Hoja de Trabajo 4.
Árboles de Decisión

INTRODUCCIÓN:

Kaggle

Kaggle es una comunidad en línea de científicos de datos, propiedad de Google LLC. Permite a los usuarios encontrar y publicar conjuntos de datos, explorar y construir modelos en un entorno de ciencia de datos basado en la web, trabajar con otros científicos de datos e ingenieros de aprendizaje automático, y participar en competencias para resolver los desafíos de la ciencia de datos. Tuvo su inicio al ofrecer competencias de aprendizaje automático y ahora también ofrece una plataforma pública de datos, una mesa de trabajo basada en la nube para la ciencia de la información y educación en IA de formato corto. El 8 de marzo de 2017, Google anunció que estaban adquiriendo Kaggle.

Conjunto de datos a utilizar

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

Notas:

- La hoja no se calificará si no pertenece a ningún grupo de los creados en canvas para esta hoja.

INSTRUCCIONES

Utilice el data set House Prices: Advanced Regression Techniques que comparte Kaggle. Utilice el análisis exploratorio que hizo en la hoja de trabajo anterior. Si considera que le faltó algo por explorar y cree que lo necesita, hágalo. Genere un informe con las explicaciones de los pasos que llevó a cabo y los resultados obtenidos. Recuerde que la investigación debe ser reproducible por lo que debe guardar el código que ha utilizado para resolver los ejercicios.

ACTIVIDADES

1. Use los mismos conjuntos de entrenamiento y prueba que usó para los modelos de regresión lineal en la hoja de trabajo anterior.
2. Elabore un árbol de regresión para predecir el precio de las casas usando todas las variables.
3. Úselo para predecir y analice el resultado. ¿Qué tal lo hizo?
4. Haga, al menos, 3 modelos más cambiando el parámetro de la profundidad del árbol. ¿Cuál es el mejor modelo para predecir el precio de las casas?
5. Compare los resultados con el modelo de regresión lineal de la hoja anterior, ¿cuál lo hizo mejor?
6. Dependiendo del análisis exploratorio elaborado cree una variable respuesta que le permita clasificar las casas en Económicas, Intermedias o Caras. Los límites de estas clases deben tener un fundamento en la distribución de los datos de precios, y estar bien explicados

7. Elabore un árbol de clasificación utilizando la variable respuesta que creó en el punto anterior. Explique los resultados a los que llega. Muestre el modelo gráficamente. Recuerde que la nueva variable respuesta es categórica, pero se generó a partir de los precios de las casas, no incluya el precio de venta para entrenar el modelo.
8. Utilice el modelo con el conjunto de prueba y determine la eficiencia del algoritmo para clasificar.
9. Haga un análisis de la eficiencia del algoritmo usando una matriz de confusión para el árbol de clasificación. Tenga en cuenta la efectividad, donde el algoritmo se equivocó más, donde se equivocó menos y la importancia que tienen los errores.
10. Entrene un modelo usando validación cruzada, prediga con él. ¿le fue mejor que al modelo anterior?
11. Haga al menos, 3 modelos más cambiando la profundidad del árbol. ¿Cuál funcionó mejor?
12. Repita los análisis usando random forest como algoritmo de predicción, explique sus resultados comparando ambos algoritmos.

EVALUACIÓN

- **(10 puntos)** Creación de la variable respuesta para árbol de clasificación. Explicación de los límites de las categorías.
- **(8 puntos)** Generación del modelo de regresión. Análisis de los resultados obtenidos
- **(8 puntos)** Comparación del árbol de regresión con el modelo de regresión lineal de la hoja anterior. Explicación de resultados
- **(8 puntos)** Tuneo del parámetro de la profundidad del árbol. Selección del mejor modelo, todo está explicado claramente.
- **(12 puntos)** Árbol de Clasificación. Representación gráfica del modelo.
- **(12 puntos)** Modelo con validación Cruzada y tuneo de la profundidad del árbol de clasificación.
- **(12 puntos)** Random Forest
- **(30 puntos)** Análisis de resultados de aplicación del algoritmo para predecir o clasificar sobre el conjunto de prueba. Comparación entre algoritmos.

MATERIAL A ENTREGAR

- Archivo .rmd, .ipynb o Google docs con el informe con lo solicitado en las instrucciones
- Script de R o de Python que utilizó debidamente organizado y comentado (Si utilizó rmd debe añadir el html que se genera)
- Link de controlador de versiones utilizado.

FECHAS DE ENTREGA

- **AVANCE:** Puntos del 1 al 4 de la sección de actividades: viernes 8 de marzo a las 23:59.
- **ENTREGA FINAL:** lunes 11 de marzo a las 23:59

NOTA: Solo se calificará el Documento Final si está entregado el avance con todo lo que se pide.