WORLD HAPPINESS REPORT PREDICTIONS

CURSO: DATA SCIENCE

AUTOR: LORENZO ALLIOT

ÍNDICE

ABSTRACT

Contexto comercial:

La ONU ha iniciado un nuevo y ambicioso proyecto, **hacer del mundo un lugar más feliz para sus habitantes.**
Mi trabajo consistirá en determinar cuáles son los puntos más importantes en los que se debe enfocar cada país y así mejorar su situación actual.
Para ahorrar costos y tiempos de encuestas la ONU decidió tomar como fuente de datos los reportes publicados por "The World Happiness Report" desde 2015 hasta 2022 porque considera que en ellos se engloban los indicadores más importantes para el desarrollo de la felicidad de los habitantes de un país.

Problema comercial:

- ¿Hay algún patrón entre los indicadores que componen el Score de felicidad?
- ¿Cuál es el indicador más importante?
- ¿En que indicadores se deberían enfocar los países para mejorar su performance?
- ¿Qué regiones son las más felices?

Objetivo:

Generar el modelo de ML más óptimo para predecir los Scores futuros de los países.

Explicación del data set

Descripción de las variables

Country : Nombre del país.

Region : Region a la que pertenece el país.

Happiness Rank : Ranking del país basado en el Score de felicidad.

Happiness Score : Una métrica medida en 2015 preguntando a las personas de la muestra: "¿Cómo calificaría su felicidad en una escala del 0 al 10, donde 10 es el más feliz".

GDP : Producto Bruto Interno (PBI) del país

Family : Ayuda social del gobierno

Health: Expectativa de vida saludable

Freedom : Libertad para tomar decisiones

Trust : Percepción de corrupción del gobierno

Generosity : percepción de generosidad de la población

Dystopia : Cada país es comparado con un país hipotético que representa el peor desempeño de cada variable, es decir, ningún país tiene peor desempeño que Dystopia, y su valor representa la suma de la distancia de cada variable del país con Dystopia.

Desempeño: Variable categórica que indica en qué categoría se encuentra el país según su desempeño.

Population: Es la población de cada país en el año indicado.

Resumen de la metada

- Hay 168 paises distintos
- Hay 8 regiones distintas
- La región con más paises es Sub-Saharan Afica con 304 Paises
- Solo el campo Region tiene 2 nulos
- Los campos GDP, Family, Life Expectancy, Freedom, Trust, Generosity, Dystopia son todos numéricos, están todos normalizados y conforman el campo Score
- El minimo de población es 0 por lo que al menos un país está null

Campo	dtype
Country	object
Region	object
Rank	int64
Score	float64
GDP	float64
Family	float64
Life Expectancy	float64
Freedom	float64
Trust	float64
Generosity	float64
Dystopia	float64
Year	int64
Desempeño	object
Population	int64

ANÁLISIS **EXPLORATORIO**

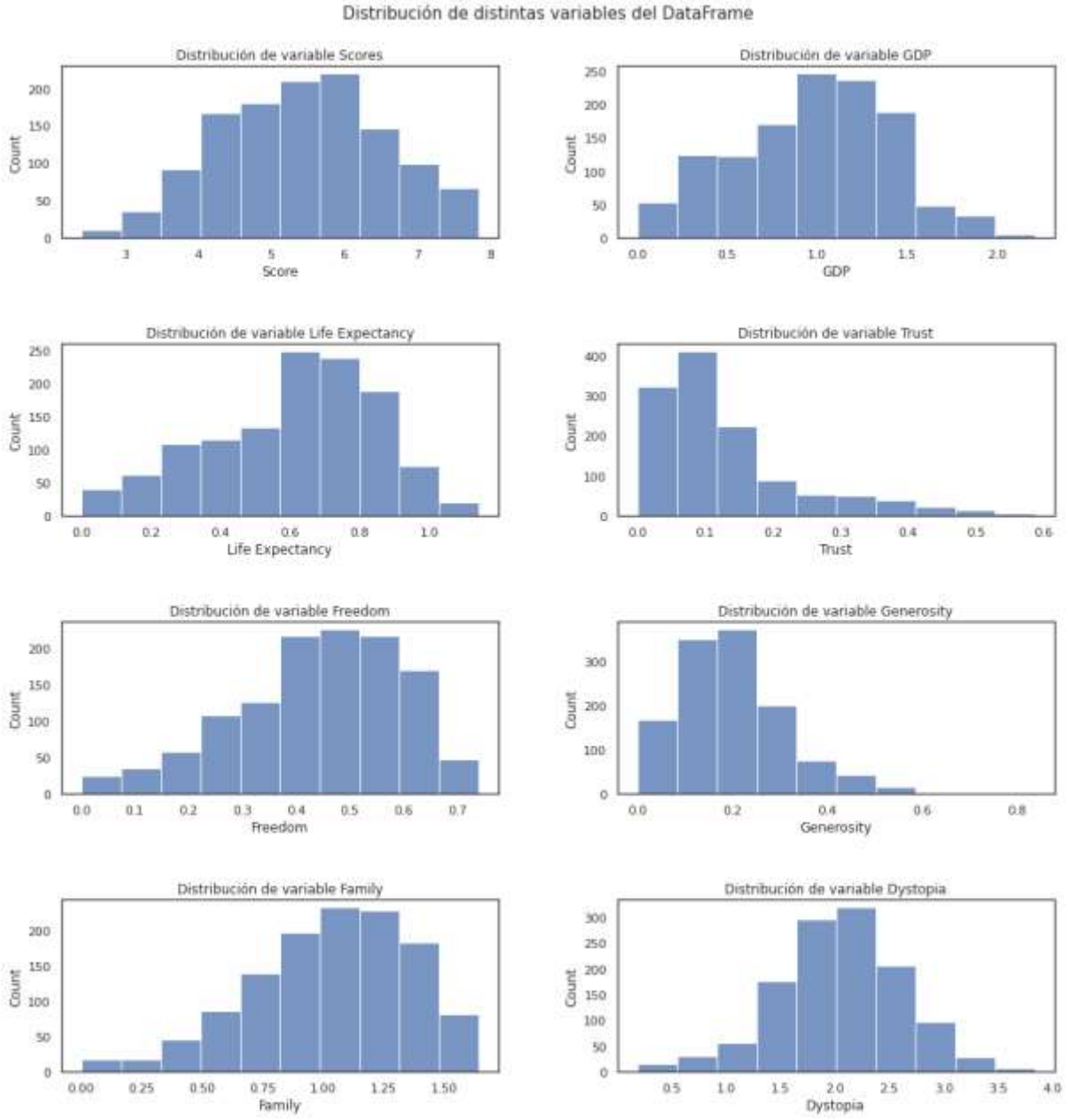
¿Hay algún patrón entre los indicadores que componen el Score de felicidad?

¿Cómo se distribuyen las variables?

En este caso no tendremos en cuenta la variable dystopia, ya que es producto del resultado de las demás variables.

En la distribución se observa que:

- GDP, Family y Life expectancy son las variables que mayor aporte hacen al Score y cabe destacar que las tres son variables cuantitativas, se pueden medir.
- Freedom y Generosity, que son las variables que menos puntaje aportan al Score, son variables cualitativas, es decir, no son medibles, sino más bien se relacionan con el sentimiento de las personas en como desarrollan su vida día a día.



¿Cómo se relacionan las variables entre sí?

Si nos concentramos en la variable Score se puede decir que:

- Las variables GDP y Life Expectancy son las que mayor influencia positiva tienen.
- Generosity que es muy baja, prácticamente nula al igual que population, no influirán en el Score.

Según se señaló en el gráfico anterior las variables que más influyen sobre el Score son GDP, Family y Life expectancy, y aquí se observa que su correlación es de las más altas entre variables.

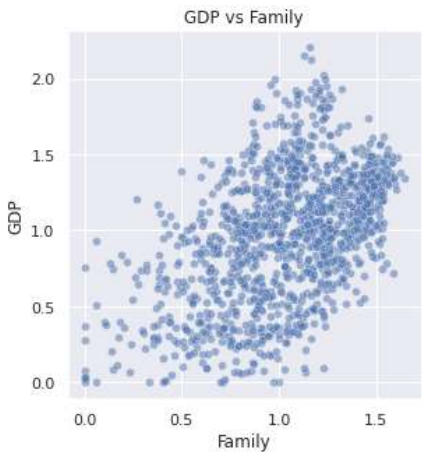
¿Son dependientes entre sí?

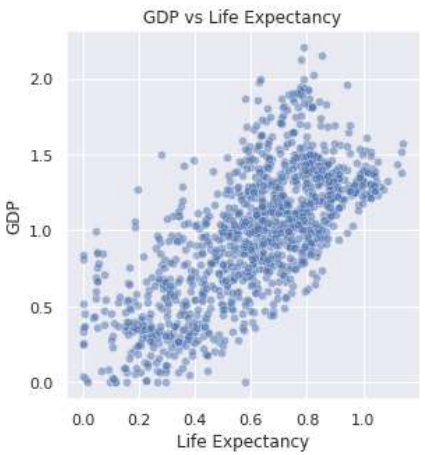
Se puede ver una clara relación con tendencia creciente entre las variables, por lo que se las podría considerar dependientes una de las otras.

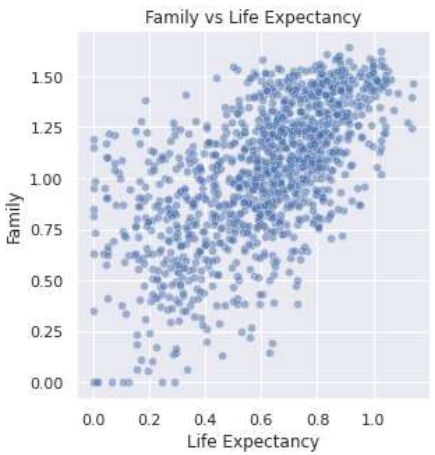
Por ejemplo:

- Cuanto mayor PBI, mayor es la asistencia que se recibe por parte del Estado.
- Cuanto mayor PBI, mayor es la esperanza de vida.
- Cuanto mayor asistencia se recibe del Estado, mayor es la esperanza de vida.

Relación entre las variable con mayor influencia en el Score



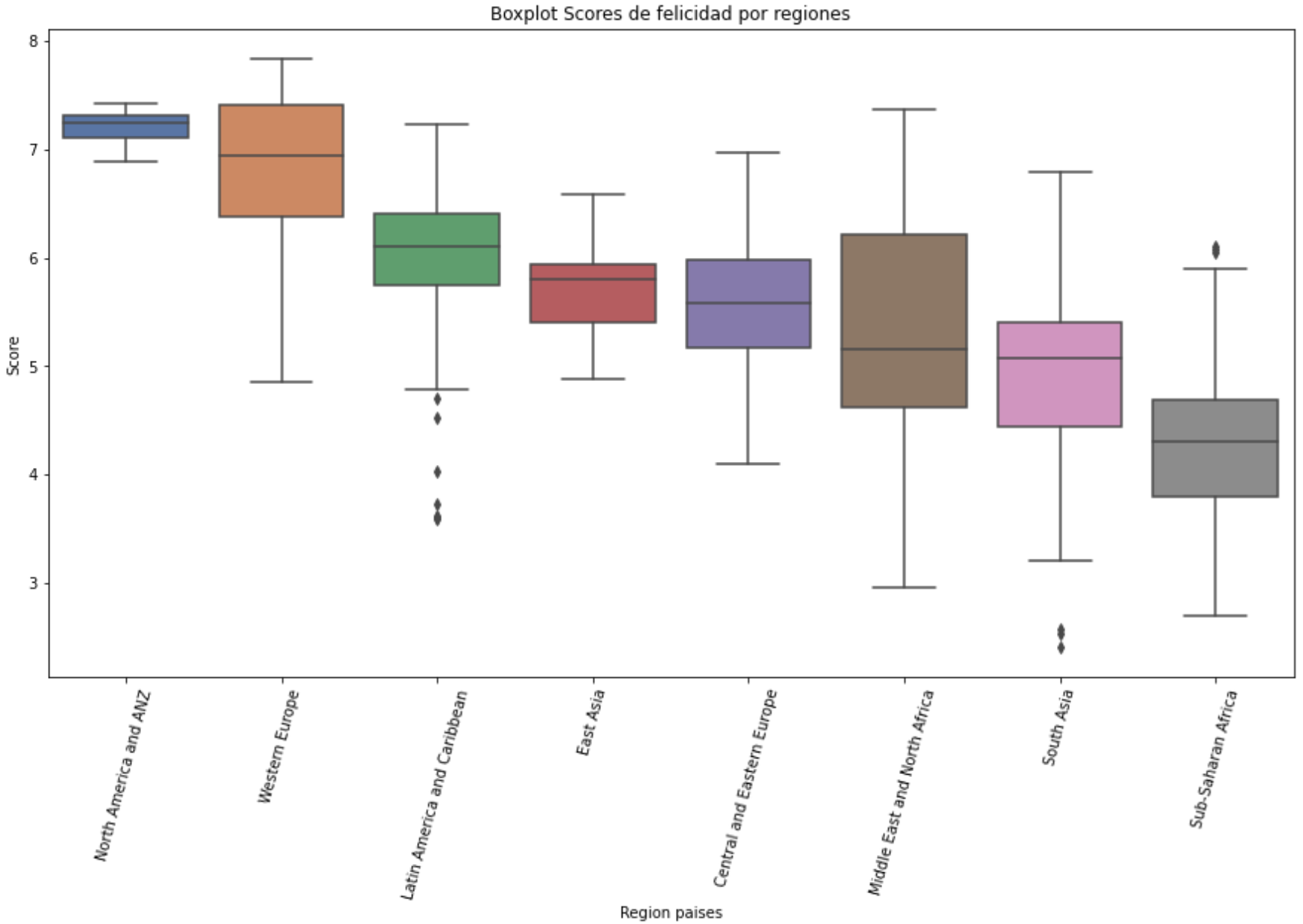




¿Qué regiones son las más felices?

¿Cómo es la distribución del Score de felicidad por regiones?

Con estos boxplots podemos ver que hay algunas regiones con valores muy concentrados y otras con valores muy dispersos, América Latina y el Caribe es la que mayor outliers tiene por debajo de primer cuartil.

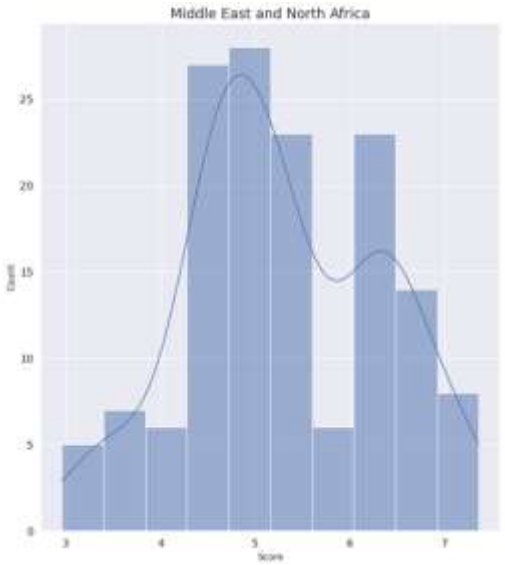


¿Es América Latina y el Caribe la región con mayor desigualdad?

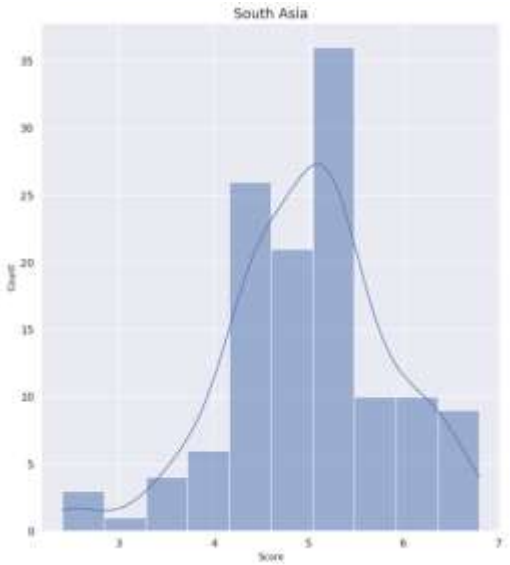
Region	Max	Min	Dif
Middle East and North Africa	7.3640	2.955	4.4090
South Asia	6.7980	2.404	4.3940
Latin America and Caribbean	7.2260	3.582	3.6440
Sub-Saharan Africa	6.1013	2.693	3.4083
Western Europe	7.8420	4.857	2.9850
Central and Eastern Europe	6.9650	4.096	2.8690
East Asia	6.5840	4.874	1.7100
North America and ANZ	7.4270	6.886	0.5410

Se observa que la región con mayor desigualdad es Medio Oriente y Africa del Norte con 4.4 puntos de diferencia entre el máximo y el mínimo, América Latina y el Caribe se encuentra en tercer lugar.

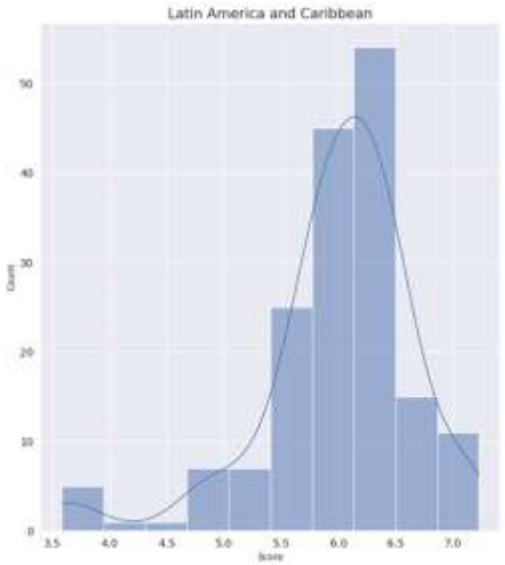
Entonces, ¿Por qué América Latina y el Caribe tiene tantos valores fuera de rango y por qué Medio Oriente y África del norte no tiene ninguno?



En la región de Medio Oriente y África del Norte su distribución está concentrada en el centro y un poco sesgada a la izquierda, pero los valores más chicos tienen una cantidad que les permite ser representativos y no quedar fuera de rango.



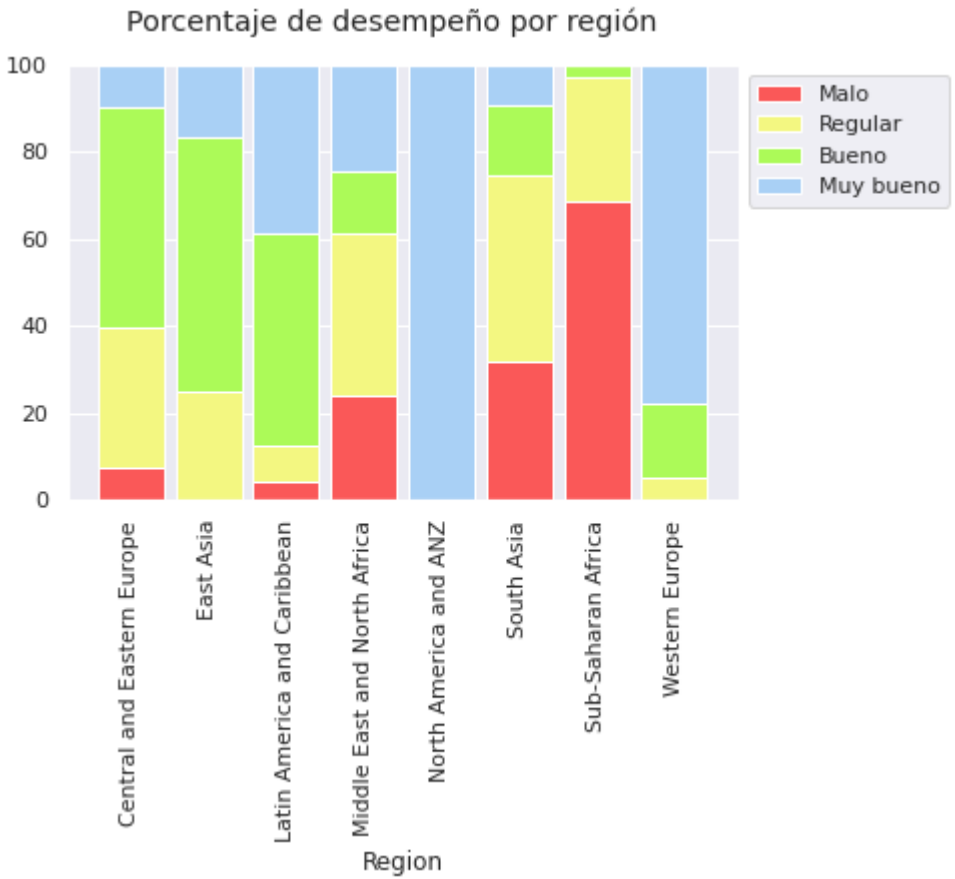
En la región del Sur Asiático su distribución es en cierta forma normal, con un pico en el centro y un pequeño sesgo a la izquierda, pero los valores más pequeños son muy bajos como para estar dentro del rango intercuartílico, de hecho son los mas bajos del dataset, por lo que se identifican como outliers.



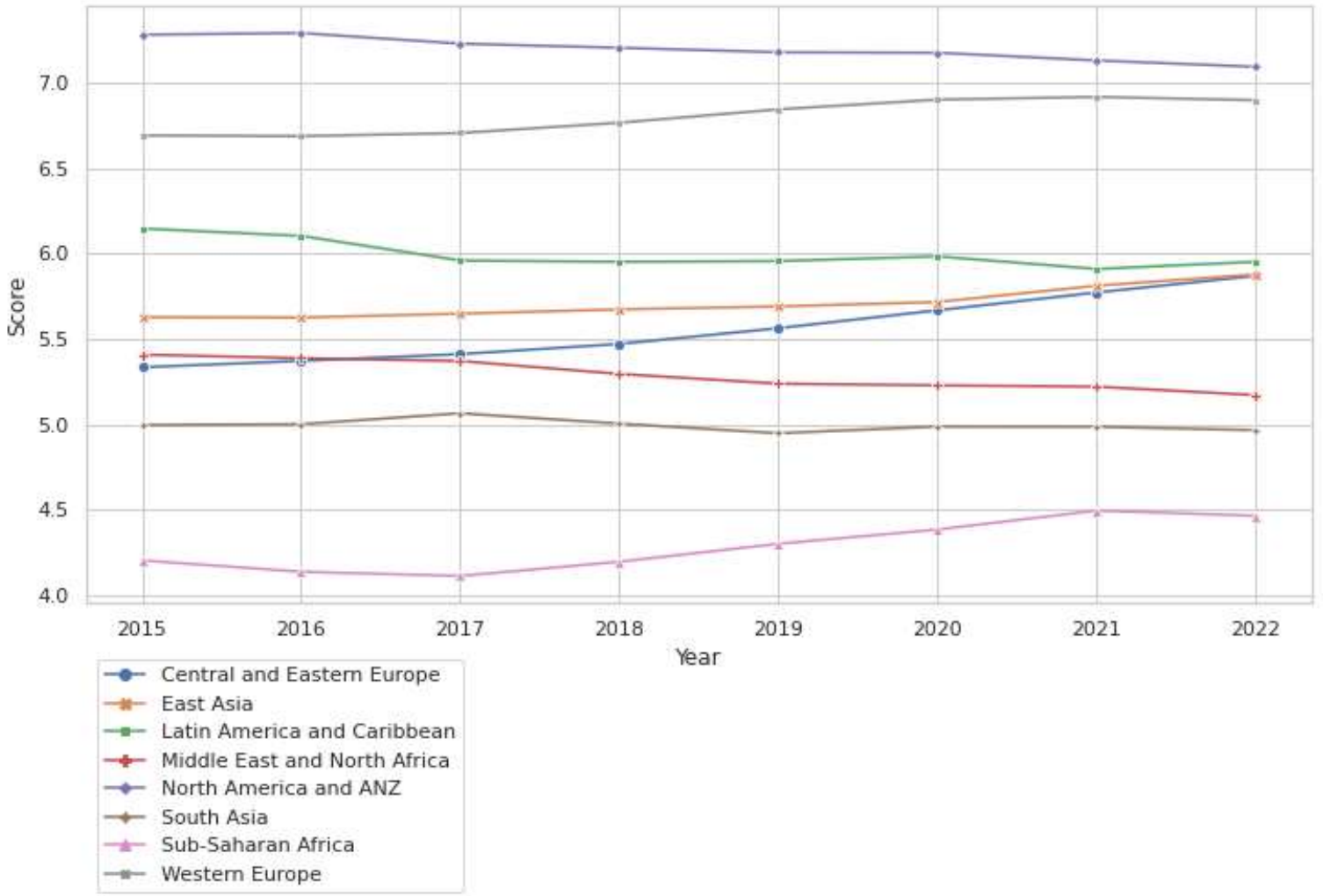
En la región de América Latina y el Caribe identificamos una gráfica totalmente sesgada hacia la izquierda con su pico corrido hacia un lado, es por eso que al tener la mayoría de los valores concentrados entre 5 y 7 los que estén por debajo se tomarán como outliers.

¿Cuál es el desempeño de las Regiones?

Cada país tiene su desempeño particular dependiendo de su propio Score, por lo que podemos ver que porcentaje de países se encuentra en cada categoría y como se conforma la región.



¿Cómo han evolucionado los Scores de las regiones durante los años?



Conclusiones EDA

- **Europa Central y del Este** tiene un desempeño variado en donde encontramos todos los desempeños, el de mayor frecuencia es **Bueno**, seguido por más del 30% de desempeño **Regular**. Su desempeño a lo largo del tiempo ha sido muy bueno realmente, es la región con mayor crecimiento durante todo el período y no ha tenido ninguna variación interanual negativa.
- **Asia del Este** no tiene desempeño **Malo**, la mayoría de los países tienen un desempeño **Bueno**, seguido por **Regular** con un 25% aprox, y por último el de menor frecuencia es **Muy bueno.** Se ha mantenido estable la mayor parte del tiempo, excepto los últimos años que ha crecido su Score.
- **América Latina y el Caribe** tiene muy poco desempeño **Malo** y **Regular**, de hecho entre ambos apenas completan el 10% de los países, el resto se divide entre **Bueno** con un 50% aprox y **Muy bueno** prácticamente un 40%. Ha bajado su desempeño, se ha mantenido estable y ha vuelto a bajar, el último año subió un poco su Score, pero está lejos de tener un buen desempeño a lo largo del tiempo.
- **Medio Oriente y África del Norte** tiene una distribución muy variada con más de 10% por cada tipo de desempeño, el de mayor frecuencia es el **regular** y podría decirse que **Muy bueno** y **Malo** tienen la misma proporción, lo cuál es muy extraño porque no se logra marcar una tendencia en la región. Nunca tuvo un desempeño interanual positivo, se ha logrado mantener algunos años, pero su tendencia es claramente negativa.
- **Norte América y Australia y Nueva Zelanda** sin dudas es la región de mejor desempeño, todos sus países tienen un desempeño **Muy bueno**, pero a lo largo del tiempo ha ido bajando su score considerablemente, si bien ha sido la mejor región durante todos los períodos ha perdido su ventaja de casi 0,5 puntos a 0,2 puntos aprox contra la segunda mejor región.
- **El Sur Asiático** también es una región heterogénea donde encontramos todas las categorías de desempeño donde casi el 30% de sus países tienen un desempeño **Malo**, **Regular** es el de mayor frecuencia con un 40% aproximadamente, y el resto se completa entre **Bueno y Muy bueno.** Prácticamente tiene el mismo Score que al comienzo, ha subido y bajado, pero a lo largo del tiempo su desempeño es neutro.
- **Sub-Sahara África** es la peor región tiene más del 60% de su desempeño como **Malo** y prácticamente el resto de su desempeño es **Regular**, no tienen ningún país con un desempeño **Muy bueno.** Si bien durante todo el período es la región con pero desempeño, ha ido aumentando su Score aunque está a 0,5 puntos de la segunda peor región.
- **Europa Occidental** no tiene ningún país con desempeño **Malo** y el 75% de sus países tiene un desempeño **Muy bueno**, el resto se completa mayormente con desempeño **Bueno** y aprox un 5% de desempeño **Regular.** Es la segunda mejor región, ha aumentado su desempeño de manera constante a lo largo del tiempo y disminuyendo su brecha con la mejor región. Tiene una tendencia positiva que se ha estancando los últimos años.

BUSCANDO EL **MEJOR MODELO**

Explicación de la búsqueda del mejor modelo.

En este proyecto se diferenciarán tres tipos de modelos distintos en los cuales todos evaluarán los algoritmos LinearRegression, Lasso, ElasticNet, DecisionTree, KNeighbors, SVR, RandomForest y se comparará por la métrica de r2 Score.

Modelo default

En este modelo se entrenarán todos los algoritmos sin hacerle ninguna modificación al conjunto de datos. Tendrá todas las variables y los parámetros e hiperparametros por default.

Modelo SFS (Sequential Feature Selection)

En este modelo se entrenarán todos los algoritmos pasando distintas cantidades de variables y como resultado se obtendrá el r2 Score de cada uno con las respectivas variables. Se mantendrán todos los parámetros e hiperparametros por default.

Modelo Hiperparamertos

En este modelo se buscará el mejor corss validation para cada algoritmo, luego se entrenarán pasando distintas cantidades de variables y se buscará nuevamente el mejor cross validation del conjunto con las variables seleccionadas para así hacer un último entrenamiento con los hiperparametros definidos anteriormente en un GridSearch. Como resultado se obtendrá el r2 Score y cross validation score de cada algoritmo con las respectivas variables y mejores hiperparametros.

Modelo default

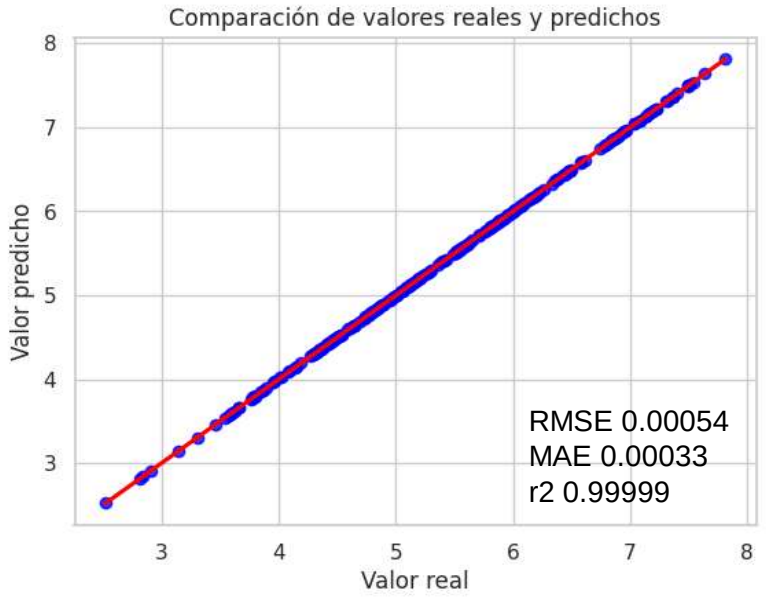
Resultado del entrenamiento

Model	Metric	BestScore	Tiempo
DecisionTreeRegressor	r2	1.000.000	0.035598
LinearRegression	r2	1.000.000	0.020441
RandomForestRegressor	r2	0.995914	1.349.367
KNeighborsRegressor	r2	0.432334	0.045662
SVR	r2	0.025637	0.265151
ElasticNet	r2	0.002234	0.009613
Lasso	r2	0.002234	0.009846

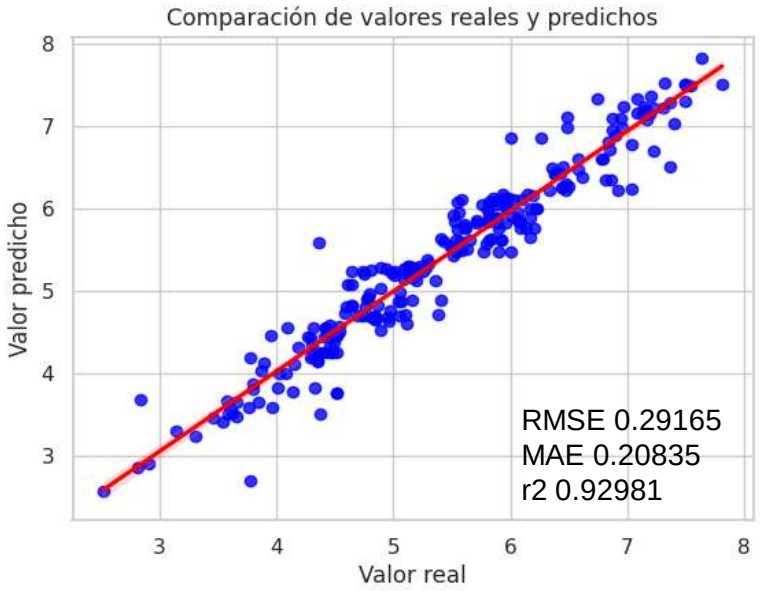
Podemos ver que la regresion lineal sigue manteniendo su 99.99% de acierto en el r2, es decir que no tiende al overfitting.

En este caso se observa que el decision tree baja de un 100% a 92% de acierto en el r2, es decir que hace overfitting, por lo que no seria recomendable su utilización

LinearRegression



DecisionTreeRegressor

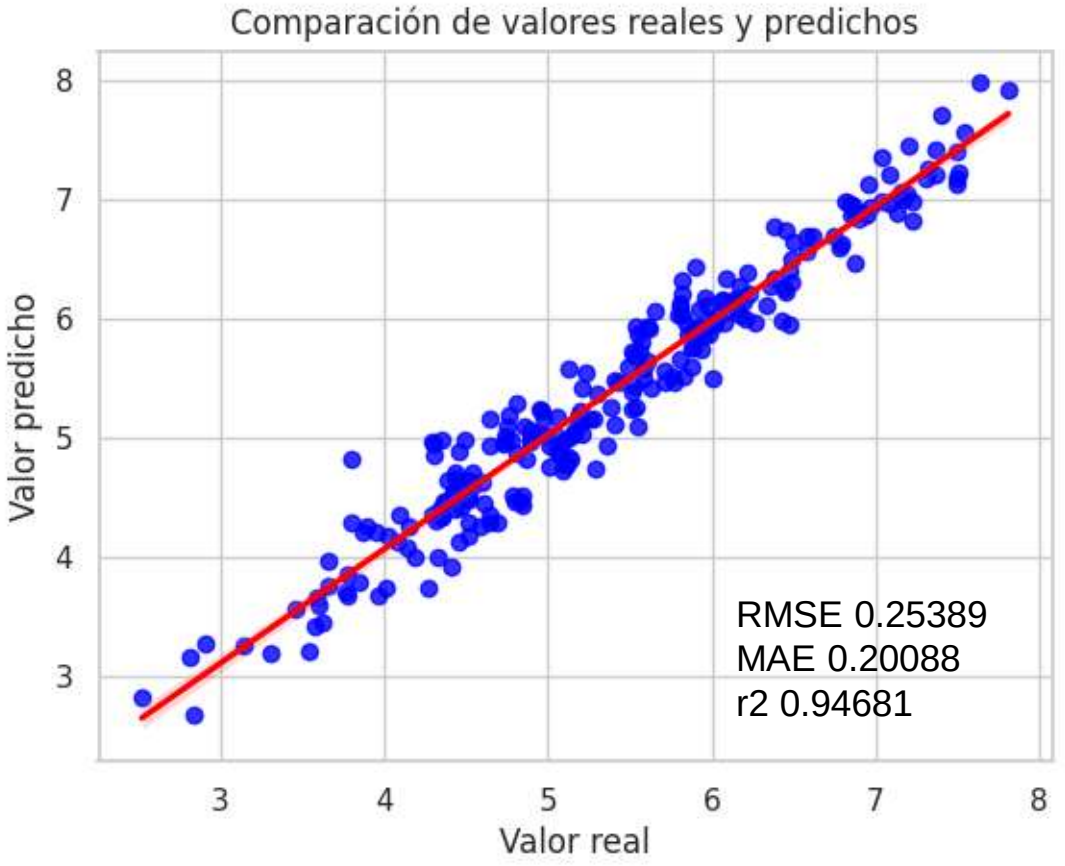


Modelo SFS

Resultado del entrenamiento

Model	Features	BestScore	Tiempo
LinearRegression	(GDP, Family, Trust, Dystopia)	0.956510	0.394110
SVR	(GDP, Family, Dystopia, Desempeño)	0.953885	11.168.062
RandomForestRegressor	(Country, Region, Desempeño, population)	0.953111	44.417.521
RandomForestRegressor	(Country, Region, Desempeño)	0.952177	32.551.484
RandomForestRegressor	(Country, Desempeño)	0.951202	22.035.964
DecisionTreeRegressor	(Country, Region, Desempeño)	0.948091	0.603368
DecisionTreeRegressor	(Country, Desempeño)	0.948046	0.293300
DecisionTreeRegressor	(Country, Region, Desempeño, population)	0.938761	1.039.432
KNeighborsRegressor	(Region, Trust, Generosity, Desempeño)	0.926267	1.004.753
LinearRegression	(GDP, Family, Dystopia)	0.920082	0.275455
KNeighborsRegressor	(Region, Trust, Desempeño)	0.919892	0.750151
SVR	(Family, Dystopia, Desempeño)	0.914723	14.682.564
SVR	(Family, Desempeño)	0.902126	10.098.293
KNeighborsRegressor	(Region, Desempeño)	0.898969	0.758557
LinearRegression	(GDP, Dystopia)	0.797657	0.217848
Lasso	(Region, GDP, Family, population)	-0.002471	0.372921
Lasso	(Region, GDP, population)	-0.002471	0.413690
Lasso	(Region, population)	-0.002471	0.182200
ElasticNet	(Region, population)	-0.002471	0.187239
ElasticNet	(Region, GDP, population)	-0.002471	0.280617
ElasticNet	(Region, GDP, Family, population)	-0.002471	0.344596

LinearRegression



En este caso de una regresión lineal con 4 variables en vez de todas se observa que el score del r2 score disminuye un poco, pero sigue siendo un buen resultado.

Si se compara el train con el test también disminuye un poco el score, pero es un valor cercano al 1% lo cuál no considero relevante para su observación

Modelo Hiperparametros

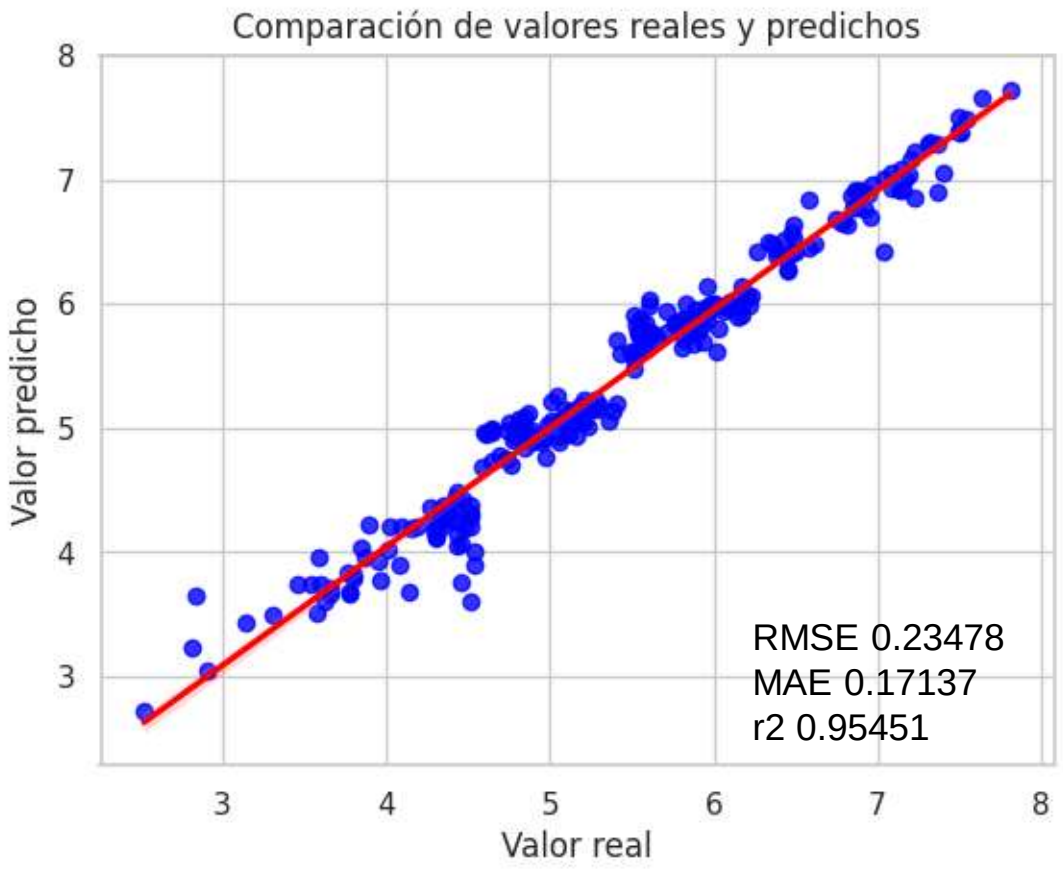
Model	Features	NumFeatures	Forward	BestCV	BestCVScore	BestRState	BestR2Score	BestParams	r2_score - cv_score	Tiempo
RandomForestRegressor	[Country, Region, Year, Desempeño]	4	True	13	0.959658	97	0.972085	{'bootstrap': True	0.012427	2.903.648.381
RandomForestRegressor	[Country, Region, Desempeño]	3	True	13	0.955622	97	0.971744	{'bootstrap': True	0.016122	2.451.726.654
DecisionTreeRegressor	[Country, Region, Desempeño]	3	True	13	0.952928	97	0.970072	{'ccp_alpha': 0.0,	0.017145	31.953.608
RandomForestRegressor	[Country, Desempeño]	2	True	12	0.954232	22	0.968422	{'bootstrap': True	0.014190	2.308.085.491
DecisionTreeRegressor	[Country, Desempeño]	2	True	12	0.951402	97	0.967275	{'ccp_alpha': 0.0,	0.015874	30.078.283
LinearRegression	[GDP, Family, Trust, Dystopia]	4	True	2	0.956717	17	0.966382	{'copy_X': True, 'f	0.009665	2.482.586
SVR	[GDP, Family, Dystopia, Desempeño]	4	True	4	0.960310	49	0.965996	{'C': 10, 'cache_s	0.005685	72.443.533
DecisionTreeRegressor	[Country, Region, Year, Desempeño]	4	True	18	0.944889	11	0.962619	{'ccp_alpha': 0.0,	0.017730	42.104.774
KNeighborsRegressor	[Region, Freedom, Trust, Desempeño]	4	True	12	0.933253	50	0.952055	{'algorithm': 'auto'	0.018801	15.182.405
KNeighborsRegressor	[Region, Trust, Desempeño]	3	True	13	0.922817	50	0.943145	{'algorithm': 'auto'	0.020328	13.905.402
LinearRegression	[GDP, Family, Dystopia]	3	True	4	0.920082	33	0.935292	{'copy_X': True, 'f	0.015210	2.392.739
SVR	[Family, Dystopia, Desempeño]	3	True	8	0.920788	13	0.932895	{'C': 10, 'cache_s	0.012108	57.302.321
KNeighborsRegressor	[Trust, Desempeño]	2	True	5	0.909926	17	0.921233	{'algorithm': 'auto'	0.011307	13.652.531
SVR	[Family, Desempeño]	2	True	2	0.904385	21	0.918861	{'C': 10, 'cache_s	0.014476	50.182.881
LinearRegression	[GDP, Dystopia]	2	True	2	0.797893	13	0.840226	{'copy_X': True, 'f	0.042333	4.000.203
ElasticNet	[Region, GDP, Family, population]	4	True	2	0.573089	87	0.007418	{'alpha': 0.1, 'copy	-0.565671	65.981.127
ElasticNet	[Region, GDP, population]	3	True	2	0.494501	87	0.007418	{'alpha': 0.1, 'copy	-0.487083	47.032.995
ElasticNet	[Region, population]	2	True	4	0.007424	87	0.007418	{'alpha': 0.1, 'copy	-0.000007	48.100.248
Lasso	[Region, GDP, population]	3	True	2	0.497122	87	0.007418	{'alpha': 0.1, 'copy	-0.489704	18.706.262
Lasso	[Region, GDP, Family, population]	4	True	4	0.550755	87	0.007418	{'alpha': 0.1, 'copy	-0.543337	20.686.575
Lasso	[Region, population]	2	True	4	0.006081	87	0.007418	{'alpha': 0.1, 'copy	0.001337	20.442.358

En este punto debemos decidir que algoritmo tomar teniendo en cuenta su r2Score, cvScore y el tiempo

- El mejor r2 es con un RandomForest y 4 variables, pero el tiempo que implica su ejecución es demasiado alto.
- Se podría utilizar un DecisionTree con 3 variables ya que su r2Score está muy cercano al del RandomForest y su tiempo de ejecución es mucho menor, aunque la diferencia entre su r2Score y su cvScore es de las más altas, por lo que puede ser un indicio de overfitting.
- También podemos tomar como opción una Regresión Lineal con 4 viarables, donde su r2Score es un poco menor, pero su tiempo de ejecución es un poco menor al del DecisionTree, y su diferencia entre r2Score y cvScore es mucho menor

Modelo Hiperparametros

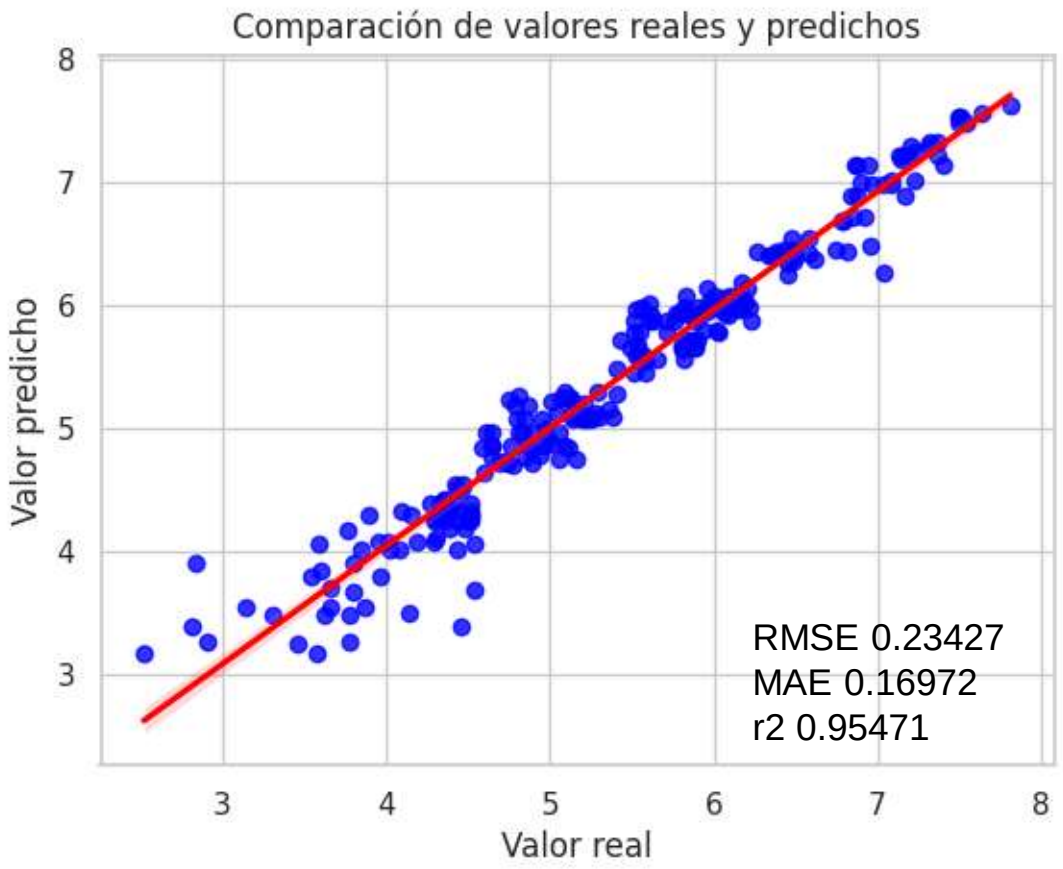
RandomForestRegressor



En este caso se entrenó un RandomForestRegressor con 4 variables ['Country', 'Region','Year', 'Desempeño']
y los hiperparametros modificados
max_depth=29, n_estimators=700, random_state=97
Respecto al train el test disminuyó su r2 score casi un 2% lo cuál no es una gran disminución, pero observando el gráfico se puede deducir que tiene ciertos problemas para predecir valores bajos y que además realiza ciertos agrupamientos que llaman la atención y podrían tender hacia el underfitting.
También una gran desventaja es su tiempo de ejecución muy alto frente a otros algoritmos.

Modelo Hiperparametros

DecisionTreeRegressor



En este caso se entrenó un DesicionTreeRegressor con 3 variables ['Country','Region','Desempeño']

y los hiperparametros modificados

max_depth=23, random_state=97

Es un resultado bastante similar al RandomForestRegressor.

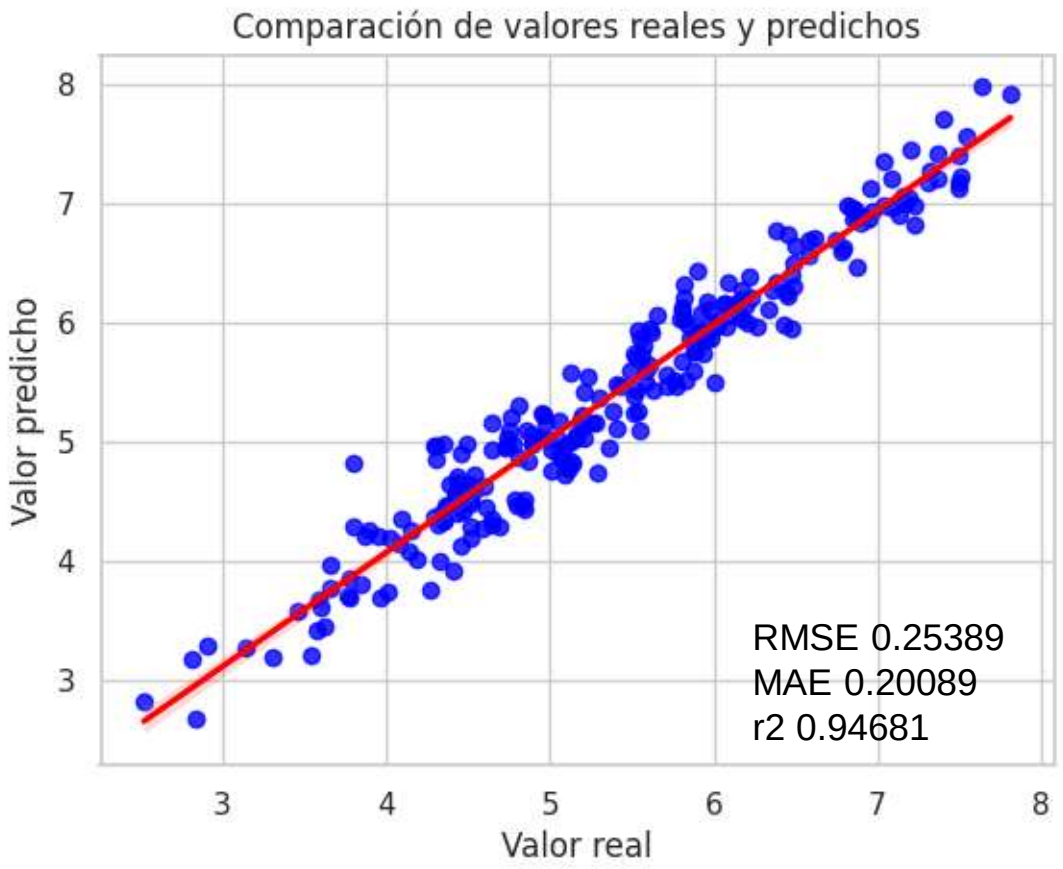
Respecto al train el test disminuyó su r2 score casi un 1.5% lo cuál no es una gran disminución, pero observando el gráfico se puede deducir que tiene ciertos problemas para predecir valores bajos y que además realiza ciertos agrupamientos que llaman la atención y podrían tender hacia el underfitting.

Los grupos no son tan marcados como el RandomForest, tiene una mayor dispersión lo que hace que sus métricas sean minimamente mejores.

Su tiempo de ejecución relativamente bajo frente a otros algoritmos. Se puede tomar como una ventaja

Modelo Hiperparametros

LinearRegression



En este caso se entrenó un LinearRegression con 4 variables ['GDP','Family','Trust','Dystopia']

y al ser regresión lineal no tiene hiperparametros para ser modificados.

Respecto al train el test disminuyó su r2 score un 2% aproximadamente lo cuál no es una gran disminución, pero en el gráfico se observa a diferencia de los otros algoritmos, que no tiene problemas para predecir los valores más bajos y no realiza agrupamientos. Incluso se puede ver que el rango entre los valores por encima y por debajo de la recta son similares y prácticamente constantes excepto por algún valor fuera de rango.

Al igual que el DecisionTree su tiempo de ejecución relativamente bajo frente a otros algoritmos se puede tomar como una ventaja y además sumarle a esto su simpleza y facilidad de ejecución

Conclusiones modelos

Entrenamiento Default:

PROS:

- Es de fácil implementación.
- Al tener todas las variables su r2 Score es muy alto.
- Funciona muy bien con conjunto de datos chicos.

CONTRAS:

- Puede ser sensible a overfitting.
- Puede llegar a ser muy básico para algunos conjuntos de datos.
- En conjuntos grandes puede demorar mucho tiempo.

Entrenamiento SFS:

PROS:

- Disminuye el tamaño del conjunto.
- Es más versátil frente a conjuntos grandes.
- Se pueden setear distintas opciones para buscar la mejor combinación.

CONTRAS:

- Disminuye el r2 Score del conjunto.
- Su tiempo de ejecución es más alto.

Entrenamiento Hiperparametros:

PROS:

- Busca la mejor opción para cada algoritmo.
- Optimiza el algoritmo y mejora su r2 Score.

CONTRAS:

- Su aplicación es más complicada.
- El tiempo de ejecución es muy alto.

Conclusiones finales

- Luego de haber ejecutado distintos modelos y conocer el conjunto de datos, siendo este un conjunto chico y con pocas variables considero que la mejor opción para este caso es un entrenamiento default de una regresión lineal con todas la variables.

- En el caso de que el conjunto incremente su cantidad de variables recomendaría realizar una selección de variables y búsqueda de hiperparametros para su optimización.

- Y por último, en el caso de que el conjunto incremente sus variables y también la cantidad de datos haría una selección de variables, pero sería cauteloso en la búsqueda de hiperparametros por una cuestión de tiempos de ejecución.

Recomendaciones

Considero que tal vez el reporte y el Score de felicidad esta muy influenciado por indicadores que no representan los sentimientos de las personas que viven allí. Poco se tiene en cuenta el contexto del país y de los habitantes.

Según World Happiness Report, la muestra varía entre 2000 y 3000 personas por país, y sostiene que tiene un intervalo de confianza del 95%, pero, ¿En qué situación coyuntural se encuentran las personas a las que se le realiza la encuesta?

No hay un contexto sobre las relaciones sociales, el clima, las costumbres, los accesos a vivienda digna, trabajo, educación, salud, alimentación.

El reporte menciona que “no aparecen porque aún no se dispone de datos internacionales comparables para la muestra completa de países”, pero se podrían ampliar las variables tomando datos desde otros reportes, consultoras u organizaciones como el Banco Mundial.