Examen II.II - Herramientas de Datos I - I25

Universidad de Costa Rica
Material del curso CA0204 II-2025
Examen I: Parte II
Encargado:

- Potoy Juárez Luis - luis.juarez@ucr.ac.cr
- Joshua Cervantes - joshua.cervantes@ucr.ac.cr

⚠ Indicaciones

1. El examen debe llevar el nombre: Examen_I_GX - Carnet - NombreApellidos. Donde X equivale al grupo 1 o 2.
2. El examen es de forma individual.
3. Puede usar todo material visto en clases, tareas y material de apoyo virtual. Todo material no visto en clases debe ser citado.
4. Duración 24 horas para entregar la prueba a partir del sábado 29 de noviembre 2025 a las 7:00 pm. Toda prueba entregada después de la fecha establecida debe ser por correo al profesor encargado y se califica con base a 70% de la nota final.
5. Valor de la prueba es de 40% de la nota final del segundo parcial.
6. Al no cumplir con el punto 2 anular la prueba.
7. No se permite el uso de material de apoyo como puede inteligencia artificial generativa, stack overflow, consulta a otras personas, uso de celular, etc.
8. Entregue un zip donde se encuentre el Excel, las carpetas usadas para R y los git. Considerar punto 1.

Ejercicio 1

Contexto.

Para este trabajo use el archivo `moras.txt` el cual contiene información de asegurados en este caso se ubican las siguientes variables:

Aprendiendo de créditos

- Identificación: Identificador único del asegurado.

- Sexo: Indica si es hombre o mujer.
- Salario: Salario bruto del asegurado.
- TIPO_ASEGURAMIENTO: El tipo de seguro que tiene el asegurado.
- SECTOR: El sector al que pertenece la persona.
- INDICADOR_ACTIVO: Indica si se encuentra activo el seguro.
- INDICADOR.EXTRANJERO: Indica si el cliente es extranjero o no.
- INDICADOR.MOROSO: Indica si el cliente es moroso o no.
- EDAD: Edad del asegurado.

El objetivo será identificar si existen diferencias para poder identificar si una persona tiene más probabilidades de ser morosa o no. Para ello se va hacer uso de dos herramientas: Excel y R.
Con la tabla de datos realice lo siguiente en Excel y sin hacer uso de tablas dinámicas:

1. (Excel) Cargue el conjunto de datos haciendo uso de Power Query de Excel. Adjunte la captura de pantalla de la carga en el Excel.
2. (Excel) Realice un resumen de 5 números incluyendo el promedio. Para las columnas numéricas
3. (Excel) Realice una columna donde se identifiquen valores atípicos haciendo uso del Z-Score, de tal manera que si el valor absoluto del Z-Score es mayor a 1,96 se considera un valor atípico. Esto para la columna de salario y edad.
4. (Excel) Detecte el porcentaje de valores faltantes en cada una de las columnas.
5. (Excel) Impute los valores faltantes haciendo uso de algún método de su preferencia.
6. (Excel) Realice gráficos donde se vea la cantidad de individuos por cada categoría. Analice los resultados.
7. (Excel) Realice un gráfico para cada una de las variables categóricas en donde se vea el porcentaje de personas morosa por cada clase. Analice los resultados.
8. (Excel) Para las variables numéricas realice gráficos de barras donde se ponga un color de acuerdo a si es moroso o no. Analice los resultados.
9. (R) Replique lo anterior, items 2-8, en el lenguaje de programación R y haga uso de Git, agregue un `.gitignore` conveniente. Cada vez que termine un punto realice un commit. El commit final debe ser el punto que se muestra a continuación. Usar commits convencionales.
10. Se busca encontrar cuáles son las variables que podrían tener mayor relación con que la persona sea morosa o no, para ello se decide realizar un árbol de decisión por su interpretabilidad y su fácil explicación a los altos mandos. A continuación se adjunta un código ejemplo de cómo se construye un árbol y cómo se muestra

```
#| eval: false
library(rpart)
library(rpart.plot)
tree <- rpart(y~., data = df)
rpart.plot(tree)
```

En este caso la variable `y` es la variable que se quiere explicar con las otras y `.` indica que se usen todas las variables, recuerde eliminar el identificador único del `df`.
Podría requerir codificar de una manera diferente algunas clases en las variables categóricas, esto para una mejor visualización.
Explique los resultados del árbol para alguien que no tiene tanta formación cuantitativa y busca tomar decisiones con base en sus resultados.