Universidad Católica San Pablo

Informe del laboratorio 1

Computer Science — Tópicos en Inteligencia Artificial

Harold Alejandro Villanueva Borda

---

1. Introducción

Este informe presenta el análisis del dataset escogido para realizar la práctica de laboratorio, en donde se explican diferentes conceptos como ruido en los datos, normalización y correlación. El conjunto de datos utilizado es *Credit Card Fraud Detection Dataset 2023*, extraído del repositorio Kaggle [3].

2. Dataset

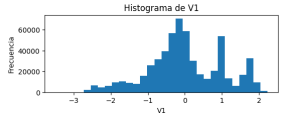
Este conjunto de datos contiene transacciones con tarjeta de crédito realizadas por titulares de tarjetas europeos en el año 2023. Comprende más de 550.000 registros

2.1. Características Clave

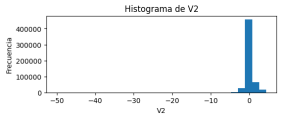
- id: Identificador único para cada transacción
- V1-V28: Características anónimas que representan varios atributos de transacción (por ejemplo, hora, ubicación, etc.)
- Cantidad: El monto de la transacción
- Clase: Etiqueta binaria que indica si la transacción es fraudulenta (1) o no (0)

3. Análisis de datos

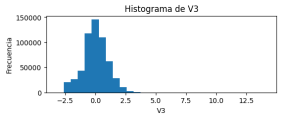
En este análisis se evaluará la cantidad de elementos, máximo, mínimo, promedio y una gráfica para observar el comportamiento de los mismos, este paso se realizara con cada columan del dataset.



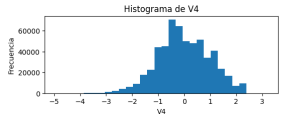
1: Gráfico 1



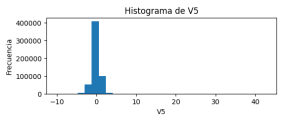
2: Gráfico 2



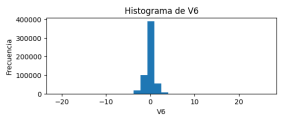
3: Gráfico 3



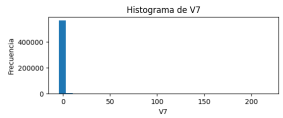
4: Gráfico 4



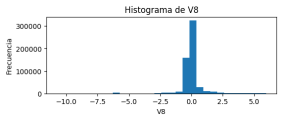
5: Gráfico 5



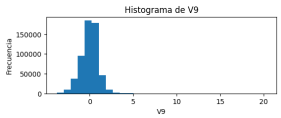
6: Gráfico 6



7: Gráfico 7



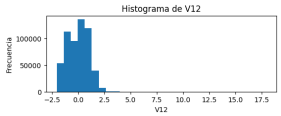
8: Gráfico 8



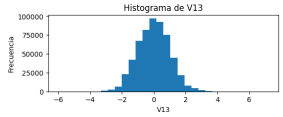
9: Gráfico 9

10: Gráfico 10

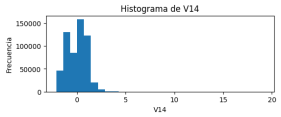
11: Gráfico 11



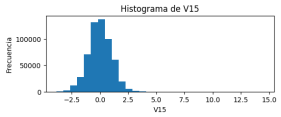
12: Gráfico 12



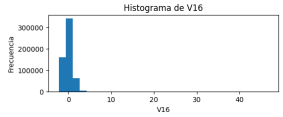
13: Gráfico 13



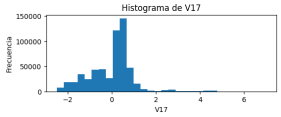
14: Gráfico 14



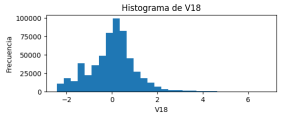
15: Gráfico 15



16: Gráfico 16

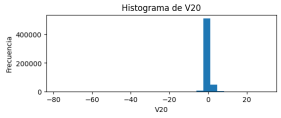


17: Gráfico 17

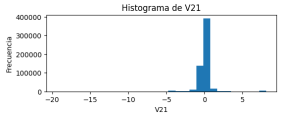


18: Gráfico 18

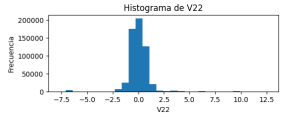
19: Gráfico 19



20: Gráfico 20

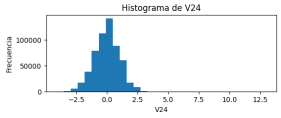


21: Gráfico 21

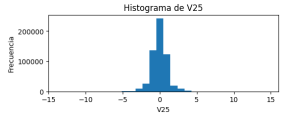


22: Gráfico 22

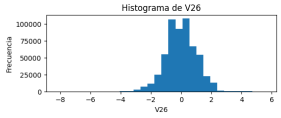
23: Gráfico 23



24: Gráfico 24



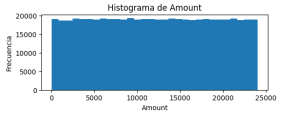
25: Gráfico 25



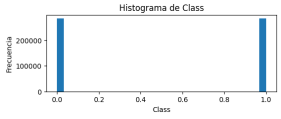
26: Gráfico 26

27: Gráfico 27

28: Gráfico 28



29: Gráfico 29



30: Gráfico 30

Figura 1: Gráfica de cada una de las columnas

Variable	Cantidad de Elementos	Máximo	Mínimo	Promedio
V1	568,630.0	2.22904613004356	-3.495583516386668	-5.638057829783114e-17
V2	568,630.0	4.361865196721416	-49.96657153869079	-1.3195454495237075e-16
V3	568,630.0	14.125833911866232	-3.1837603416948093	-3.518787865396553e-17
V4	568,630.0	3.201535546069201	-4.951222429093022	-2.879008253506271e-17
V5	568,630.0	42.716890639914205	-9.952785617741023	7.99724514862853e-18
V6	568,630.0	26.168402294404643	-21.11110792759147	-3.958636348571122e-17
V7	568,630.0	217.873038474627	-4.351839315074907	-3.198898059451412e-17
V8	568,630.0	5.958040147327273	-10.75634230545734	2.1092734079507747e-17
V9	568,630.0	20.270062075837107	-3.751918738076145	3.998622574314265e-17
V10	568,630.0	31.72270910795672	-3.163275761885778	1.9913140420085041e-16
V11	568,630.0	2.5135727491214537	-5.954723293982779	-1.1835922819970225e-16
V12	568,630.0	17.913556111364983	-2.020399318328518	-5.758016507012542e-17
V13	568,630.0	7.187485954748435	-5.955226700040189	-5.698037168397828e-18
V14	568,630.0	19.169544406102982	-2.1074168038580363	-4.07859502580055e-17
V15	568,630.0	14.532202180325108	-3.8618127653411154	2.6490874554832007e-17
V16	568,630.0	46.6529060440468	-2.214512888665661	-1.719407706955134e-17
V17	568,630.0	6.994124024684426	-2.484938386554947	-3.398829188167125e-17
V18	568,630.0	6.783716009168727	-2.4219487219161318	-5.837988958498827e-17
V19	568,630.0	3.8316716979071006	-7.80498794807604	2.4791459960748445e-17
V20	568,630.0	29.872812160323736	-78.14783856605457	-1.5794559168541347e-17
V21	568,630.0	8.087080028016498	-19.382523087206284	4.758360863433976e-17
V22	568,630.0	12.63251122579015	-7.734798174224937	3.948639792135337e-18
V23	568,630.0	31.707626578253517	-30.295450154840687	6.194741066300928e-18
V24	568,630.0	12.965638661146754	-4.067967795102357	-2.7990358020199858e-18
V25	568,630.0	14.621509105774306	-13.612633178980907	-3.178904946579841e-17
V26	568,630.0	5.623285408193404	-8.226969338778424	-7.497417326839247e-18
V27	568,630.0	113.23109281324442	-10.498633077959608	-3.5987603168828385e-17
V28	568,630.0	77.25593674214326	-39.03524318743944	2.609101229740058e-17
Amount	568,630.0	24,039.93	50.01	12,041.957634577848
Class	568,630.0	1.0	0.0	0.5

Cuadro 1: Estadísticas de las variables

- **Cantidad de Elementos**: Todas las variables tienen 568,630 elementos, indicando un tamaño uniforme del conjunto de datos.
- **Rango de Valores**: Las variables tienen rangos variados. 'V7' presenta el rango más amplio, mientras que 'V1' tiene un rango más estrecho. 'Amount' muestra un rango mucho mayor que otras variables.
- **Promedio Cercano a Cero**: Muchas variables tienen promedios cercanos a cero, sugiriendo datos distribuidos simétricamente alrededor de un valor central muy pequeño.
- **Variable 'Class'**: Tiene un promedio de 0.5, indicando una distribución equilibrada entre dos categorías (0 y 1).
- **Datos Atípicos**: La presencia de valores extremos en algunas variables sugiere posibles datos atípicos o distribuciones no simétricas.

4. Limpieza de datos

Respecto al conjunto de datos utilizado, todas las columnas y filas contienen un valor, por lo que se afirma que el si es un conjunto de datos completo sin celdas vacías.

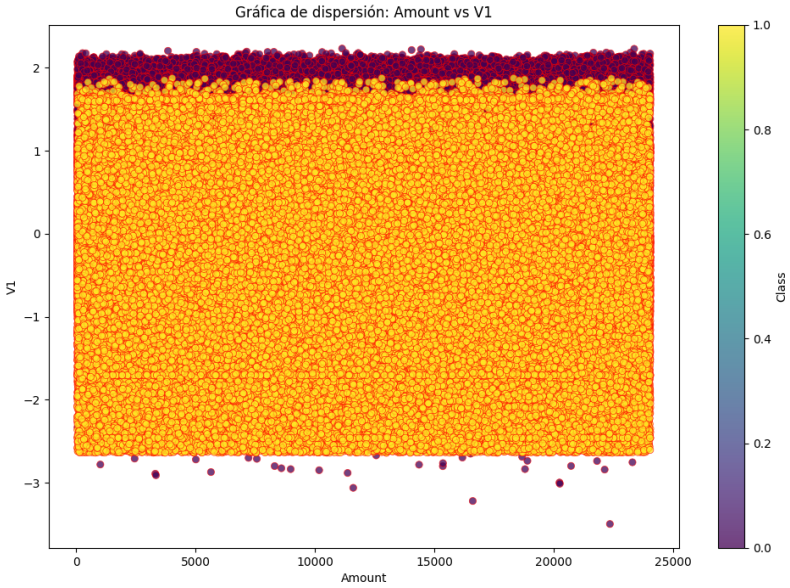


Figura 2: Gráfica de dispersión para la verificación del ruido dentro de los datos

4.1. Distribución de los datos

- La mayoría de los datos están concentrados en la parte superior de la gráfica, con valores de "V1çercanos a 2 y valores de .ªmount"que cubren un amplio rango.
- Existe una clara concentración de puntos en la región superior, que parece estar mejor definida y agrupada, lo que podría indicar la presencia de un patrón en los datos.

4.2. Posible ruido

- En la parte inferior de la gráfica, se observan algunos puntos dispersos con valores de "V1ïnferiores a -2, que parecen no seguir el mismo patrón que la mayoría de los datos.
- Estos puntos dispersos en la parte inferior podrían considerarse como ruido o outliers (valores atípicos), ya que no siguen el patrón predominante.

La gráfica sugiere que la mayoría de los datos están bien agrupados en un rango definido de V1 y Amount. Sin embargo, los puntos dispersos en la parte inferior podrían estar afectando el análisis general y podrían requerir una inspección adicional para determinar si son outliers genuinos, errores de medición o simplemente variaciones normales.

5. Normalización de los datos

La normalización de datos es crucial porque ayuda a que los algoritmos de aprendizaje automático funcionen de manera más eficiente y precisa. Normalizar los datos asegura que todas las características contribuyan de manera equitativa al modelo, evitando que las características con valores más grandes dominen el proceso de aprendizaje [2].

5.1. Técnicas para la normalización de datos

- Escalado Min-Max: Transforma los datos para que estén dentro de un rango específico, generalmente entre 0 y 1.
- Estandarización (Z-score): Transforma los datos para que tengan una media de 0 y una desviación estándar de 1.
- Normalización robusta: Utiliza la mediana y el rango intercuartílico para escalar los datos, útil para datos con outliers [2].

En ente trabajo se ha normalizado los datos usando la técnica Min-Max.

6. Matriz de correlación

La matriz de correlación es una tabla que muestra los coeficientes de correlación entre múltiples variables. Cada celda en la tabla muestra la correlación entre dos variables. Esta matriz sirve para identificar relaciones lineales entre variables, lo que puede ayudar a entender la estructura de los datos y a seleccionar características relevantes para modelos de aprendizaje automático [1].

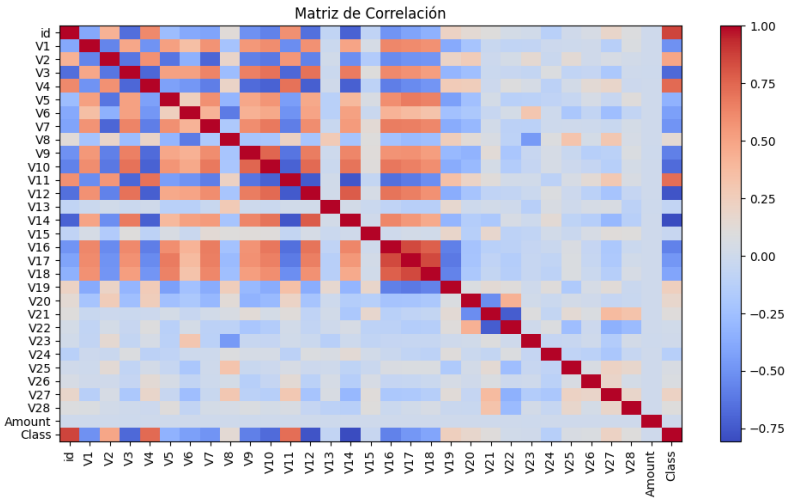


Figura 3: Gráfico de la matriz de correlación

- Correlaciones Fuertes: Algunas variables como V17 y V18 están fuertemente correlacionadas entre sí, lo que podría indicar colinealidad.
- Baja Correlación con Çlass": La mayoría de las variables no muestran una fuerte correlación con la variable objetivo Çlass", lo que sugiere que ninguna variable individual tiene un impacto dominante en la predicción.
- Relevancia en Modelado: Las fuertes correlaciones entre variables pueden redundar en el modelo, y las variables con baja correlación con Çlass"podrían ser menos relevantes en el análisis predictivo.

7. Conclusiones

- El conjunto de datos analizado es exhaustivo y no presenta valores faltantes, sin embargo, se identificaron valores atípicos que podrían influir negativamente en la precisión del análisis.
- La normalización de los datos, implementada mediante la técnica Min-Max, resultó esencial para garantizar que todas las características contribuyan equitativamente en los procesos de modelado y análisis.
- La matriz de correlación reveló la presencia de colinealidad entre ciertas variables, mientras que la mayoría mostró una baja correlación con la variable objetivo. Este hallazgo sugiere que algunas características podrían tener una influencia limitada en la efectividad del modelo predictivo.

8. Repositorio

La implementaci+on se encuentra en el siguiente respositorio: GitHub.

Referencias

[1] Edutin Academy. *Curso de Inteligencia Artificial*. Accessed: 2024-08-30. 2024. URL: https://edutin.com/curso-de-inteligencia-artificial.

[2] Xataka. *Cursos Gratis de Inteligencia Artificial: 32 cursos online para aprender y dominar la IA y explotar todas sus posibilidades*. Accessed: 2024-08-30. 2024. URL: https://www.xataka.com/basics/cursos-gratis-inteligencia-artificial-32-cursos-online-para-aprender-dominar-ia-explotar-todas-sus-posibilidades.

[3] Nelgiriya Yewithana. *Credit Card Fraud Detection Dataset 2023*. Accessed: 2024-08-30. 2023. URL: https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023.