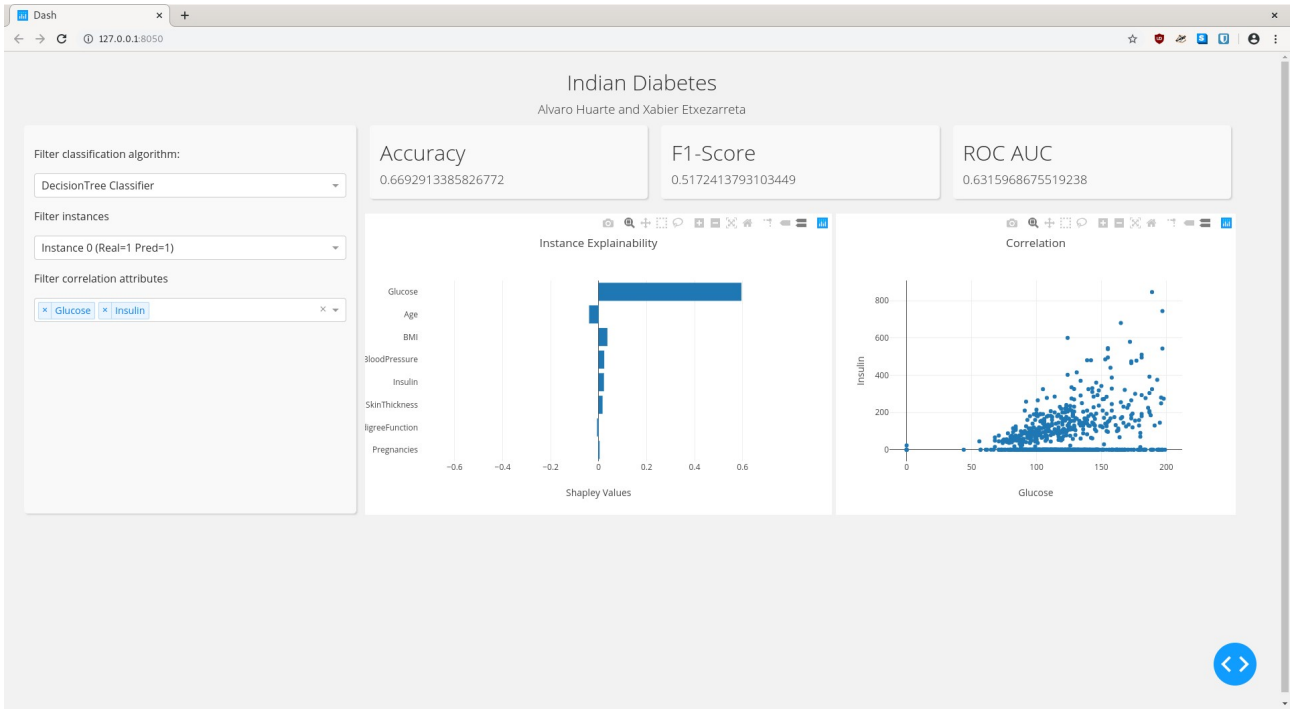
 # VD_plotly_dash
Alvaro Huarte y Xabier Etxezarreta



Introducción
El objetivo de este dashboard ha sido representar con el dataset de diabetes los resultados de la clasificación con diferentes algoritmos, la interpretabilidad de las predicciones y la correlaciones entre las variables.

Conjunto de datos
El objetivo del dataset es predecir si un paciente tiene o no diabetes, basándose en ciertas mediciones médicas que están incluidas en el conjunto de datos. Es originalmente del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales

El dataset consta de varios indicadores médicos (datos de entrada) y una variable de salida que clasifica las instancias:

Datos de entrada
- Pregnancies: Número de veces que ha estado embarazada.
- Glucose: Concentración de glucosa en plasma.
- BloodPressure: Presión arterial (mm Hg).
- SkinThickness: Grosor del pliegue de la piel del tríceps (mm).
- Insulin: Insulina (mu U/ml).
- BMI: Índice de masa corporal.
- DiabetesPedigreeFunction: Función de pedigrí de la diabetes.
- Age: Edad (años).

Datos de salida
- Class: (0 o 1)

Librerías
Todas las librerías utilizas están definidas en el fichero "requirements.txt".
```bash
pip install -r requirements.txt
```
Filtros
A continuación se detallan las diferentes opciones o filtros que se han desarrollado para la modificación del dashboard.

Filtro de algoritmos (dropdown simple)
El objetivo de este filtro es dar la posibilidad al usuario de elegir el tipo de algoritmo de CLASIFICACIÓN, pudiendo de esta forma comparar los resultados obtenidos con cada uno de ellos. Se han introducido tres algoritmos, los tres basados en árboles: DecisionTree, RandomForest y XGBoost.

Para que el usuario pueda realizar las comparaciones se han utilizado tres métricas: Accuracy, F1-Score y ROC-AUC. Estos indicadores son actualizados cuando el usuario cambia el filtro. Hemos decidido representar estos tres indicadores con valores numéricos debido a la facilidad de lectura, comprensión y comparación que conlleva utilizarlos. Cuando se selecciona un algoritmo se actualiza el dropdown de instancias.

Filtro de instancias (dropdown simple)
EL objetivo es dar la opción al usuario de elegir que instancia quiere representar en la gráfica de barras horizontal para que pueda entender la predicción realizada. El dropdown está compuesto por una lista de instancias, indicando en cada una de ellas el valor real de la clase y la predicción realizada por el algoritmo seleccionado. Con esto el usuario puede identificar mejor que predicciones se han realizado de forma correcta o de forma errónea.

Se ha discutido la opción de introducir colores en las barras dependiendo si el valor es positivo o negativo, facilitando ver que variables afectan positivamente o negativamente a la predicción. Se pensó en utilizar el color verde en los valores positivos y el color rojo en los valores negativos. Esta posibilidad fue desechada debido a que el color rojo representa algo negativo y el verde algo positivo cuando en lo representado un valor negativo no significa algo negativo y viceversa. Se llego a la conclusión, después de consultar a un experto (Dani), de que utilizar otros colores no aportan ninguna información extra al usuario por lo que esta idea fue desechada.

En esta gráfica de barras se ha decidido representar de forma horizontal y posicionando el valor 0 en el centro del eje X para poder visualizar de forma más clara que variables aportan positivamente y negativamente. Las variables se han ordenado de mayor a menor (valor absoluto), en función a la aportación a la predicción.

Filtro de correlaciones (dropdown múltiple)
Su objetivo es mostrar la relación de dos variables a través de un gráfico de dispersión. Las opciones utilizadas en este filtro corresponden a cada una de las variables del dataset. Se ha utilizado un dropdown de tipo "multi", permitiendo al usuario realizar todas las combinaciones posibles entre dos variables.

Como restricciones, la gráfica de correlaciones se mostrará únicamente si se seleccionan dos variables, en el caso de seleccionar un número mayor o menor no aparecerá ninguna

gráfica. Se ha utilizado el mismo color para representar todas las correlaciones. No se ha encontrado sentido el uso de los colores.

Preprocesamiento
Se han utilizado tres algoritmos basados en árboles: DecisionTree, RandomForest y XGBoost. El dataset se ha dividido en train y test, utilizando la parte de train para entrenar y la parte de test para el cálculo de las métricas y la interpretabilidad de las instancias.

Para el cálculo de la importancia de las variables en la predicción, se ha utilizado la libreria "SHAP". Esta librería calcula los "shapeley values" de cada variable representando la importancia y la tendencia de cada variable.