MBA USP ESALQ

Previsão do Desempenho Acadêmico de Estudantes do Ensino Superior Utilizando Técnicas de Aprendizado de Máquina

Alan Marques da Rocha
Caroline Belisário Zorzal

SUMÁRIO

Introdução

Introdução

- Desempenho acadêmico de alunos de cursos superiores;
- Desafios das instituições para lidar com diversos tipos de aprendizagem;
- Promover uma experiência de aprendizagem satisfatória.

Introdução

- Informações e coleta de dados;
- Busca de *insights;*
- Algoritmos de Aprendizado de Máquina;
- Redução da evasão e predição do desempenho acadêmico.

Material e Métodos

Dados

- "Predict Students' Dropout and Academic Success", Valoriza (2020);
- 17 cursos de graduação;
- 4424 instâncias;
- 36 atributos;
- 03 classes ("Dropout", "Graduate" e "Enrolled").

Dados

Tabela 01 – Exemplo de representação dos atributos na base de dados.

N°	Nome do Atributo	Tipo	Representação
1	Estado civil	Categórica	1 – solteiro 2 – casado 3 – viúvo 4 – divorciado 5 – união estável 6 – separado judicialmente
7	Qualificação anterior (grau)	contínua	Nota da qualificação anterior (entre 0 e 200)

Dados

Tabela 02 – Quantitativo de atributos em cada classe da base de dados.

Classe	**“Graduate”**	**“Dropout”**	**“Enrolled”**
Número de atributos	2209	1421	794
	49,9%	32,1%	17,9%
Representação	1	-1	0

Algoritmo para Desbalanceamento das Classes

- "Synthetic Minority Over-sampling Technique (SMOTE)" (Chawla e Bowyer, 2002);
- Encontra exemplos de vizinhos da classe minoritária no espaço de atributos, sintetizando um novo exemplo no espaço entre os seus vizinhos.

SMOTE

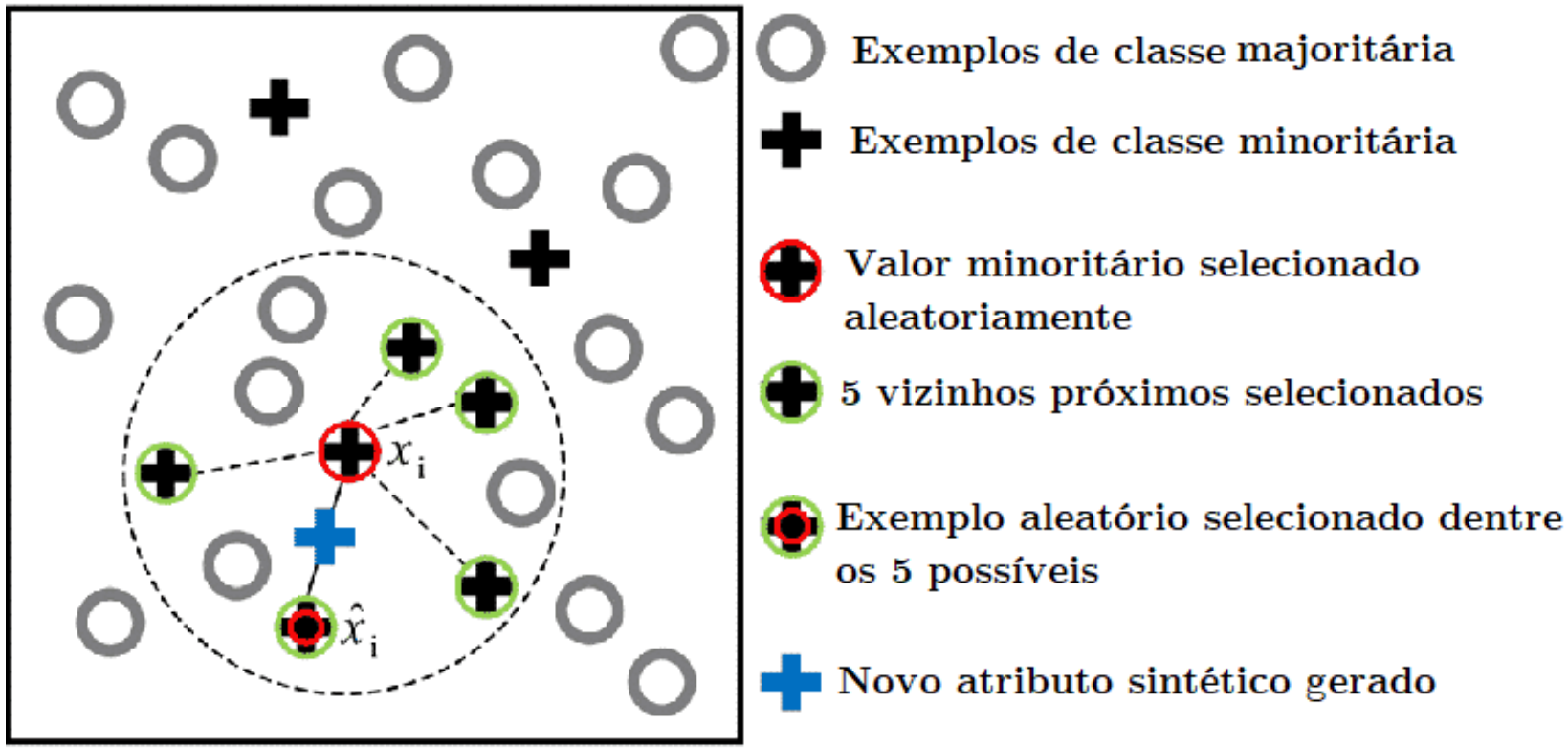


Figura 1. Geração de novos atributos sintéticos.

Algoritmos de Aprendizado de Máquina

- Rede Neural Artificial do tipo Perceptron Multicamadas (RNA-MLP);
- Decision Tree (DT);
- Random Forest (RF).

Algoritmos de Aprendizado de Máquina

- Baixo custo computacional;
- Fácil implementação;
- Hiperparâmetros definidos através de uma busca otimizada;
- MATLAB (Versão acadêmica R2022a).

Técnica de Validação dos Modelos

Dois experimentos realizados:

- 1° - Divisão do conjunto de treinamento e teste em 80% e 20%, respectivamente, após a geração de dados sintéticos com a técnica SMOTE.

Técnica de Validação dos Modelos

2° - Validação cruzada estratificada ("Stratified k-fold"):

- Leva em consideração a distribuição das classes durante a divisão dos "folds";
- Útil quando o conjunto de dados apresenta um desbalanceamento significativo entre as classes.

Métricas de Avaliação

Os algoritmos foram avaliados através das métricas:

- Acurácia:

$$acc = \frac{VP_G + VP_D + VP_E}{S_T}$$

$$S_T = VP_G + \frac{FP_G}{D} + \frac{FP_G}{E} + VP_D + \frac{FP_D}{G} + \frac{FP_D}{E} + VP_E + \frac{FP_E}{G} + \frac{FP_E}{D}$$

Métricas de Avaliação

Os algoritmos foram avaliados através das métricas:

- F-Score (F1):

$$F1_G = \frac{2 \times P_G \times Ss_G}{Ss_G + P_G}$$

$$Ss_G = \frac{VP_G}{VP_G + \frac{FP_D}{G} + \frac{FP_E}{G}}$$

$$P_G = \frac{VP_G}{VP_G + \frac{FP_G}{D} + \frac{FP_G}{E}}$$

Resultados e Discussão

Padronização dos Dados

- Garante que todas as variáveis estejam na mesma escala, evitando que algumas variáveis com valores mais altos dominem as outras durante o treinamento do modelo;
- Z-Score.

Padronização dos Dados

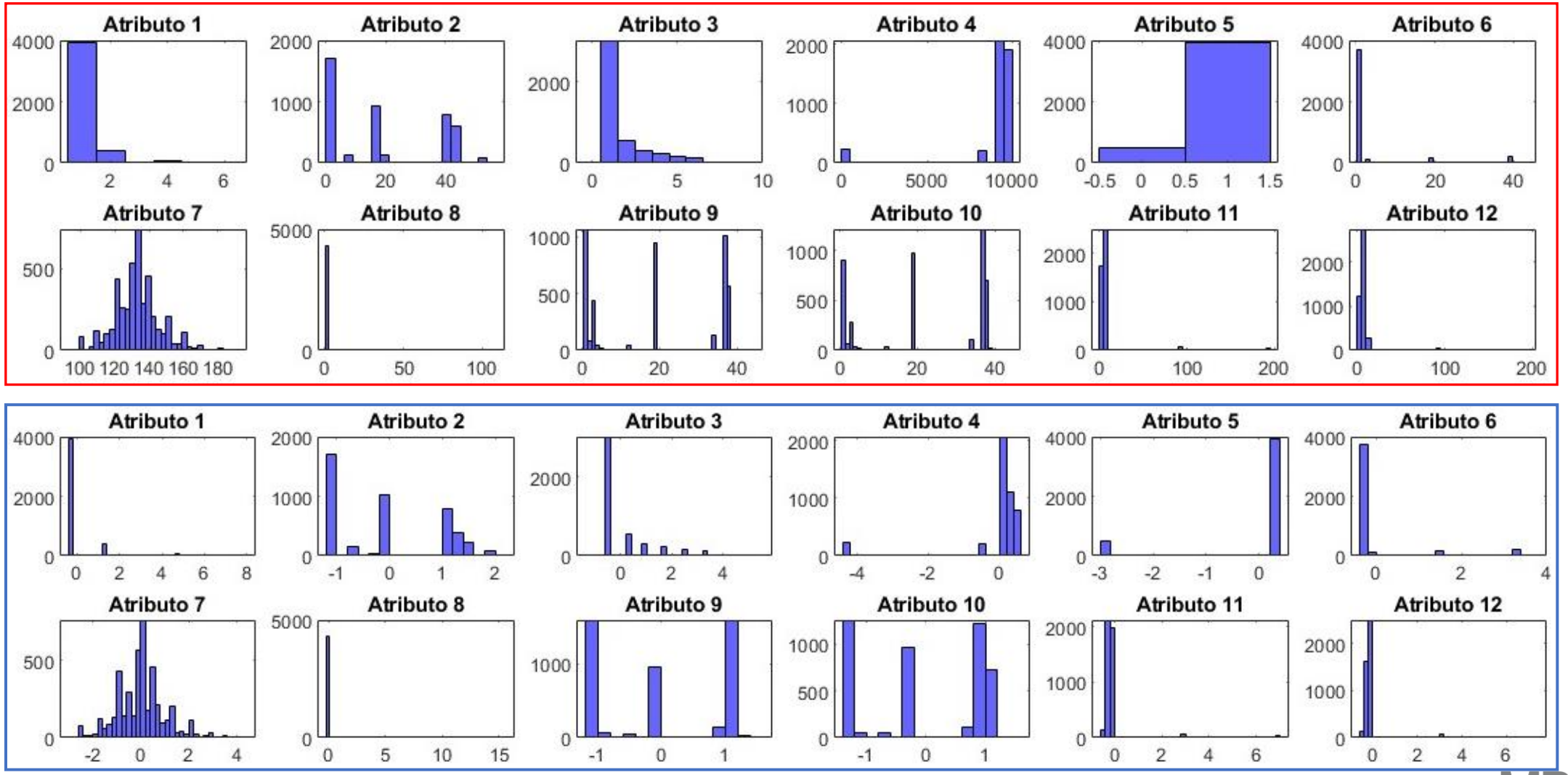


Figura 2. Histograma das frequências dos valores de cada atributo.

Balanceamento das Classes com SMOTE

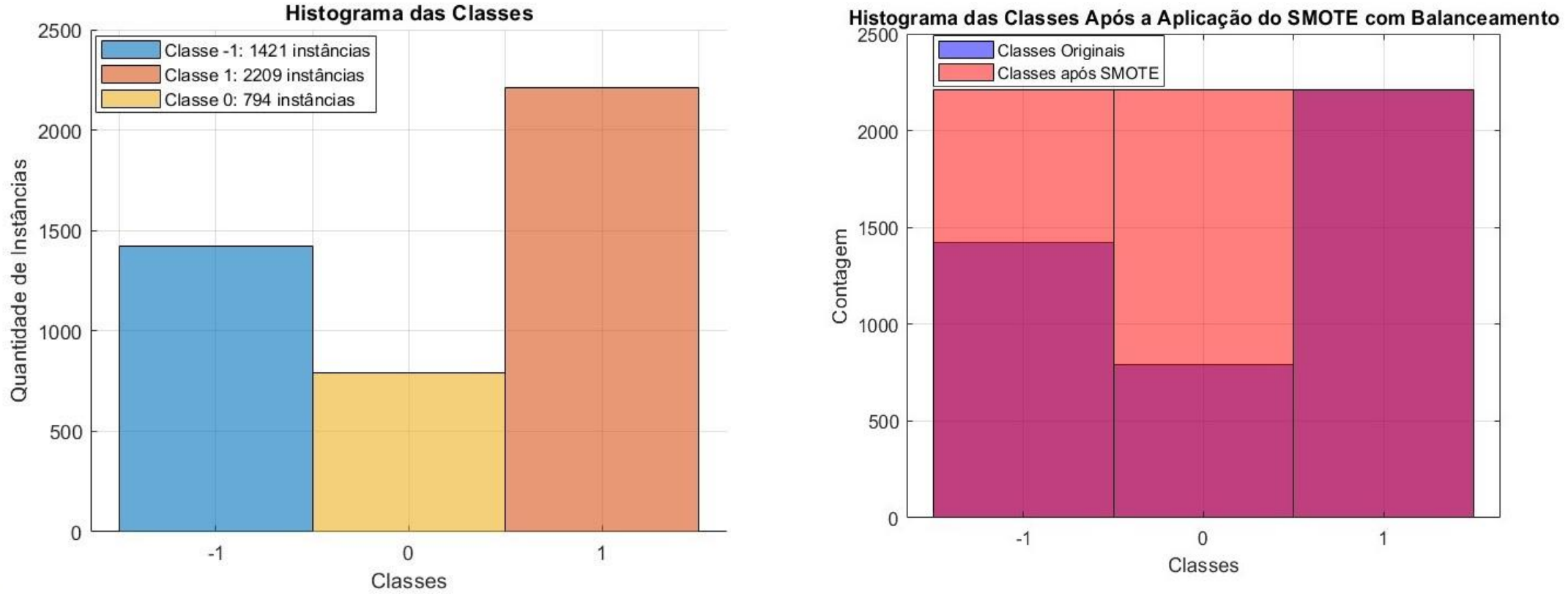


Figura 3. Histograma do número de instâncias das classes antes e depois do balanceamento.

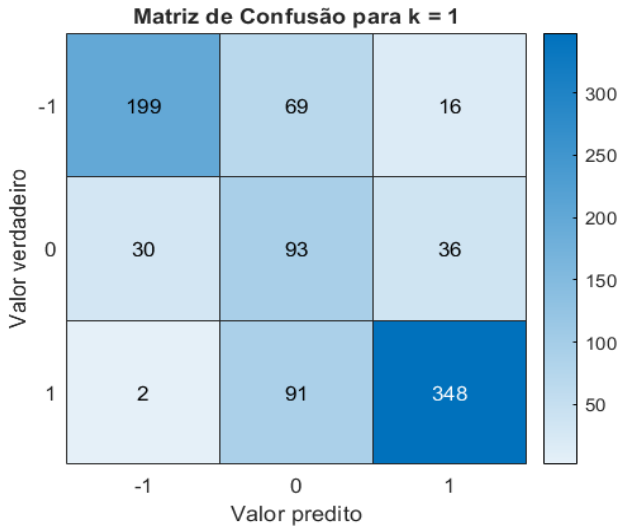
Validação Cruzada Estratificada

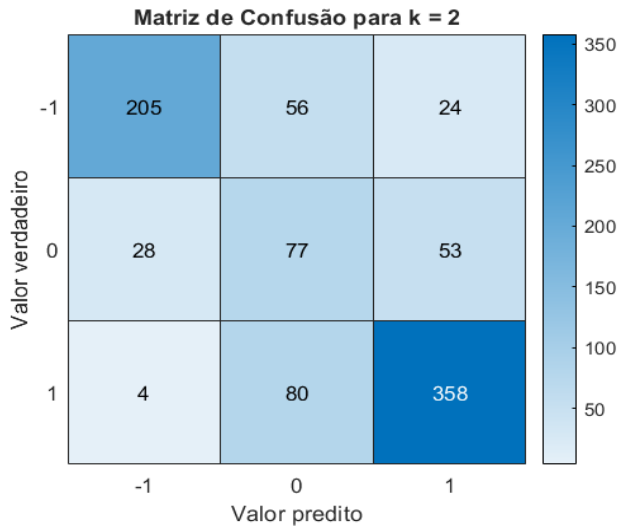
Tabela 03 – Média das métricas dos modelos sem balanceamento com validação cruzada estratificada.

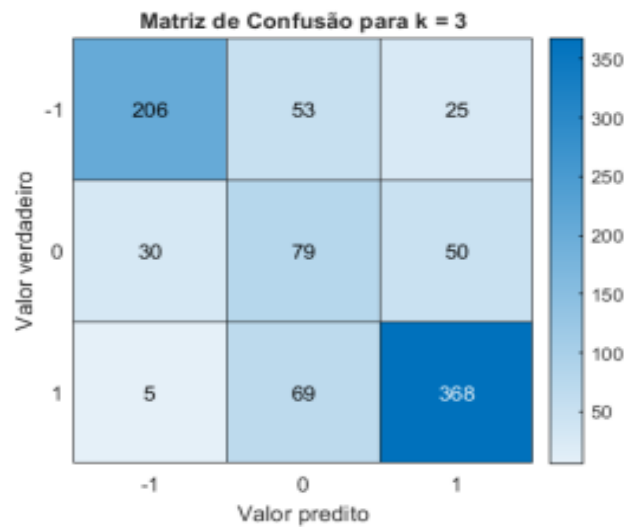
Modelo	**Acurácia**	**F1**
RNA-MLP	73,15%	85,87%
Random Forest	78,03%	67,30%
Decision Tree	70,30%	80,12%

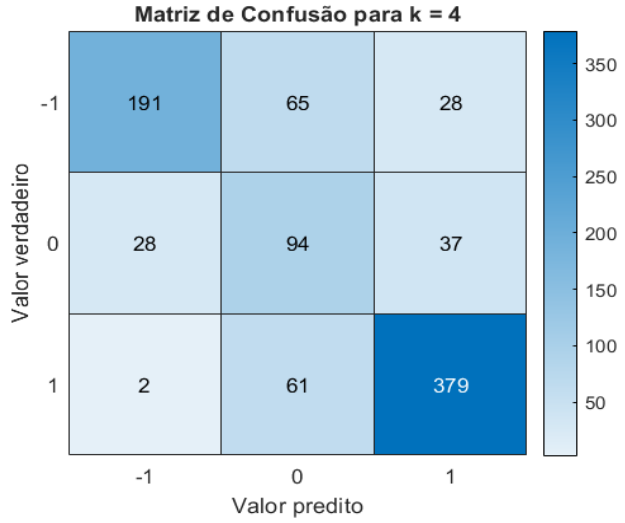
22

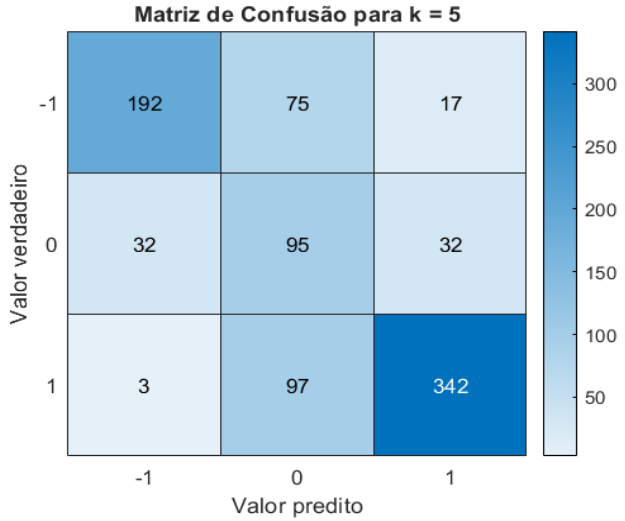
Matriz de Confusão do Algoritmo RNA-MLP





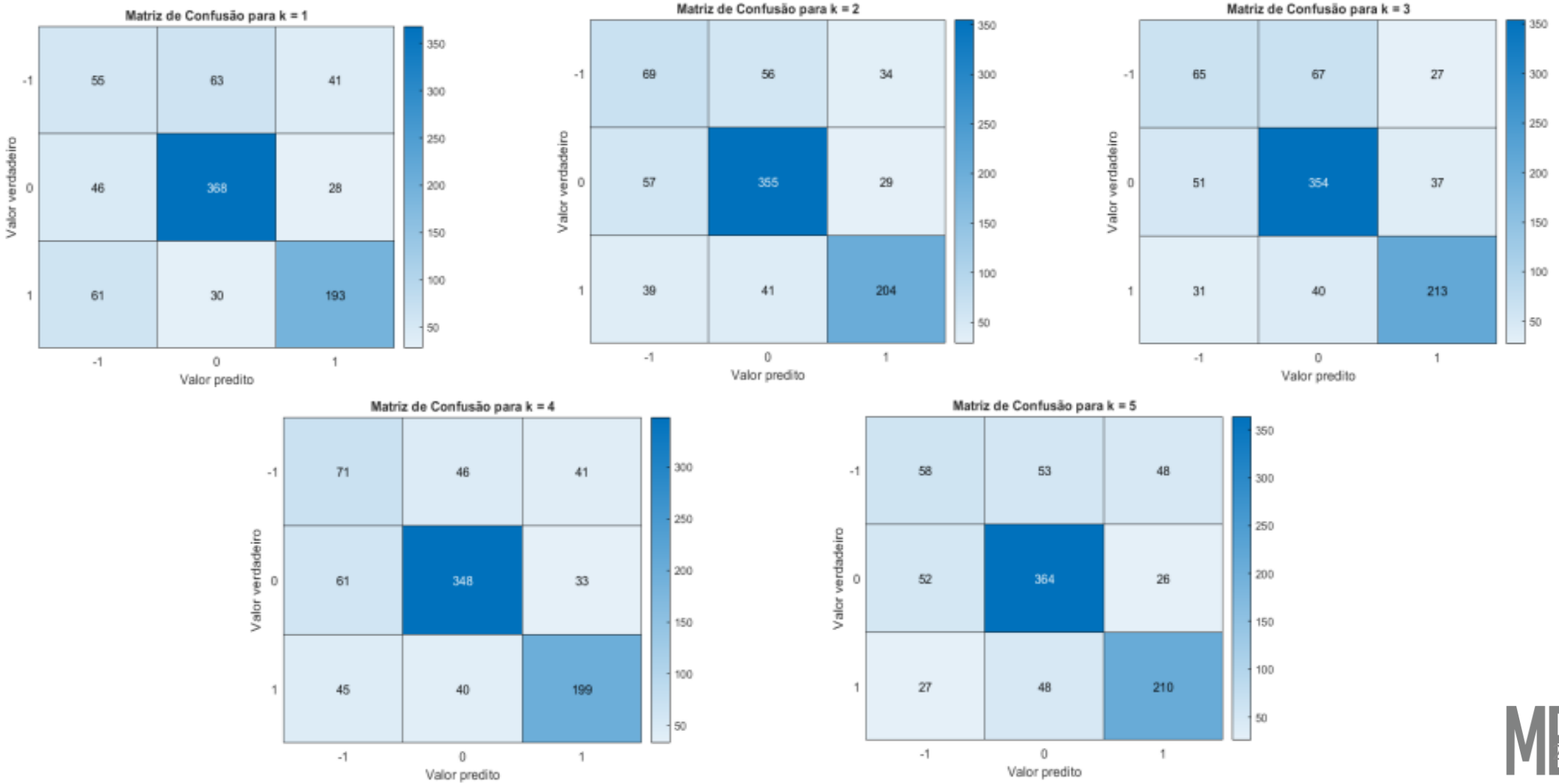






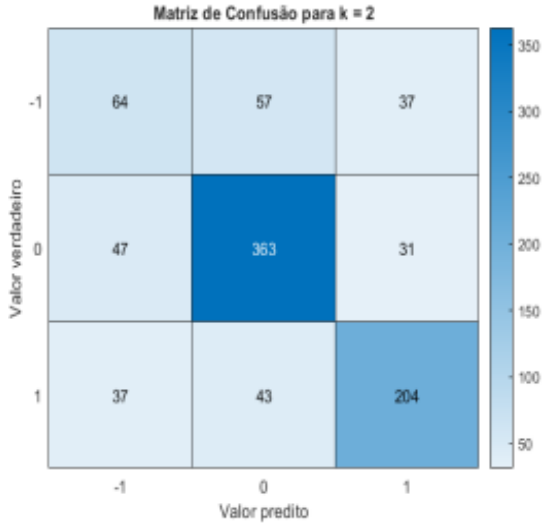
23

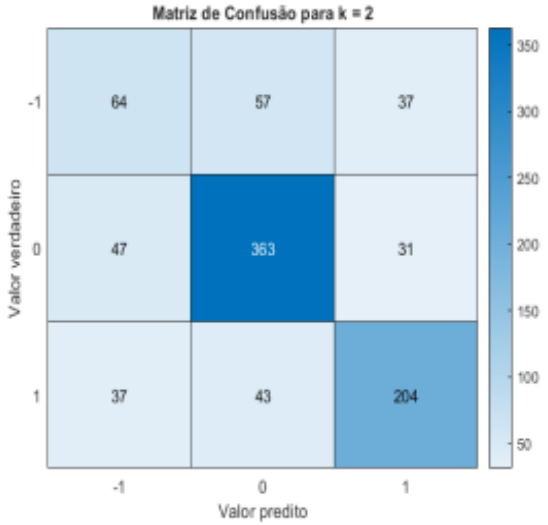
Matriz de Confusão do Algoritmo Decision Tree

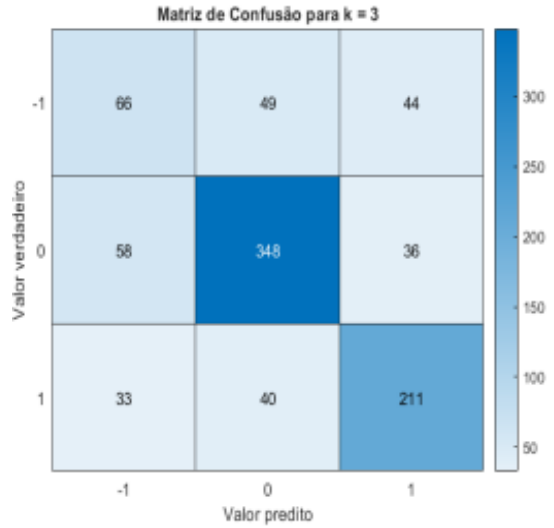


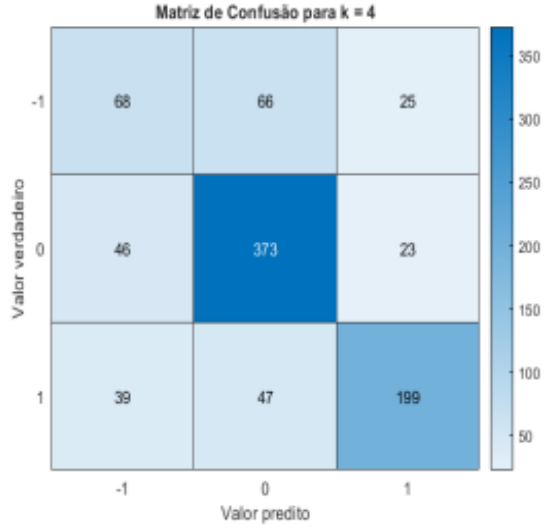
24

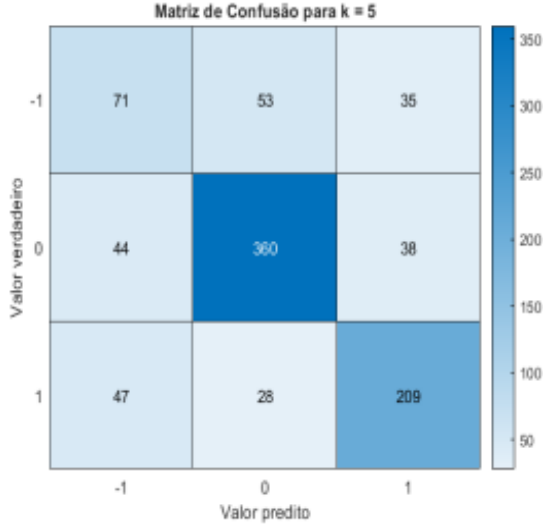
Matriz de Confusão do Algoritmo Random Forest









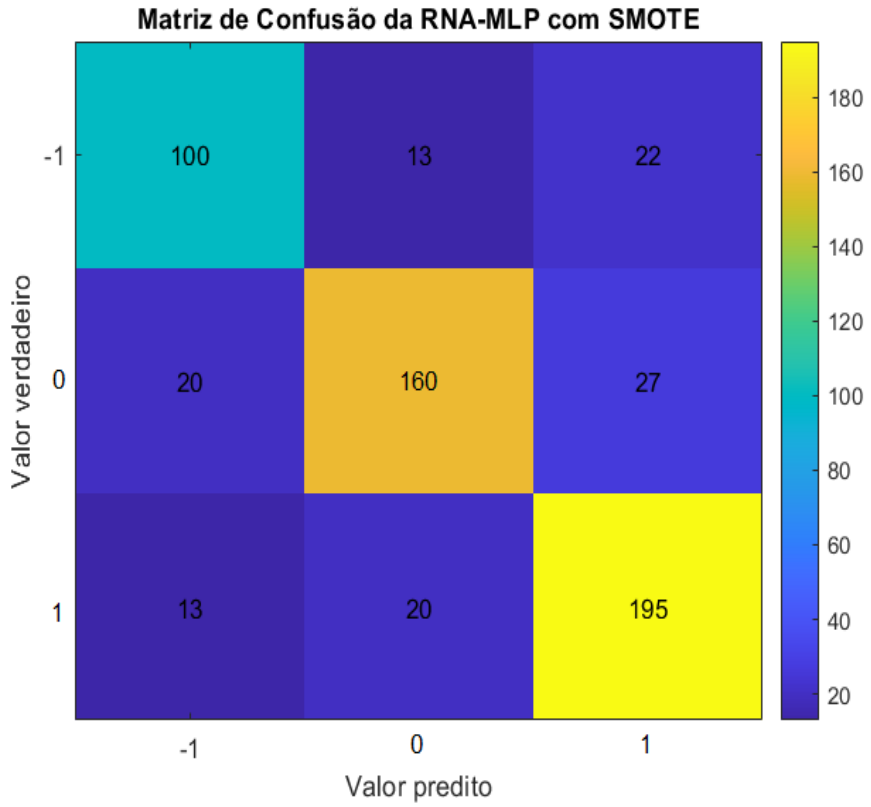


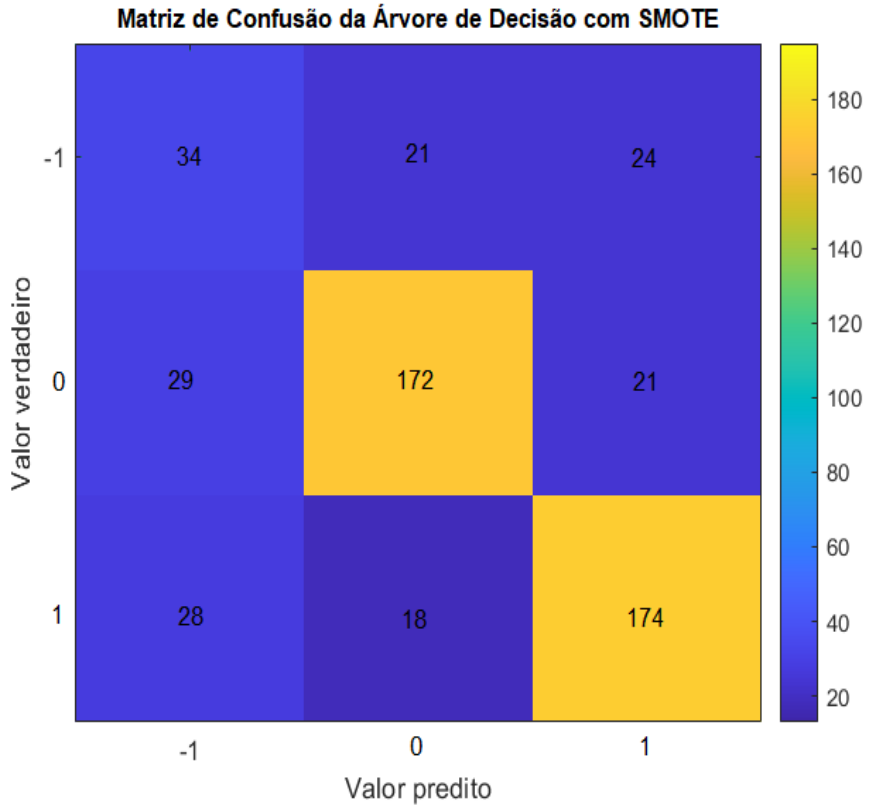
Validação com SMOTE

Tabela 04 – Métricas dos modelos com balanceamento das classes através da técnica SMOTE.

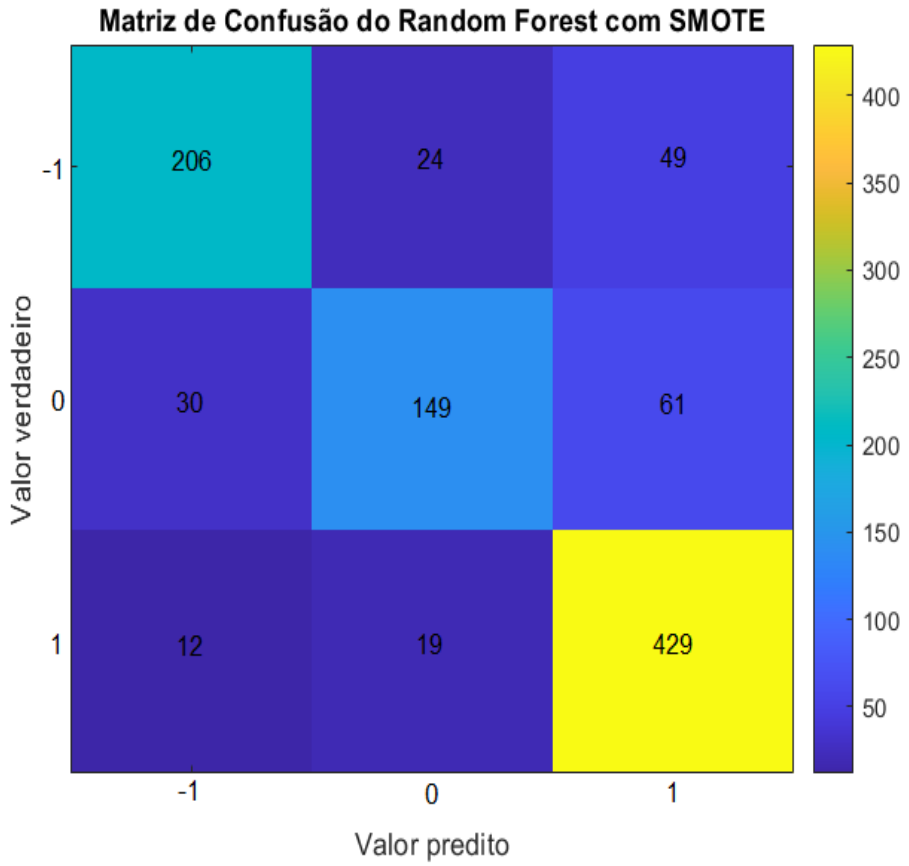
Modelo	**Acurácia**	**F1**
RNA-MLP	80,42%	87,60%
Random Forest	80,68%	87,78%
Decision Tree	74,04%	83,26%

Validação Utilizando os Dados Balanceados com a Técnica SMOTE





Validação Utilizando os Dados Balanceados com a Técnica SMOTE



28

Considerações Finais

Considerações Finais

- Foram investigados a utilização de algoritmos de aprendizado de máquina para predizer o desempenho acadêmico de alunos de cursos superiores;
- Dois experimentos foram investigados utilizando-se a validação cruzada estratificada e a técnica SMOTE para gerar dados sintéticos;

30

Considerações Finais

- Verifica-se que o RF apresentou o melhor desempenho global, obtendo a maior acurácia média, quando comparado com o RNA-MLP e DT;
- A técnica SMOTE obteve uma melhora significativa da acurácia em ambos os modelos.

31

MBA
USP
ESALQ

Obrigado!

eng.alanmarquesrocha@gmail.com