Hoja de Trabajo 4.
Modelos de Regresión Lineal

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.

Competencias de Kaggle

Desde sus inicios se ha dedicado a organizar competencias para estimular a los científicos de datos a resolver problemas reales sirviendo a grandes empresas y a organizaciones sociales.

Opciones de competencias para esta hoja de trabajo

1. PetFinder.my Adoption Prediction. How cute is that doggy in the shelter? (https://www.kaggle.com/c/petfinder-adoption-prediction) "*En esta competencia, desarrollarás algoritmos para predecir la capacidad de adopción de las mascotas, específicamente, ¿con qué rapidez se adopta una mascota? Si tienen éxito, se adaptarán a las herramientas de inteligencia artificial que guiarán a los refugios y rescatadores de todo el mundo para mejorar el atractivo de los perfiles de sus mascotas, reducir el sufrimiento y la eutanasia de los animales*."

Notas:

- La hoja de trabajo se realizará en parejas.
- 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.

ACTIVIDADES

1. Use los mismos conjuntos de entrenamiento y prueba para probar el algoritmo.
2. Elabore un modelo de regresión lineal 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.
3. 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.

4. 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.
5. 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.
6. Compare la eficiencia del algoritmo con el resultado obtenido con el árbol de decisión. ¿Cuál es mejor para predecir? ¿Cuál se demoró más en procesar?
7. Actualice el kernel en kaggle donde ponga el código generado en esta tarea para que compita por uno de los premios.

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 (Si aplica). Explicación de los resultados obtenidos
- **(20 puntos)** Comparación del método de regresión lineal con el árbol de decisión.

MATERIAL A ENTREGAR

- Archivo .r o .py con el código y hallazgos comentados
- Archivo .pdf con las conclusiones y hallazgos encontrados. (Opcional, puede incluir comentarios en el archivo de código)
- Link del kernel creado en kaggle.