Universidad de Valparaíso
Facultad de Ingeniería
Escuela de Ingeniería Informática

BASE DE DATOS III

Tarea N°2 (20%)

Prof: Ana Aguilera Faraco
Ayudante: Fernanda Fuentes
fernanda.fuentesp@estudiantes.uv.cl
Octubre 2025

Introducción:

En el análisis de bases de datos, las técnicas de clasificación permiten asignar etiquetas a registros en función de sus características, facilitando la predicción de comportamientos futuros. En esta tarea, utilizaremos un dataset de establecimientos públicos hospitalarios de Chile para explorar cómo los modelos de clasificación pueden ayudarnos a entender mejor los factores asociados a la eficiencia hospitalaria, permitiendo anticipar qué hospitales presentan un desempeño más óptimo en función de variables.

Objetivo de la tarea:

El objetivo de esta tarea es construir modelos de clasificación para predecir si un hospital es **eficiente** o **no** basándose en sus características. Para ello, se utilizarán diferentes algoritmos de clasificación, comparando su rendimiento y seleccionando el modelo más adecuado para realizar predicciones precisas sobre la eficiencia hospitalaria. El dataset utilizado para esta tarea contiene información sobre indicadores clave del proceso de hospitalización, tales como camas ocupadas y disponibles, traslados, egresos, días de estadía, letalidad, entre otros.

Obtención y preparación del dataset

El dataset de establecimientos hospitalarios de Chile está disponible en el portal de datos abiertos del gobierno chileno.

- Fuente oficial: Indicadores de Hospitalización – Portal de Datos Abiertos

Nota importante: El dataset original **NO incluye** la etiqueta EFICIENCIA precalculada. Deberá ser creada como parte fundamental de esta tarea.

a. Creación de la variable objetivo (EFICIENCIA)

Antes de realizar el análisis de clasificación, se debe **definir y calcular** la variable objetivo EFICIENCIA para el dataset. Para ello se sugiere revisar en literatura médica, estándares del Minsal u otros criterios técnicos para:

- Justificar claramente su elección.
- Citar fuentes si corresponde.
- Demostrar que su definición es coherente con el contexto hospitalario.

La variable eficiencia tendrá un valor binario como (EFICIENCIA = 1) (NO EFICIENTE=0)

b. Validación de la variable EFICIENCIA

Después de crear la variable EFICIENCIA, se debe:

1. **Analizar la distribución de clases:**

- ¿Cuántos hospitales son eficientes vs no eficientes?

2. **Evaluar el balanceo:**
   - Si hay desbalance mayor a 70-30, considera aplicar técnicas de balanceo.
   - Opciones: SMOTE, Random Undersampling, Random Oversampling, etc.
3. **Documentar decisiones:**
   - Explica cualquier técnica de balanceo aplicada.
   - Justifica por qué la aplicaste o por qué no fue necesaria.

c. Columnas del dataset:

Al trabajar con el dataset de establecimientos hospitalarios de Chile, las siguientes columnas son clave para análisis de clasificación:

- Variable objetivo (target):
  - **EFICIENCIA**: Esta columna indica si un hospital es eficiente o no (1 = Sí, 0 = No).
- Variables independientes (características):
  - **TIPO_PERTENENCIA**: Código numérico que identifica la pertenencia del establecimiento (por ejemplo, público, privado).
  - **GLOSA_SSS**: Nombre del Servicio de Salud (por ejemplo: "Servicio de Salud Metropolitano Norte").
  - **PERIODO**: Año del registro .
  - **ESTABLECIMIENTO**: Nombre del establecimiento.
  - **AREA_FUNCIONAL**: Nombre del área funcional (como “Medicina”,”Urgencia”, etc).
  - **DIAS_CAMAS_OCUPADAS:** Total de las camas ocupadas durante el periodo.
  - **DIAS_CAMAS_DISPONIBLES**: Total de días que las camas estuvieron disponibles durante el periodo.
  - **DIAS_ESTADA:** Suma de los días de estadía de todos los pacientes hospitalarios durante el periodo.
  - **NUMERO_EGRESOS:** Total de pacientes que egresaron del hospital.
  - **MES: Mes** que se realizó el registro.
  - **EGRESO_FALLECIDOS:** Número de pacientes que fallecieron durante hospitalización.
  - **TRASLADOS:** Cantidad de egresos que corresponden a pacientes trasladados a otro centro.
  - **INDICE_OCUPACIONAL:** Proporción de camas ocupadas respecto a las disponibles.
  - **PROMEDIO_CAMAS_DISPONIBLES:** Promedio de camas disponibles.
  - **PROMEDIO_DIAS_ESTADA:** Días que un paciente permanece hospitalizado.
  - **LETALIDAD:** Porcentaje de fallecidos respecto al total de egresos.
  - **INDICE_ROTACION (**p: Número promedio de egresos por cama durante el periodo.
  - **COD_SSS:** Código numérico que identifica al Servicio de Salud al que pertenece un establecimiento hospitalario.
  - **CODIGO_ESTABLECIMIENTO:** Código único que identifica a cada

establecimiento de salud dentro del sistema.

- **COD_AREA_FUNCIONAL:** Código numérico que corresponde al área funcional del hospital o centro de salud.
- **OCUPACION:** Este indicador mide la proporción de camas ocupadas en relación con las camas disponibles en un hospital durante un período determinado.
- **PROMEDIO_ESTADIA:** El promedio de estadía es el tiempo promedio que los pacientes permanecen en el hospital durante un período específico.

Instrucciones generales:

- La tarea N°2 es grupal y en caso de copia se aplicarán las sanciones correspondientes.
- Utilice todos los recursos que ofrece Python para realizar un trabajo completo (librerías).
- Puntaje total: 100 puntos. Nota 4,0: 60 puntos.

Recomendaciones:

Recuerde aplicar las métricas estadísticas vistas en clase, como las medidas de tendencia central (media, mediana), las medidas de dispersión (varianza, desviación estándar) y las técnicas de reducción de dimensionalidad. También es importante que los resultados sean acompañados por gráficos que faciliten la comprensión de los datos.

Trabajo a realizar:

1. **Conjunto de datos**
   a. **Descripción utilizando métodos estadísticos:** Debe proporcionar una descripción estadística detallada del dataset, identificando claramente todas las variables que contiene, especificando el tipo de dato asociado a cada una (por ejemplo, numérico, categórico, texto, etc.) y destacando la relevancia de cada variable en el contexto del análisis. Además, se debe calcular y analizar medidas descriptivas como la media, mediana, moda, desviación estándar, percentiles, entre otros, para todas las variables numéricas. Es importante identificar y describir posibles outliers (valores atípicos) y analizar su impacto en el conjunto de datos.

   b. **Visualización de los datos:** Incluir gráficos para visualizar la distribución de las variables y su relación con la variable objetivo (EFICIENCIA). Esto incluye histogramas para observar la distribución de las variables numéricas, gráficos de caja (boxplot) para detectar valores atípicos y gráficos de dispersión para evaluar relaciones entre variables. Se debe prestar especial atención a la visualización de los outliers y cómo estos afectan las distribuciones.

   c. **Exploración, limpieza y transformación de datos:**
      i. Realiza una exploración inicial del dataset enfocándose en la carga y visualización de los datos, observando las primeras filas para

entender su estructura. Luego, genera una descripción estadística de las variables numéricas para identificar patrones y posibles valores anómalos.

ii. Maneja adecuadamente los valores anómalos o nulos en el conjunto de datos, creando un subconjunto que contenga solo las filas con datos significativos. Luego, vuelve a calcular las métricas estadísticas utilizando este nuevo conjunto, que servirá como base para los análisis y tareas posteriores.

iii. Haga las transformaciones de los datos que considere necesarias para facilitar el análisis posterior, esto es, transformación de datos categóricos, estandarizaciones, etc.

2. **Clasificación y análisis**

   a. **Aplicación de algoritmos de clasificación:**

      i. Aplica tres algoritmos de clasificación diferentes y asegúrese de utilizar al menos un meta-algoritmo en los modelos. Algunos ejemplos de algoritmos de clasificación incluyen:

         - KNeighborsClassifier (K-Vecinos más cercanos)
         - AdaBoostClassifier (Meta-algoritmo)
         - RandomForestClassifier (Meta-algoritmo)
         - LogisticRegression (Regresión logística)
         - GaussianNB (Naive Bayes Gaussiano)
         - GradientBoostingClassifier (Meta-algoritmo)
         - DecisionTreeClassifier (Árbol de decisión)

   b. **Selección de modelos óptimos:** Para seleccionar los modelos más óptimos, utiliza el análisis de la curva ROC (Receiver Operating Characteristic), que permitirá comparar el rendimiento de los modelos en términos de la tasa de falsos positivos y verdaderos positivos.

   c. **Ajuste de hiperparámetros:** Realizar ajuste de hiperparámetros (grid search y random search) en los clasificadores para mejorar su rendimiento. Explicar el impacto de los ajustes realizados.

   d. **Entrenamiento y validación de modelos:**

      i. Crear, entrenar y validar tres modelos de clasificación para predecir si un hospital es eficiente o no.

      ii. Dividir los datos en conjuntos de entrenamiento y prueba, y valida los resultados de los modelos utilizando métricas de evaluación como precisión, recall, F1-score, y la curva ROC.

   e. **Evaluación de modelos:**

      i. Aplicar las curvas ROC para evaluar el rendimiento de los modelos.

ii. Presentar un reporte de clasificación, que incluya las métricas de evaluación como precisión, recall, F1-score, y matriz de confusión para cada modelo.

f. Predicciones:

i. Realizar predicciones, puede utilizar el modelo con mejor rendimiento (mejor score).
ii. Analizar los resultados obtenidos y cómo los diferentes factores (Días camas ocupadas, número de egresos, etc) influyen en la eficiencia (EFICIENCIA) entre hospitales.

g. Preguntas

i. ¿Qué conclusiones puedes obtener sobre los métodos utilizados?
ii. ¿Cuál recomendarías utilizar y por qué?
iii. ¿Qué algoritmo de clasificación mostró el mejor rendimiento según las métricas utilizadas (precisión, recall, F1-score)?
iv. ¿Cómo afectó el ajuste de hiperparámetros al rendimiento de los modelos?
v. ¿Qué ventaja se observa en el uso de algoritmos de clasificación para predecir la compatibilidad entre los participantes frente a los métodos tradicionales?

3. Selección y reducción de variables

i. Aplique 1 método de selección y 1 método de reducción de variables.
ii. Evalúe el impacto de aplicar estos métodos en los algoritmos de clasificación utilizados en la etapa anterior.

4. Conclusiones

a. Comparar los resultados obtenidos entre los diferentes algoritmos de clasificación aplicados en el análisis de la compatibilidad entre los participantes.
b. Discutir qué tan efectivos fueron los modelos de clasificación para predecir las decisiones en comparación con las predicciones esperadas.
c. Reflexiona sobre qué enfoques resultaron más adecuados en función del contexto del análisis y considerar cómo podrían mejorar las predicciones si se incorporan más variables o diferentes técnicas.

Entrega:

- El nombre del archivo debe ser “T2-NombreApellido.ipynb”, cada NombreApellido debe ir separado con un guión “-”
- La tarea debe ser enviada por medio del aula virtual hasta el 22 octubre del 2025.

Rúbrica:

Presenta	Aspectos a evaluar	No aplica (0%)	Deficiente (30%)	Regular (60%)	Bueno (80%)	Destacado (100%)	Puntaje máximo del ítem
Calidad en la presentación del cuaderno	• Orden de código (permite al corrida secuencial del cuaderno) • Buena ortografía y redacción (comentarios claros y entendibles) • Documentación (si utiliza librerías, debe especificarlas)	No incluye estos aspectos.	Solo incluye comentarios y el orden de código no es adecuado.	2 aspectos no están presentes o no están detallados.	Todos los aspectos están claros y detallados	Todos los aspectos están claros y detallados.	**5**
Definición y Creación de la Etiqueta de Eficiencia	• Se creó la variable EFICIENCIA con base en umbrale razonables. • Se definieron y calcularon las métricas relevantes. • Proponen definición de eficiencia basándose en literatura médica, estándares del Minsal o criterios técnicos.	No incluye estos aspectos.	Solo incluye comentarios y el orden de código no es adecuado.	2 aspectos no están presentes o no están detallados.	Todos los aspectos están claros y detallados	Todos los aspectos están claros y detallados	**10**
Descripción de dataset utilizando métodos estadísticos	• Explica el dataset (objetivo y mapa general) • Describe las variables • Obtiene información del dataset (variables, tamaño, etc.)	No incluye estos aspectos.	No se detalla bien o no cumple con al menos 2 aspectos.	2 aspectos no están presentes o no están detallados.	1 aspecto no queda claro o no está presente.	Todos los aspectos están claros y detallados.	**10**
Primer análisis de Clasificación	• Realiza el análisis de clasificación • Obtiene resultados claros • Incluye métricas y temas planteados. • Estudio de hiperparámetros	No incluye estos aspectos.	No queda claro el análisis o solo incluye 1 aspecto.	2 aspectos no están presentes o no están detallados.	1 aspecto no queda claro o no está presente.	Todos los aspectos están claros y detallados.	**12**
Segundo análisis de Clasificación	• Realiza el análisis de clasificación • Obtiene resultados claros • Incluye métricas y temas planteados • Estudio de hiperparámetros	No incluye estos aspectos.	No queda claro el análisis o solo incluye 1 aspecto.	2 aspectos no están presentes o no están detallados.	1 aspecto no queda claro o no está presente.	Todos los aspectos están claros y detallados.	**12**
Tercer análisis de Clasificación	• Realiza el análisis de clasificación • Obtiene resultados claros • Incluye métricas y temas planteados • Estudio de hiperparámetros	No incluye estos aspectos.	No queda claro el análisis o solo incluye 1 aspecto.	2 aspectos no están presentes o no están detallados.	1 aspecto no queda claro o no está presente.	Todos los aspectos están claros y detallados.	**12**
Estudio de selección y reducción de variables	• Incluye 1 método de selección • Incluye un método de reducción • Ejecuta los clasificadores • Evalúa y discute el impacto	No incluye estos aspectos.	No queda claro el análisis o solo incluye 1 aspecto.	2 aspectos no están presentes o no están detallados.	1 aspecto no queda claro o no está presente.	Todos los aspectos están claros y detallados.	**9**

Preguntas (discusión de resultados)	• Responde a todas las preguntas de manera clara	No incluye estos aspectos.	No responde en su totalidad las preguntas y las respuestas son muy deficientes.	Responde parcialmente las preguntas con conclusiones deficientes.	Responde parcialmente las preguntas con o con conclusiones deficientes.	Responde todas las preguntas con claridad y detalle.	**10**
Conclusiones	• Detalla claramente sus conclusiones • Domina el tema de clasificación y regresión	No incluye estos aspectos.	No responde en su totalidad la conclusión, siendo una respuesta superficial, sin un análisis profundo.	Responde de forma superficial sin un análisis profundo.	Responde parcialmente la conclusión, pero con respuestas que reflejan una buena comprensión	Las conclusione s están claras, bien detalladas y reflejan una buena comprensió n.	**10**
Predicciones	• Presenta predicciones • Utiliza bien las etiquetas y obtiene y_pred como resultado	No incluye estos aspectos.	No queda claro	2 aspectos no están presentes o no están detallados.	1 aspecto no queda claro o no está presente.	Todos los aspectos están claros y detallados.	**10**