**Ingeniería de Sistemas**

**Computación Científica y Analítica**

**Taller 3**

Enlace de entrega: [aquí](https://aulavirtual.unbosque.edu.co/mod/assign/view.php?id=336900)

Fecha máxima de entrega: **Junio 5, 2022 11:59 PM**

Revise cada uno de los scripts dispuestos en [este](https://github.com/datascience-unbosque/ml-tutorial) repositorio y responda las siguientes preguntas:

1. **Del script de forest fires:**
   1. ¿Se desea resolver el problema utilizando aprendizaje supervisado o no supervisado?

El poblema es de aprendizaje supervisado porque se parte de un conjunto de datos etiquetado previamente, es decir, se conoce el valor del atributo objetivo para el conjunto de datos.

* 1. ¿Es un problema de clasificación o de regresión?

Es un problema de regresión porque se utiliza para predecir valores continuos.

* 1. ¿Qué interpretación le puede dar a los resultados obtenidos?

De acuerdo con el error absoluto medio (11,70) y el error cuadrático (13,92), los valores de predicción se alejan bastante de los valores reales. Además, el coeficiente de determinación (0.056%) indica que alto grado de imperfección en las predicciones.

1. **Del script de recursos humanos (rrhh):**
   1. ¿Cuál es la clase para la que el modelo más se equivoca? ¿Por qué?

La clase para la cual el modelo más se equivoca es medium porque de acuerdo con la matriz de confusión se clasificaron 1724 valores como low cuando en realidad eran medium.

Chart

Description automatically generated

* 1. ¿Cuál cree que es el propósito del parámetro *max\_depth* usado al momento de instanciar el modelo de árbol de decisión?

La profundidad máxima teórica que puede alcanzar un árbol de decisión es uno menos que el número de muestras de entrenamiento, pero ningún algoritmo llegará a este punto por el sobreajuste. Hay que tener en cuenta que es la cantidad de muestras de entrenamiento y no la cantidad de atributos porque los datos se pueden dividir en el mismo atributo varias veces. En general, cuanto más profundo un árbol, más complejo se volverá el modelo porque tendrá más divisiones y capturará más información sobre los datos; esta es una de las causas fundamentales del sobreajuste dado que el modelo no podrá generalizar bien en el conjunto de prueba. También es malo tener una profundidad muy baja porque el modelo no se ajustará bien, así que no hay una fórmula para encontrar el mejor valor para todos, solo a través de la experimentación.

* 1. Para este caso particular, ¿por qué cree que es difícil obtener un buen clasificador? Dado que hay demasiados datos, un solo árbol de decisión puede tener mucha profundidad y provocar un sobreajuste, por lo que se recomienda en estos casos utilizar un random forest.

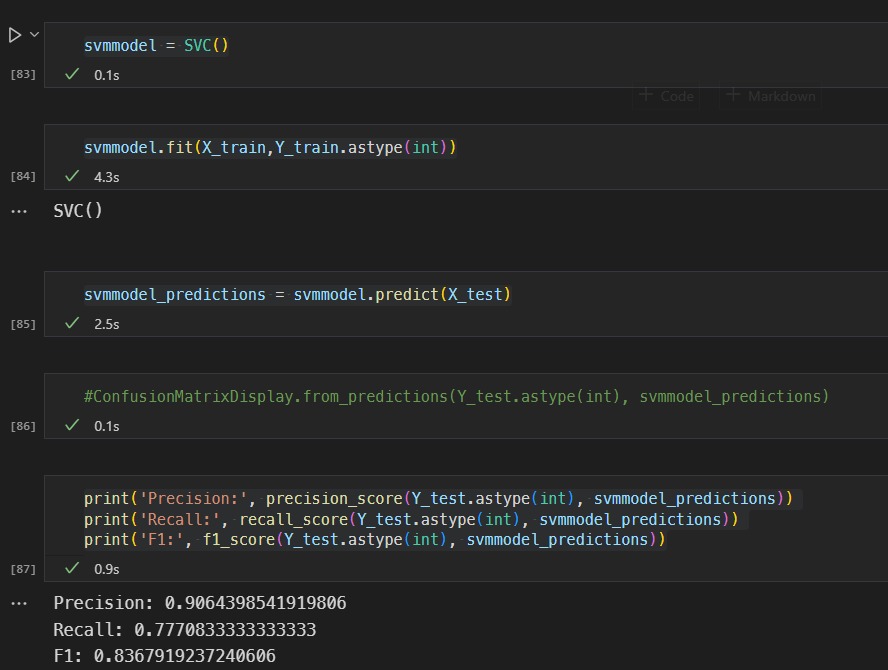
1. **Identificación de géneros musicales:** Tenga en cuenta que hay dos scripts: *music.ipynb* y *music-multiclass.ipynb*. En el primero se intenta crear un modelo clasificador solo para dos clases (caso binario) y en el segundo se entrena uno para todas las clases (géneros musicales) del dataset.
   1. Para el caso binario (*jazz and blues* vs. *soul and reggae*) ¿Es posible obtener mejores métricas entrenando un modelo basado en Random Forest? La implementación del random forest no es sustancialmente mejor a los otros modelos, que están alrededor del 0.8 de precisión, recall y f1.

Text

Description automatically generated

* 1. Escoja otro par de géneros, entrene un conjunto de modelos y documente los resultados del mejor que se haya obtenido.

Los géneros escogidos fueron foll y punk. El mejor modelo fue el de Support vector machine.



* 1. Para el caso multi-clase, ¿cuál es la clase para la que el modelo más se equivoca? ¿Por qué?

El modelo se equivoca más para las clases classic pop rock y folk. La primera cuenta el 40% de los datos y la segunda el 22%, de modo que los datos están centrados alrededor de estas dos clases.

A screenshot of a computer

Description automatically generated with low confidence

El modelo tiene una precisión, un recall y un f1 alrededor del 58%.

Text

Description automatically generated

* 1. Para el caso multi-clase, el modelo basado en red neuronal parece estar mayoritariamente sesgado hacia un género particular ¿Cuál género cree que es?

El género es classic pop rock.

1. **Segmentación de cajas de compensación familiar (subsidio):**
   1. ¿Qué cajas de compensación parecen ser mayoritariamente diferentes a las demás?

Los cluster más pequeños son el cero y el dos. Para el cero, las cajas de compensación diferentes son la caja de compensación familiar de Antioquia, Colsubsidio y Compensar.

Graphical user interface, text

Description automatically generated

Para el cluster dos, son Cafam y Caja de compensación del Valle del Cauca.

A picture containing text, monitor, screenshot, black

Description automatically generated

* 1. ¿A partir de qué características utilizadas para el entrenamiento del modelo se podría explicar la razón por la que las cajas anteriores fueron agrupadas en clusters tan pequeños?

Dado que esas cajas de compensación tienen una cantidad mucho mayor de empresas afiliadas, se deduce que son mucho más grandes y manejan más afiliados.

Text

Description automatically generated

* 1. ¿Se pueden obtener resultados más homogéneos utilizando cantidades diferentes de clusters para el entrenamiento? Entienda homogeneidad como clusters con cantidades similares de instancias de datos.

Usando el método del codo, el punto en el que se observa e cambio brusco en la inercia dirá el número óptimo de clusters a seleccionar para ese data set; el punto que representaría al codo del brazo será el número óptimo.

Text

Description automatically generated

Chart, line chart

Description automatically generated

Chart

Description automatically generated

Para este caso, el número de cluster óptimo es de dos.

**Mecanismo de entrega:** [aquí](https://datascience-unbosque.github.io/index.html#faq)

**Rúbrica:**

* 80% - Cumplimiento de todos los requerimientos.
* 20% - Calidad y documentación de código, uso de GitHub y contribución equitativa del grupo de trabajo. Video subido a YouTube.