología de enseñanza, departamento y nivel de formación.

https://snies.mineducacion.gov.co/1778/articles-391574_recurso.xlsb



Conjunto de Datos

Colums	Dtype	Descipción
IES	object	Nombre de la intitucion de educacion superior.
PrincipalSeccional	object	Si la intitucion corresponde a principal o seccional
SectorIES	object	Si la intitucion es oficial o privada.
CaracterIES	object	Tipo de institucion (Universidad, instituto, etc)
ProgramaAcademico	object	Nombre del programa academico.
NivelFormacion	object	Nivel de formacion del programa academico.
MetodologiaPrograma	object	Metodologia usada (presencial, virtual)
AreaConocimiento	object	Area de conocimiento del programa academico.
DepartamentoOfertaPrograma	object	Departamento donde se ofrece el programa academico.
MunicipioOfertaPrograma	object	Municipio donde se ofrece el programa academico.
Hombre 2000-1	in64	Hombres registrados en ese semestre.
Mujer 2000-1	in64	Mujeres registrados en ese semestre.
Total 2000-1	int64	Total de estudiantes matriculados ese semestre.
Hombre 2000-2	int64	Hombres registrados en ese semestre.
Mujer 2000-2	int64	Mujeres registrados en ese semestre.
Total 2000-2	int64	Total de estudiantes matriculados ese semestre.
***	int64	**
Hombre 2013-2	int64	Hombres registrados en ese semestre.
Mujer 2013-2	int64	Mujeres registrados en ese semestre.
Total 2013-2	int64	Total de estudiantes matriculados ese semestre.
dtypes: int(84), object(10)		

Los datos abarcan cerca de 23891 filas con un total de 103 columnas. De las cuales se reparten entre tipo object y int64, teniendo en su mayoría esta última.

Conjunto de Datos

Visión general del dataset con el uso de:

Datos.head()

	CodigoInstitución	IES	PrincipaloSeccional	SectorIES	CaracterIES	CodigoDepartamento	DepartamentoDomicilio(IES)	CodigoMunicipio	MunicipioDomicilio	CodigoPrograma(SNIES)	...	Total 2012-1	Hombre 2012-2	Mujer 2012-2	Total 2012-2	Hombre 2013-1	Mujer 2013-1	Total 2013-1	Hombre 2013-2
0	1101	UNIVERSIDAD NACIONAL DE COLOMBIA	PRINCIPAL	OFICIAL	UNIVERSIDAD	11	BOGOTA D.C.	11001	BOGOTA D.C.	1	...	120	81	26	107	98	31	129	48
1	1101	UNIVERSIDAD NACIONAL DE COLOMBIA	PRINCIPAL	OFICIAL	UNIVERSIDAD	11	BOGOTA D.C.	11001	BOGOTA D.C.	2	...	3	26	25	51	24	28	52	31
2	1101	UNIVERSIDAD NACIONAL DE COLOMBIA	PRINCIPAL	OFICIAL	UNIVERSIDAD	11	BOGOTA D.C.	11001	BOGOTA D.C.	3	...	54	30	21	51	32	21	53	25
3	1101	UNIVERSIDAD NACIONAL DE COLOMBIA	PRINCIPAL	OFICIAL	UNIVERSIDAD	11	BOGOTA D.C.	11001	BOGOTA D.C.	4	...	30	19	10	29	29	6	35	21
4	1101	UNIVERSIDAD NACIONAL DE COLOMBIA	PRINCIPAL	OFICIAL	UNIVERSIDAD	11	BOGOTA D.C.	11001	BOGOTA D.C.	5	...	49	30	14	44	34	19	53	24

5 rows × 103 columns

Conjunto de Datos

Visión de las estadísticas generales, tales como:

count: Cantidad de valores no nulos en la columna.
mean: Media (promedio) de los valores en la columna.
std: Desviación estándar, que mide cuanto varían los datos respecto a la media.
min: Valor Mínimo en la columna.
25% (Q1 – Primer Cuartil): El 25% de los datos son menores o iguales a este valor.
50% (Q2 – Mediana): el 50% de los datos están por debajo de este valor (valor central).
75% (Q3 – Tercer Cuartil): El 75% de los datos son menores o iguales a este valor.
max: Valor máximo en la columna.

Conjunto de Datos

datos.describe()

	CodigoInstitución	CodigoDepartamento	CodigoMunicipio	CodigoPrograma(SNIES)	Hombre 2000-1	Mujer 2000-1	Total 2000-1	Hombre 2000-2	Mujer 2000-2	Total 2000-2	...	Total 2012-1	Hombre 2012-2	Mujer 2012-2	Total 2012-2	Hombre 2013-1	Mujer 2013-1
count	23891.000000	23891.000000	23891.000000	23891.000000	23891.000000	23891.000000	23891.000000	23891.000000	23891.000000	23891.000000	...	23891.000000	23891.000000	23891.000000	23891.000000	23891.000000	23891.000000
mean	3450.977690	21.384999	21455.847097	35536.657402	2.603030	2.548449	5.151480	1.990917	1.795237	3.786154	...	16.089908	6.333054	6.687916	13.020970	7.173496	7.657863
std	2777.157558	22.616680	22621.036999	35413.543821	13.804849	15.588165	26.680996	11.576605	11.650808	21.161347	...	60.878547	24.860115	34.543917	54.943589	25.085670	35.825390
min	1101.000000	5.000000	5001.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1711.000000	11.000000	11001.000000	6130.500000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	2102.000000	11.000000	11001.000000	15054.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	3812.000000	17.000000	17001.000000	55143.500000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	13.000000	3.000000	3.000000	8.000000	4.000000	3.000000
max	9906.000000	91.000000	91001.000000	102826.000000	506.000000	1107.000000	1283.000000	570.000000	672.000000	1008.000000	...	4394.000000	1516.000000	3157.000000	4673.000000	1185.000000	2167.000000

8 rows × 88 columns

Modelos de Predicción

Objetivo: Predecir el número de matrículas utilizando diferentes modelos de Machine Learning.

A. Regresión Lineal

La Regresión Lineal busca encontrar una relación entre las matrículas pasadas y futuras, ajustando una línea recta a los datos.

Métricas de Evaluación:

MAE (Error Absoluto Medio)
MSE (Error Cuadrático Medio)
R^2 Score (Indica qué tan bien el modelo explica la variabilidad)

Modelos de Predicción – Regresión Lineal

Código utilizado:

```
columnas_entrada = [col for col in datos.columns if 'Total' in col and '2013' not in col]
X = datos[columnas_entrada]
y = datos['Total 2013-2']

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

modelo = LinearRegression()
modelo.fit(X_train, y_train)

y_pred = modelo.predict(X_test)
```

Modelos de Predicción – Regresión Lineal

Comparativa entre los datos de predicción y los datos reales:

```
comparativa = {"Prediccion": y_pred, "Valor Real": y_test}
pd.DataFrame(comparativa)
```

Resultados:

	Prediccion	Valor Real
16341	3.411440	2
18596	4.222653	0
15483	3.468040	0
17177	3.411440	0
11974	77.277208	52
...	...	...
10011	3.705756	1
11142	5.143408	0
20964	24.208589	0
10430	5.038676	3
5627	159.029712	205

Modelos de Predicción – Regresión Lineal

Evaluación del modelo:

```
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f' Evaluación del Modelo:')
print(f' MAE: {mae:.2f}')
print(f' MSE: {mse:.2f}')
print(f' R² Score: {r2:.4f}')
```

Resultados de la Evaluación del modelo:

```
Evaluación del Modelo:
MAE: 15.08
MSE: 3116.77
R² Score: 0.2886
⚠ El modelo no es confiable para predecir 2014-1, se requiere ajuste.
```

Modelos de Predicción

Objetivo: Predecir el número de matrículas utilizando diferentes modelos de Machine Learning.

A. Random Forest Regressor

Un algoritmo de aprendizaje automático basado en múltiples árboles de decisión. Es más robusto que la regresión lineal y maneja mejor la variabilidad de los datos.

Métricas de Evaluación:

MAE (Error Absoluto Medio)
MSE (Error Cuadrático Medio)
R^2 Score (Indica qué tan bien el modelo explica la variabilidad)

Modelos de Predicción – Random Forest Regressor

Código utilizado:

```
datos = datos.select_dtypes(include=['number'])

Y_2013_1 = datos['Total 2013-1']
Y_2013_2 = datos['Total 2013-2']

X = datos.drop(columns=['Total 2013-1', 'Total 2013-2'])

X_train, X_test, y_train_2013_1, y_test_2013_1 = train_test_split(X, Y_2013_1, test_size=0.2, random_state=42)
X_train, X_test, y_train_2013_2, y_test_2013_2 = train_test_split(X, Y_2013_2, test_size=0.2, random_state=42)

modelo_rf_2013_1 = RandomForestRegressor(n_estimators=200, random_state=42)
modelo_rf_2013_1.fit(X_train, y_train_2013_1)

modelo_rf_2013_2 = RandomForestRegressor(n_estimators=200, random_state=42)
modelo_rf_2013_2.fit(X_train, y_train_2013_2)

datos['Predicción 2013-1'] = modelo_rf_2013_1.predict(X)
datos['Predicción 2013-2'] = modelo_rf_2013_2.predict(X)
```

Modelos de Predicción – Random Forest Regressor

Comparativa entre los datos de predicción y los datos reales:

```
comparativa_2013_1 = pd.DataFrame({
    "Predicción 2013-1": modelo_rf_2013_1.predict(X_test),
    "Valor Real 2013-1": y_test_2013_1
})
pd.DataFrame(comparativa_2013_1)
```

Resultados:

	Predicción 2013-1	Valor Real 2013-1
16341	0.00	0
18596	0.00	0
15483	0.00	0
17177	6.01	6
11974	74.19	73
...	...	...
10011	1.00	1
11142	2.00	2
20964	0.00	0
10430	1.00	1
5627	240.86	242

Modelos de Predicción – Random Forest Regressor

Comparativa entre los datos de predicción y los datos reales:

```
comparativa_2013_2 = pd.DataFrame({
    "Predicción 2013-2": modelo_rf_2013_2.predict(X_test),
    "Valor Real 2013-2": y_test_2013_2
})
pd.DataFrame(comparativa_2013_2)
```

Resultados:

	Predicción 2013-2	Valor Real 2013-2
16341	2.000	2
18596	0.000	0
15483	0.000	0
17177	0.000	0
11974	52.150	52
...	...	...
10011	1.000	1
11142	0.000	0
20964	0.000	0
10430	3.000	3
5627	201.725	205

Modelos de Predicción – Random Forest Regressor

Evaluación del modelo:

```
mae_2013_1 = mean_absolute_error(Y_2013_1, datos['Predicción 2013-1'])
mse_2013_1 = mean_squared_error(Y_2013_1, datos['Predicción 2013-1'])
r2_2013_1 = r2_score(Y_2013_1, datos['Predicción 2013-1'])

mae_2013_2 = mean_absolute_error(Y_2013_2, datos['Predicción 2013-2'])
mse_2013_2 = mean_squared_error(Y_2013_2, datos['Predicción 2013-2'])
r2_2013_2 = r2_score(Y_2013_2, datos['Predicción 2013-2'])

print(f' Evaluación del modelo para Total 2013-1:')
print(f' MAE: {mae_2013_1:.2f}')
print(f' MSE: {mse_2013_1:.2f}')
print(f' R² Score: {r2_2013_1:.4f}')
print('-------------------------------------')
print(f' Evaluación del modelo para Total 2013-2:')
print(f' MAE: {mae_2013_2:.2f}')
print(f' MSE: {mse_2013_2:.2f}')
print(f' R² Score: {r2_2013_2:.4f}')
```

Modelos de Predicción – Random Forest Regressor

Resultados de la Evaluación del modelo:

```
 Evaluación del modelo para Total 2013-1:
 MAE: 0.23
 MSE: 31.30
 R² Score: 0.9898
--------------------------------------
 Evaluación del modelo para Total 2013-2:
 MAE: 0.30
 MSE: 67.52
 R² Score: 0.9869
Archivo 'predicciones_2013.xlsx' guardado con éxito.
```

Visualización de Resultados

Objetivo: Comparar predicciones con valores reales.

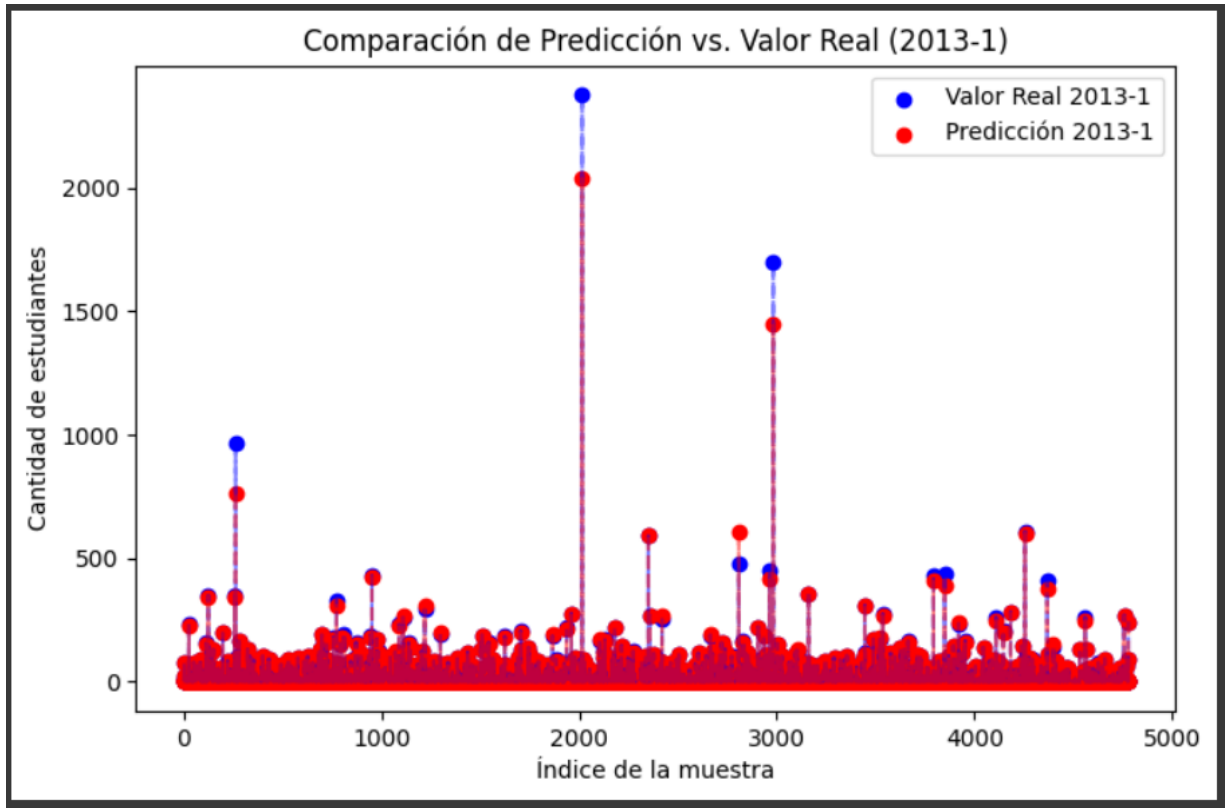
Código de generación de gráficos de dispersión (matplotlib)

```
import matplotlib.pyplot as plt

plt.figure(figsize=(8, 5))
plt.scatter(range(len(y_test_2013_1)), y_test_2013_1, color='blue', label='Valor Real 2013-1')
plt.scatter(range(len(y_test_2013_1)), modelo_rf_2013_1.predict(X_test), color='red', label='Predicción 2013-1')
plt.plot(range(len(y_test_2013_1)), y_test_2013_1, color='blue', linestyle='dashed', alpha=0.5)
plt.plot(range(len(y_test_2013_1)), modelo_rf_2013_1.predict(X_test), color='red', linestyle='dashed', alpha=0.5)
plt.xlabel("Índice de la muestra")
plt.ylabel("Cantidad de estudiantes")
plt.title("Comparación de Predicción vs. Valor Real (2013-1)")
plt.legend()
plt.show()
```

Visualización de Resultados

Comparación de valores reales vs predicciones para el semestre 2013-1



Visualización de Resultados

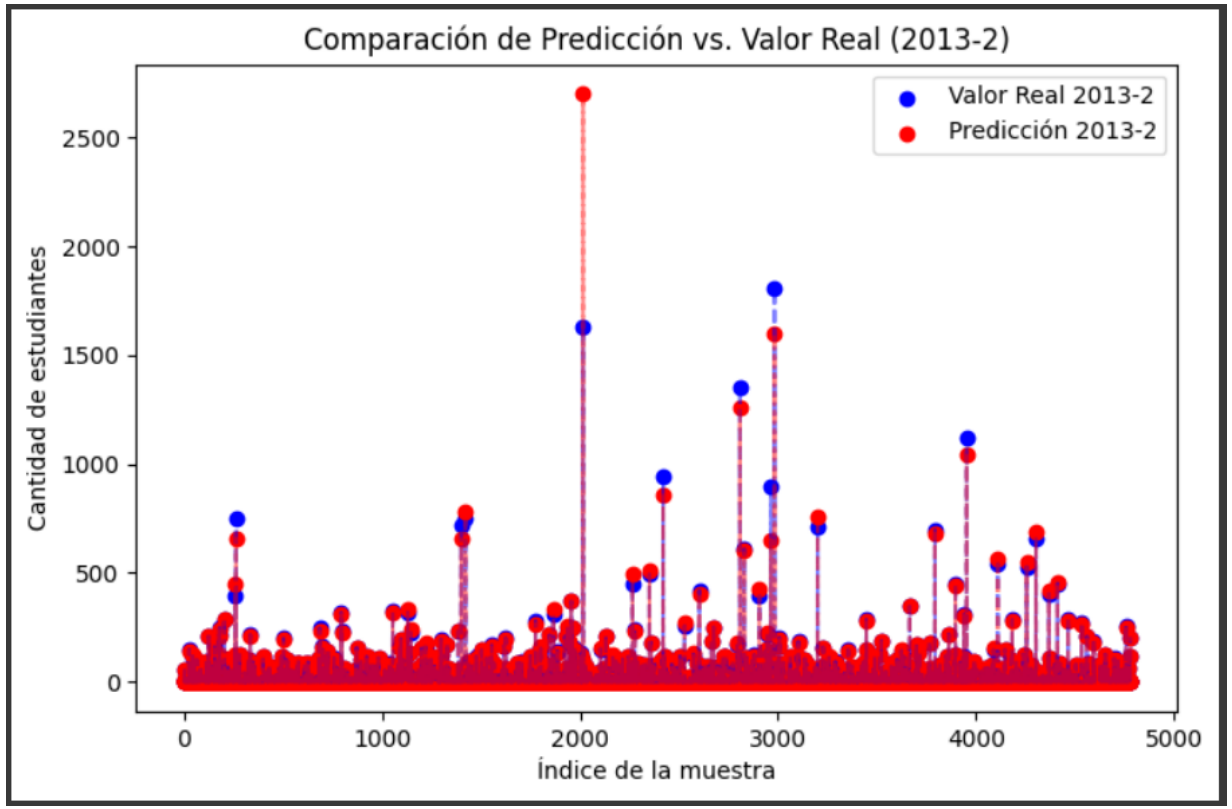
Objetivo: Comparar predicciones con valores reales.

Código de generación de gráficos de dispersión (matplotlib)

```
plt.figure(figsize=(8, 5))
plt.scatter(range(len(y_test_2013_2)), y_test_2013_2, color='blue', label='Valor Real 2013-2')
plt.scatter(range(len(y_test_2013_2)), modelo_rf_2013_2.predict(X_test), color='red', label='Predicción 2013-2')
plt.plot(range(len(y_test_2013_2)), y_test_2013_2, color='blue', linestyle='dashed', alpha=0.5)
plt.plot(range(len(y_test_2013_2)), modelo_rf_2013_2.predict(X_test), color='red', linestyle='dashed', alpha=0.5)
plt.xlabel("Índice de la muestra")
plt.ylabel("Cantidad de estudiantes")
plt.title("Comparación de Predicción vs. Valor Real (2013-2)")
plt.legend()
plt.show()
```

Visualización de Resultados

Comparación de valores reales vs predicciones para el semestre 2013-2



Predicción de Matrículas

Objetivo: Usar el modelo para predecir datos futuros.

En este proceso se usa el modelo entrenado y evaluado para estimar Total 2014-1 y Total 2014-2.

```
columnas_modelo = modelo_rf_2013_1.feature_names_in_

X_prediccion = datos.drop(columns=['Total 2013-1', 'Total 2013-2'], errors='ignore')

datos['Total 2014-1 (Predicho)'] = modelo_rf_2013_1.predict(X_prediccion[columnas_modelo])

X_prediccion['Total 2014-1'] = datos['Total 2014-1 (Predicho)']

columnas_modelo_2013_2 = modelo_rf_2013_2.feature_names_in_

datos['Total 2014-2 (Predicho)'] = modelo_rf_2013_2.predict(X_prediccion[columnas_modelo_2013_2])

predicciones_2014 = datos[['Total 2014-1 (Predicho)', 'Total 2014-2 (Predicho)']]
print(predicciones_2014)

predicciones_2014.to_excel("Predicciones_2014.xlsx", index=False)
print(" Archivo 'Predicciones_2014.xlsx' generado con éxito.")
```

Resultados de la Predicción

	Total 2014-1 (Predicho)	Total 2014-2 (Predicho)
0	128.580	72.785
1	51.735	47.345
2	52.985	40.900
3	35.150	29.855
4	52.665	42.935
...	...	...
23886	69.050	40.205
23887	0.000	5.000
23888	16.960	0.000
23889	7.000	1.000
23890	60.840	48.925

Conclusiones

Análisis final del proyecto:

Lo que se logró:

- Modelar y analizar tendencias de matrícula en universidades.
- Comparar modelos de predicción y determinar cuál funciona mejor.
- Generar predicciones para el 2014 basadas en datos reales.

Limitaciones:

- La Regresión Lineal no es precisa para este tipo de datos.
- El modelo Random Forest tiene mejor rendimiento, pero depende de la calidad de los datos.

Referencias

- Ministerio de Educación Nacional. (2021). Sistema Nacional de Información de la Educación Superior (SNIES) https://snies.mineducacion.gov.co/1778/articles-391574_recurso.xlsb
- Pérez, M., & Rodríguez, L. (2020). Análisis predictivo en la educación superior: retos y oportunidades. Editorial Académica.
- Ramírez, J. (2019). Big Data y educación: un enfoque innovador para la toma de decisiones. Universidad Nacional de Colombia.

Corporación
UNIVERSITARIA
AUTONOMA
DE NARIÑO
AUNAR VILLAVICENCIO

VIGILADA MINEDUCACIÓN

¡Gracias!

/ @Aunarvillavo
www.aunarvillavicencio.edu.co

#AUNAR Villavicencio