UNIVERSIDADE D COIMBRA

deer.uc
DEPARTAMENTO DE ENGENHARIA ELETROTÉCNICA E DE COMPUTADORES

Licenciatura em Engenharia Eletrotécnica e de Computadores

Aprendizagem Computacional
Desafio 2023/2024

Trabalho submetido por:
Gabriel Gabriel, 2021216541
Pedro Fernandes, 2021215490
Turma PL3

16 de Março de 2024

1 Processamento de dados

O processamento de dados é uma etapa crucial no desenvolvimento de modelos de aprendizado de máquina. Envolve a preparação e transformação dos dados brutos em um formato adequado para análise. Este capítulo abordará as técnicas e práticas comuns utilizadas para limpar, normalizar e manipular dados, garantindo que estejam prontos para serem utilizados em modelos de aprendizado de máquina. A qualidade do processamento de dados tem um impacto significativo na precisão e desempenho dos modelos, tornando esta fase fundamental para o sucesso de qualquer projeto de ciência de dados.

1. Verificamos a percentagem de valores iguais em várias colunas para remover aquelas que não iriam variar muito a coluna "Accident_severity".
2. Utilizamos a moda para as colunas que tinham valores nulos e substituímos pelo respetivo valor.
3. Analisámos o dataset para remover os duplicados.
4. Inicialmente removemos as linhas que tinham mais que 3 valores nulos, mas verificamos que alterava muito o dataset, então revertimos essa decisão.
5. Fizemos um check de "outlier values", no entanto, não foram identificados resultados anormais, portanto decidimos manter os mesmos.
6. Para a classificação binária, definimos 0 para "Serious_Injury" e "Fatal_Injury", e 1 para "Slight_Injury".
7. Para a classificação multiclasse, atribuimos o valor 0 para "Slight_Injury", 1 para "Serious_Injury" e 2 para "Fatal_Injury".
8. Convertimos a coluna "Time" para "Hour" com o objetivo de analisar em que horas ocorreram mais acidentes e convertemos essa variável para int64.
9. Separámos as variáveis do tipo object das do tipo int64 para facilitar a análise do dataset.

2 Análise da dados

2.1 Classificação binária

Nesta matriz correlação podemos ver a relação de inúmeras variáveis relacionadas com os acidentes do nosso dataset, onde a nossa variável "y" é "Accident_severity". E com ela podemos ver de uma maneira geral, variáveis que se relacionam muito com "Accident_severity":

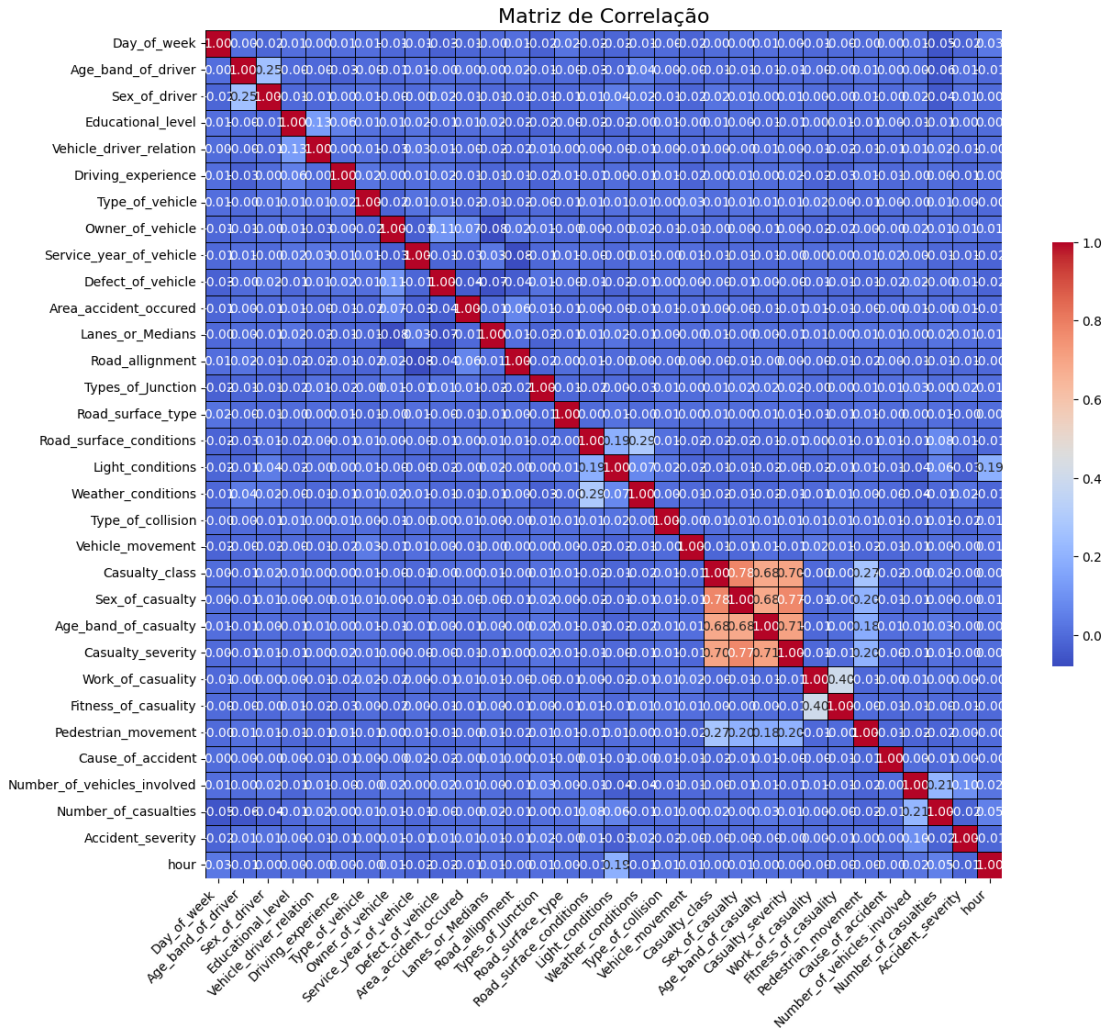

Figure 1: Matriz confusão

1. Existe uma correlação positiva relativamente forte (˜0.78) entre o número de veículos envolvidos e o número de vítimas. Isso é intuitivo, pois mais veículos envolvidos podem levar a mais vítimas. Correlação entre "Vehicle_movement" e "Type_of_collision":

2. Há uma correlação moderada (˜0.66) entre o movimento do veículo e o tipo de colisão, o que faz sentido, já que a maneira como um veículo está se movendo pode influenciar o tipo de colisão que ocorre. Correlação entre "Casualty_severity" e "Age_band_of_casualty":

3. Uma correlação de (˜0.62) sugere que a gravidade das vítimas pode estar relacionada à faixa etária das vítimas, o que é esperado, pois a idade pode influenciar a gravidade dos ferimentos em acidentes. Outras Correlações Notáveis "Weather_conditions" e "Light_conditions":

4. A correlação (˜0.60) entre as condições climáticas e as condições de iluminação indica que essas duas variáveis estão relacionadas, o que é razoável, pois o clima pode afetar a visibilidade e as condições de luz. "Casualty_class" e "Vehicle_movement":

5. Existe uma correlação (˜0.71) entre a classe de vítimas e o movimento do veículo. Isso sugere que diferentes classes de vítimas (motorista, passageiro, pedestre) podem ser afetadas de maneira diferente dependendo do movimento do veículo no momento do acidente.

O mapa de calor fornece uma visão abrangente das correlações entre pares, destacando que, embora a maioria das variáveis não apresente uma correlação forte com a ”gravidade do acidente”, o número de veículos envolvidos em um acidente tem um impacto notável no resultado da gravidade. Esta perceção foi nos valiosa para decidir o que não remover, e o que remover no nosso dataset

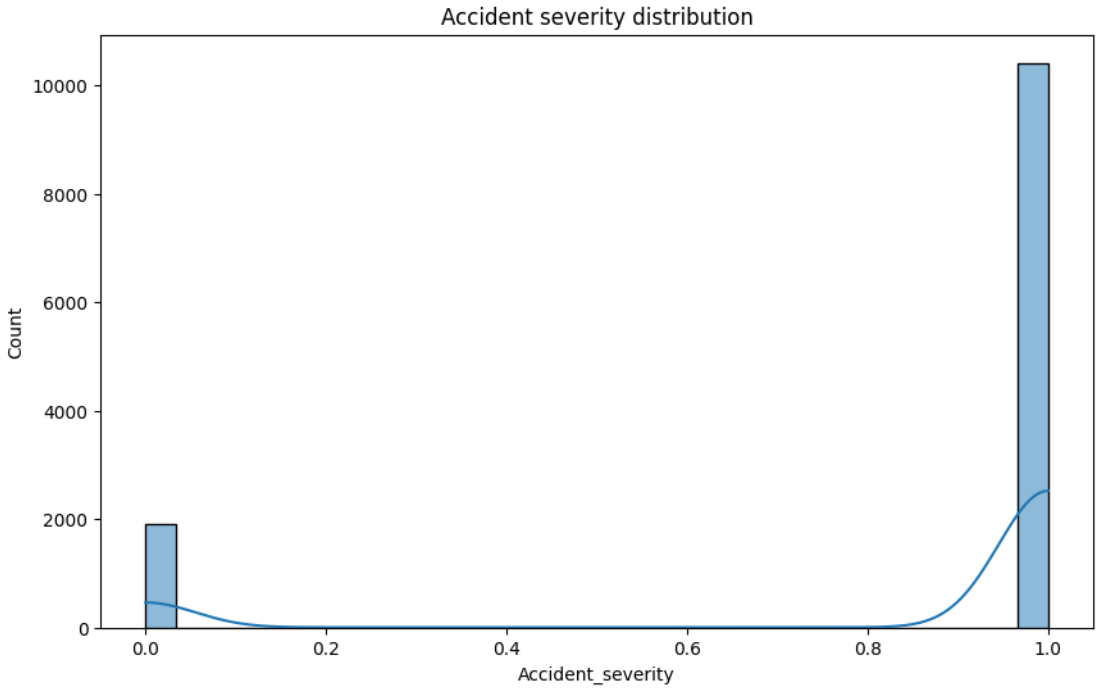

Figure 2: histograma da accident_severity_distribution

O histograma da ”Accident_severity” revela uma distribuição altamente assimétrica, com a maioria dos acidentes concentrados no extremo inferior da escala de gravidade. A maioria dos acidentes tem um valor de gravidade próximo de 1, indicando que resultam em ferimentos leves ou nenhuns, como evidenciado pela barra alta que representa mais de 10.000 casos. Em contraste, os acidentes de alta gravidade, com um valor de gravidade próximo de 0, são muito menos frequentes, formando um pequeno pico no lado esquerdo do gráfico. A linha kernel density estimate (KDE) ilustra ainda mais esse padrão, mostrando uma concentração elevada de acidentes de baixa gravidade e uma cauda longa que se estende para uma gravidade mais elevada. Esta distribuição sugere que, embora os acidentes graves sejam raros, são ainda significativos o suficiente para justificar intervenções de segurança focadas e alocação de recursos para gerir e prevenir tais incidentes de forma eficaz.

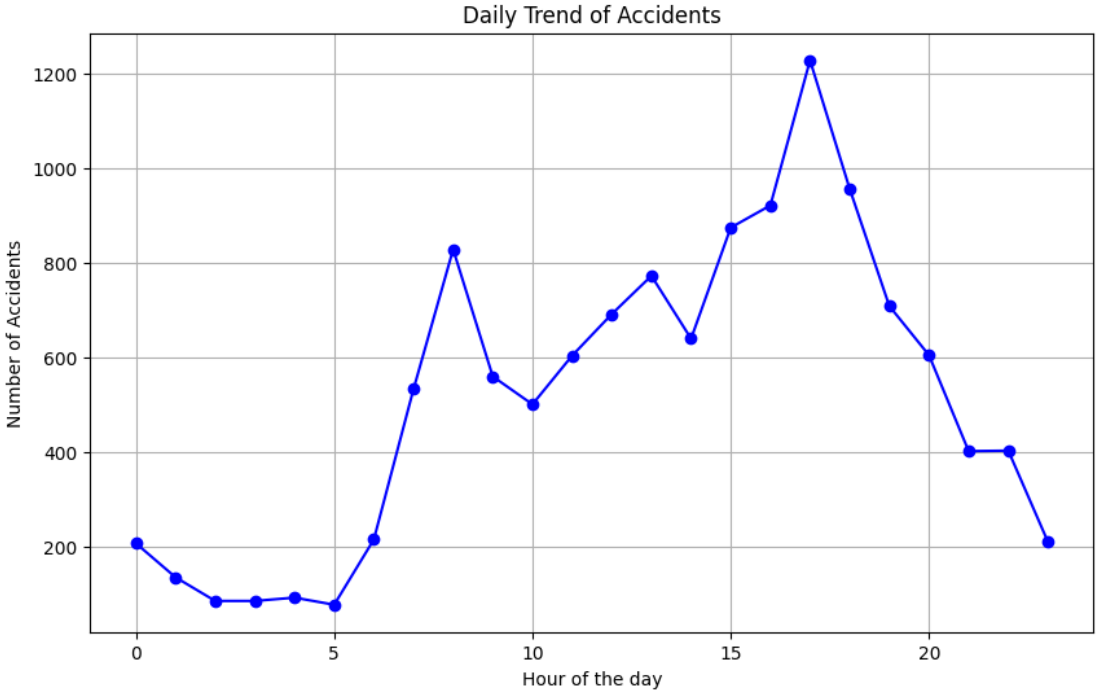

Figure 3: Hour_trend_of_accidents

Analisando as horas do dia com a ajuda da nova coluna ”hour”, podemos analisar dois picos: nas 8 horas da manhã e nas 17 horas. Pode-se especular que os acidentes atingem tais máximos justamente porque é a hora mais comum onde as pessoas vão trabalhar e saem do trabalho, respetivamente. Às 17 o número de acidentes atinge um pico com mais 1200 acidentes especulamos nós pelo cansaço proveniente do trabalho.

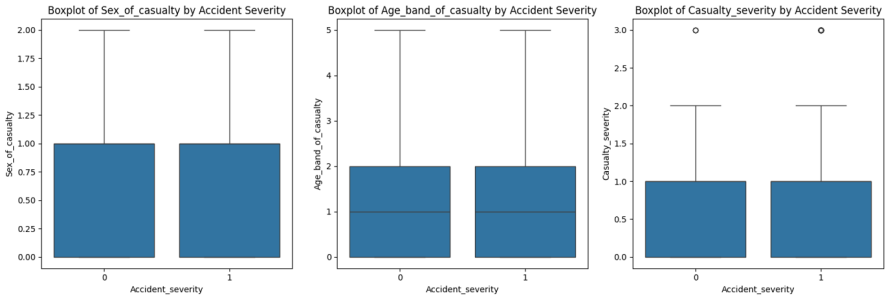

Figure 4: Boxplot

Os boxplots fornecem uma visão geral de como várias as variáveis se relacionam com ”Accident_severity”. A maioria das variáveis, como Hour,Day_of_week, Age_band_of_driver, Sex_of_driver, Educational_level, Vehicle_driver_relation, Driving_experience, Type_of_vehicle, Owner_of_vehicle, Service_year_of_vehicle, Area_accident_occurred, Lanes_or_Medians, Road_alignment, Types_of_Junction, Road_surface_type, Road_surface_conditions, Light_conditions, Weather_conditions, Type_of_collision, Vehicle_movement, Sex_of_casualty, Age_band_of_casualty, Work_of_casualty, Fitness_of_casualty, and Pedestrian_movement, apresentam distribuições semelhantes para ambos os níveis de gravidade. Isso indica que esses fatores não diferenciam fortemente entre acidentes de baixa e alta gravidade.

No entanto, algumas variáveis apresentam diferenças. ”Number_of_vehicles_involved e Number_of_casualties são maiores em acidentes de alta gravidade, indicando uma correlação com a gravidade do acidente. ”Casualty_class” e ”Casualty_severity” também mostram diferenças significativas, com acidentes de alta gravidade envolvendo vítimas mais graves e classes de vítimas variadas.

Em resumo, fatores específicos como o número de veículos envolvidos e o número e gravidade das vítimas desempenham um papel mais crítico na determinação da gravidade do aci-

dente.

2.2 Classificação Multiclass

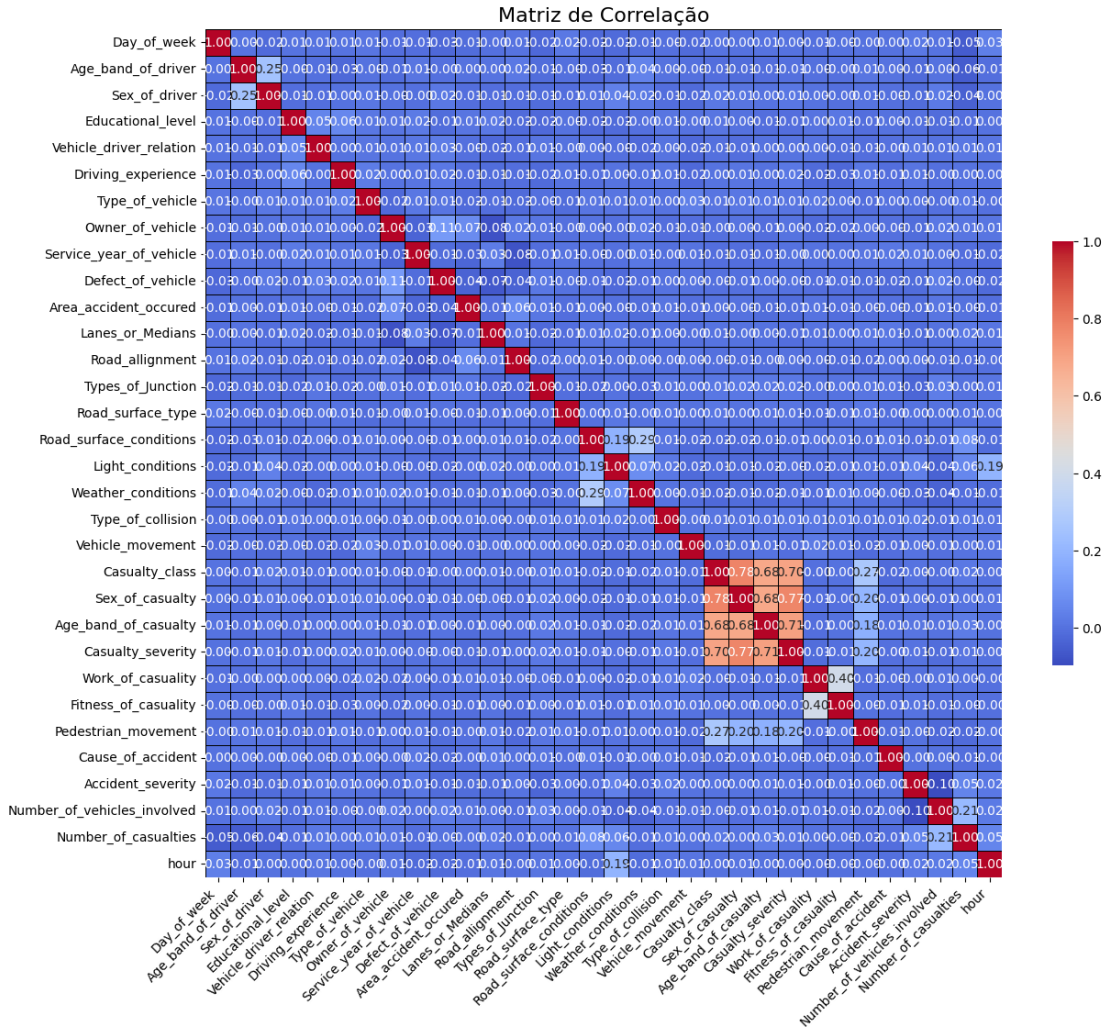

Figure 5: Matriz confusão multiclass

Esta matriz correlação agora uma accuracy melhor para ”Sight_Injury”, no entanto, tem problemas com ”Fata_Inury” e ”Serious_Injury”, mostrando que estas precisam de ser melhoradas de alguma maneira para classificar mais casos severos.

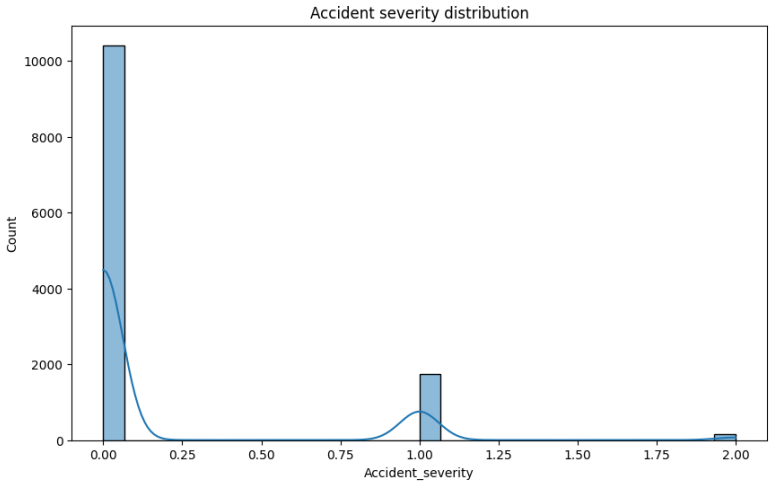

Figure 6: Accident severity distribution

Este histograma revela que há uma escassez casos, tanto para "Serious_Injury" como para "Fatal_Injury"."Slight_Injury" ainda continua com o maior número de casos, com mais de 10000. Serious fica com, aproximadamente 2000 e fatal com aproximadamente 250.

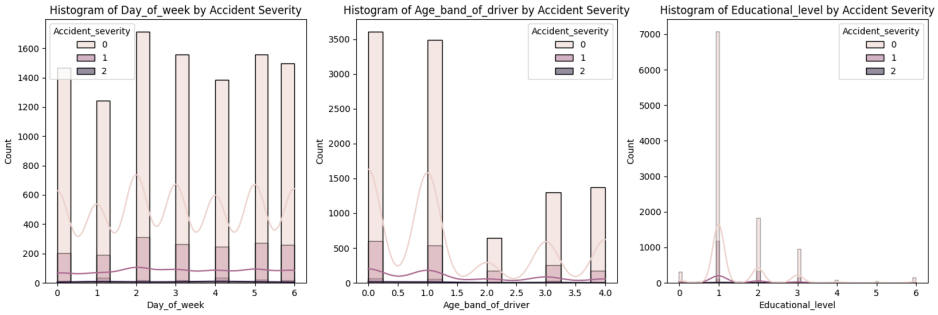

Figure 7: Histograma multiclass

Falta apenas uma descrição aqui para esta parte estar completa.

3 Binary algorithms

3.1 Neural Networks

Em NN (Neural Newtorks) recorremos a undersampling (RandomClassifier) e OverSampling (SMOTE). Obtivemos uma melhor accuracy quando recorremos ao SMOTE, embora tenha tido uma má percentagem no recall (0.34). Com recurso a Undersampling (RandomClassifier), a accuracy e o recall tiveram valores semelhantes, um pouco mais que 0.50 (0.57 e 0.56, respetivamete). Com a ajuda de mais 3 modelos MLP, embora nos tenha demorado mais tempo a executar o código, tanto para o valor 0, como para o valor 1, com o uma accuracy de 0.73, obtivemos valores de recall 0.75 e 0.74 (0 e 1 respetivamente).

```
Accuracy: 0.5723951285520974
              precision    recall  f1-score   support

           0       0.19      0.56      0.29       570
           1       0.88      0.57      0.69      3125

    accuracy                           0.57      3695
   macro avg       0.54      0.57      0.49      3695
weighted avg       0.77      0.57      0.63      3695
```

Figure 8: $\mathrm{NN}_{B}inary_{U}nderSample$

```
Accuracy: 0.7066305818673884
              precision    recall  f1-score   support

           0       0.22      0.34      0.27       570
           1       0.87      0.77      0.82      3125

    accuracy                           0.71      3695
   macro avg       0.54      0.56      0.54      3695
weighted avg       0.77      0.71      0.73      3695
```

Figure 9: $\mathrm{OverSample}_{N}N_{B}innary$

```
Class labels: [0 1]
Misclassified samples: 1589
Accuracy: 0.73
              precision    recall  f1-score   support

           0       0.73      0.75      0.74      2992
           1       0.74      0.72      0.73      3003

    accuracy                           0.73      5995
   macro avg       0.74      0.73      0.73      5995
weighted avg       0.74      0.73      0.73      5995

[[2256  736]
 [ 853 2150]]
```

Figure 10: Mais $\mathrm{MLP}_{l}ayers$

3.2 Naive Bayes

Naive Bayes é uma família de algoritmos de classificação baseados no teorema de Bayes com a suposição de independência ingênua (daí o nome "naive") entre cada par de características. Os classificadores Naive Bayes são amplamente utilizados devido à sua simplicidade, eficiência e capacidade de lidar com grandes volumes de dados.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.17      0.46      0.25       570
           1       0.86      0.58      0.69      3125

    accuracy                           0.56      3695
   macro avg       0.51      0.52      0.47      3695
weighted avg       0.75      0.56      0.62      3695

Confusion Matrix:
 [[ 265  305]
 [1313 1812]]
Accuracy Score: 0.5621109607577808
```

Figure 11: Oversampling$_{b}ayes$

```
Classification Report:
              precision    recall  f1-score   support

           0       0.20      0.48      0.28       570
           1       0.87      0.66      0.75      3125

    accuracy                           0.63      3695
   macro avg       0.54      0.57      0.52      3695
weighted avg       0.77      0.63      0.68      3695

Confusion Matrix:
 [[ 271  299]
 [1077 2048]]
Accuracy Score: 0.6276048714479026
```

Figure 12: Undersampling$_{b}ayes$

3.3 Support Vector Machines (SVM)

O Support Vector Machine (SVM) é um algoritmo de aprendizado de máquina supervisionado utilizado principalmente para tarefas de classificação, mas também pode ser aplicado em regressão e detecção de outliers. O objetivo principal do SVM é encontrar um hiperplano em um espaço de alta dimensão que separe os dados em diferentes classes da melhor forma possível. No contexto de classificação, SVM tenta maximizar a margem entre as classes, onde a margem é definida como a distância entre o hiperplano e os pontos de dados mais próximos de qualquer classe, conhecidos como vetores de suporte.

Kernel: poly, C: 0.1

Accuracy: 0.73

Confusion Matrix

$$\begin{bmatrix} 104 & 466 \\ 519 & 2606 \end{bmatrix}$$

Classification Report

	Precision	Recall	F1-Score	Support
Class 0	0.17	0.18	0.17	570
Class 1	0.85	0.83	0.84	3125
Accuracy			0.73	3695
Macro Avg	0.51	0.51	0.51	3695
Weighted Avg	0.74	0.73	0.74	3695

4 Classificação Multiclasse

4.1 Neural Networks

- **Accuracy:** 0.6771

	Precisão	**Recall**	**F1-Score**	**Suporte**
Classe 0	0.86	0.75	0.80	3125
Classe 1	0.19	0.32	0.24	523
Classe 2	0.04	0.09	0.05	47
Accuracy			0.68	
Média Macro	0.36	0.38	0.36	3695
Média Ponderada	0.76	0.68	0.71	3695

Table 1: Relatório de Classificação Oversampling

- **Accuracy:** 0.4154

	Precision	Recall	F1-Score	Support
Class 0	0.88	0.43	0.57	3125
Class 1	0.14	0.32	0.19	523
Class 2	0.04	0.72	0.07	47
Accuracy			0.42	3695
Macro Avg	0.35	0.49	0.28	3695
Weighted Avg	0.76	0.42	0.51	3695

Table 2: Relatório de Classificação Undersampling

4.2 Support Vector Machines

Melhores valores obtidos em Multiclass: **Kernel: rbf, C: 100**

Accuracy: 0.82

Confusion Matrix:

$$\begin{bmatrix} 2999 & 124 & 2 \\ 503 & 19 & 1 \\ 45 & 2 & 0 \end{bmatrix}$$

Classification Report:

	precision	recall	f1-score	support
0	0.85	0.96	0.90	3125
1	0.13	0.04	0.06	523
2	0.00	0.00	0.00	47
accuracy			0.82	3695
macro avg	0.33	0.33	0.32	3695
weighted avg	0.73	0.82	0.77	3695

4.3 Naive Bayes

O Naive Bayes é particularmente eficaz em problemas de classificação de texto e mineração de texto, como classificação de documentos e análise de sentimentos. É amplamente utilizado em aplicativos do mundo real, como filtragem de spam de e-mail, classificação de documentos, análise de sentimento em redes sociais e muito mais.

Uma das principais vantagens do Naive Bayes é sua simplicidade e rapidez de treinamento, especialmente em conjuntos de dados de grande escala. Além disso, mesmo que a suposição de independência não seja verdadeira na prática para muitos conjuntos de dados, o Naive Bayes ainda pode fornecer resultados bastante precisos em muitos casos.

No entanto, o Naive Bayes pode ser limitado em sua capacidade de capturar interações complexas entre as características. Além disso, se uma categoria de classe (ou resultado) tiver uma probabilidade de zero para uma determinada característica, isso pode causar distorções significativas nos resultados do modelo. Em tais casos, é necessário aplicar técnicas de suavização para evitar esses problemas.

Ao avaliarmos o desempenho do nosso modelo de Naive Bayes para classificação binária, observamos os seguintes valores de precisão, recall e F1-score para cada classe:

	Precision	Recall	F1-Score	Support
Class 0	0.86	0.48	0.62	3125
Class 1	0.13	0.14	0.14	523
Class 2	0.02	0.57	0.04	47
Accuracy			0.44	3695
Macro Avg	0.34	0.40	0.26	3695
Weighted Avg	0.75	0.44	0.54	3695

Confusion Matrix

$$\begin{bmatrix} 1509 & 460 & 1156 \\ 229 & 72 & 222 \\ 17 & 3 & 27 \end{bmatrix}$$

5 Métricas de avaliação

Métricas de avaliação são medidas quantitativas usadas para analisar a performance de modelos de machine learning. Com estas métricas conseguimos avaliar se os algoritmos que usámos foram uma ótima escolha(ou não).

5.1 Matriz confusão

Para entendermos melhor as próximas métricas de avaliação, precisamos de conhecer a matriz confusão. É uma tabela capaz de visualizar o desempenho de um algoritmo de classificação com os seus quatro componentes principais:

- True Positives (TP): O modelo previu corretamente a classe positiva.
- True Negatives (TN): O modelo previu corretamente a classe negativa.
- False Positives (FP): O modelo previu incorretamente a classe positiva.(Erro tipo I)
- False Negatives (FN): O modelo previu incorretamente a classe negativa,(Erro tipo II)

5.2 Accuracy

"Accuracy" mede a proporção de previsões corretas feitas pelo modelo de classficação. A sua fórmula é a seguinte:

$$\text{Accuracy} = \frac{Numerodeprevisoescorretas}{Numerototaldeprevisoes} \tag{1}$$

5.3 Precision

”Precision” é uma métrica de avaliação que tem a capacidade do classificador de não rotular como positiva uma amostra que é negativa.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

5.4 Recall

”Recall” ou ”Sensivity” mede a abilidade que o modelo tem, de encontrar todos os casos relevantes dentro de um dataset.

A formula é a seguinte:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

5.5 F1 score

F1 score particularmente é útil em situações onde a data não é balanceada Combina a precision e o recall para garantir uma única medida de qualidade

$$\text{F1} = \frac{2TP}{(2TP) + FN + FP} \tag{4}$$

6 Conclusão

Ao final da análise comparativa entre a classificação binária e a classificação multiclass utilizando três algoritmos distintos — Máquina de Vetores de Suporte (SVM), Redes Neurais (NN) e Critério de Bayes — observamos uma diferença significativa nos desempenhos entre os dois tipos de classificação com uma melhor avaliação na classe binária!