

Presentación de resultados del proyecto

José Pablo Kiesling Lange - 21581
Erick Stiv Junior Guerra Muñoz - 21781
Diego Andrés Morales Aquino - 21762
Pablo Andrés Zamora Vásquez - 21780

Guatemala, 23 de mayo de 2024

Determine cuál será su variable respuesta, explique la elección de la misma

Al realizar el análisis exploratorio, se halló que el acceso a condiciones óptimas de vida tales como agua potable, saneamiento, electricidad, tecnología, entre otros, son fundamentales para el progreso educativo y laboral en campos industriales y de servicio en los países; es por ello que el propósito principal del proyecto es determinar cómo se ve afectado el acceso a la educación en diferentes países, en base a sus condiciones sociales, económicas y laborales; así como el acceso a diferentes servicios a lo largo de las últimas décadas. Por lo tanto, la variable respuesta del experimento será **“años esperados de escolaridad”**.

La variable seleccionada nos permitirá determinar la expectativa de años de escolaridad para un niño en edad de ingresar a la escuela en cada país, en base a las variables que representan las condiciones mencionadas anteriormente, tales como el Índice de Desarrollo Humano (IDH), porcentaje de la población total con acceso a diferentes servicios, la tasa de fertilidad adolescente, el promedio de años de escolaridad, las tasas de desigualdad en ingresos y educación, la esperanza de vida, entre otros.

Explicación del método utilizado para obtener los conjuntos de entrenamiento y prueba

Para evaluar la capacidad de generalización de los modelos empleados, únicamente se requirió separar el conjunto de datos en dos subconjuntos: entrenamiento y prueba; se llegó a esta conclusión puesto que se empleó validación cruzada, con la cual se dividió repetidamente el conjunto de entrenamiento en diferentes subconjuntos para entrenar y validar el modelo, proporcionando una evaluación robusta del rendimiento del modelo. El proceso descrito causa que resulte innecesario un conjunto de validación separado, ya que la validación cruzada ya satisface el propósito de evaluar el modelo de manera confiable, además de ayudar a prevenir el sobreajuste, con lo que podemos tomarnos la libertad de utilizar más datos para el entrenamiento en función de mejorar la capacidad del modelo.

Para obtener los conjuntos de entrenamiento y prueba se utilizó el método *“train_test_split”* de la biblioteca *“scikit-learn”*. Se especificó que el tamaño del conjunto de prueba fuera un tercio del total de los datos, garantizando que el modelo se entrene con el 67% de los datos y, al finalizar su ajuste, se evalúe con el 33% restante.

Utilice los algoritmos que determinó que serían los más útiles para aplicar a sus datos. Explique las transformaciones que tuvo que hacer para aplicarlos correctamente.

Los algoritmos de regresión que se definieron como los más útiles para determinar la expectativa de años de escolaridad fueron: **regresión lineal múltiple, árbol de decisión, random forest y SVR**.

Antes de entrenar cualquier modelo, fue necesario ordenar y limpiar los datos. El conjunto de datos proviene de dos fuentes: indicadores de países relacionados a los ODS, provenientes del Banco Mundial; y dimensiones del IDH para cada país, provenientes del Programa de Desarrollo de las Naciones Unidas. Debido a esto, fue necesario combinar inicialmente estos datos en un solo conjunto.

Originalmente, el *dataset* de los indicadores relacionados a los ODS contaba con 106,488 observaciones y 36 variables, mientras que el *dataset* de las dimensiones del IDH contaba con 206 observaciones y 1,076 variables.

Imagen 1: Cantidad de variables y observaciones del primer dataset

```
num_rows, num_columns = datos.shape

print("Observaciones:", num_rows)
print("Variables:", num_columns)
```

✓ 0.0s

Observaciones: 106488
Variables: 36

Imagen 2: Cantidad de variables y observaciones del segundo dataset

```
num_rows, num_columns = indicators_data.shape

print('Observaciones:', num_rows)
print('Variables:', num_columns)
```

✓ 0.0s

Observaciones: 206
Variables: 1076

Al eliminar las variables que no son de interés del primer *dataset* y agrupar las observaciones por país y año, se obtienen 8,083 observaciones y 18 variables.

Imagen 3: Primer dataset agrupado por país y año

| Indicator Name | Country Code | Year | Access to clean fuels and technologies for cooking (% of population) | Access to electricity (% of population) | Adolescent fertility rate (births per 1000 women ages 15-19) | Adolescents out of school (% of lower secondary school age) | Employment in agriculture (% of total employment) (modeled ILO estimate) | Employment in industry (% of total employment) (modeled ILO estimate) | Employment in services (% of total employment) (modeled ILO estimate) | GDP per capita (current US\$) | Households and NPISHs Final consumption expenditure (annual % growth) |
|----------------|--------------|------|----------------------------------------------------------------------|-----------------------------------------|--------------------------------------------------------------|-------------------------------------------------------------|--------------------------------------------------------------------------|-----------------------------------------------------------------------|-----------------------------------------------------------------------|-------------------------------|-----------------------------------------------------------------------|
| 0 | ABW | 1990 | NaN | 100.000000 | 50.2700 | NaN | NaN | NaN | NaN | 12305.387910 | NaN |
| 1 | ABW | 1991 | NaN | 99.153656 | 49.7000 | NaN | NaN | NaN | NaN | 13494.685160 | NaN |
| 2 | ABW | 1992 | NaN | 99.197128 | 49.1300 | NaN | NaN | NaN | NaN | 14048.347270 | NaN |
| 3 | ABW | 1993 | NaN | 99.239914 | 48.7648 | NaN | NaN | NaN | NaN | 14942.274960 | NaN |
| 4 | ABW | 1994 | NaN | 100.000000 | 48.3996 | NaN | NaN | NaN | NaN | 16241.575590 | NaN |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 8078 | ZWE | 2016 | 29.8 | 42.561729 | 90.6766 | 2.261470 | 66.879997 | 7.05 | 26.070000 | 1464.588957 | -10.501114 |
| 8079 | ZWE | 2017 | 29.8 | 44.178635 | 86.1350 | 2.356790 | 66.480003 | 6.90 | 26.629999 | 1235.189032 | 1.396143 |
| 8080 | ZWE | 2018 | 29.9 | 45.572647 | 83.2486 | 6.226070 | 66.019997 | 6.75 | 27.230000 | 1254.642265 | -22.393220 |
| 8081 | ZWE | 2019 | 30.1 | 46.781475 | 80.3622 | 7.544310 | 66.190002 | 6.57 | 27.240000 | 1316.740657 | NaN |
| 8082 | ZWE | 2020 | 30.4 | 52.747669 | 77.4758 | 21.875412 | NaN | NaN | NaN | 1214.509820 | NaN |

8083 rows x 18 columns

De igual manera, al realizar el mismo procedimiento con el segundo *dataset*, se obtienen 6,798 observaciones y 8 variables.

Imagen 4: Segundo dataset agrupado por país y año

| variable | Country Code | country | Year | hdi | eys | le | mys | pop_total |
|----------|--------------|-------------|------|-------|-----------|-----------|----------|-------------|
| 0 | AFG | Afghanistan | 1990 | 0.284 | 2.936460 | 45.967000 | 0.871962 | 10.694796 |
| 1 | AFG | Afghanistan | 1991 | 0.292 | 3.228456 | 46.663000 | 0.915267 | 10.745167 |
| 2 | AFG | Afghanistan | 1992 | 0.299 | 3.520452 | 47.596000 | 0.958573 | 12.057433 |
| 3 | AFG | Afghanistan | 1993 | 0.307 | 3.812448 | 51.466000 | 1.001878 | 14.003760 |
| 4 | AFG | Afghanistan | 1994 | 0.300 | 4.104445 | 51.495000 | 1.045184 | 15.455554 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 6793 | ZZK.WORLD | World | 2018 | 0.735 | 12.651182 | 72.816178 | 8.483187 | 7649.567694 |
| 6794 | ZZK.WORLD | World | 2019 | 0.739 | 12.726141 | 73.012121 | 8.574383 | 7730.660623 |
| 6795 | ZZK.WORLD | World | 2020 | 0.736 | 12.860674 | 72.257355 | 8.686374 | 7806.602793 |
| 6796 | ZZK.WORLD | World | 2021 | 0.735 | 12.911840 | 71.365524 | 8.682265 | 7874.877719 |
| 6797 | ZZK.WORLD | World | 2022 | 0.739 | 12.989766 | 72.004070 | 8.661598 | 7940.608797 |

6798 rows × 8 columns

Finalmente, al combinar ambos *datasets* en el mismo conjunto de datos, se obtienen 6,045 observaciones y 24 variables.

Imagen 5: Conjunto de datos con ambos datasets iniciales

| Country Code | Year | Access to clean fuels and technologies for cooking (% of population) | Access to electricity (% of population) | Adolescent fertility rate (births per 1000 women ages 15-19) | Adolescents out of school (% of lower secondary school age) | Employment in agriculture (% of total employment) (modeled ILO estimate) | Employment in industry (% of total employment) (modeled ILO estimate) | Employment in services (% of total employment) (modeled ILO estimate) | GDP per capita (current US\$) | ... | People using safely managed drinking water services (% of population) | Primary completion rate total (% of relevant age group) | Unemployment total (% of total labor force) (modeled ILO estimate) | Urban population (% of total population) | |
|------------------------|------|----------------------------------------------------------------------|-----------------------------------------|--------------------------------------------------------------|-------------------------------------------------------------|--------------------------------------------------------------------------|-----------------------------------------------------------------------|-----------------------------------------------------------------------|-------------------------------|-------------|-----------------------------------------------------------------------|---------------------------------------------------------|--------------------------------------------------------------------|------------------------------------------|--------|
| 0 | AFG | 1990 | NaN | NaN | 162.8354 | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN | 21.177 | |
| 1 | AFG | 1991 | NaN | NaN | 163.3822 | NaN | 63.430000 | 12.43 | 24.139999 | NaN | ... | NaN | NaN | 21.266 | |
| 2 | AFG | 1992 | NaN | NaN | 163.9290 | NaN | 63.650002 | 12.21 | 24.139999 | NaN | ... | NaN | NaN | 21.355 | |
| 3 | AFG | 1993 | NaN | NaN | 164.1888 | NaN | 64.410004 | 11.57 | 24.030001 | NaN | ... | NaN | 27.575171 | 21.444 | |
| 4 | AFG | 1994 | NaN | NaN | 164.4486 | NaN | 64.440002 | 11.48 | 24.080000 | NaN | ... | NaN | NaN | 21.534 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 6040 | ZWE | 2016 | 29.8 | 42.561729 | 90.6766 | 2.261470 | 66.879997 | 7.05 | 26.070000 | 1464.588957 | ... | 30.183683 | 99.265839 | 4.788 | 32.296 |
| 6041 | ZWE | 2017 | 29.8 | 44.178635 | 86.1350 | 2.356790 | 66.480003 | 6.90 | 26.629999 | 1235.189032 | ... | 29.997683 | 95.476372 | 4.785 | 32.237 |
| 6042 | ZWE | 2018 | 29.9 | 45.572647 | 83.2486 | 6.226070 | 66.019997 | 6.75 | 27.230000 | 1254.642265 | ... | 29.827913 | 92.195152 | 4.796 | 32.209 |
| 6043 | ZWE | 2019 | 30.1 | 46.781475 | 80.3622 | 7.544310 | 66.190002 | 6.57 | 27.240000 | 1316.740657 | ... | 29.673569 | 88.508812 | 4.833 | 32.210 |
| 6044 | ZWE | 2020 | 30.4 | 52.747669 | 77.4758 | 21.875412 | NaN | NaN | NaN | 1214.509820 | ... | 29.535890 | 90.017349 | 5.351 | 32.242 |
| 6045 rows × 24 columns | | | | | | | | | | | | | | | |

Como siguiente paso, se determinó la cantidad de datos faltantes en el nuevo conjunto formado.

Imagen 6: Cantidad de datos faltantes por variable

| | |
|--------------------------------------------------------------------------|------|
| Valores pre limpieza | |
| Valores NaN: | |
| Country Code | 0 |
| Year | 0 |
| Access to clean fuels and technologies for cooking (% of population) | 2059 |
| Access to electricity (% of population) | 787 |
| Adolescent fertility rate (births per 1000 women ages 15-19) | 310 |
| Employment in agriculture (% of total employment) (modeled ILO estimate) | 825 |
| Employment in industry (% of total employment) (modeled ILO estimate) | 825 |
| Employment in services (% of total employment) (modeled ILO estimate) | 825 |
| GDP per capita (current US\$) | 247 |
| Individuals using the Internet (% of population) | 336 |
| People using at least basic drinking water services (% of population) | 2025 |
| People using at least basic sanitation services (% of population) | 2034 |
| Primary completion rate total (% of relevant age group) | 2633 |
| Unemployment total (% of total labor force) (modeled ILO estimate) | 645 |
| Urban population (% of total population) | 0 |
| country | 0 |
| hdi | 622 |
| ey | 245 |
| le | 0 |
| mys | 539 |
| pop_total | 0 |
| dtype: int64 | |

Una vez hecho esto, se decidió eliminar las variables con una cantidad de datos faltante mayor al 40% de la totalidad de los datos; estas son: “Porcentaje de la población utilizando servicios de agua potable manejados de manera segura”, “Porcentaje de adolescentes fuera de la escuela” y “Consumo final de hogares e instituciones privadas sin fines de lucro al servicio de los hogares”. Para completar los datos faltantes de las variables con un porcentaje de estos menor o igual al 40% de la totalidad de los datos se decidió utilizar el promedio.

Una vez realizado esto, el conjunto de datos ya era apto para entrenar los modelos.

Imagen 7: Conjunto de datos final

| | Country Code | Year | Access to clean fuels and technologies for cooking (% of population) | Access to electricity (% of population) | Adolescent fertility rate (births per 1000 women ages 15-19) | Employment in agriculture (% of total employment) (modeled ILO estimate) | Employment in industry (% of total employment) (modeled ILO estimate) | Employment in services (% of total employment) (modeled ILO estimate) | GDP per capita (current US\$) | Individuals using the Internet (% of population) | ... | People using at least basic sanitation services (% of population) | Primary completion rate total (% of relevant age group) | Unemployment total (% of total labor force) (modeled ILO estimate) | Urban population (% of total population) |
|------|--------------|------|----------------------------------------------------------------------|-----------------------------------------|--------------------------------------------------------------|--------------------------------------------------------------------------|-----------------------------------------------------------------------|-----------------------------------------------------------------------|-------------------------------|--------------------------------------------------|-----|-------------------------------------------------------------------|---------------------------------------------------------|--------------------------------------------------------------------|------------------------------------------|
| 0 | AFG | 1990 | 62.983721 | 79.109812 | 162.8354 | 30.544245 | 19.891328 | 49.565843 | 11208.654030 | 0.000000 | ... | 71.981983 | 85.880304 | 8.220484 | 21.177 |
| 1 | AFG | 1991 | 62.983721 | 79.109812 | 163.3822 | 63.430000 | 12.430000 | 24.139999 | 11208.654030 | 0.000000 | ... | 71.981983 | 85.880304 | 10.649000 | 21.266 |
| 2 | AFG | 1992 | 62.983721 | 79.109812 | 163.9290 | 63.650002 | 12.210000 | 24.139999 | 11208.654030 | 0.000000 | ... | 71.981983 | 85.880304 | 10.821000 | 21.355 |
| 3 | AFG | 1993 | 62.983721 | 79.109812 | 164.1888 | 64.410004 | 11.570000 | 24.030001 | 11208.654030 | 0.000000 | ... | 71.981983 | 27.575171 | 10.723000 | 21.444 |
| 4 | AFG | 1994 | 62.983721 | 79.109812 | 164.4486 | 64.440002 | 11.480000 | 24.080000 | 11208.654030 | 0.000000 | ... | 71.981983 | 85.880304 | 10.726000 | 21.534 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 6040 | ZWE | 2016 | 29.800000 | 42.561729 | 90.6766 | 66.879997 | 7.050000 | 26.070000 | 1464.588957 | 23.119989 | ... | 37.529193 | 99.265839 | 4.788000 | 32.296 |
| 6041 | ZWE | 2017 | 29.800000 | 44.178635 | 86.1350 | 66.480003 | 6.900000 | 26.629999 | 1235.189032 | 24.400000 | ... | 36.941673 | 95.476372 | 4.785000 | 32.237 |
| 6042 | ZWE | 2018 | 29.900000 | 45.572647 | 83.2486 | 66.019997 | 6.750000 | 27.230000 | 1254.642265 | 25.000000 | ... | 36.357160 | 92.195152 | 4.796000 | 32.209 |
| 6043 | ZWE | 2019 | 30.100000 | 46.781475 | 80.3622 | 66.190002 | 6.570000 | 27.240000 | 1316.740657 | 25.100000 | ... | 35.774337 | 88.508812 | 4.833000 | 32.210 |
| 6044 | ZWE | 2020 | 30.400000 | 52.747669 | 77.4758 | 30.544245 | 19.891328 | 49.565843 | 1214.509820 | 29.300000 | ... | 35.192363 | 90.017349 | 5.351000 | 32.242 |

6045 rows × 16 columns

Cree varios modelos, con los algoritmos seleccionados. Varíe los parámetros buscando el mejor modelo sin caer en un modelo desajustado o sobreajustado. Construya un modelo final con los mejores parámetros. Explique con datos como llegó a ellos.

Para los algoritmos de árbol de decisión, random forest y SVR, se implementó la comparación automática de múltiples modelos mediante *GridSearch*, combinando distintas variaciones de los hiperparámetros de interés. El objetivo fue determinar el modelo que maximiza el rendimiento según el error cuadrático medio para cada algoritmo. Además, se utilizó validación cruzada con 5 *folds*, lo cual asegura una evaluación robusta y confiable del rendimiento de cada combinación de hiperparámetros.

Imagen 8. Parámetros utilizados para construir modelos con Árboles de decisión

```
param_grid = {  
    'max_depth': [None, 10, 20, 30],  
    'min_samples_split': [2, 10, 20],  
    'min_samples_leaf': [1, 5, 10],  
    'max_features': [None, 'auto', 'sqrt', 'log2'],  
    'ccp_alpha': [0.0, 0.01, 0.1, 1.0]  
}
```

Luego de evaluar 576 combinaciones de modelos con 5 pliegues para validación cruzada para el algoritmo de árbol de decisión, el modelo que optimiza su error cuadrático medio posee los parámetros por defecto: `ccp_alpha = 0.0`, `max_depth = None`, `max_features = None`, `min_samples_leaf = 1` y `min_samples_split = 2`.

Imagen 9. Parámetros utilizados para construir modelos con Random Forest.

```
param_grid = {  
    'n_estimators': [100, 200, 300, 1000],  
    'max_depth': [None, 10, 20, 30],  
    'min_samples_split': [2, 10, 20],  
    'min_samples_leaf': [1, 5, 10],  
    'max_features': ['auto', 'sqrt', 'log2'],  
    'bootstrap': [True, False]  
}
```

Imagen 10. Parámetros utilizados para construir modelos con SVR.

```
param_grid = {  
    'C': [0.1, 1, 10, 100, 500, 1000],  
    'epsilon': [0.01, 0.1, 0.2, 0.5],  
    'kernel': ['linear', 'rbf', 'poly'],  
    'gamma': ['scale', 'auto', 0.1, 1, 10],  
}
```

Como se muestra en las *imágenes 9 y 10*, en el caso de los algoritmos random forest y SVR, no fue posible evaluar la combinación completa de todos los hiperparámetros, debido a la alta demanda de recursos computacionales que esto requiere, por lo que se evaluó sobre cada uno de los parámetros con sus posibles valores de forma individual, para luego determinar el modelo con el mejor rendimiento.

El mejor modelo para el algoritmo de Random Forest incluye los parámetros `n_estimators = 1000`, `bootstrap=False`, `max_features='log2'`. Mientras que, el mejor modelo para el algoritmo SVT corresponde a la combinación de parámetros `C=500`, `epsilon=0.01`, `kernel='rbf'`, `gamma='scale'`.

Una vez se obtuvieron los modelos y sus mejores parámetros, se hicieron predicciones con los mismos con una observación existente en el conjunto de datos y una observación con valores de prueba que no están en el conjunto de datos. Para el dato existente, se utilizó el registro de Guatemala en 2020, el cual contiene los siguientes datos:

Imagen 11. Datos del país seleccionado del conjunto de datos

| | |
|--------------------------------------------------------------------------|-------------|
| Country Code | GTM |
| Year | 2020 |
| Access to clean fuels and technologies for cooking (% of population) | 50.1 |
| Access to electricity (% of population) | 97.055267 |
| Adolescent fertility rate (births per 1000 women ages 15-19) | 67.408 |
| Employment in agriculture (% of total employment) (modeled ILO estimate) | 30.544245 |
| Employment in industry (% of total employment) (modeled ILO estimate) | 19.891328 |
| Employment in services (% of total employment) (modeled ILO estimate) | 49.565843 |
| GDP per capita (current US\$) | 4603.339643 |
| Individuals using the Internet (% of population) | 49.97 |
| People using at least basic drinking water services (% of population) | 94.006428 |
| People using at least basic sanitation services (% of population) | 67.928581 |
| Primary completion rate total (% of relevant age group) | 80.736687 |
| Unemployment total (% of total labor force) (modeled ILO estimate) | 3.553 |
| Urban population (% of total population) | 51.836 |
| country | Guatemala |
| hdi | 0.638 |
| eys | 10.767204 |
| le | 71.797 |
| mys | 5.668648 |
| pop_total | 17.362718 |

Como se puede observar en la *Imagen 11*, el valor de la variable objetivo, **Años esperados de escolaridad**, es 10.77, lo que indica que se espera obtener este valor en las predicciones de los modelos.

Cuadro 1. Predicciones de años esperados de escolaridad de Guatemala en 2020, utilizando los modelos seleccionados con los mejores parámetros

| Modelo (algoritmo) | Predicción | % error |
|---------------------------|------------|---------|
| Regresión lineal múltiple | 11.43 | 6.13 |
| Árbol de decisión | 10.79 | 0.19 |
| Random Forest | 11.48 | 6.59 |
| SVR | 10.89 | 1.11 |

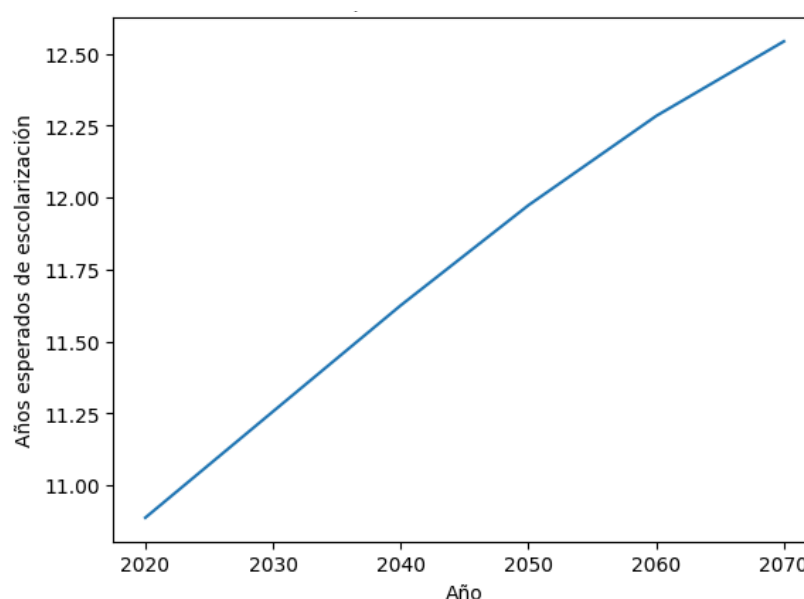
Según lo que se observa en el *cuadro 1*, la mejor predicción fue la del Árbol de decisión, teniendo un valor de 10.79 y un % de error de 0.19. Por otro lado, la peor predicción fue la de Regresión lineal múltiple con un valor de 11.43 y un % de error de 6.13.

Para la predicción de la observación con valores de prueba que no están en el conjunto de datos, se utilizaron los mismos valores que tenía Guatemala en 2020 y se fue incrementando el año de 10 en 10 para predecir cómo cambiarían los años esperados de escolaridad en el país. Cabe destacar que, al ser una predicción de un dato que no estaba en el conjunto de datos, no se calculó % de error. Asimismo, en los modelos de **Árbol de decisión** y **Random Forest** no hubo cambios entre años en la predicción, dada la naturaleza del algoritmo.

Por consiguiente, se optó por usar SVR como único modelo de predicción ya que este, como se discutirá más adelante, es el mejor modelo de regresión según el coeficiente de determinación.

Tomando en cuenta lo anterior, se muestra la siguiente gráfica para ilustrar el crecimiento de años esperados de escolaridad en Guatemala de 2020 a 2070.

Gráfico 1. Predicción de años esperados de escolarización en Guatemala.



Como se puede observar, los años esperados de escolaridad en Guatemala en el futuro aumentan conforme avanzan los años; específicamente, la regresión polinomial que siguen los años de escolaridad esperados contra el tiempo es:

$$11.8 + 0.84x - 0.09x^2$$

donde x es el año de interés de Guatemala con los valores de 2020.

Discuta los resultados de los algoritmos incluyendo la matriz de confusión obtenida, en caso de que sea un problema de clasificación, o la métrica de error que considere mejor si el problema es de regresión.

Para evaluar los modelos de regresión se utilizó como métrica el coeficiente de determinación, en el que los valores más cercanos a 1 positivo reflejan una mayor capacidad del modelo para obtener resultados de mayor precisión.

Al evaluar dicha métrica con el conjunto de datos de prueba, se obtienen los siguientes resultados:

Cuadro 2. Coeficiente de determinación (R^2) obtenido para cada modelo utilizado.

| Modelo (algoritmo) | Coeficiente de determinación (R^2) |
|---------------------------|----------------------------------------|
| Regresión lineal múltiple | 0.9248 |
| Árbol de decisión | 0.9452 |
| Random Forest | 0.9790 |
| SVR | 0.9821 |

Como se aprecia en el *cuadro 2*, el modelo con el mejor coeficiente de determinación es SVR, con un valor de 0.9821. Como se mencionó anteriormente, al aplicar *GridSearch* se seleccionaron los mejores parámetros para cada modelo. Cabe destacar que este método ya aplica validación cruzada, por lo que se revisó automáticamente el ajuste de los modelos para finalmente obtener en cada uno de ellos un coeficiente de determinación bastante elevado y válido para el estudio realizado.

Link al repositorio de GitHub: https://github.com/Aq202/uvg_mineria_proyecto1.git