Ranking Mundial de Universidades

Proyecto final de Data Science para CoderHouse

Integrantes: *Victoria Gobbi, Candela Esquivel, Sebastian De Cunto*

Introducción

El dataset corresponde al ranking mundial de las universidades Times Higher Education de 2012 a 2015, el mismo cuenta con 2200 registros, con un total de 1024 universidades. Las universidades que se encuentran dentro del ranking pertenecen a los continentes de América, África, Asia, Europa y Oceanía, y cada una de ellas cuenta con la información acerca de su posicionamiento en el ranking nacional y mundial, y otros tipos datos que influyen en ambos rankings.

Cabe mencionar, que el ranking mundial de universidades de Times Higher Education es un conocido ranking mundial de clasificación de universidades,que es considerado como una de las medidas universitarias más influyentes y ampliamente observadas a la hora de elegir una universidad.

Objetivo

El objetivo del proyecto es predecir cuáles serán las universidades que se encuentren dentro del Top 500 del ranking mundial de universidades de Times Higher Education.

Exploración del Dataset

Las variables disponibles son de diferente tipo : en el caso de la variable institution y country son variables nominales y el resto son variables cuantitativas.El dataset cuenta con dos columnas del tipo *float*, dos del tipo *objects* , una columna de tipo datetime y el resto con valores enteros.

A continuación se describen las variables con las que cuenta el dataset:

- **World_rank:** Ranking mundial de la universidad. Se trata de una variable cuantitativa numérica.

- **Institution:** Nombre de la universidad. Se trata de una variable cualitativa categórica.
- **Country :** País al cual pertenece la universidad. Se trata de una variable cualitativa categórica.
- **National_rank :** Ranking nacional de la universidad. Se trata de una variable cuantitativa numérica.
- **Quality_of_education:** Ranking sobre la calidad de la educación, mide el número de ex alumnos que han ganado distinciones académicas importantes en relación con el tamaño de la universidad, es decir mide el peso de cada universidad respecto a las distinciones ganadas. Se trata de una variable cuantitativa numérica.
- **Alumni_employment :** Ranking sobre la cantidad de ex alumnos que cuenta con los mejores empleos, mide la cantidad de ex alumnos con empleos en altos cargos ejecutivos en las empresas más grandes del mundo en relación con el tamaño de la universidad. Se trata de una variable cuantitativa numérica.
- **Quality_of_faculty** : Ranking sobre la calidad de la universidad, mide el número de alumnos de la facultad que han ganado distinciones académicas importantes.
- **Publications:** Ranking que muestra las universidades que cuentan con mayores cantidades de publicaciones. Se trata de una variable cuantitativa numérica.
- **Influence:** Ranking sobre la Influencia de la universidad, es medido por el número de artículos de investigación que aparecen en las revistas con mayor influencia. Se trata de una variable cuantitativa numérica.
- **Citations:** Ranking sobre la cantidad de investigaciones de cada universidad. Se trata de una variable cuantitativa numérica.
- **Broad_impact** : Ranking sobre el impacto de la universidad en el mundo. Solo está disponible para los años 2014 y 2015.Se trata de una variable cuantitativa numérica.
- **Patents:** Ranking que muestra las universidades que más patentaron. Se trata de una variable cuantitativa numérica.

- **Score:** Puntaje sobre las universidades, es usado para determinar el ranking donde cada una de las variables nombradas anteriormente influyen sobre el score. Se trata de una variable cuantitativa numérica.
- **Year:** Año del ranking.

Es importante destacar, que todas las variables son medidas año a año.

El dataset fue extraído de Kaggle.

El link es el siguiente:
https://www.kaggle.com/datasets/mdelrosa/cwur-university-rankings-201920

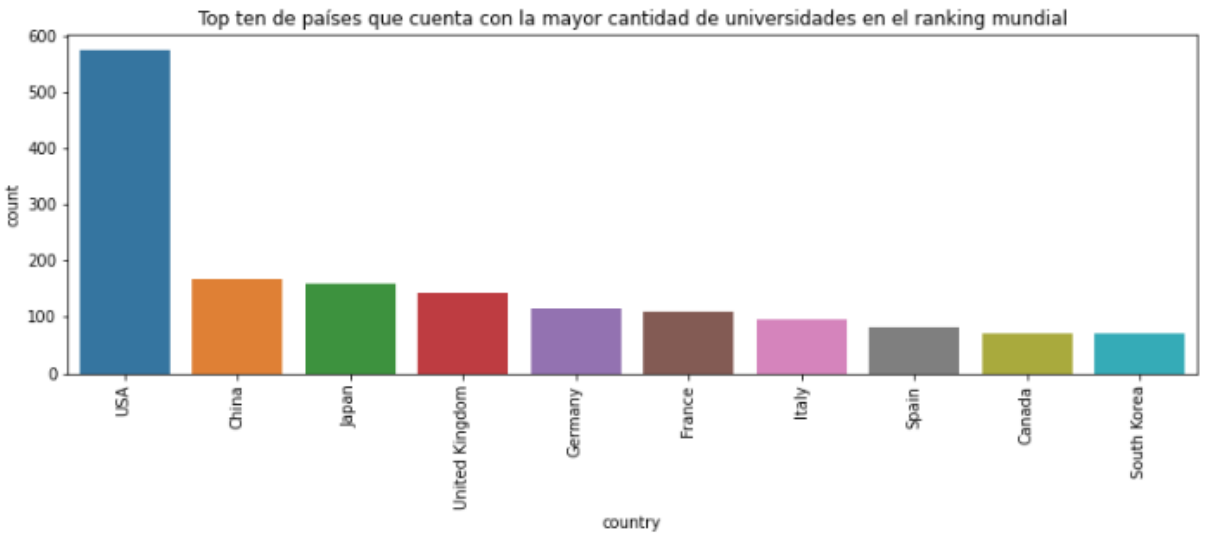
Preparación de los datos

El dataset posee datos nulos en la columna broad_Impact para los años 2012 y 2013 ya que en ese momento no existía información acerca de esa variable, por lo tanto los mismos fueron reemplazados con el valor 0. Del total de datos de la columna que son 2200, 200 son nulos. Es decir existía un total de 9.09 de datos nulos en la columna

Otra modificación que se realizó en el dataset fue cambiar el tipo de dato de la columna year, ya que la misma aparecía como entero cuando le corresponde el tipo de dato datetime correspondiente a fechas..

Top ten de países que cuentan con la mayor cantidad de universidades en el ranking

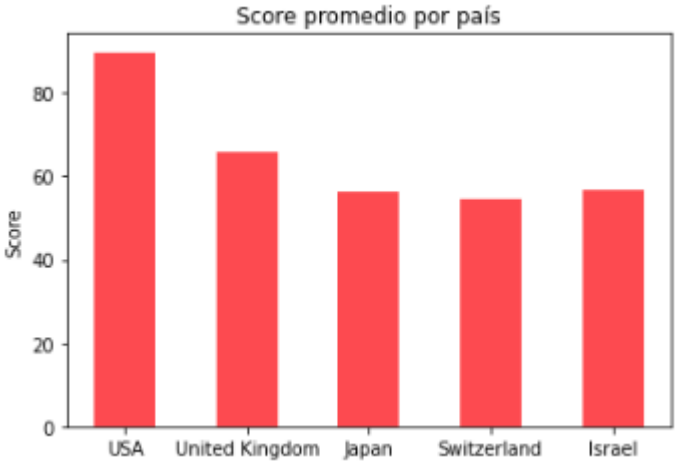
En este gráfico podemos observar los 10 países que cuentan con la mayor cantidad de universidades ubicadas dentro del ranking, donde podemos observar que EEUU es el país que cuenta con mayor cantidad de universidades. Tanto en EE. UU. como China tienen muchas empresas de alta tecnología, lo que probablemente sea el resultado del talento formado en esas universidades.



Score promedio por país (TOP 5)

En el siguiente gráfico podemos observar un histograma respecto a la variable score.

Tal como se puede observar USA se encuentra adelante del resto de los países con score promedio arriba de los 80 puntos



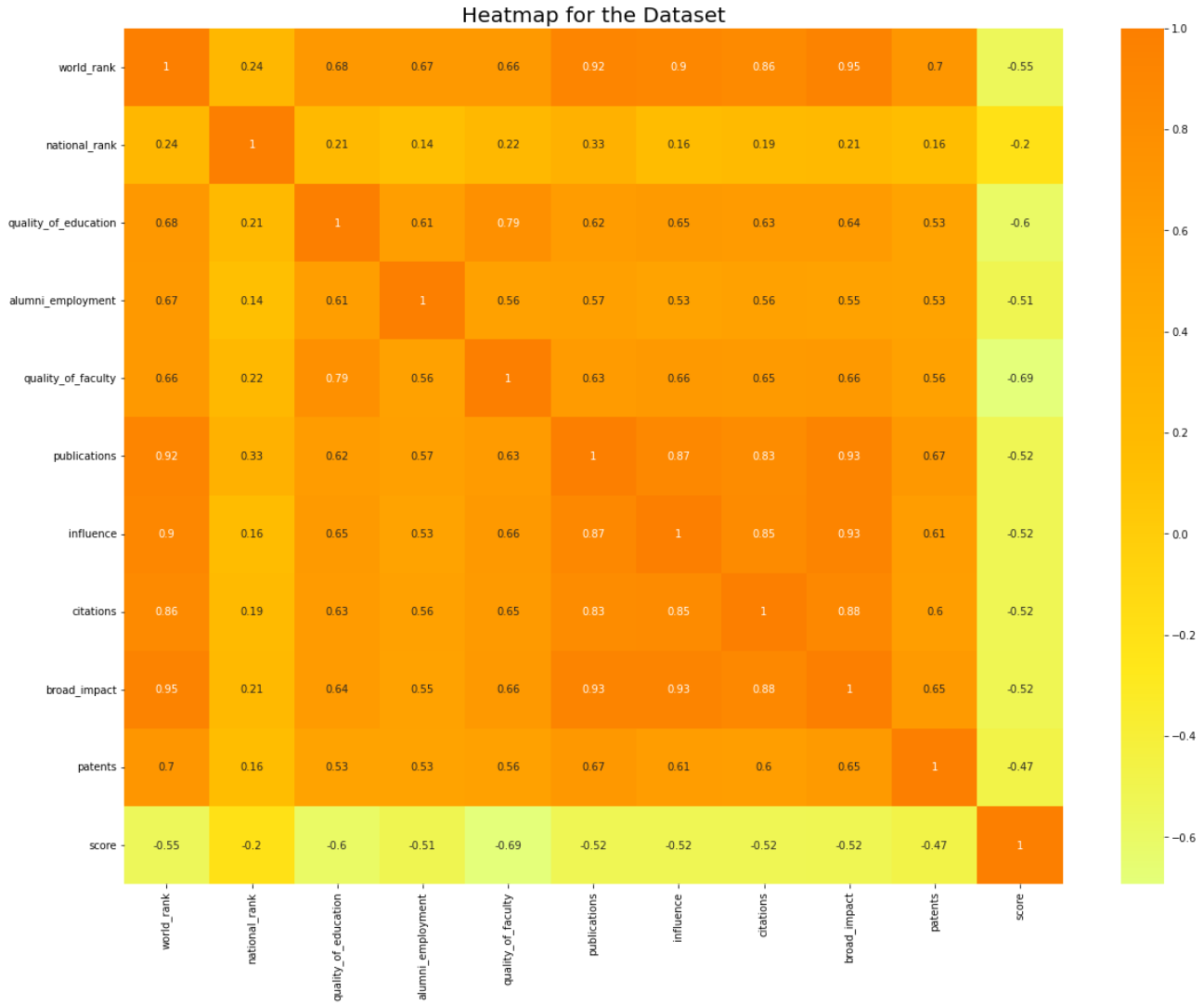
Correlación entre las variables

En el heatmap de correlaciones vemos la existencia de una alta correlación positiva entre la variable word_rank y las variables publications, influence, citations y broad impact.

Esto quiere decir que las universidades que cuentan con mayor cantidad de publicaciones, de artículos de investigación que aparecen en las revistas con mayor influencia a nivel mundial y que cuentan con un gran impacto a nivel mundial son las universidades mejor rankeadas.

Por otra parte también podemos que las variables quality_of_education, alumni_employment y quality_of_faculty, estas variables se encargan de medir la calidad de la educación y de la universidad por lo cual, tienen una relación positiva con la variables world rank, no tan alta como en de las variables comentadas anteriormente, pero si podemos ver que ante mejor ranking en algunas de estas variables mejor será el ranking mundial de la universidad.

Por último, también podemos ver que hay una baja correlación negativa entre las variables world rank y score, esto se debe que ante mayor score, menor ranking, es decir mejor rankeadas están las universidades.

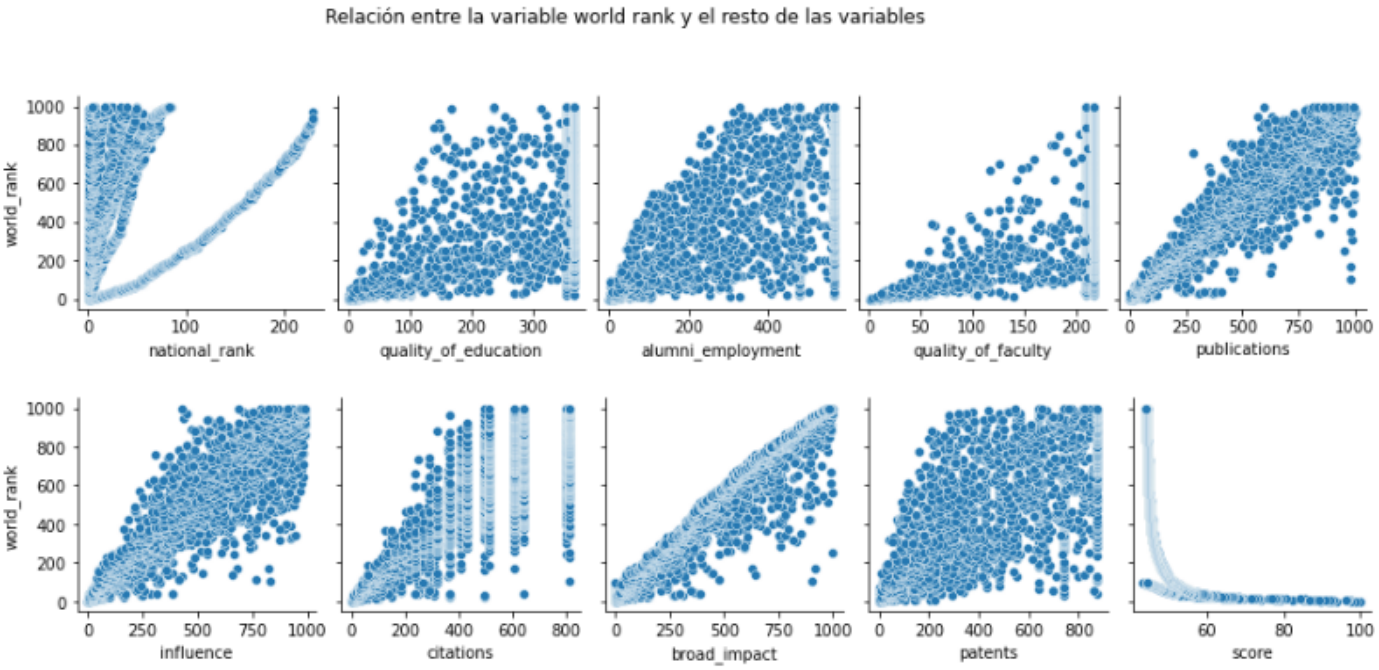


Relación entre la variable world rank y el resto de las variables

Tal como venimos comentando anteriormente podemos ver que las variables citations, publications, influence, broad_impact están muy relacionadas con la variable world rank,donde las universidades que cuentan mejor ranking, están mejor rankeadas en estas variables. En algunos casos podemos ver algunas universidades que no cumplen con esta relación, sobre todo en las variables citations, pero en general la regla se cumple.

Por otra parte, podemos observar que cuanto mejor ranking nacional tienen las universidades, mejor es el ranking mundial, pero tal como dijimos anteriormente hay algunas universidades que no cumplen con esto.

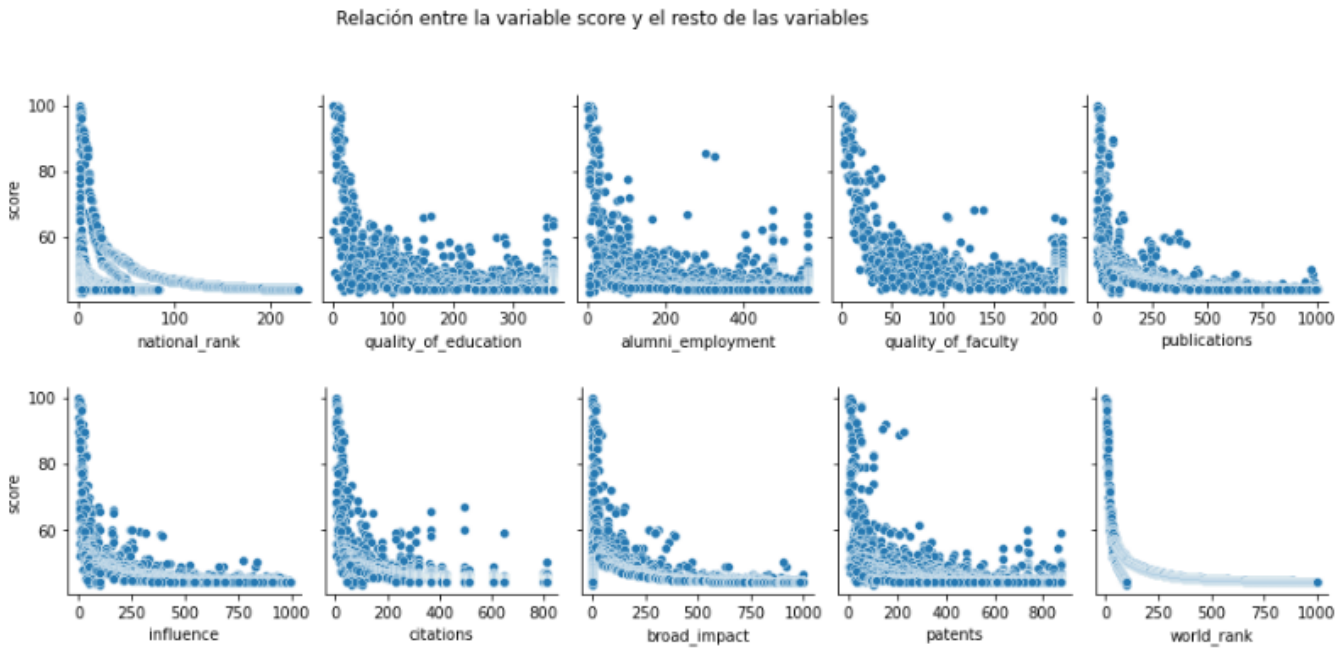
También podemos observar que las variables quality_of_education, alumni_employment, queality_of facultative, patents, influyen de forma positiva en la variable world rank pero en menor medida, ya que podemos ver algunas universidades que tienden a tener mejor ranking mundial pero tiene mal ranking en estas variables.



Relación entre la variable score y el resto de las variables

Se puede observar que ante mayor score mejor va a ser el ranking de la universidad tal como lo vimos anteriormente. Vemos que la variable score se relaciona de forma negativa con la variable world rank.

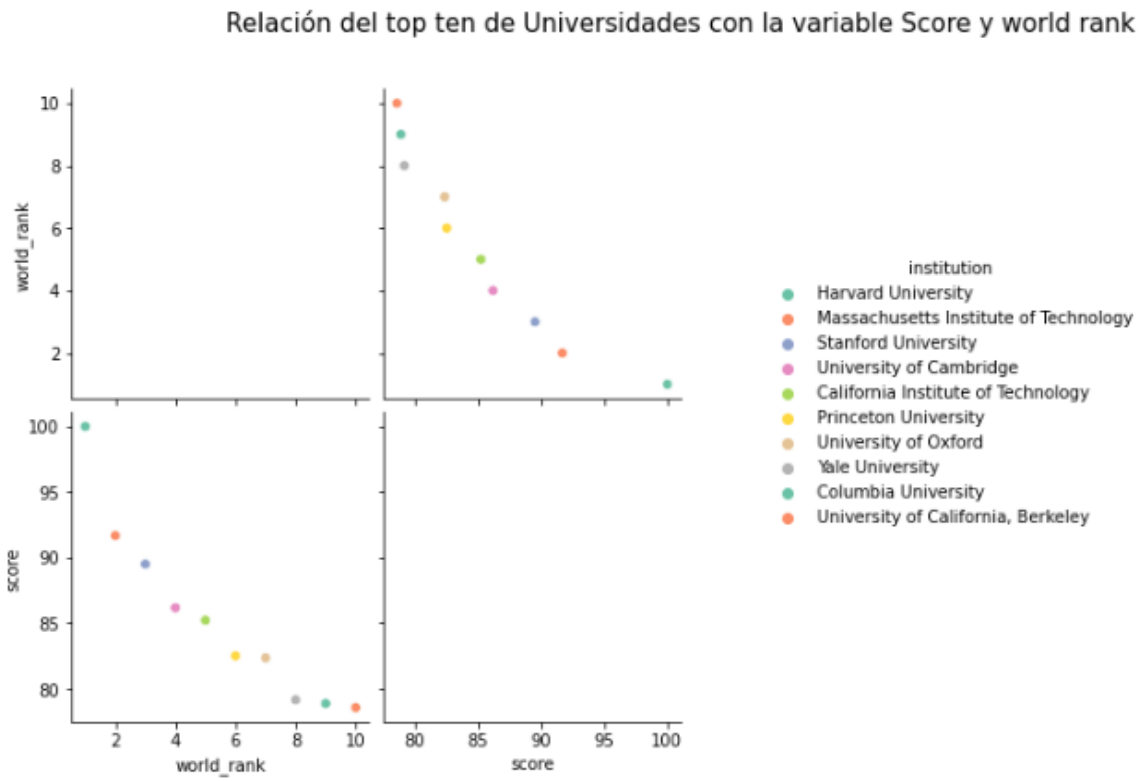
Por otro lado, sucede algo similar con demás variables donde a mayor score mejor ranking tienen las universidades en esas variables, como publications,influence, citations, broad_impact,quality_of_education, alumni_employment, queality_of facultative, patents, pero podemos observar que algunas universidades cuentan con bajos scores y altos rakings en estas variables, por lo que podemos ver que la variable score tiene una baja correlación negativa con las variables que mencionamos.



Relación del top ten de Universidades con la variable Score y world rank

En los siguientes gráficos haremos un análisis de las mejores 10 universidades, donde podremos observar que el score no baja de 75 puntos, y esto hace que el ranking de las universidades en estos puestos sea tan alto

Por lo tanto a mayor score, el ranking tiende a ser mejor , esto tiene sentido al repetir lo anteriormente mencionado, el score se calcula teniendo en cuenta los rankings.



Relación de las últimas 10 Universidades con la variable Score y world rank

Todo lo contrario sucede en este gráfico, las 10 universidades con menor ranking mundial poseen un score que no supera los 45 puntos.

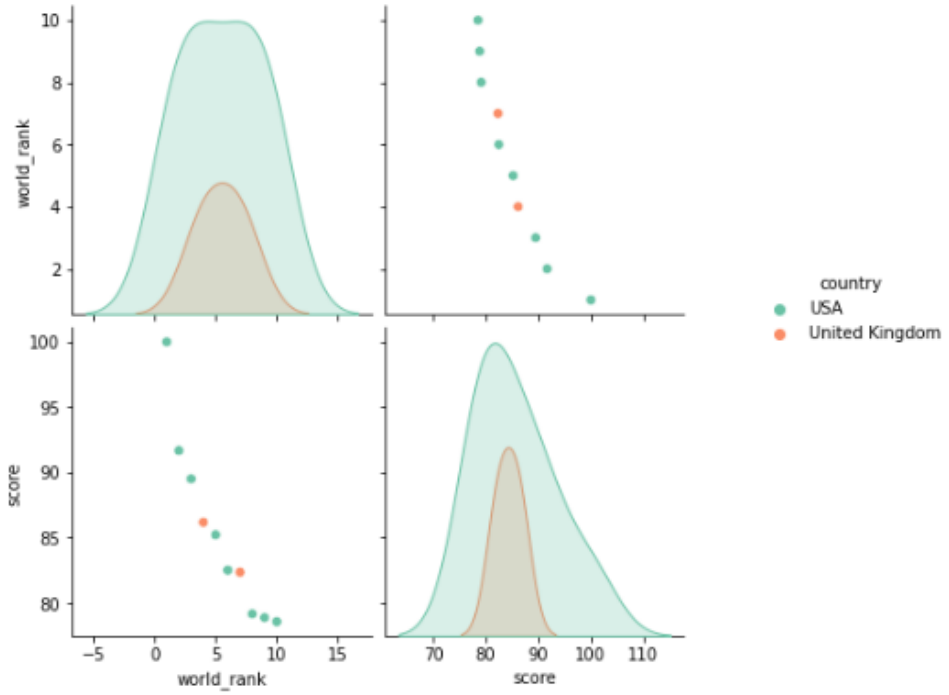
En resumen, podemos observar que un score por debajo de los 80 puntos no te permite entrar al top 10.

Score y World Rank de los países United Kingdom y USA

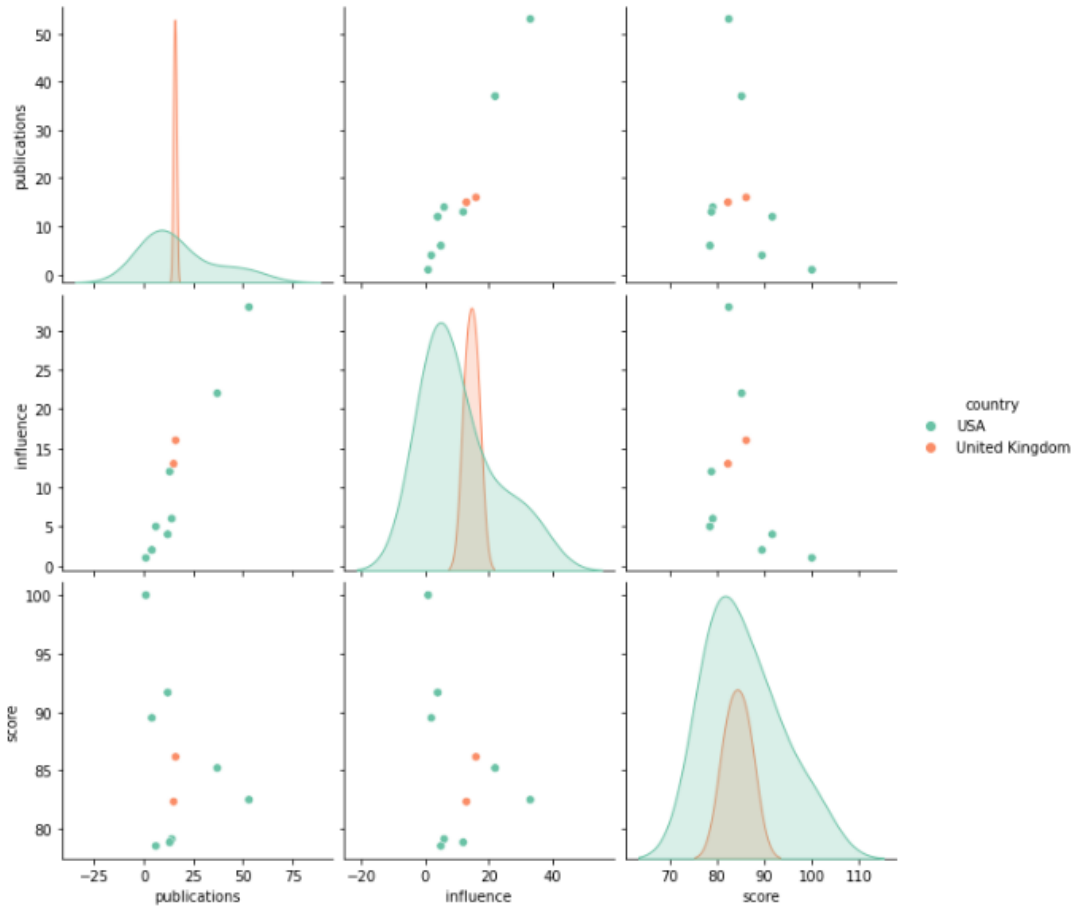
En concordancia con los gráficos anteriores, comparamos a USA y UK con las variables score y World Rank

Podemos observar que USA tiene una acumulación mayor que UK respecto a la variable Score en las primeras 10 posiciones**.** Dándole la ventaja para obtener los primeros puestos en varias categorías y en el ranking mundial.

Score y World Rank de los países United Kingdom y USA

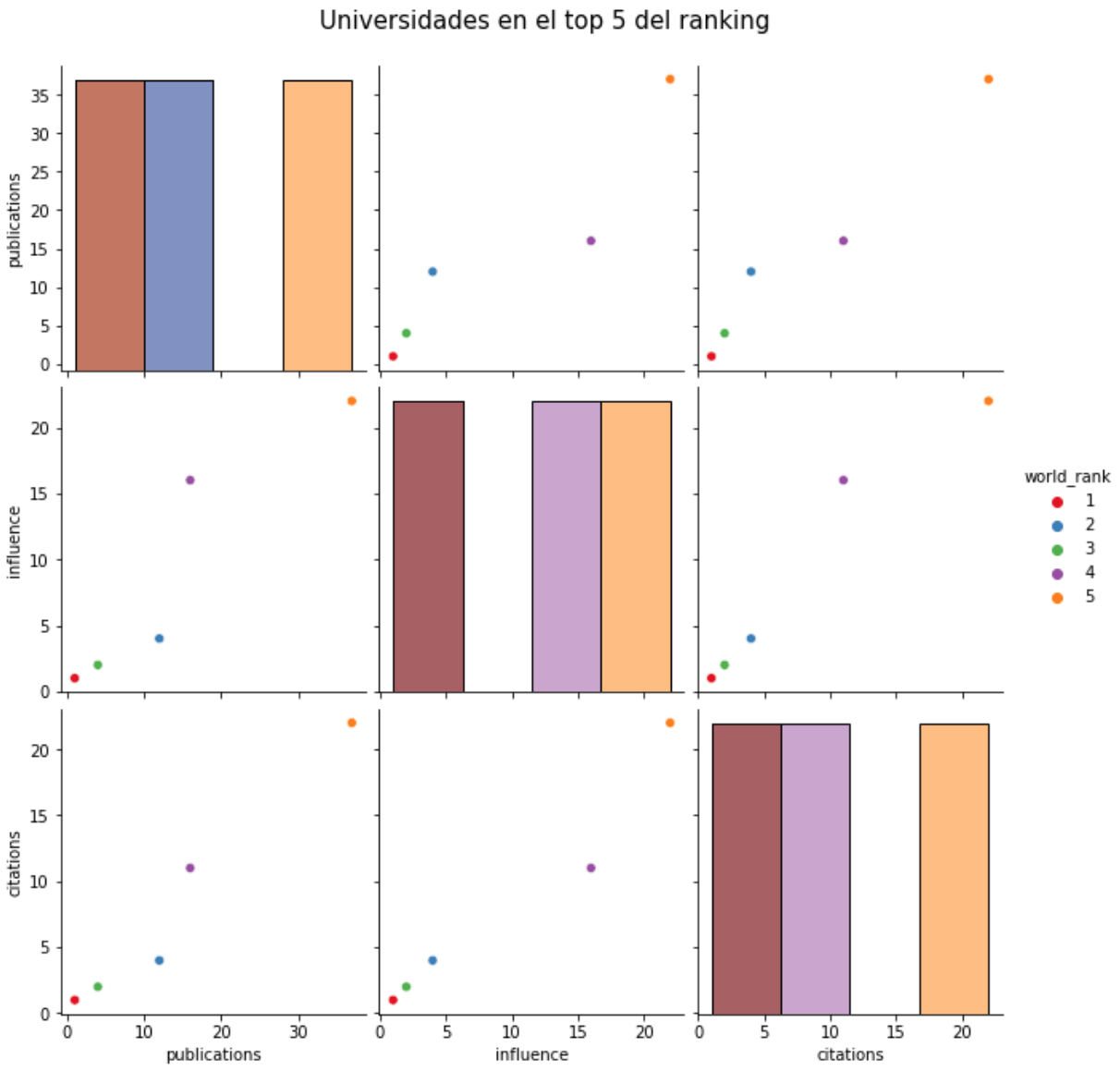


Publications, Influence y Score de los países United Kingdom y USA



Análisis de los 5 primeros puestos en los distintos rankings individuales

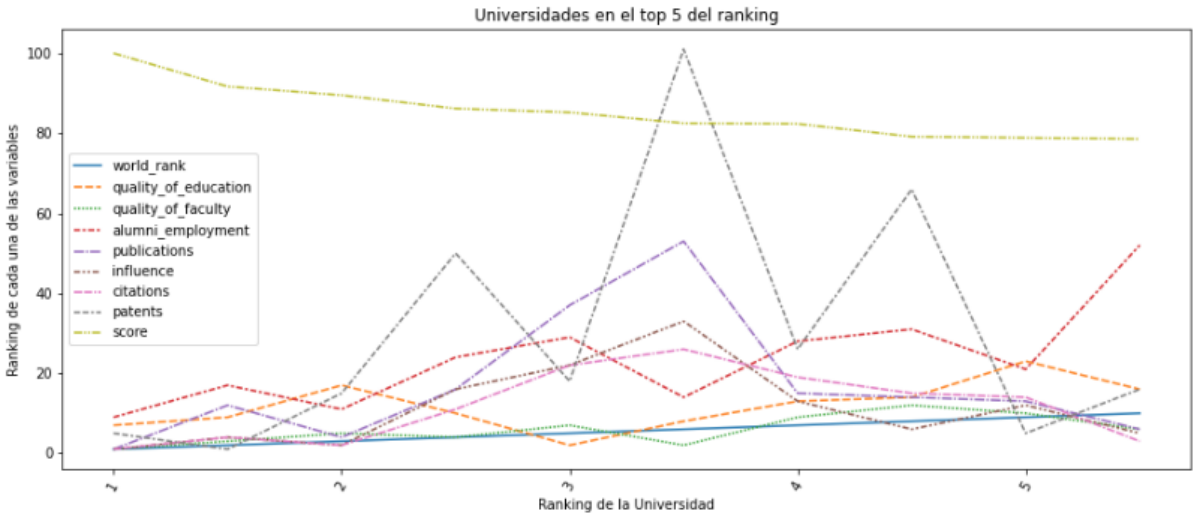
Finalmente en este gráfico observamos cómo el ranking va bajando a medida que baja el ranking de las demás variables. Y que el puesto 5 a pesar de ser top 5, está muy detrás de sus puestos superiores.



Finalmente en este gráfico observamos cómo el ranking va bajando a medida que bajan los demás rankings. Y que el puesto 5 a pesar de ser top 5, está muy detrás de sus puestos superiores.

Ranking Top 10 de universidades del año 2012

En el siguiente gráfico podemos observar, tal como venimos mencionando anteriormente que las universidades mejor rankeadas cuentan con los mejores rankings en las demás variables, lo mismo sucede con la variable score, a mejor ranking, mayor score.



Modelos

La variable a predecir se trata de la variable World Rank. Para poder trabajar con modelos de clasificación generamos una nueva variable llamada Top 500 del tipo binaria, donde clasificamos con 1 si se encuentra dentro del Top 500 y con 0 si no se encuentra allí.

Teniendo en cuenta que vamos a utilizar un modelo de clasificación, no vamos a tener en cuenta las siguientes variables que se tratan de variables cualitativas:

- Institución y Country (Por cardinalidad)
- Year (Por falta de información relevante)
- World Rank(Por lo mencionado anteriormente, se reemplazó con la nueva variable Top 500)
- Score (Se la eliminó ya que influencia en el modelo, ya que un gran score como se muestra en los gráficos y en los análisis realizados anteriormente, al tener un score alto, esa universidad va a ser parte de las mejores universidades. Caso contrario ocurre con las universidades con menor score, no poseen un buen ranking (en el gráfico se observan las universidades que con peor ranking tiene menos de 50 puntos de score)). Otro punto a considerar dentro de esta variable, es que está “formada” por las demás variables, por lo tanto podemos no considerarla y utilizar las demás, para no ser redundantes.

Es importante destacar, que la variable broad impact, fue tenida en cuenta para predecir debido a que se trata de una variable importante para el dataset, ya que la misma rankea a las universidades de acuerdo al impacto que tienen mundialmente y esta variable es influyente en la variable score, es decir cuanto mejor ranking mejor score y es la variable score la que define luego el ranking mundial.

Otro punto aclaratorio, es que no hay desbalance en la variable Top 500, ya que cuenta 1198 registros de la clase 1 y 1002 registros de la clase 0.

Modelos evaluados:

Los modelos que decidimos evaluar son los siguientes:

- **Árbol de decisión,** trata de clasificar en función del árbol, es decir los nodos intermedios (las ramas) representan soluciones. Los nodos finales (las hojas) nos dan la predicción que vamos buscando.
- **Random Forest,** se trata de un modelo que predice de acuerdo a árboles de decisión combinados con bagging, esto quiere decir que distintos árboles ven distintas porciones de los datos, ningún árbol ve todos los datos de entrenamiento. Esto hace que cada árbol se entrene con distintas muestras de datos para un mismo problema. De esta forma, al combinar sus resultados, unos errores se compensan con otros y tenemos una predicción que generaliza mejor.
- **Regresión logística,** es un método de regresión que permite estimar la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa, que clasifica las observaciones en un grupo u otro dependiendo del valor que tome la variable empleada como predictor. Para este modelo se decidió normalizar las variables.

Métricas de los Modelos evaluados

A todos los modelos se les aplicó gridsearch para dar con los mejores hyperparametros y mejorar las métricas del mismo, además de aplicarle el cross validation correspondiente, para procurar que el modelo tiene buenas métricas no solamente por haber tomado un train determinado.

1. Modelo de árbol de decisión

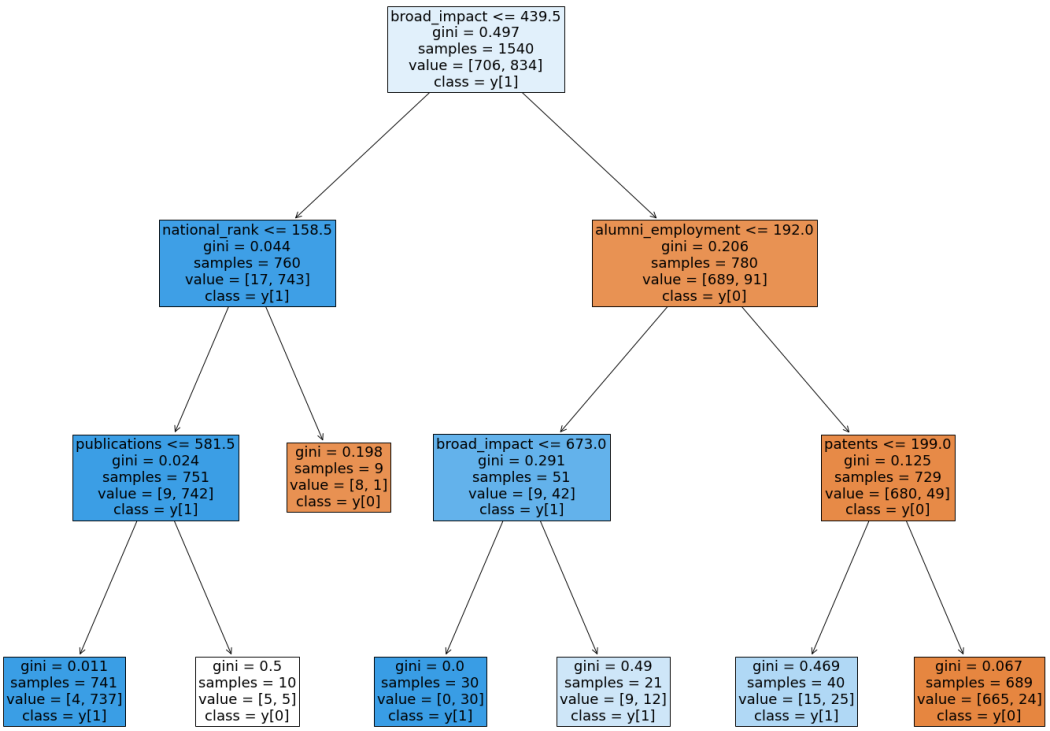
Por lo que observamos en el dataset, la mayoría de las variables son categóricas, por lo que definimos una variable binaria para poder clasificar, Top 500, que se trata de la variable a predecir. Es por ello que un árbol de decisiones surge como un modelo atractivo para intentar

predecir si una universidad se encuentra dentro del Top 500 del ranking mundial de universidades.

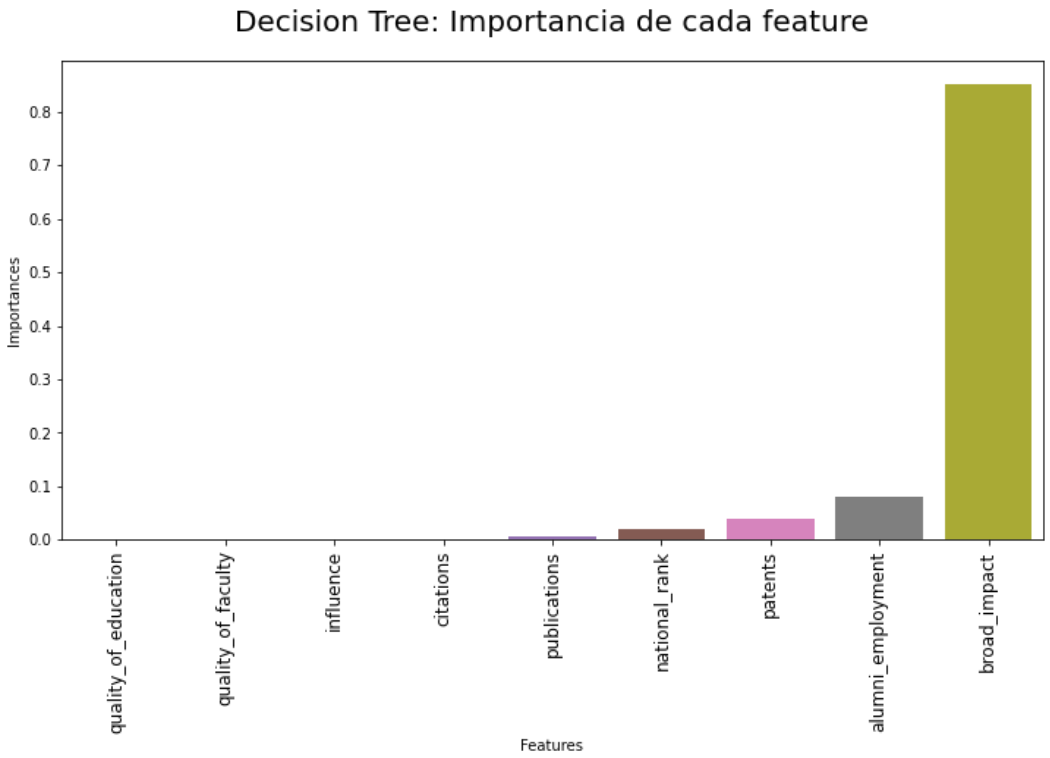
La variable que tiene mayor peso al tomar la decisión es la variable broad impact.

Para encontrar los mejores hiper parámetros para el modelo se utilizó Grid Search CV de SKlearn y se obtuvo que para el árbol óptimo la mejor profundidad es 3 y la mejor cantidad mínima de muestras por hoja es 5.

Se ajustó el modelo con los hiperparámetros optimizados y se obtuvo la siguiente estructura



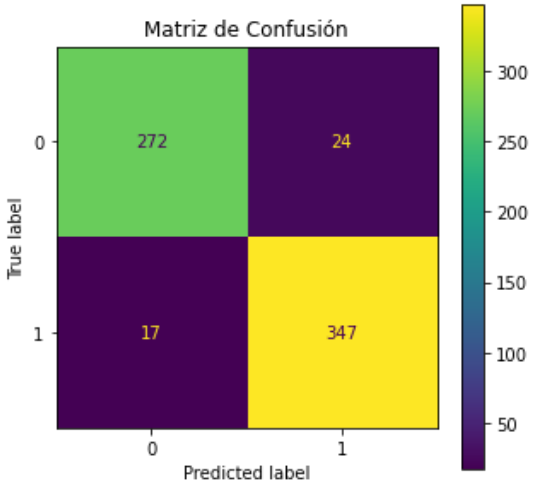
En el siguiente gráfico podemos observar que la variable más importante es la variable broad_impact, seguido por alumni_employment.



En el siguiente cuadro podremos observar algunas métricas acerca del modelo:

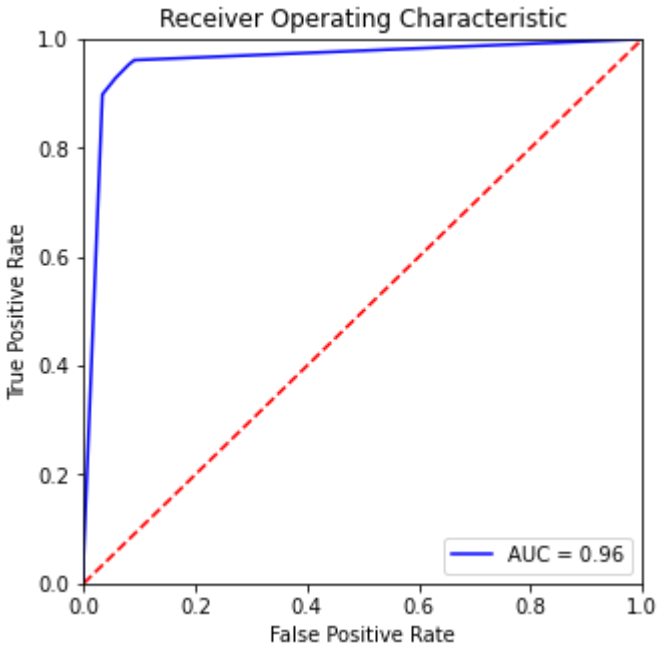
Modelo Árbol de Decisión			
Accuracy	Precisión	Recall	f1
0.937879	0.935310	0.953297	0.944218

Podemos observar que las métricas del modelo son muy buenas, por lo cual el modelo predice bien la variable elegida.



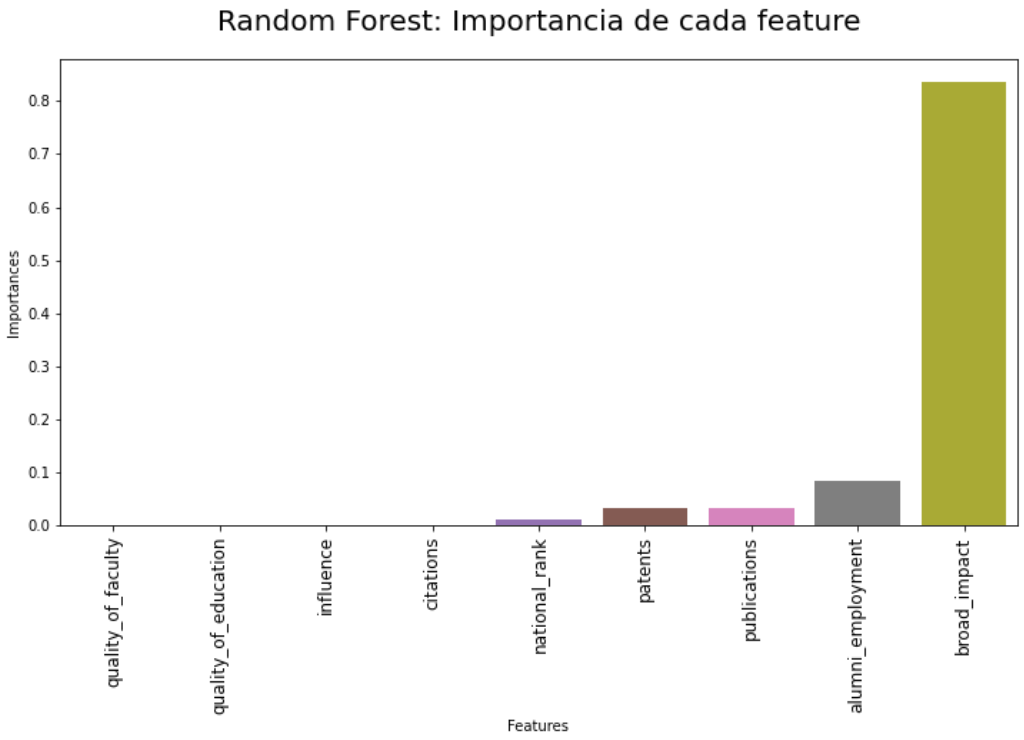
La matriz de confusión, nos muestra que el modelo clasificó 17 universidades fuera del top 500 cuando en realidad la misma si se encontraba allí. Por otro lado también clasificó 24 universidades dentro del top500 cuando en realidad no estaba allí.

La curva AUC es: 0.96 (Es decir es mucho mejor tener el modelo que no tenerlo)



2. Random Forest

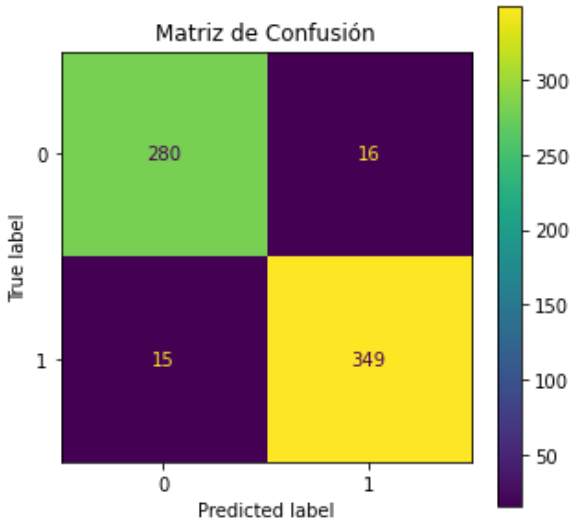
Si bien el Modelo de Árbol de decisión modela bien nuestros datos, vamos a probar con Random Forest para ver si podemos mejorar la precisión de la predicción.



Aquí se muestran las importancias que da a cada feature el modelo de random forest. Vemos que la variable más relevante sigue siendo la variable broad_impact.

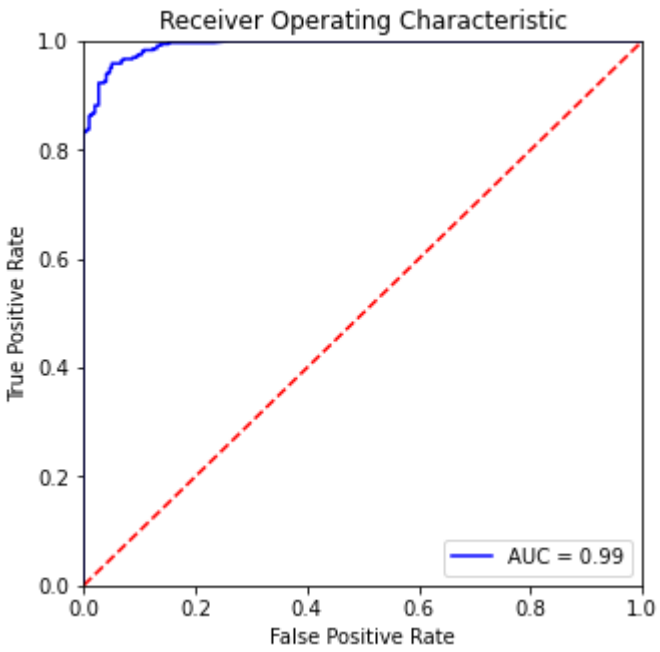
En el siguiente cuadro podremos observar algunas métricas acerca del modelo:

Modelo Random Forest			
Accuracy	Precisión	Recall	f1
0.953030	0.953030	0.956164	0.958791



La matriz de confusión, nos muestra que el modelo clasificó 15 universidades fuera del top500 cuando en realidad estaban allí y 16 dentro del top 500 cuando en realidad no lo estaban.

La curva AUC es: 0.99

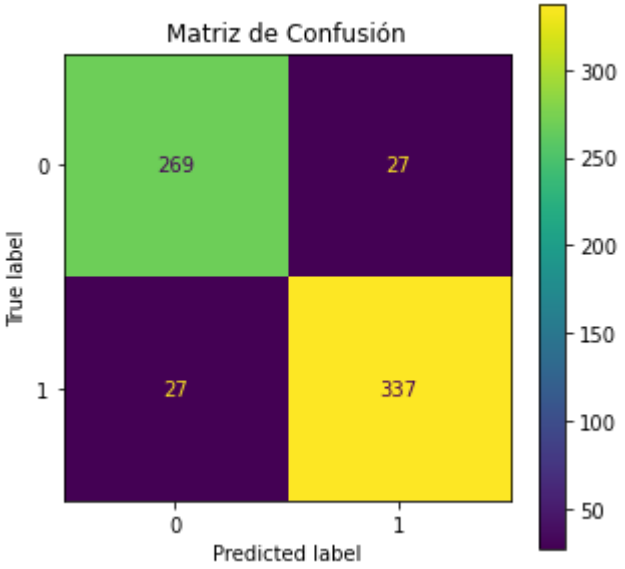


3. Regresión logística

Por último, decidimos probar un modelo más, en este caso el de Regresión Logística, donde para encontrar los mejores hiperparámetros para el modelo se utilizó RandomizedSearch y se obtuvo así los mejores parámetros a utilizar en el modelo.

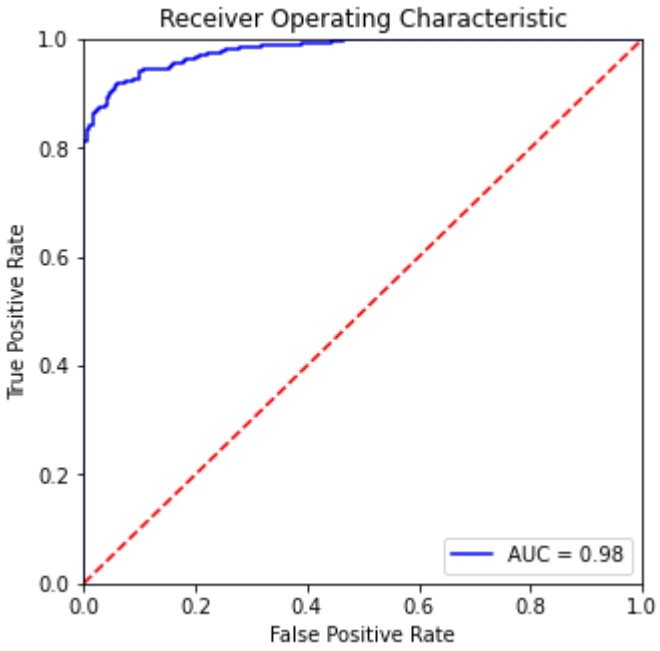
Se ajustó el modelo con los hiperparámetros optimizados y se obtuvo la siguientes métricas:

Modelo Regresión Logística			
Accuracy	Precisión	Recall	f1
0.918182	0.925824	0.925824	0.925824



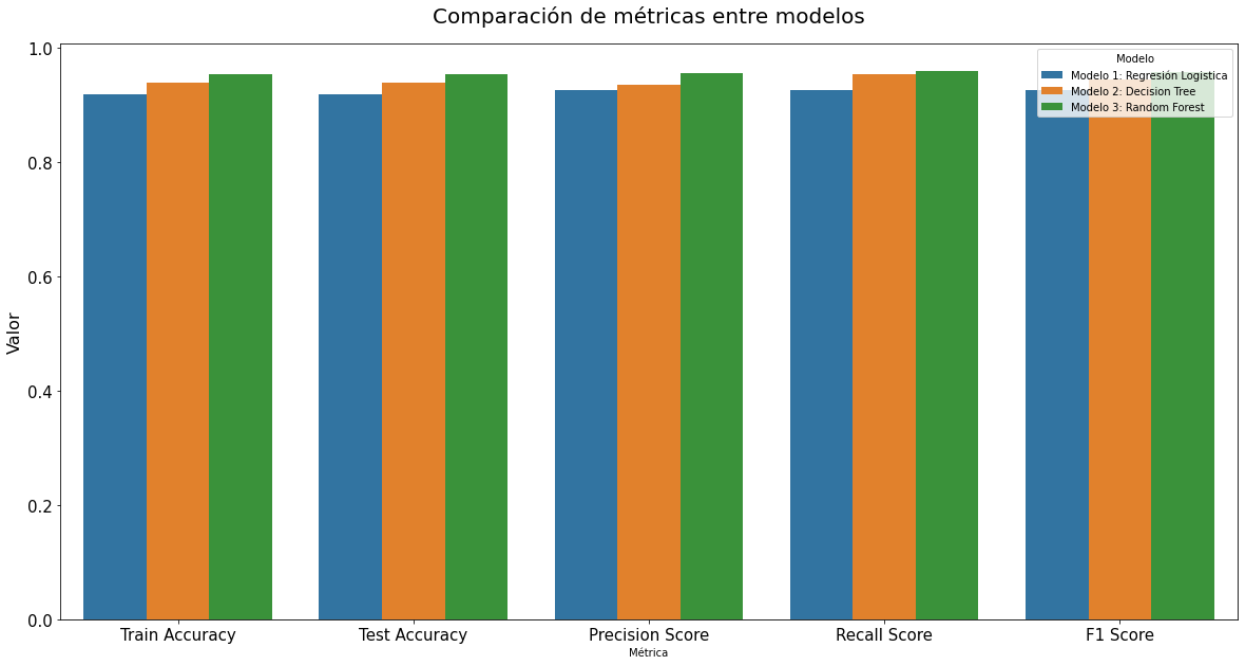
La matriz de confusión, nos muestra que el modelo clasificó 27 universidades fuera del top 500 cuando en realidad la misma si se encontraba allí. Por otro lado también clasificó 27 universidades dentro del top500 cuando en realidad no estaba allí.

La curva AUC es: 0.98



Cuadro comparativo Modelos

Modelo	Train accuracy	Test Accuracy	Precisión	Recall	F1
Modelo 1: Regresión Logística	0.918182	0.918182	0.925824	0.925824	0.925824
Modelo 2: Decision Tree	0.937879	0.937879	0.935310	0.953297	0.944218
Modelo 3: Random Forest	0.953030	0.953030	0.956164	0.958791	0.957476



Conclusión

Para concluir, consideramos que el Random Forest es el mejor de los modelos, ya que las métricas así lo indican, es claro que debe ser mejor que un Decision Tree ya que son varios trees. El 2do mejor modelo es el Decision tree ya que supera en todas las métricas al Regresión logística.

Creemos que las dificultades aparecieron al principio, porque no estábamos acostumbrados al manejo de tanta data y tuvimos que investigar más a profundidad qué es lo que estamos analizando, de qué se trataba cada variable, del mismo modo al manejar tanta data el principal problema era detectar cuál iba a ser nuestro target, a qué apuntaba nuestro trabajo y modelo. En un principio creíamos que regresión lineal era lo indicado, pero posteriormente nos pareció enriquecedor pasar el modelo a clasificación. Nos encontramos con un dataset que fue fácil de manipular y que rápidamente pudimos ocuparnos de los null y nan, eso permitió que nos podamos enfocar más en otros aspectos y hacer un análisis más profundo de la data. Pasamos por todas las etapas, arrancando con análisis univariado hasta análisis multivariado.