Hoja de Trabajo 6.
Modelos de Regresión Logística

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 de trabajo se realizará en las mismas parejas de la hoja anterior.
- Los grupos serán seleccionados por afinidad.
- 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. Debe hacer un análisis exploratorio para entender mejor los datos, sabiendo que el objetivo final es predecir los precios de las casas. Recuerde explicar bien cada uno de los hallazgos que haga. La forma más organizada de hacer un análisis exploratorio es generando ciertas preguntas de las líneas que le parece interesante investigar. Genere un informe en pdf 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 y/o cada uno de los pasos llevados a cabo si utiliza una herramienta visual.

ACTIVIDADES

1. Transforme la variable respuesta de manera que pueda aplicar el modelo de regresión logística (variable dicotómica).
2. Use los mismos conjuntos de entrenamiento y prueba que utilizó en las hojas anteriores.
3. Elabore un modelo de regresión logística utilizando el conjunto de entrenamiento y explique los resultados a los que llega. Muestre el modelo gráficamente. El experimento debe ser

reproducible por lo que debe fijar que los conjuntos de entrenamiento y prueba sean los mismos siempre que se ejecute el código.

4. Analice el modelo. Determine si hay multicolinealidad en las variables, y cuáles son las que aportan al modelo, por su valor de significación. Haga un análisis de correlación de las variables del modelo y especifique si el modelo se adapta bien a los datos. Explique si hay sobreajuste (overfitting) o no.
5. Utilice el modelo con el conjunto de prueba y determine la eficiencia del algoritmo para clasificar o predecir, en dependencia de las características de la variable respuesta.
6. Haga un análisis de la eficiencia del algoritmo usando una matriz de confusión. Tenga en cuenta la efectividad, donde el algoritmo se equivocó más, donde se equivocó menos y la importancia que tienen los errores.
7. Compare la eficiencia del algoritmo con el resultado obtenido al aplicar a los datos modelos de árbol de decisión, random forest, naive bayes y regresión lineal (opcional). ¿Cuál es mejor para predecir? ¿Cuál se demoró más en procesar?

EVALUACIÓN

- **(25 puntos)** Análisis del modelo generado. Recuerde explicar los razonamientos.
- **(25 puntos)** Análisis de las variables a incluir en el modelo. Pruebas de normalidad, correlación, etc.
- **(10 puntos)** Aplicación del modelo al conjunto de prueba.
- **(20 puntos)** Matriz de confusión de cada modelo. Explicación de los resultados obtenidos
- **(20 puntos)** Comparación del método de regresión logística con el resto de los algoritmos estudiados.

MATERIAL A ENTREGAR

- Archivo .r o .py con el código y hallazgos comentados
- Archivo .pdf o html con las conclusiones y hallazgos encontrados.