Alzheimers Disease Dataset

Henrique Z. Grande, João G. Pitol, Lucas B. Shuck, Rafael V. Serenato

Data Science – PUCPR
2025

Overview and Estatistical Description

O dataset representa um conjunto de informações sobre a saúde de 2149 pacientes idosos que estão separados em duas categorias, os que tem diagnósticos positivos para Alzheimer e os que não tem.

As principais colunas do data set são:

Age: Idade dos pacientes.

Gender: Sexo dos pacientes.

AlcoholConsumption: Nível de consumo de álcool.

Smoking: Presença do habito de fumar.

MMSE: Pontuação do teste de “Mini Exame do Estado Mental”.

MemoryComplaints: Presença de reclamações de problema de memória.

BehavioralProblems: Presença de problemas de comportamento.

FunctionalAssessment: Pontuação no teste de capacidade de realizar atividades cotidianas.

ADL: Pontuação no teste de capacidade de realizar básicas de autocuidado.

Diagnosis: Diagnóstico da doença Alzheimer.

Overview and Estatistical Description

Nosso problema é um problema de classificação representado pela nossa variável target 'Diagnosis', que indica a presença ou não da doença Alzheimer nos pacientes e está distribuída de maneira desigual pela base.

Nosso dataset possui a seguinte estrutura:

Instancias: 2149

Variáveis: 33

Classes da variável target: 2

Variáveis numéricas float: 12

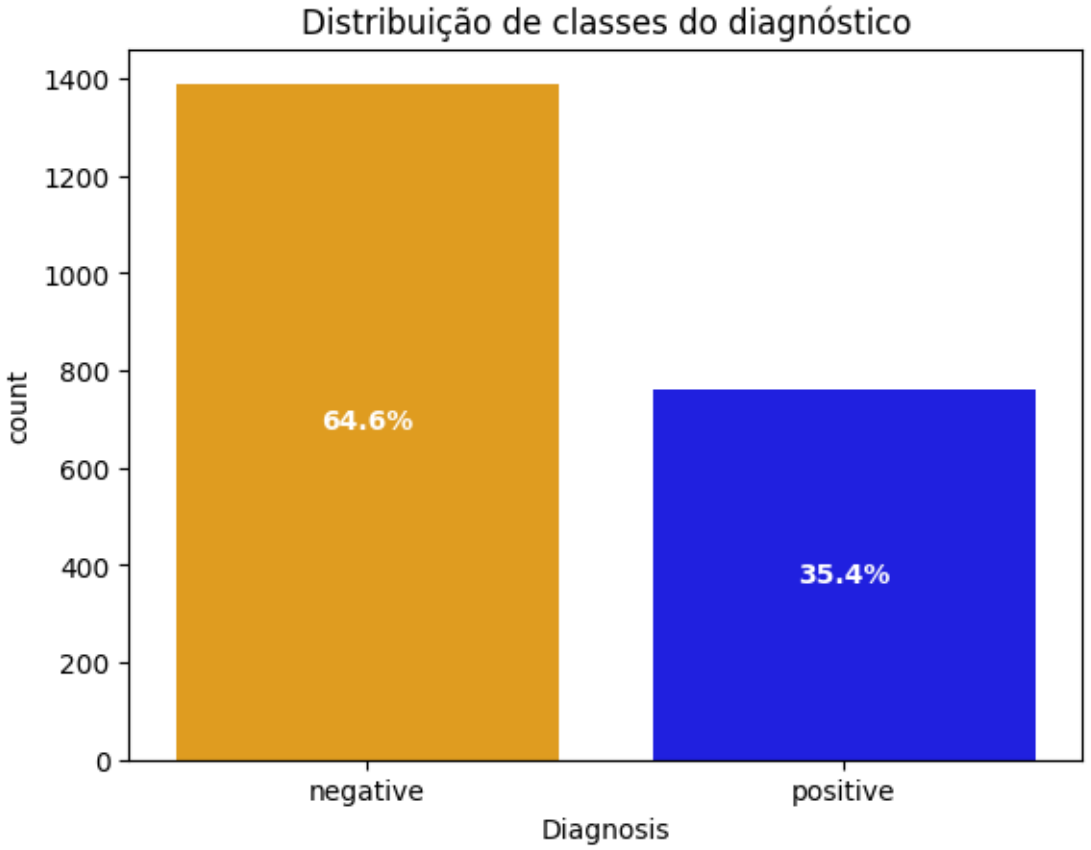
Variáveis numéricas inteiros: 3

Variáveis categóricas: 18

Univariate Data Analysis

Analise da distribuição da variável categórica ‘Diagnosis’ que indica se o paciente tem ou não a doença Alzheimer diagnosticada.

Pelo gráfico é possível notar que a distribuição desta variável é desbalanceada, sendo 64.6% dos casos um diagnóstico negativo e apenas 35.4% um diagnóstico positivo.



Univariate Data Analysis

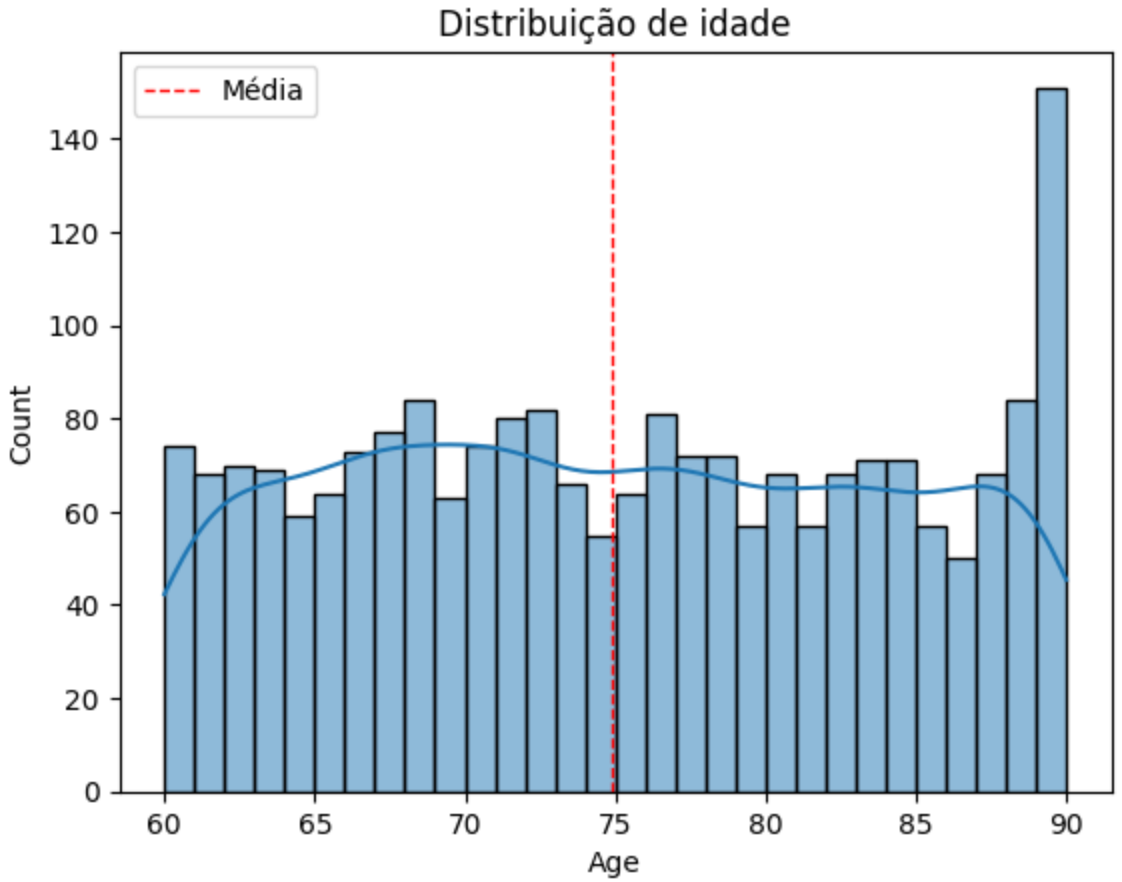
Analise da distribuição da variável numérica ‘Age’ que indica a quantidade de pacientes por idade presentes no Dataset.

Pelo gráfico é possível notar que a distribuição desta variável é bem uniforme indo dos 60 aos 90 anos sendo a média de idade um pouco abaixo de 75 anos e existindo uma leve concentração de pessoas com 90 anos.

Média: 74,91 | Mediana: 75,0 | Desvio: 8,99

Assimetria: Próximo de 0, simétrica.

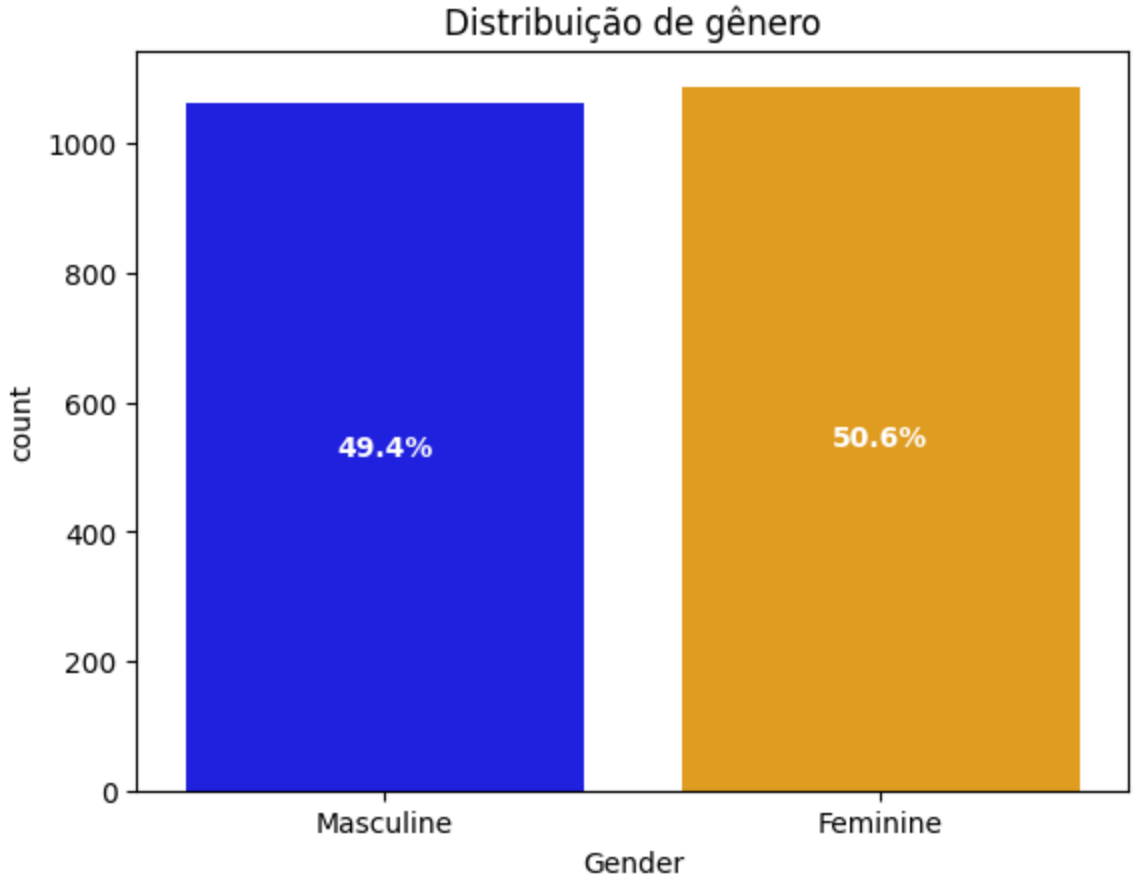
Curtose: Platicúrtica.



Univariate Data Analysis

Analise da distribuição da variável categórica ‘Gender’ que indica o sexo biológico do paciente.

Pelo gráfico é possível notar que a distribuição desta variável é praticamente igualitária, sendo 49.4% dos pacientes homens e 50.6% mulheres.



Univariate Data Analysis

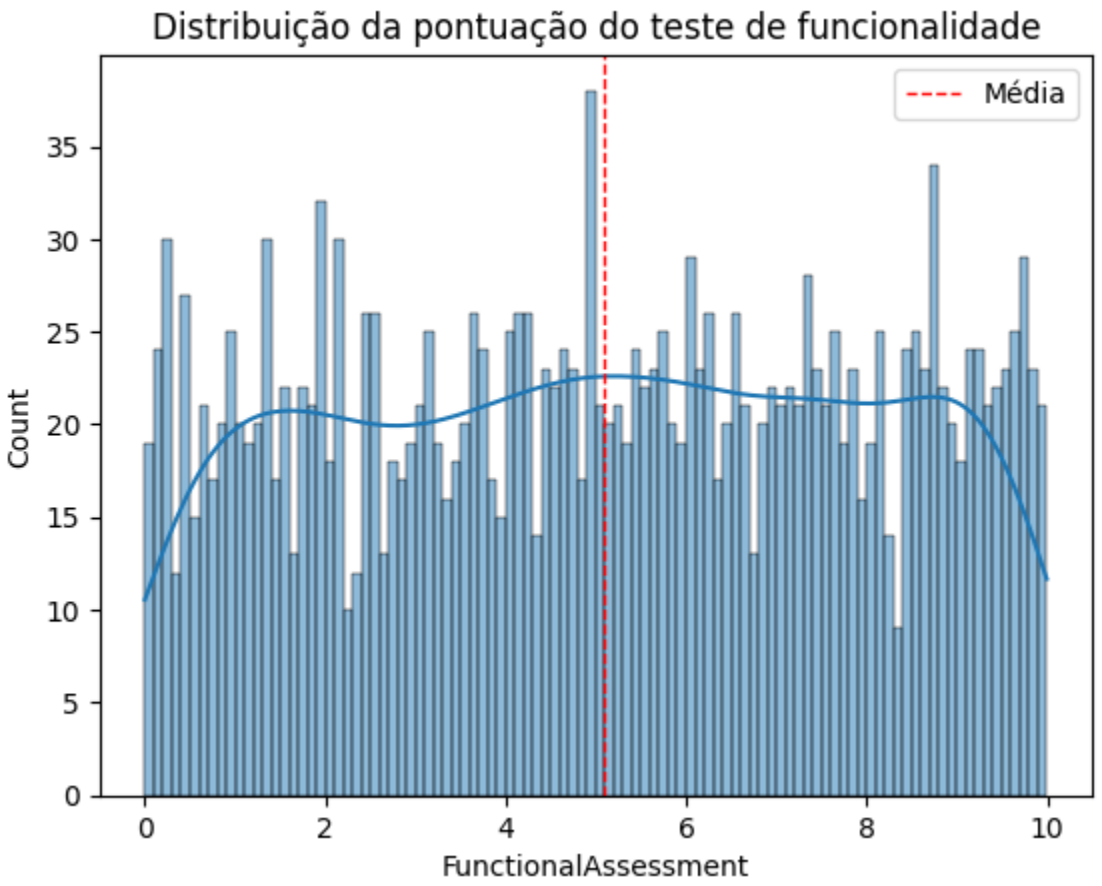
Analise da distribuição da variável numérica ‘FunctionalAssessment’ que indica a pontuação dos pacientes no teste que mede a capacidade de realizar tarefas cotidianas.

Pelo gráfico é possível notar que a distribuição desta variável é bem uniforme tendo pacientes com notas de 0 até 10 sendo a média de pontuação 5.

Média: 5,08 | Mediana: 5,09 | Desvio : 2,89

Assimetria: Próximo de 0, simétrica.

Curtose: Platicúrtica.



Univariate Data Analysis

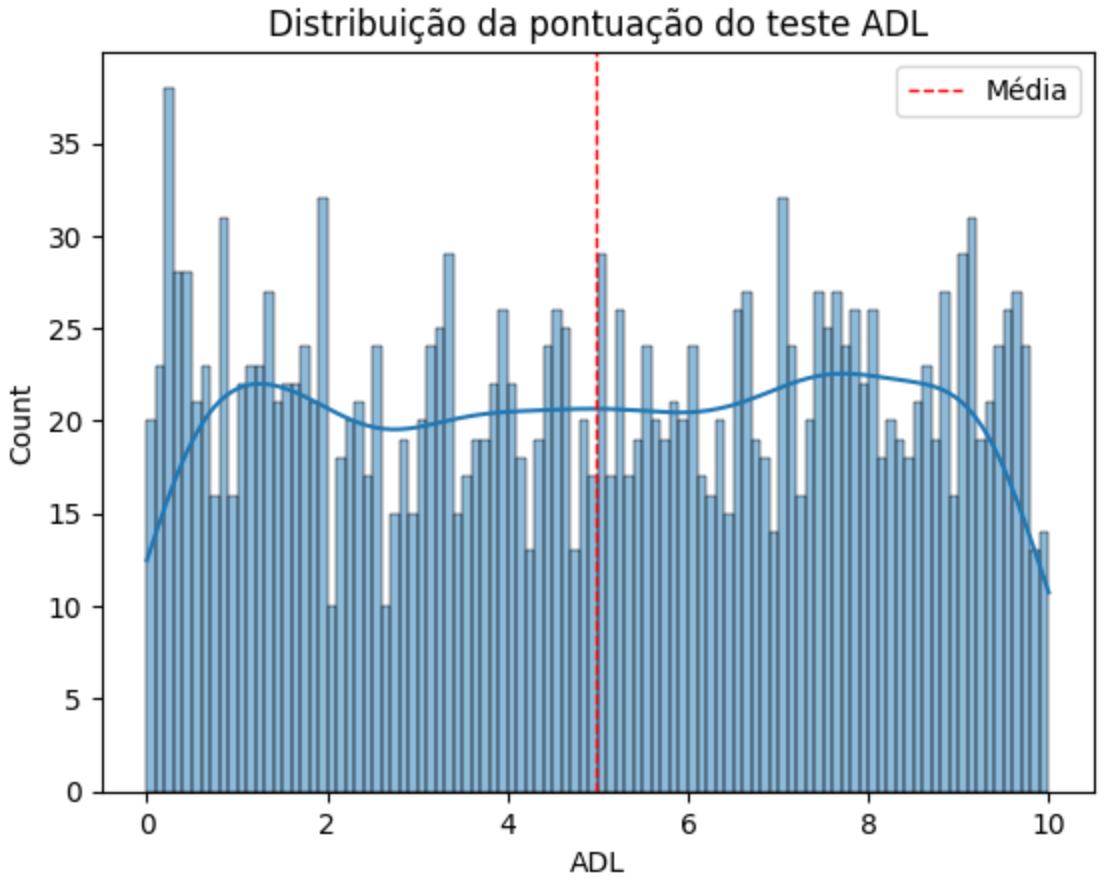
Analise da distribuição da variável numérica ‘ADL’ que indica a pontuação dos pacientes no teste que mede a capacidade de realizar tarefas fundamentais relacionadas a auto cuidado.

Pelo gráfico é possível notar que a distribuição desta variável é bem uniforme tendo pacientes com notas de 0 até 10 sendo a média de pontuação 5.

Média: 4,98 | Mediana: 5,04 | Desvio: 2,95

Assimetria: Próximo de 0, simétrica.

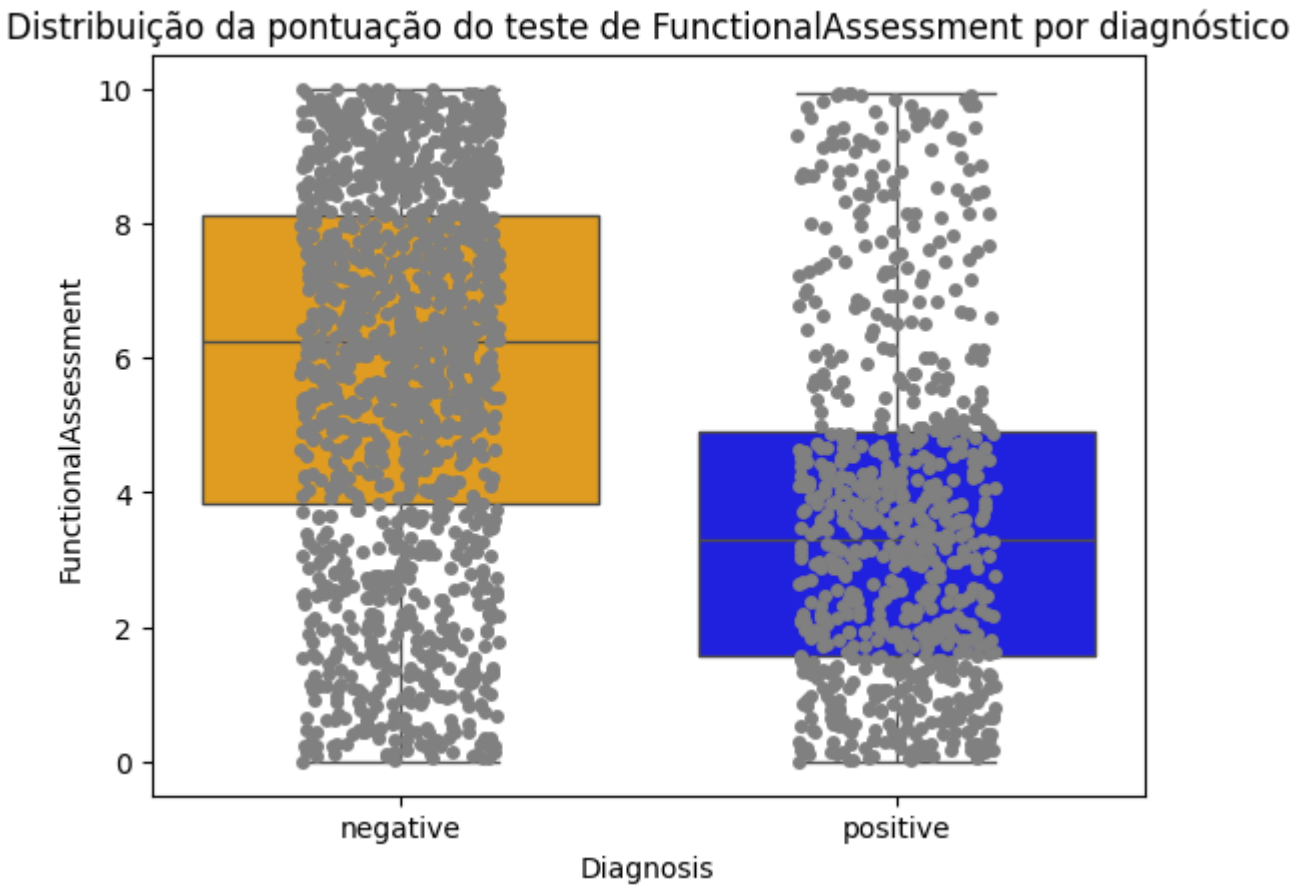
Curtose: Platicúrtica.



Multivariate Data Analysis

Analise da distribuição de pontuação dos testes de ‘FunctionalAssessment’ por classe da variável target ‘Diagnosis’.

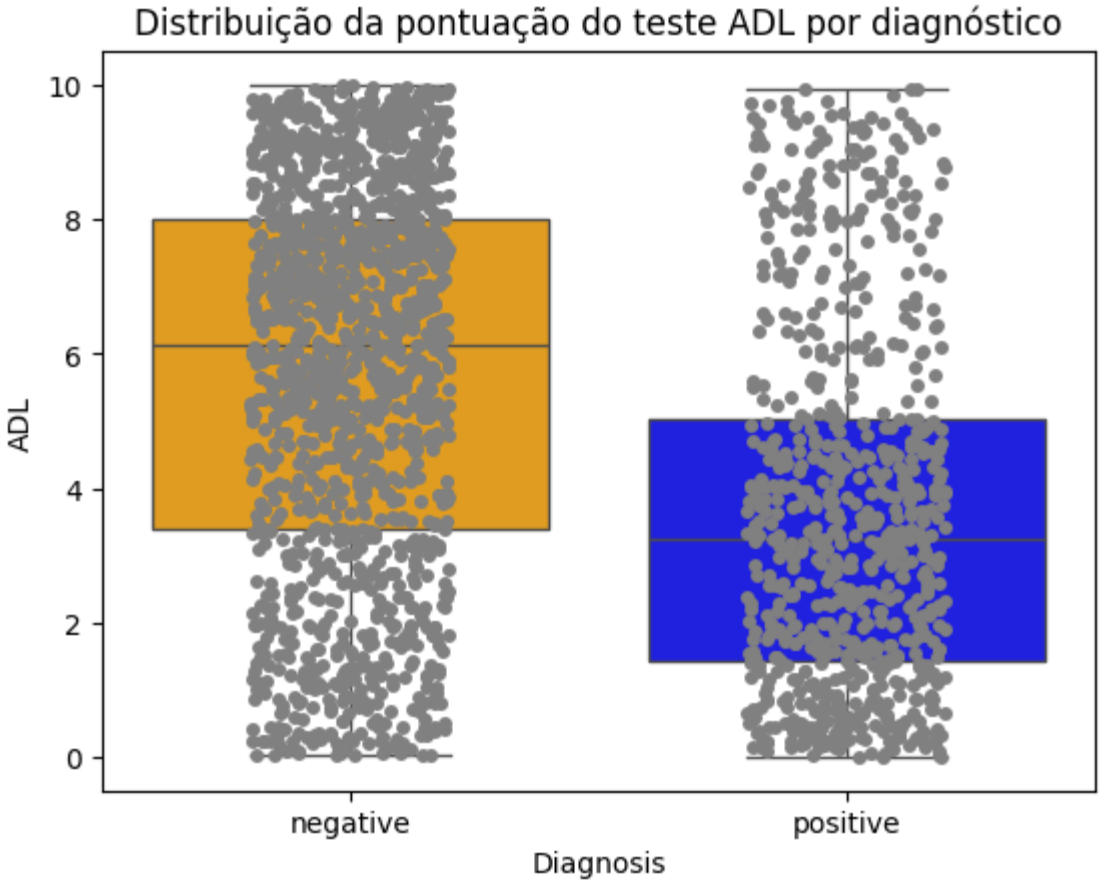
Pelo gráfico é possível notar que a distribuição da pontuação deste teste nos diagnósticos negativos tem uma concentração acima da média com muitas pontuações altas, já nos diagnósticos positivos é possível identificar uma alta concentração de pontuações abaixo da média.



Multivariate Data Analysis

Analise da distribuição de pontuação dos testes de ‘ADL’ por classe da variável target ‘Diagnosis’.

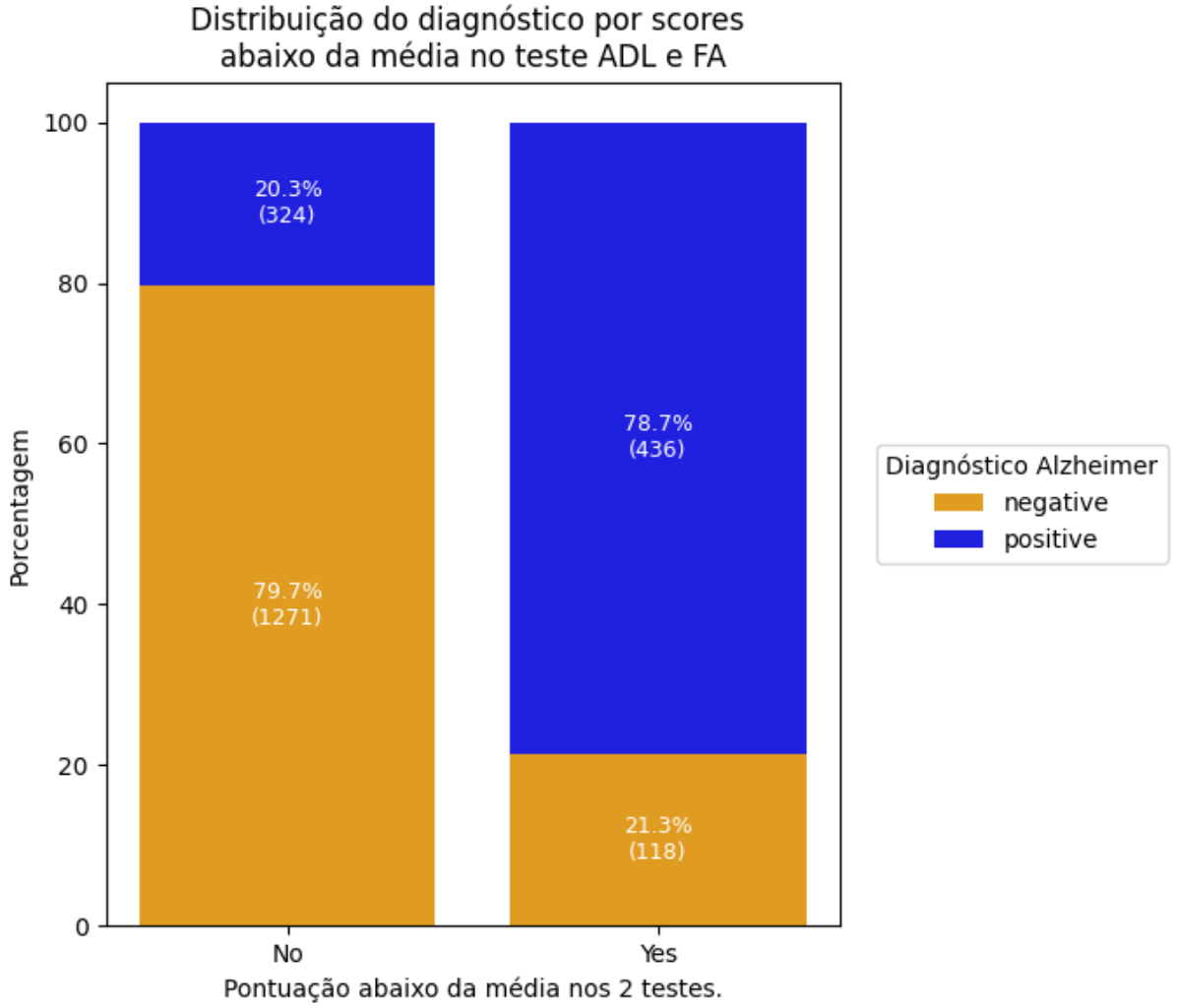
Pelo gráfico é possível notar que a distribuição da pontuação deste teste nos diagnósticos negativos tem uma concentração acima da média com várias pontuações altas, já nos diagnósticos positivos é possível identificar uma alta concentração de pontuações abaixo da média.



Multivariate Data Analysis

Analise da quantidade pacientes com pontuações abaixo da média nos testes de ‘ADL’ e ‘FunctionalAssessment’ por classe da variável target ‘Diagnosis’.

Pelo gráfico baseado em uma nova variável criada para representar um 'grupo de risco' relacionado a agregação de notas baixas é possível notar que 78.7% dos pacientes que apresentam pontuações abaixo da média nos 2 testes simultaneamente tem Alzheimer enquanto no outro caso apenas 20.3% deles apresentam diagnóstico positivo, além disso também é possível concluir que a maioria dos pacientes de toda a base com diagnósticos positivos tem notas abaixo da média nos 2 testes, evidenciando mais uma vez este 'grupo de risco'.



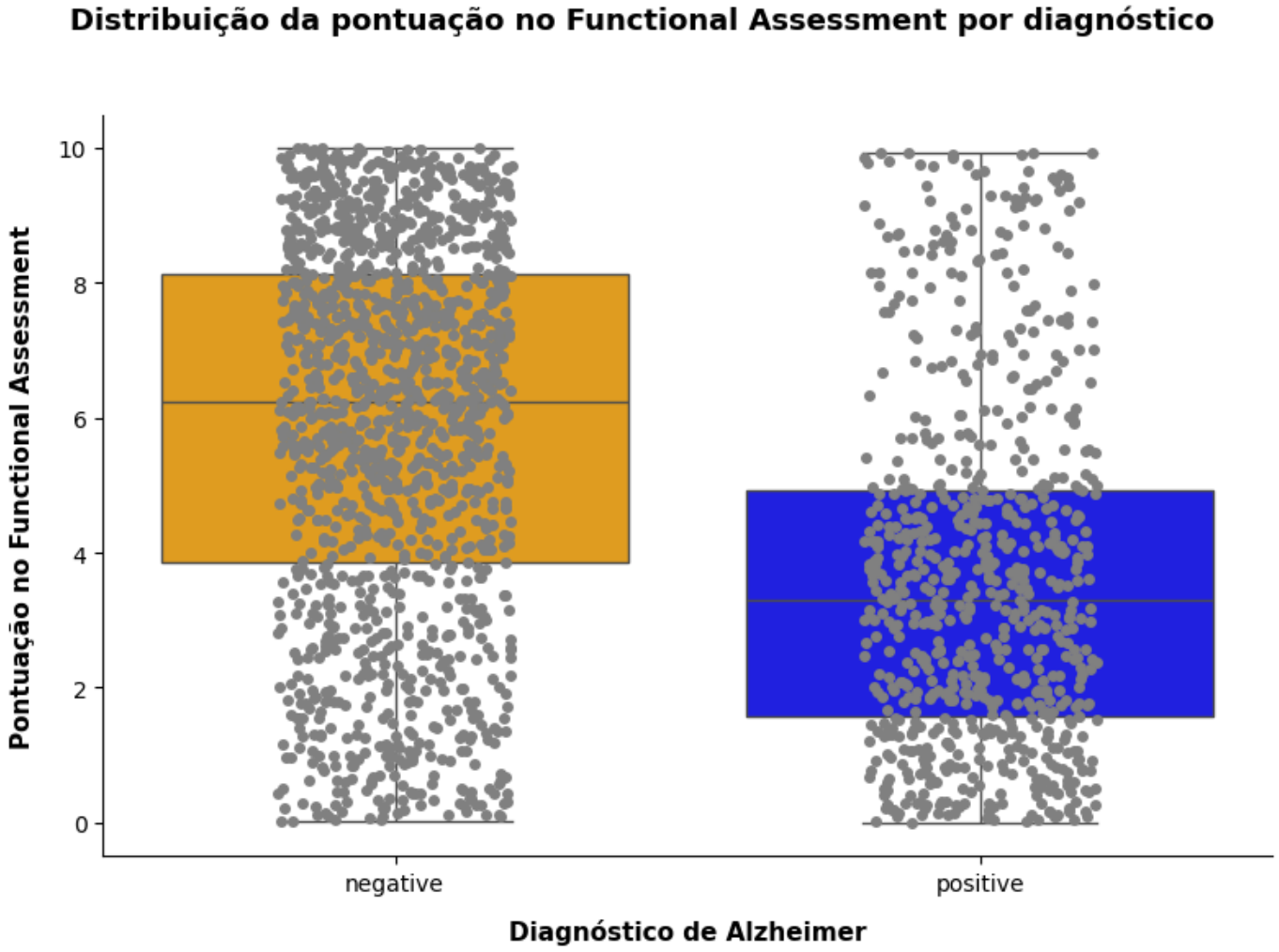
Effective Data Visualization

Mudanças feitas:

Título descrevendo a ideia do gráfico e em negrito, com formatação melhorada.

Aumento do tamanho do boxplot, aumentando a área para a distribuição dos scatterplots facilitando a visualização das concentrações.

Rótulos dos eixos foram traduzidos, ajustados para facilitar entender o que estão descrevendo e estão em negrito.



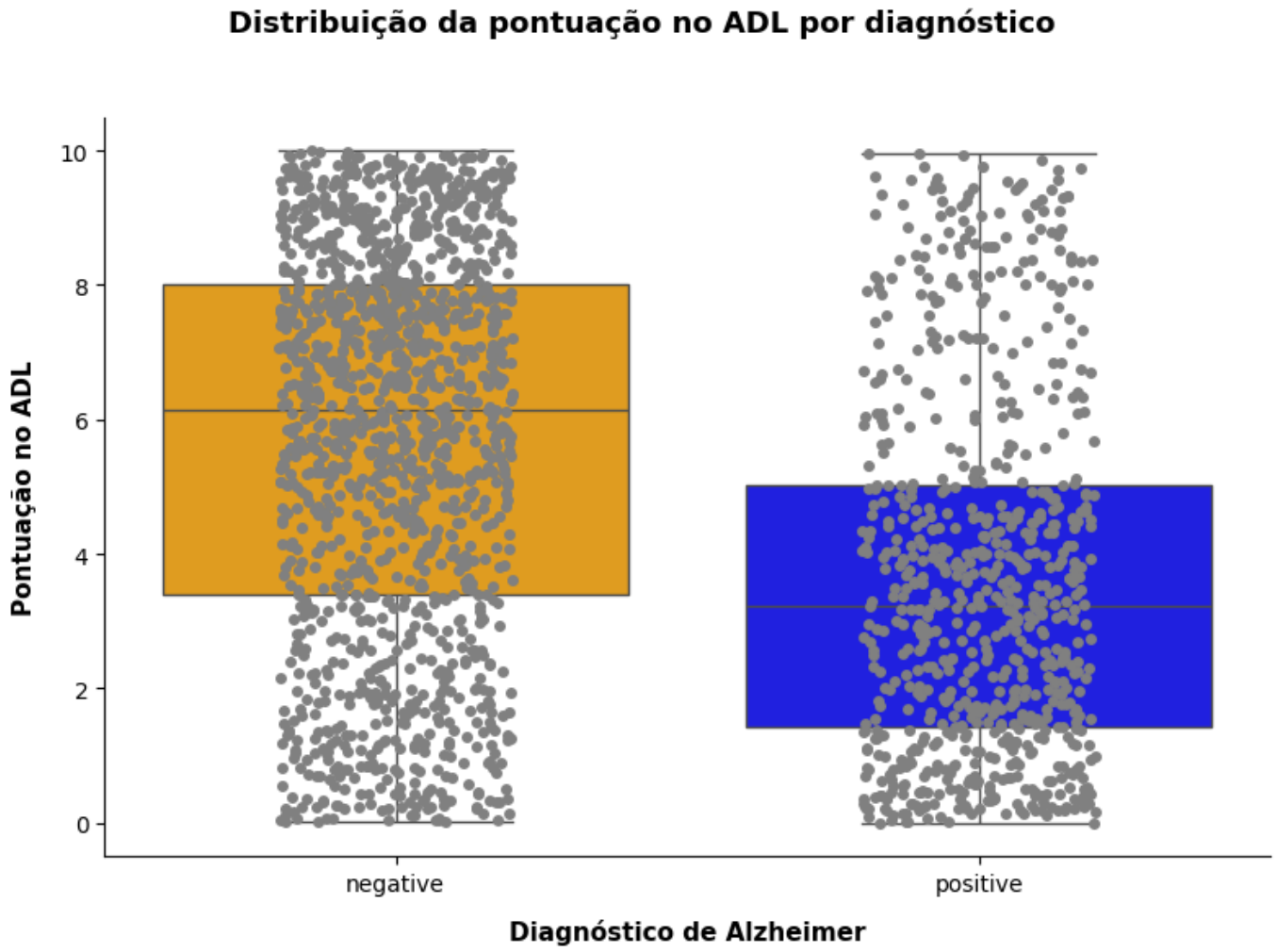
Effective Data Visualization

Mudanças feitas:

Título descrevendo a ideia do gráfico e em negrito, com formatação melhorada.

Aumento do tamanho do boxplot, aumentando a área para a distribuição dos scatterplots facilitando a visualização das concentrações.

Rótulos dos eixos foram traduzidos, ajustados para facilitar entender o que estão descrevendo e estão em negrito.



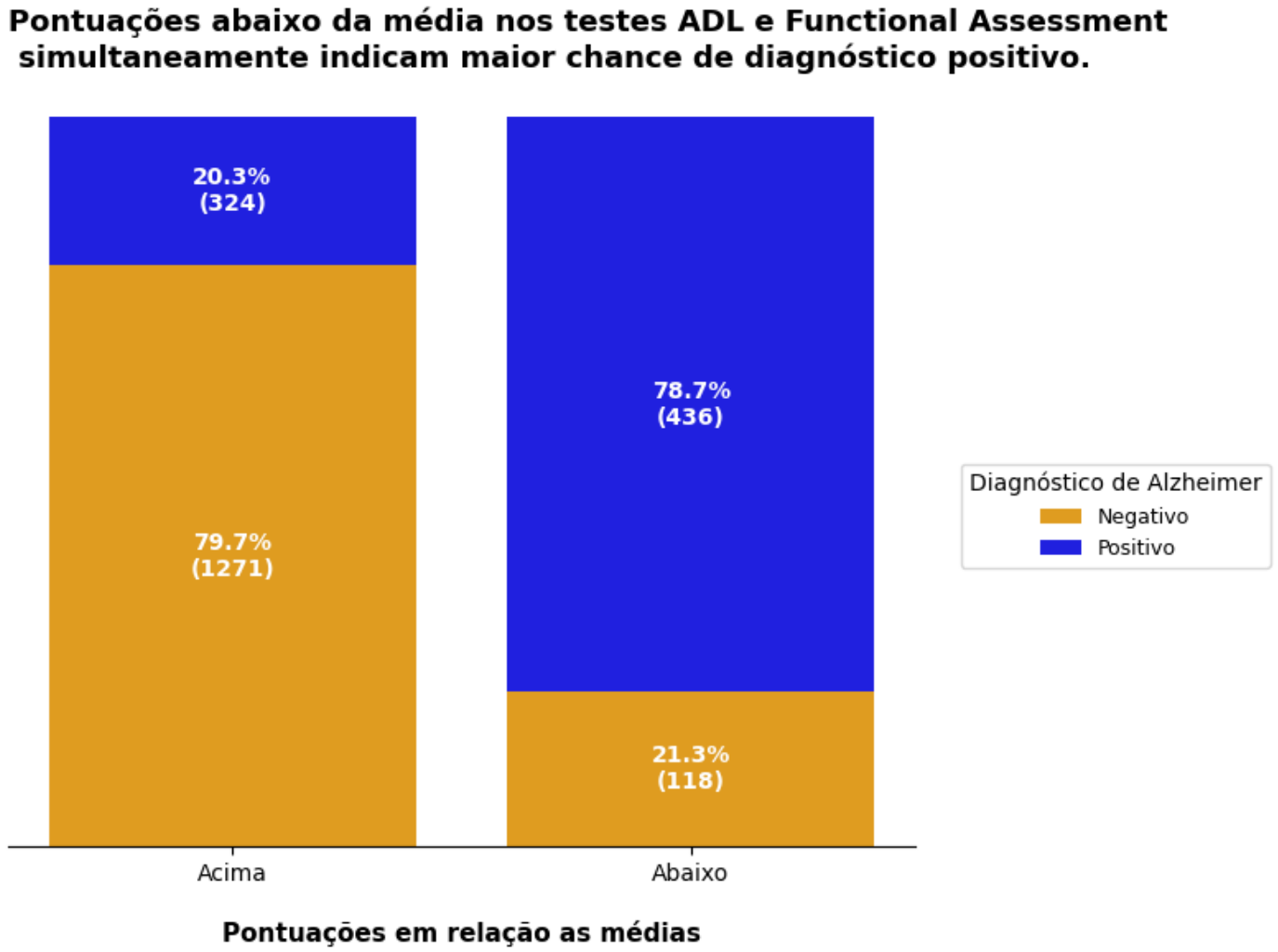
Effective Data Visualization

Mudanças feitas:

Título descrevendo a ideia do gráfico e em negrito, com formatação melhorada.

Números absolutos e porcentagens presentes nas barras foram colocados em negrito para melhor visualização.

Rótulos dos eixos foram traduzidos, ajustados para facilitar entender o que estão descrevendo e estão em negrito.



Data Selection

Não foram aplicadas técnicas de seleção de dados neste trabalho, pois a base não exibe dados que possam ser considerados antiéticos ou ilegais uma vez que apenas expõe dados médicos, socio demográficos e comportamentais de maneira anonimizada, sem expor nenhum paciente.

Também não aplicamos técnicas de redução de dimensionalidade nem de seleção de atributos, uma vez que nossa base apresenta poucas variáveis e todas possuem algum grau de relevância quando relacionadas com a variável target.

Data Processing

Nesta etapa foram aplicadas técnicas de padronização dos dados e técnicas de balanceamento da variável target que tem uma distribuição de 65% negativo e 35% positivo.

Para a variável target foi usado um LabelEncoder simples trocando o ‘negative’ para 0 e ‘positive’ para 1.

Para as variáveis numéricas foi aplicado o StandardScaler deixando os dados numéricos com uma média próxima a 0 e um desvio padrão próximo a 1.

Para as variáveis categóricas nominais utilizamos o OneHotEncoder para que nenhuma ordem artificial seja induzida e para variáveis categóricas ordinais utilizamos o OrdinalEncoder para mantermos a ordem de relevância existente neste tipo de variável.

Data Processing

Após a padronização dos dados dividimos a base em treino e teste mantendo a proporção original da variável target em cada fold utilizando o StratifiedKFolds utilizando 10 Folds e o random_state = 42.

Com a divisão feita aplicamos 4 técnicas de balanceamento abaixo, apenas na divisão de treino de cada fold.

Random Over Sampling (ROS): Que faz a duplicação aleatória de instancias da classe minoritária.

SMOTE (Synthetic Minority Over-sampling Technique): Que cria instancias sintéticas da classe minoritária em uma zona segura.

Random Under Sampling (RUS): Que faz a remoção aleatória de instancias da classe majoritária.

NearMiss (versão 1): Que seleciona instancias majoritárias com base em distancia mínima as minoritárias.

		Treino	Teste	SMOTE	RandomOver	RandomUnder	NearMiss-1
Fold	Classe						
Fold 1	0	1250	139	1250	1250	684	684
	1	684	76	1250	1250	684	684
Fold 2	0	1250	139	1250	1250	684	684
	1	684	76	1250	1250	684	684
Fold 3	0	1250	139	1250	1250	684	684
	1	684	76	1250	1250	684	684
Fold 4	0	1250	139	1250	1250	684	684
	1	684	76	1250	1250	684	684
Fold 5	0	1250	139	1250	1250	684	684
	1	684	76	1250	1250	684	684
Fold 6	0	1250	139	1250	1250	684	684
	1	684	76	1250	1250	684	684
Fold 7	0	1250	139	1250	1250	684	684
	1	684	76	1250	1250	684	684
Fold 8	0	1250	139	1250	1250	684	684
	1	684	76	1250	1250	684	684
Fold 9	0	1250	139	1250	1250	684	684
	1	684	76	1250	1250	684	684
Fold 10	0	1251	138	1251	1251	684	684
	1	684	76	1251	1251	684	684

Machine Learning

KNN

O primeiro modelo escolhido foi o KNN, ele é um dos que se beneficiam da normalização dos dados e dentre os 3 escolhido teve o pior resultado com todos os tipos de balanceamento.

Foi configurado apenas com o parâmetro K=5.

Modelo_Balanceamento	Acurácia	Precision	Recall	F1-score
KNN - RandomOver	0.6994	0.5612	0.6882	0.6174
KNN - SMOTE	0.6603	0.5129	0.7895	0.6218
KNN - RandomUnder	0.7087	0.5719	0.7158	0.6352
KNN - NearMiss-1	0.7446	0.6544	0.5908	0.6203

Machine Learning

Random-Forest

O segundo modelo escolhido foi o Random-Forest, ele não se beneficiou da etapa de normalização dos dados e dentre os 3 escolhidos teve o melhor desempenho com todas as técnicas de balanceamento.

Como nos testes iniciais este foi o melhor modelo decidimos modificar seus Hyper parâmetros para obtermos o melhor resultado possível deste modelo.

Os melhores parâmetros encontrados foram:

n_estimators=200, max_depth=10, min_samples_split=5, min_samples_leaf=2, max_features='sqrt', class_weight='balanced' e radom_state=42.

Modelo_Balanceamento	Acurácia	Precision	Recall	F1-score
RandomForest - RandomOver	0.9395	0.9402	0.8855	0.9118
RandomForest - SMOTE	0.9427	0.9481	0.8868	0.9160
RandomForest - RandomUnder	0.9418	0.9208	0.9145	0.9174
RandomForest - NearMiss-1	0.9362	0.9016	0.9211	0.9109

Machine Learning

SVM

O terceiro modelo escolhido foi o SVM, ele é um dos que se beneficiaram da etapa de normalização dos dados e dentre os 3 modelos é oque teve o resultado médio com todos os tipos de balanceamento.

Foi configurado apenas com os parâmetros SVC(kernel='rbf') e random_state=42

Modelo_Balanceamento	Acurácia	Precision	Recall	F1-score
SVM - RandomOver	0.8404	0.7671	0.7895	0.7777
SVM - SMOTE	0.8353	0.7750	0.7540	0.7639
SVM - RandomUnder	0.8311	0.7311	0.8276	0.7758
SVM - NearMiss-1	0.8125	0.7082	0.8053	0.7526

Validation

Para validação foi usado o protocolo de validação cruzada estratificada com 10 folds com o Random_state=42 para garantir que todos os modelos fossem testados com as mesmas partições dos dados.

Cada um dos 3 modelos foi treinado com os 4 tipos de balanceamentos e validados em conjuntos de testes sem o balanceamento.

A partir disto foram retiradas as métricas de acurácia, precisão, recall e F1-score para cada fold de cada combinação de modelo de ML e método de balanceamento.

Na tabela demonstramos as médias das métricas dos 10 folds de cada combinação.

	Modelo_Balanceamento	Acurácia	Precision	Recall	F1-score
0	KNN - RandomOver	0.6994	0.5612	0.6882	0.6174
1	KNN - SMOTE	0.6603	0.5129	0.7895	0.6218
2	KNN - RandomUnder	0.7087	0.5719	0.7158	0.6352
3	KNN - NearMiss-1	0.7446	0.6544	0.5908	0.6203
4	RandomForest - RandomOver	0.9395	0.9402	0.8855	0.9118
5	RandomForest - SMOTE	0.9427	0.9481	0.8868	0.9160
6	RandomForest - RandomUnder	0.9418	0.9208	0.9145	0.9174
7	RandomForest - NearMiss-1	0.9362	0.9016	0.9211	0.9109
8	SVM - RandomOver	0.8404	0.7671	0.7895	0.7777
9	SVM - SMOTE	0.8353	0.7750	0.7540	0.7639
10	SVM - RandomUnder	0.8311	0.7311	0.8276	0.7758
11	SVM - NearMiss-1	0.8125	0.7082	0.8053	0.7526

Validation

Para validar estatisticamente as diferenças entre os modelos foi aplicado o teste de Friedman, que indicou que havia uma diferença estatística entre os modelos.

Com esta confirmação foi realizado o teste de Nemenyi para que pudéssemos visualizar as distancias estatísticas entre os modelos.