Proyecto 2: Estudio sobre la dotación docente por establecimiento educacional

1 Objetivo

El objetivo principal de este proyecto es construir y validar un modelo predictor sobre un conjunto de datos reales, aplicando los conceptos básicos del área de Machine Learning utilizando el lenguaje Python.

A cada grupo (de 4 a 5 estudiantes) se le asignará dos regiónes con la que deberá trabajar para la elaboración de este proyecto.

2 Descripción

En este proyecto, se utilizará la base de datos relacionada con el **Esquema de Registro de Dotación Docente por Establecimiento para el año 2023**. Estos datos han sido extraídos del centro de estudios del MINEDUC. Para obtener información más detallada sobre cada una de las características del conjunto de datos, incluyendo el tipo de dato y una descripción detallada, se le proporcionará un archivo llamado **Anexo 1.pdf**.

3 Preguntas

3.1 Pregunta 0 (descripción de variables)

Con el objetivo de familiarizarse y entender mejor la base de datos, se requiere realizar un análisis de esta.

1. Generar un gráfico por cada una de las variables. El gráfico debe tener un título descriptivo y debe mostrar información clara de la variable.

2. Se debe indicar por cada una de las variables su tipo (cuantitativa o cualitativa) justificando su decisión. Se debe recordar que el tipo de variable depende de su naturaleza, no del formato en el que se encuentre en la base de datos.

3.2 Pregunta 1 (test de independencia)

Los docentes UTP (Unidad Técnico Pedagógica) en Chile son profesionales de la educación encargados de la coordinación pedagógica y técnica en las instituciones educativas. Se dedican a la planificación, evaluación y seguimiento del rendimiento académico de los estudiantes, ofrecen apoyo y orientación a otros docentes, participan en el desarrollo profesional y gestionan aspectos curriculares. También mantienen comunicación con los padres para informarles sobre el progreso de los estudiantes, desempeñando un papel fundamental en la mejora de la calidad educativa en el país.

1. Obtenga un gráfico apropiado que le permita estudiar la distribución del número de docentes pertenecientes a la planta UTP en los establecimientos educacionales de sus regiones. Discuta las características más relevantes presentes en sus gráficos.

2. Al Ministerio de Educación le preocupa que existan diferencias en el número de docentes UTP entre los distintos tipos de establecimientos (municipales, particulares subvencionados y particulares pagados). Implemente un test de independencia chi-cuadrado que le permita concluir si existe dependencia entre el tipo de establecimiento: (solo considere municipales, particulares subvencionados y particulares pagados) y el número de docentes UTP. En particular:

    - Reporte la tabla de frecuencias observadas y esperadas.

- Reporte el p-valor obtenido y concluya con un valor $\alpha = 0.05$. Interprete sus resultados.

3. Obtenga un gráfico apropiado para mostrar la distribución del número de docentes de planta UTP por establecimiento. Finalmente, muestre como esa distribución cambia (o no) para los distintos establecimientos: municipales, particulares suvencionados y particulares pagados. Incluya una breve discusión sobre si lo que observa en sus gráficos concuerda con lo obtenido en la implementación del test anterior.

3.3 Pregunta 2 (test de bondad de ajuste)

En Chile, las "Horas de Contrato" (variable `HH_A`) para los docentes de aula se refieren a la cantidad de horas que los profesores están contratados para trabajar en una institución educativa durante un período específico, generalmente un año escolar. La cantidad de horas de contrato puede variar según el nivel educativo y el tipo de establecimiento educacional.

En el contexto de un estudio educativo, se plantea la hipótesis de que las **Horas de Contrato de los Docentes de Aula** siguen una distribución exponencial. Para investigar esta afirmación, sigua los siguientes pasos:

1. **Análisis Visual:** Genere un histograma de densidad para la variable en cuestión y añada en este gráfico la función de densidad asociada a la distribución exponencial. Disculta sobre la factibilidad de la hipótesis planteada en el estudio utilizando el histograma generado.

   Hint: Recuerde que la densidad de la distribución exponencial esta dada por:

   $$f(x) = \lambda \exp\{-\lambda x\},$$

   y que para la construcción de su gráfico es razonable estimar el parámetro λ usando $\bar{X}^{-1}$ donde $\bar{X}$ representa el promedio de `HH_A`.

2. **Evaluación Estadística:** Implemente un test de hipótesis de bondad de ajuste con un nivel de significancia de $\alpha = 0.05$ que le permita concluir sobre la hipótesis planteada en el estudio. Recuerde que el test de bondad de ajuste le permite analizar si los datos se comportan como una distribución en específico.

3.4 Pregunta 3 (regresión lineal simple y múltiple)

1. **Regresión lineal simple:** Implemente un modelo de regresión lineal simple que le permita modelar las *Horas de Contrato de los Docentes de Aula* (variable `HH_A`) en función del *Total de Docentes de aula* (variable `DC_A`).

   Los pasos que deberá realizar para dar respuesta a esta pregunta son:

   (a) Obtenga la correlación entre las variables involucradas.

   (b) Visualize la relación entre las variables

   (c) Separe los datos en conjuntos de datos de entrenamiento y testeo. Luego ajuste el modelo de regresión.

   (d) Muestre en una tabla los coeficientes obtenidos (Intercepto y pendiente). Interprete sus resultados.

   (e) Visualize el modelo de regresión utilizando los coeficientes encontrados en el paso anterior.

   (f) Obtenga y comente el coeficiente de determinación obtenido.

   (g) Evalue su modelo utilizando los datos del set de testeo. Para esto, calcule el MSE y RMSE predictivos (error cuadrático medio y raíz del error cuadrático medio). ¿Qué significan los valores obtenidos?

2. **Regresión lineal múltiple:** En esta pregunta se le pide que implemente un modelo de regresión lineal múltiple para modelar las *Horas de Contrato de los Docentes de Aula* (variable `HH_A`) en función del *Total de Docentes de aula* (variable `DC_A`) y de **tres variables más que usted debe escoger**. Para escoger estas variables usted debe usar criterios claros y objetivos (por ejemplo, puede elegir las variables que más se correlacionan con la variable dependiente).

   Utilizando la misma separación de entrenamiento y testing realizada en el ítem anterior responda: ¿Qué impacto tienen las variable adicionales en el rendimiento de su modelo? Comente en función de los nuevas métricas obtenidas (error cuadrático medio y R2).

3.5 Pregunta 4 (clasificación y validación)

En esta pregunta se le pide que implemente dos modelos de clasificación: Support Vector Machine y Regresión Logística. El objetivo es generar un modelo que prediga si la variable *Horas de Contrato de los Docentes de Aula* (variable `HH_A`) es mayor a 160 horas. Los modelos deben ser implementados usando validación simple y K-Fold Cross validation. Los pasos para responder la pregunta son:

1. Genere una nueva variable a partir de *Horas de Contrato de los Docentes de Aula* (variable `HH_A`) que tome el valor de 1 si supera las 160 horas, y tenga valor 0 si es menor o igual a este valor.
2. Seleccione la variable *Total de Docentes de aula* (variable `DC_A`) y **las tres mejores variables**. Para escoger estas variables usted debe justificar usando criterios claros y objetivos.
3. Genere un nuevo dataframe que contenga: la variable *Total de Docentes de aula*, las 3 variables que escogió y la nueva variable generada.
4. Con el data frame obtenido, aplique el modelo SVM utilizando la técnica de validación simple. Obtenga los valores de las métricas: accuracy, recall, precision y f1 score. Explique cómo interpretaría los resultados de estas métricas.
5. Con el data frame obtenido, aplique el modelo Regresión Logística utilizando la técnica de validación simple. Obtenga los valores de las métricas: accuracy, recall, precision y f1 score. Explique cómo interpretaría los resultados de estas métricas.
6. Mencione la principal deficiencia de aplicar validación simple aplicada para este caso.
7. Con el data frame obtenido, aplique el modelo SVM utilizando la técnica de validación cruzada. Obtenga los valores de las métricas: accuracy, recall, precision y f1 score. Determine si su modelo se encuentra sobre entrenado.
8. Con el data frame obtenido, aplique el modelo Regresión Logística utilizando la técnica de validación cruzada. Obtenga los valores de las métricas: accuracy, recall, precision y f1 score. Determine si su modelo se encuentra sobre entrenado.

4 Entregables

- Un informe grupal que debe contener la respuesta a cada pregunta. Además, el informe debe seguir la siguiente estructura:
    1. **Resumen:** Indica lo que el lector encontrará en el documento, el estudio y conclusiones más importantes.
    2. **Introducción:** Abarca la temática, objetivos y relevancia del estudio, proporcionando una visión concisa de la estructura y estableciendo una base para las secciones subsiguientes.
    3. **Metodología:** Indicar la metodología utilizada, detallando pasos y forma de trabajar.
    4. **Preguntas:** Responder de forma clara las preguntas escogidas. Es fundamental responder de manera clara a las preguntas seleccionadas. **Las respuestas deben ser exhaustivas, detallando los pasos seguidos y el análisis realizado para llegar a la conclusión final.**
    5. **Conclusión:** En este punto debe indicar las principales conclusiones del estudio.
- Un archivo ipynb grupal que respalde sus respuestas. Debe incluir el análisis que realizó para llegar a las conclusiones de su informe. El archivo debe estar ordenado, recuerde que puede añadir texto en formato markdown para facilitar la comprensión.
- Diapositivas y un video de 8 a 10 minutos, donde presente su trabajo.

5 Evaluación de Pares

Nota entre 1 y 100 que cada alumno asigna a sus compañeros de grupos y a sí mismo de acuerdo al aporte realizado por el mismo en el trabajo. El promedio de estas calificaciones será el factor por el cual se multiplica la nota obtenida por el grupo. Si un alumno no realiza esta evaluación será calificado en el proyecto con la nota mínima.