



Tarea de Clasificación: Predicción del nivel salarial según género y características laborales

Identificación

Asignatura: Ciencia de Datos (Electiva)
Curso: TQ

Objetivos

Construir un modelo de clasificación supervisada para predecir el nivel salarial mensual de personas empleadas, utilizando como variables predictoras el género, la educación, la ocupación y otras características del trabajo.

Variables sugeridas

Variable objetivo (y):

- nivel_salarial (categórica: *bajo, medio, alto*).
Se obtiene a partir de la variable continua del salario neto, dividiendo en terciles o cuartiles:
`df['nivel_salarial'] = pd.qcut(df['salario_neto'], q=3, labels=['bajo','medio','alto'])`

Variables predictoras (X):

- genero
- edad
- nivel_educativo
- ocupacion o rama_actividad
- horas_trabajadas
- zona (urbana/rural)
- tipo_empleo (público, privado, independiente)
- antigüedad o experiencia (si existe)
- departamento



Preparación de los Datos

1. Filtrar personas con empleo y salario declarado.
2. Seleccionar las variables relevantes.
3. Limpiar y tratar valores faltantes.
4. Codificar las variables categóricas (*One-Hot Encoding* o *Label Encoding*).
5. Crear la variable objetivo nivel_salarial.
6. Dividir en *train* (70 %) y *test* (30 %).

Modelos a implementar

Pueden probar al menos **dos clasificadores**, por ejemplo:

- Regresión logística multinomial (modelo base).
- Árbol de decisión o Random Forest (modelo no lineal).
- (Opcional) Gradient Boosting o SVM para comparación.

Evaluación de los modelos

- Reportar **accuracy, precision, recall, F1-score**.
- Matriz de confusión.
- Gráfico de importancia de variables.
Análisis interpretativo: ¿Qué peso tiene el sexo frente a la educación o la ocupación en la predicción del salario?

Interpretación y Discusión

- Comparar los hallazgos del modelo con la prueba t inicial:
¿El género sigue siendo un factor que influye significativamente en el nivel salarial?
- Discutir la **brecha salarial** desde una perspectiva de datos y equidad.
- Identificar otras variables que también influyen en los ingresos.
- Reflexionar sobre limitaciones y consideraciones éticas (evitar reforzar sesgos).

Entregable

- Notebook con el código comentado.
- Gráficos (matriz de confusión, importancia de variables).
- Conclusión que responda:
 - ¿Qué tan bien predice el modelo?
 - ¿Qué factores parecen más determinantes?