TRABALHO INTELIGÊNCIA COMPUTACIONAL - COC 786

USO DE INTELIGÊNCIA ARTIFICIAL PARA CLASSIFICAÇÃO DO PADRÃO DE COMPRAS DE CLIENTES BANCÁRIOS

Amanda Isabela de Campos

Prof. Alexandre G. Evsukoff

USO DE INTELIGÊNCIA ARTIFICIAL PARA CLASSIFICAÇÃO DO PADRÃO DE COMPRAS DE CLIENTES BANCÁRIOS

- **Apresentação do problema**
- *Bank Marketing Data Set*
  - Problema de Classificação (classificar se um determinado cliente compra ou não o depósito a prazo ofertado por ligações);
  - 17 atributos para 11162 registros;
  - Tipo de variáveis (numéricas e categóricas);
  - Variável de saída: Binária (O cliente efetuou um depósito a prazo? 'sim' ou 'não');
  - Problema "balanceado".

Caracterização e Visualização de Dados

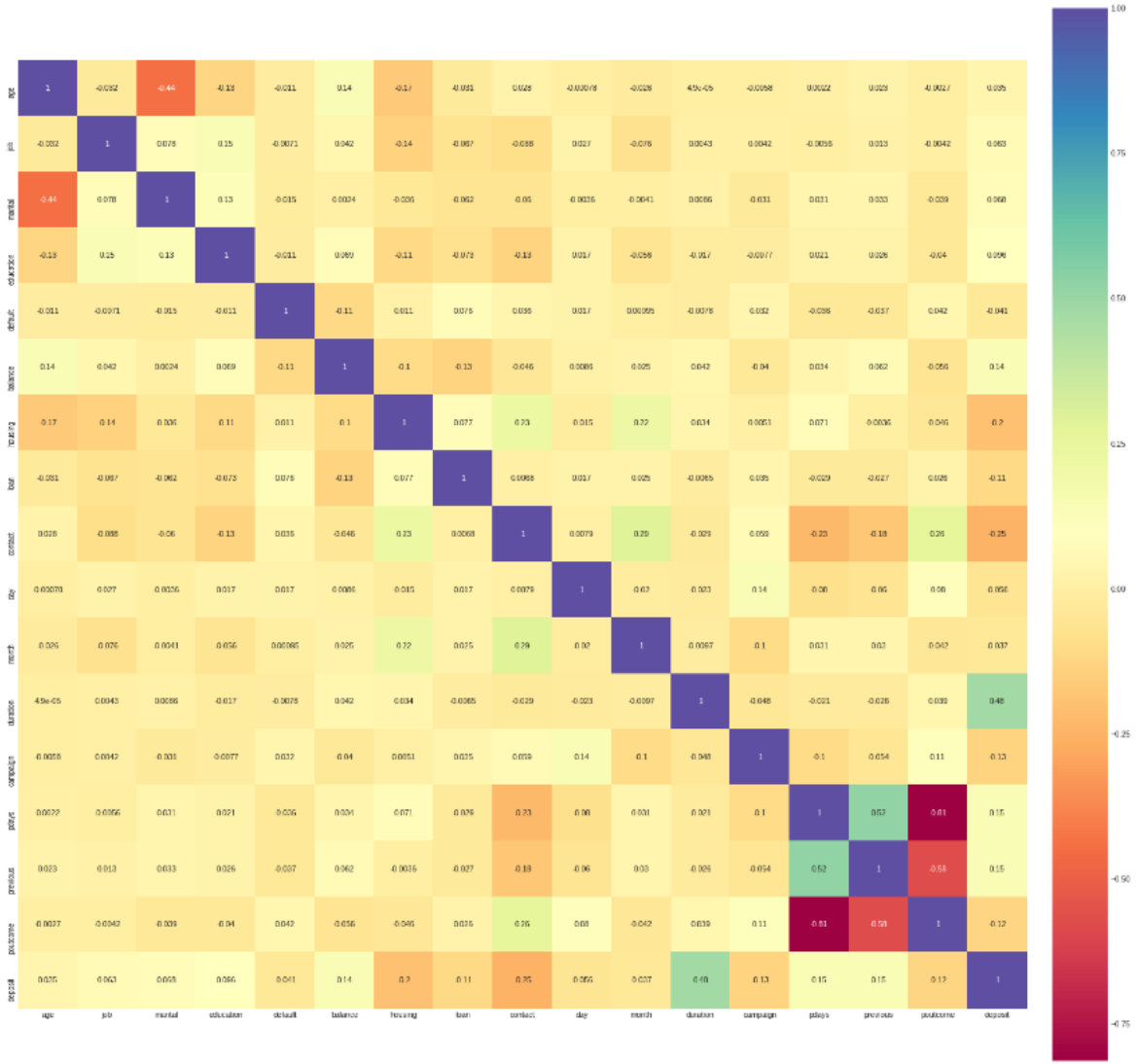


Figura 1. Matriz de correlação com mapa de cores

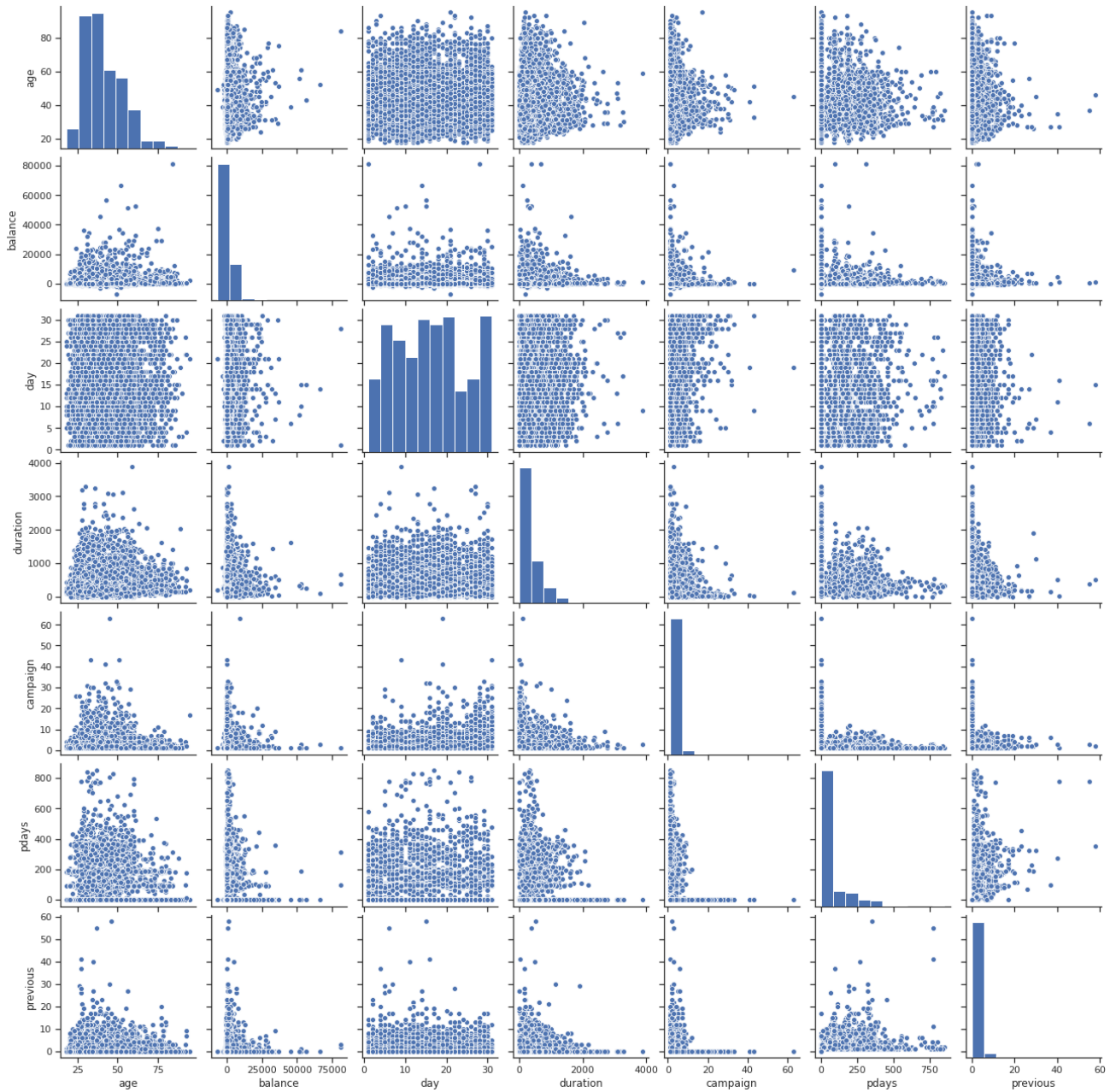


Figura 2. Gráfico de Projeção dos atributos numéricos

•Pré-processamento

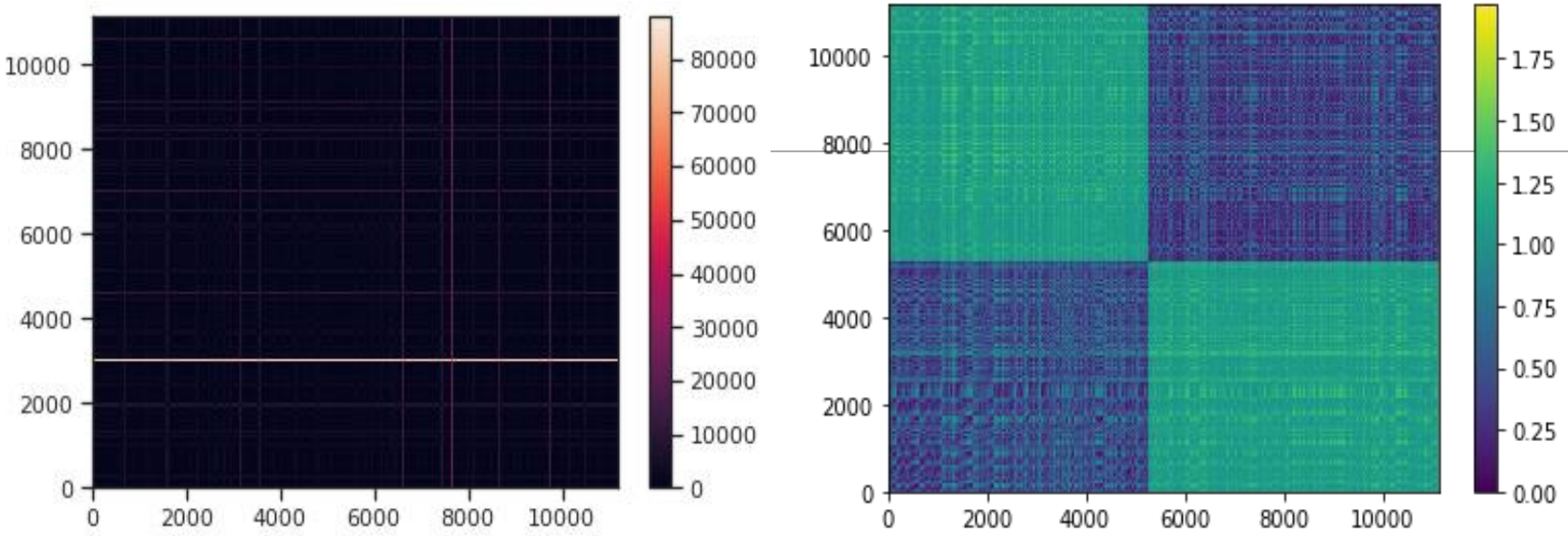


Figura 3. Matriz de distâncias (a) com os dados não padronizados e (b) com os dados padronizados

- Padronização das variáveis: *MinMaxScaler* [0,1]
- Não existem valores ausentes.

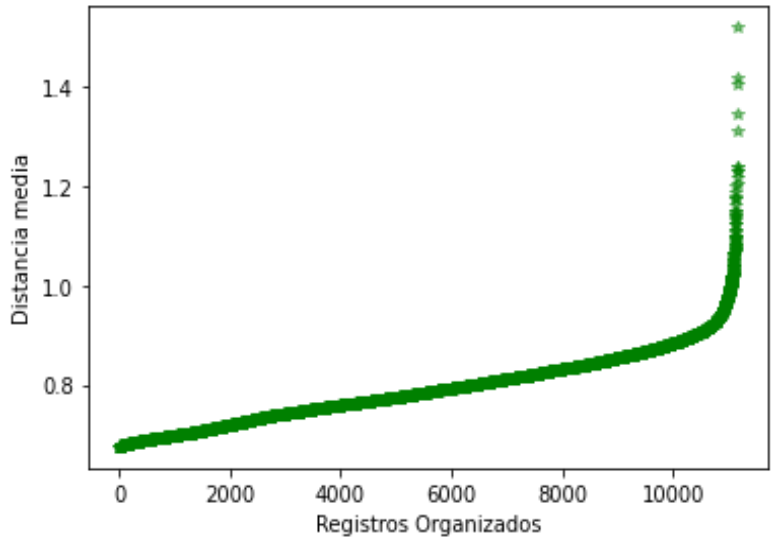


Figura 4. Detecção de outliers a partir da distância

•Modelos de classificação adotados

- **Naive Bayes** $P(C_i|\mathbf{x}) = \frac{p(\boldsymbol{x}|C_i)P(C_i)}{p(\boldsymbol{x})}$

- **Regressão Logística** $J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(\hat{p}^{(i)}\right) + \left(1-y^{(i)}\right)\log\left(1-\hat{p}^{(i)}\right)\right]$

- **SVM**

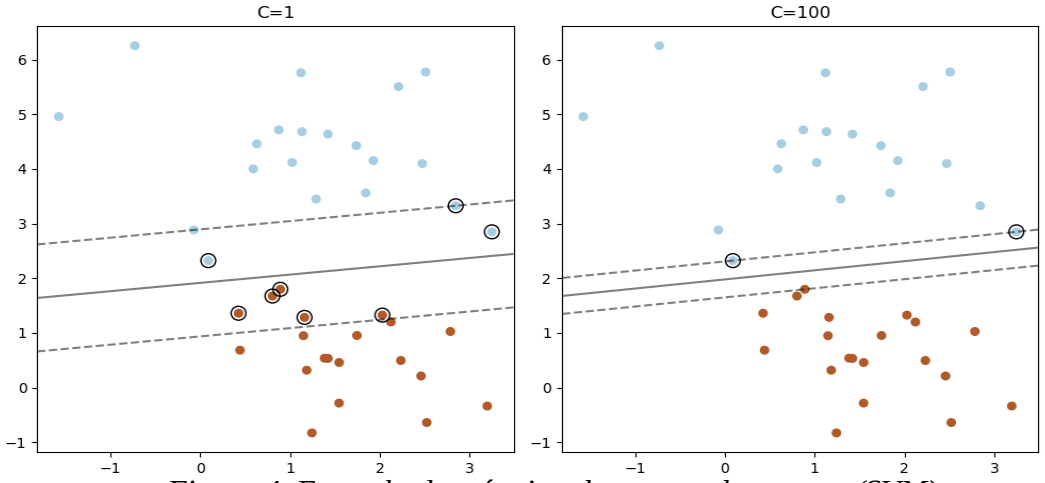


Figura 4. Exemplo de máquina de vetores de suporte (SVM)

•Modelos de classificação adotados

- KNN
- Árvore de decisão
- Gradient Boosting
- Floresta aleatória

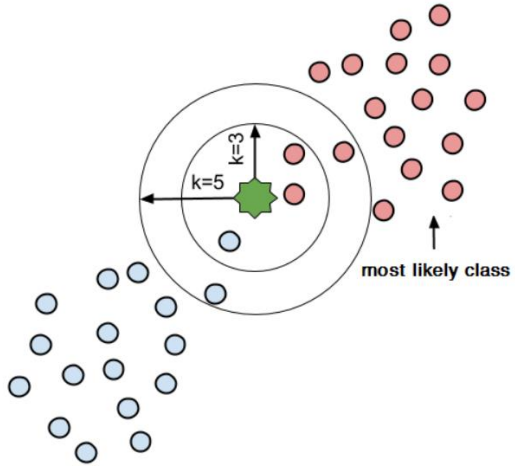


Figura 5. Exemplo de classificação KNN para um problema de duas classes

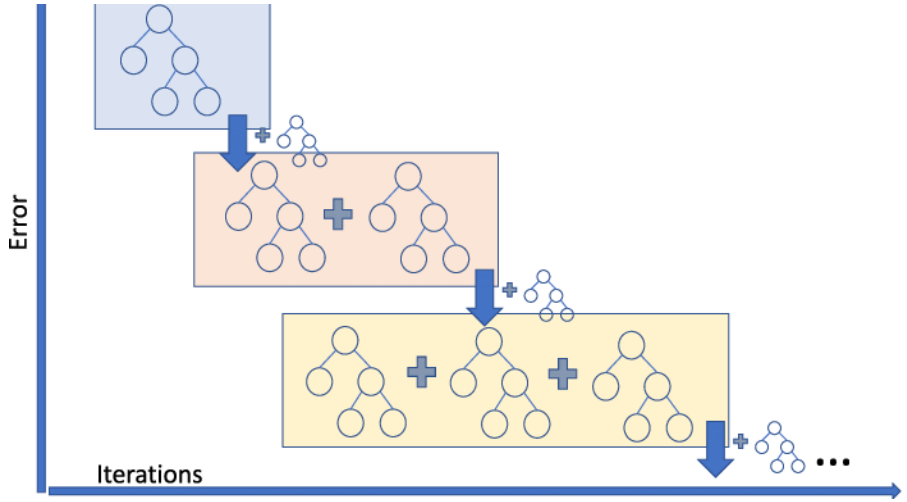


Figura 6. Exemplo de Gradient Boosting

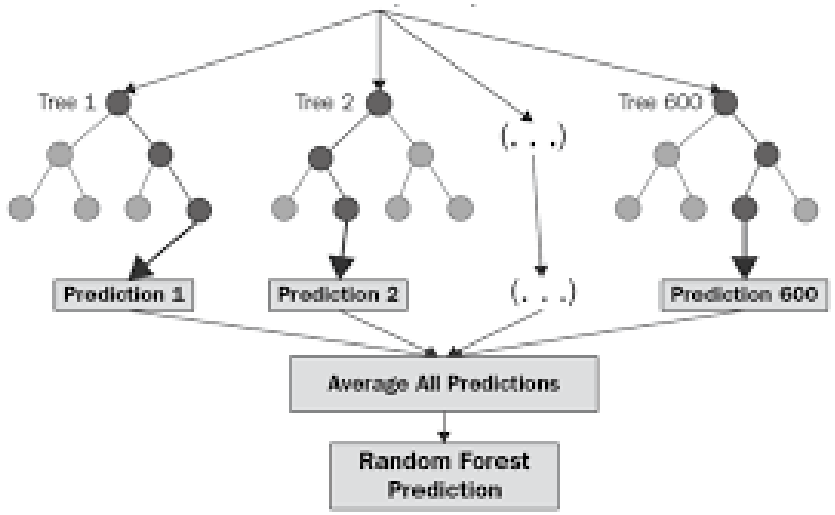


Figura 7. Exemplo de Floresta aleatória

•Metodologia

- **Validação cruzada (cv = 10)**
- ***OneHotEncoder***
- **Avaliação: Recall, Precisão, F_1, Acurácia e AUC**

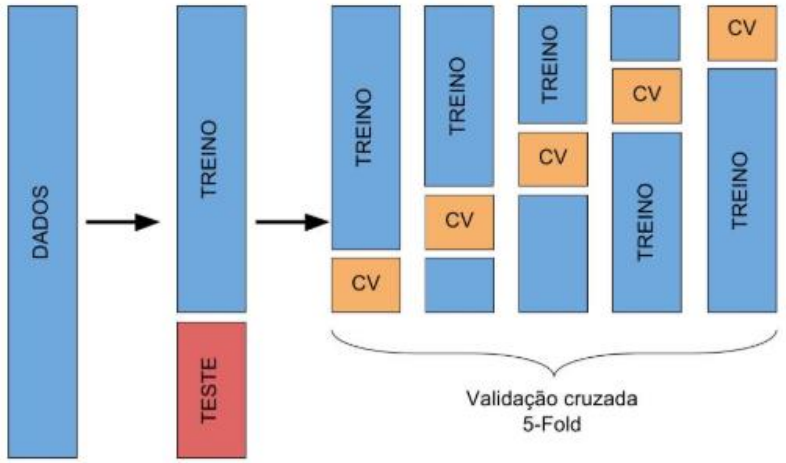


Figura 8. Técnica de validação cruzada

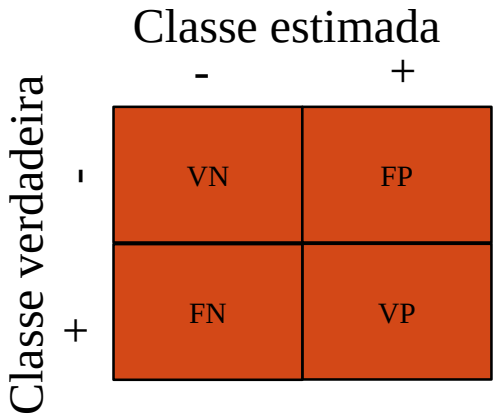


Figura 9. Matriz de confusão

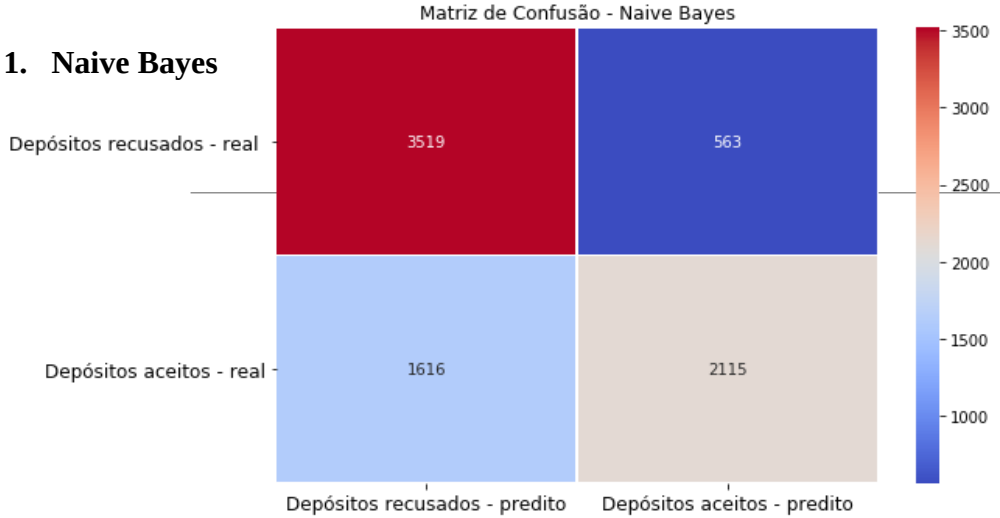
$$precisão = \frac{VP}{VP+FP} \qquad (1)$$

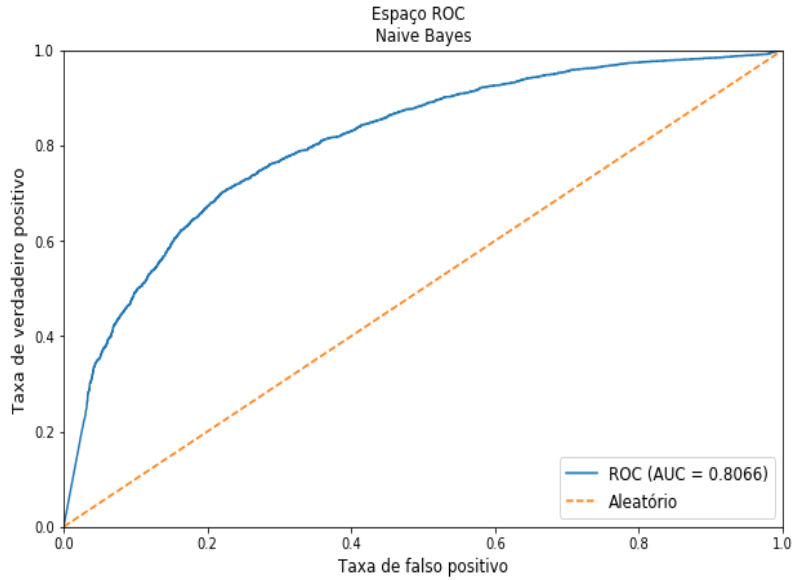
$$revocação = \frac{VP}{VP+FN} \qquad (2)$$

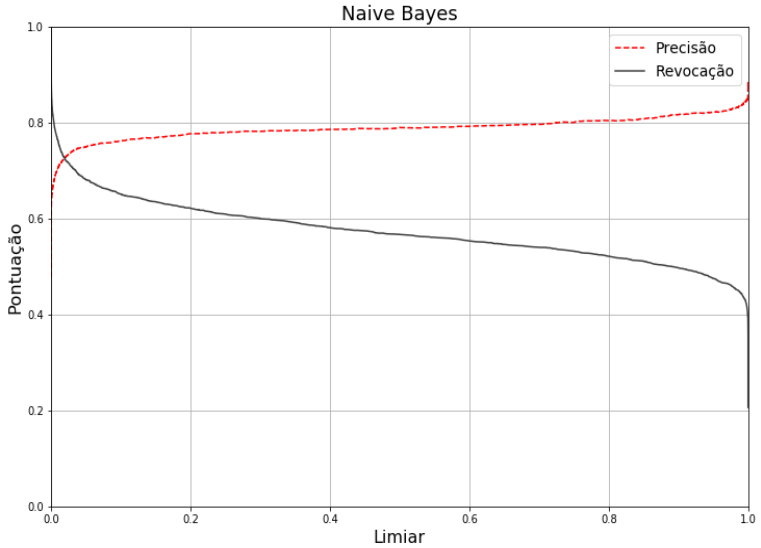
$$F_1 = \frac{2}{\frac{1}{precisão}+\frac{1}{revocação}} = \frac{VP}{VP+\frac{FN+FP}{2}} \qquad (3)$$

•Resultados dos modelos lineares

1. **Naive Bayes**

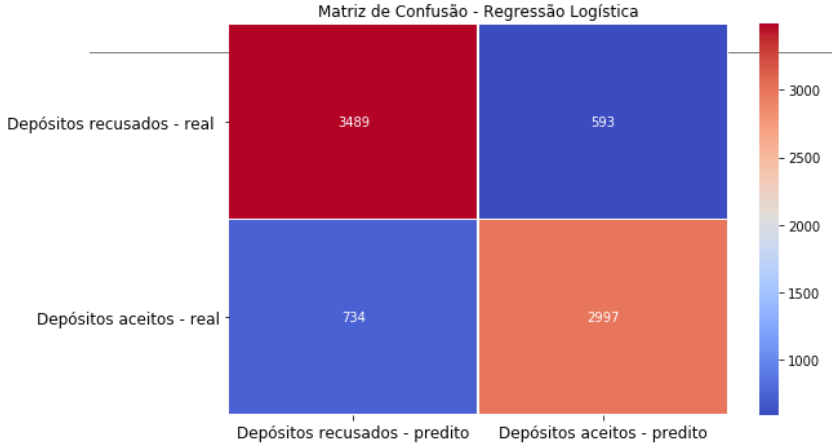


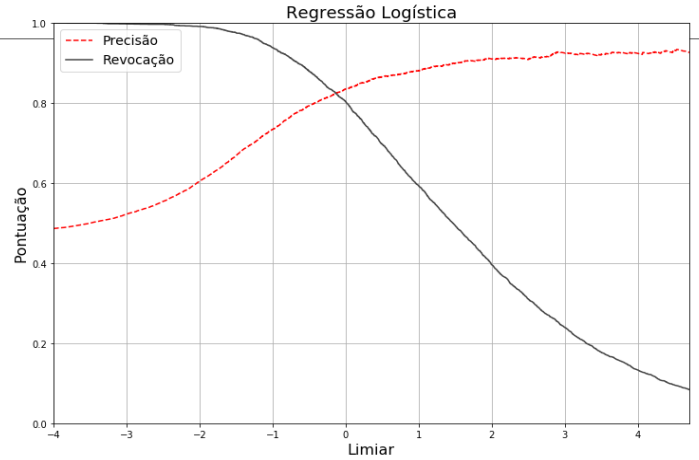


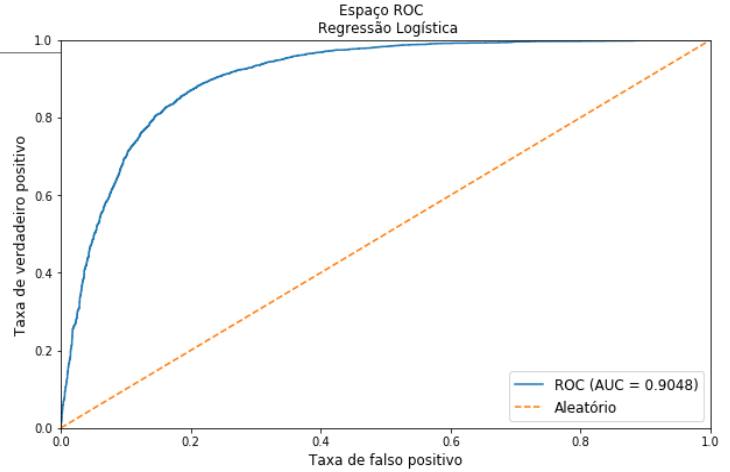


•Resultados dos modelos lineares

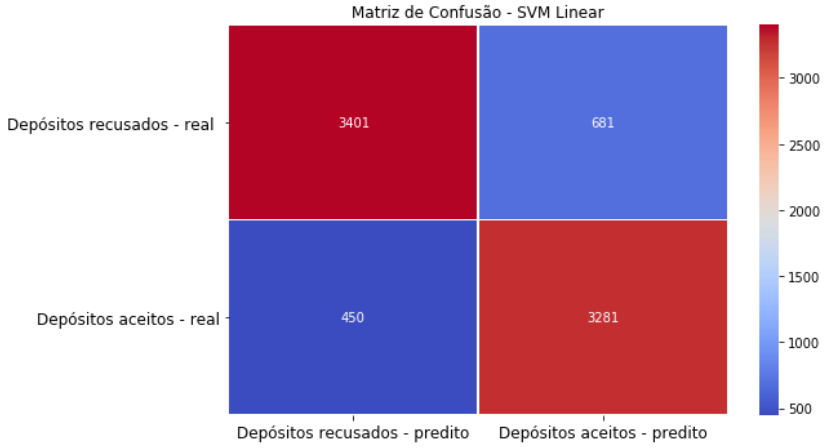
2. Regressão Logística

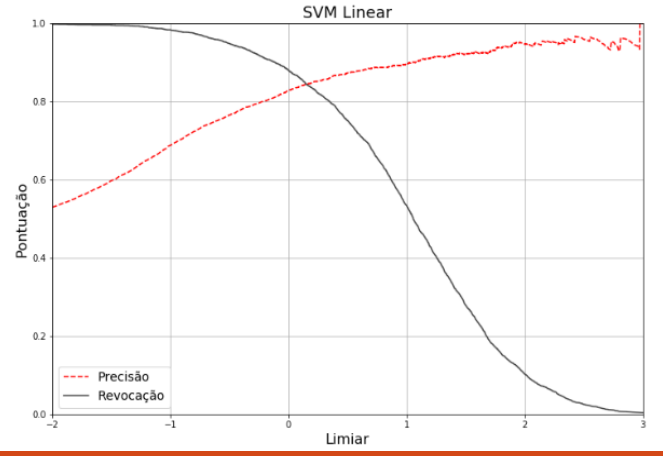


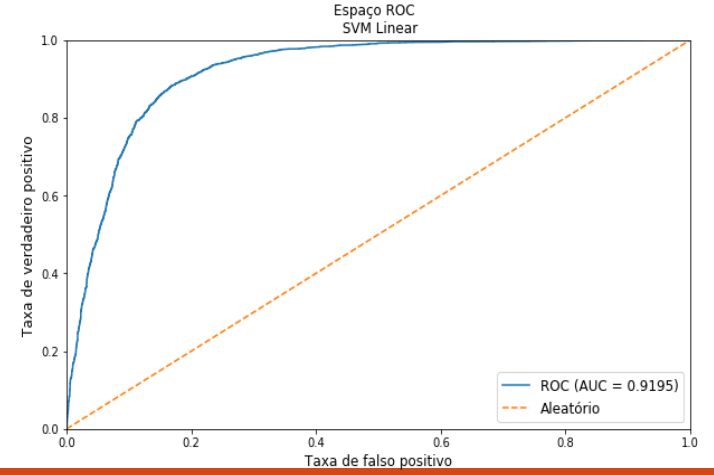




3. SVM

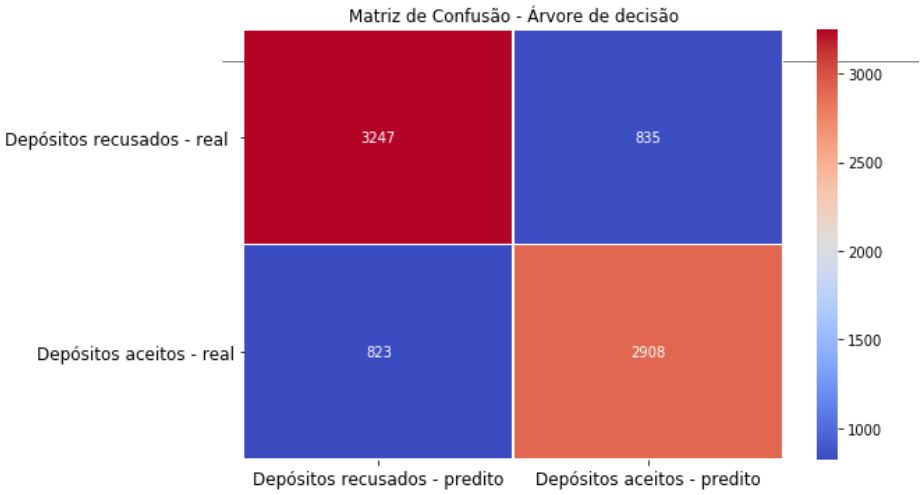


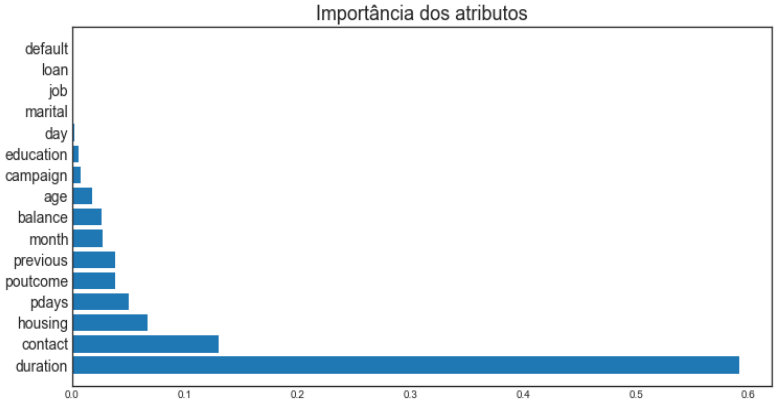




•Resultados dos modelos não- lineares

1. Árvore de decisão





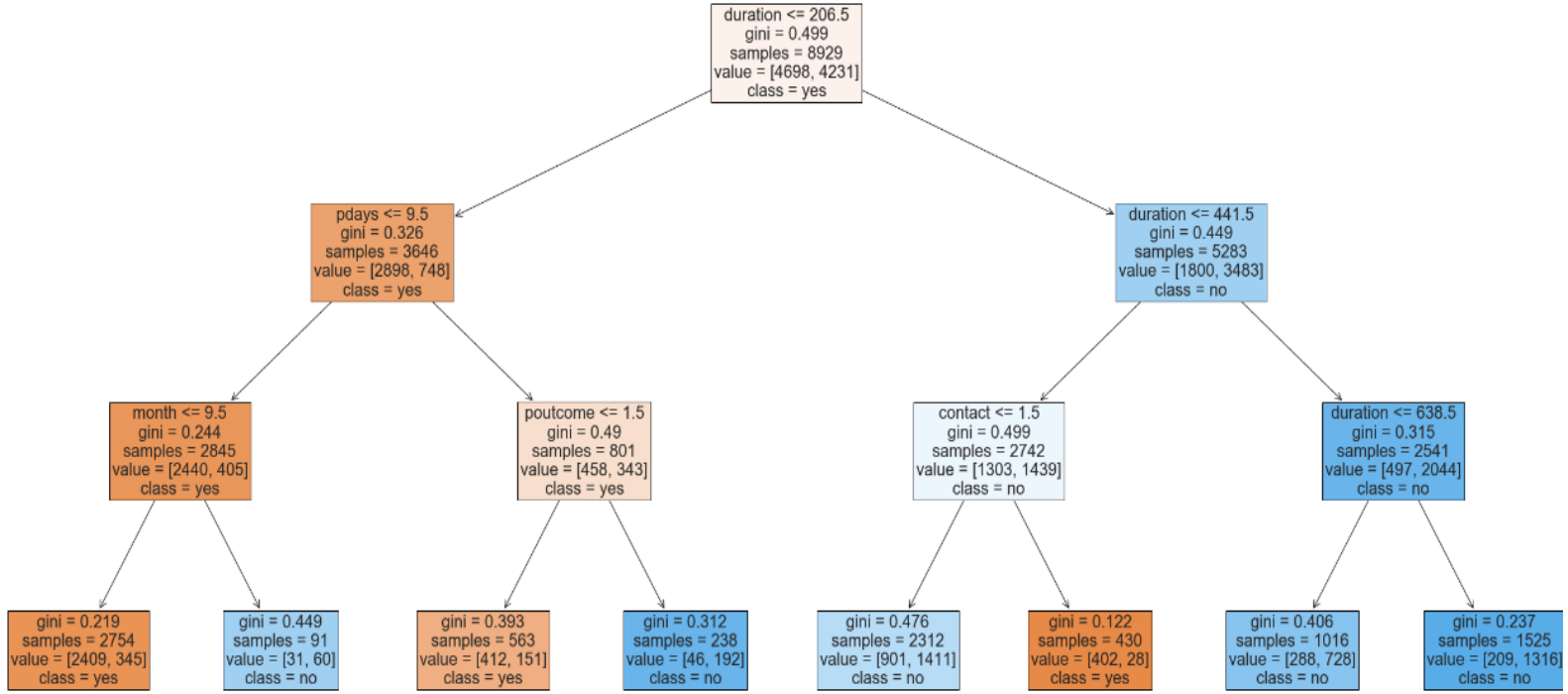
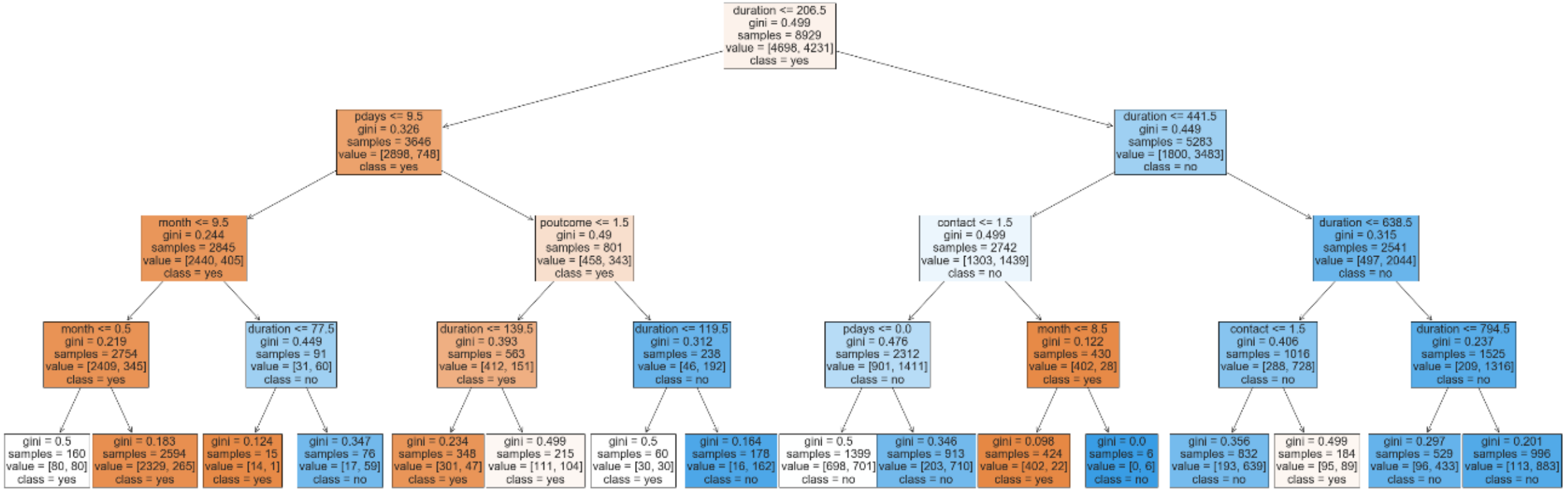


Figura 10. Importância dos atributos no modelo de Árvore de decisão.

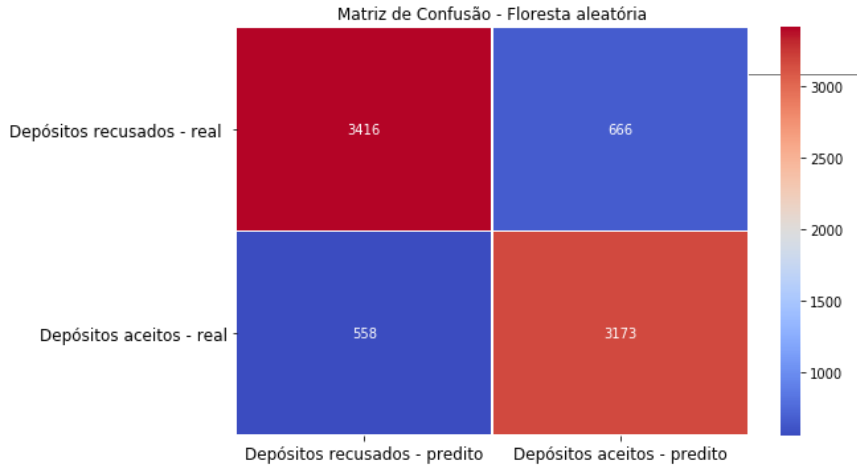
•Resultados dos modelos não- lineares

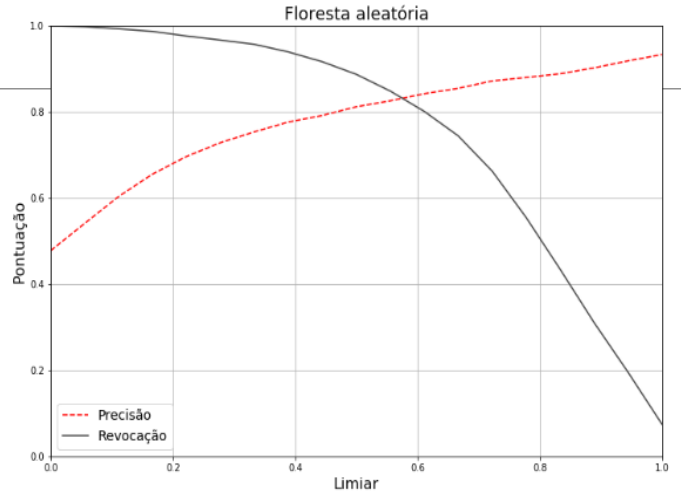
1. Árvore de decisão

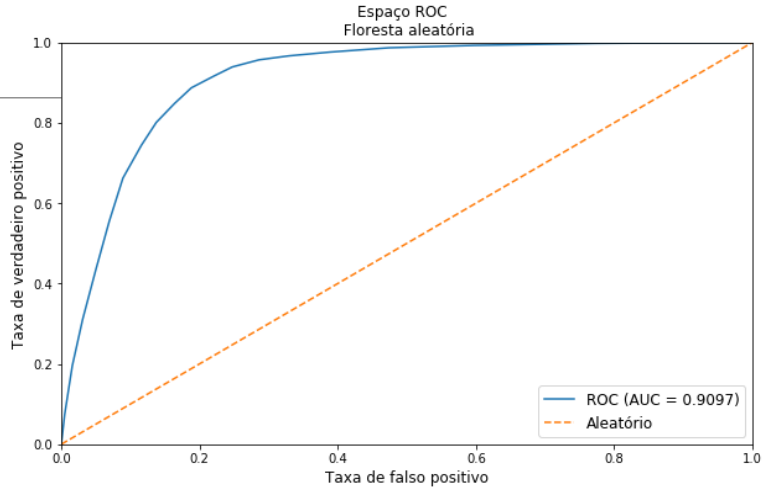


•Resultados dos modelos não lineares

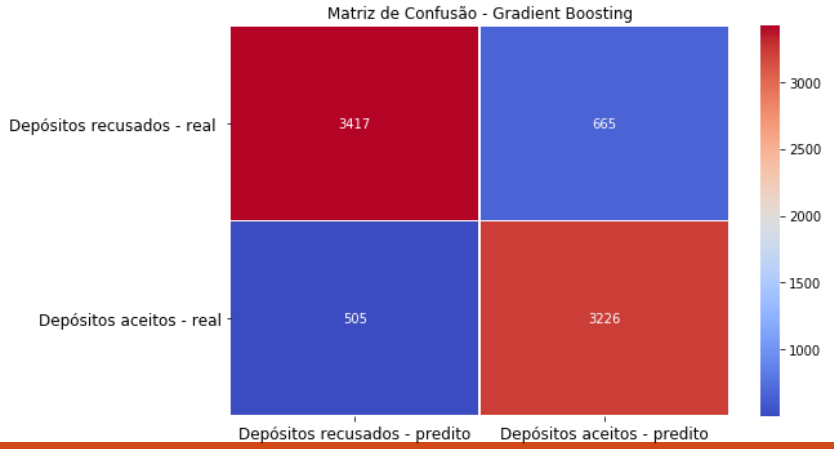
2. Floresta aleatória

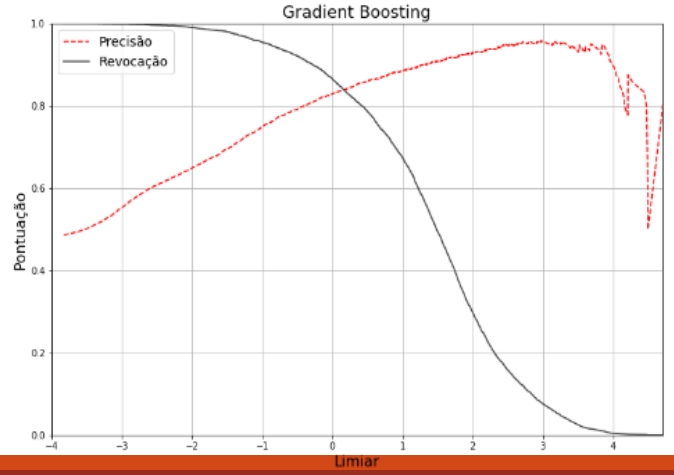


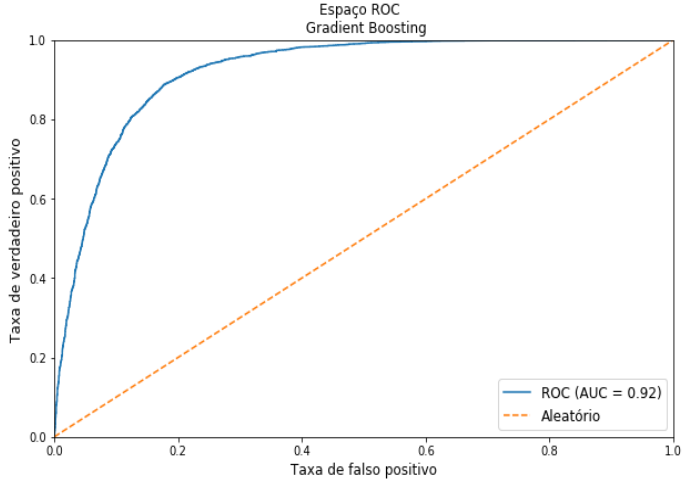




3. Gradient Boosting







- **Comparação dos resultados**

Tabela 1. Comparação dos resultados de todos os modelos

Modelo	Recall	Precisão	F1	Acurácia	AUC
Regressão Logística	0.8033	0.8348	0.8187	0.8302	0.9048
SVM	0.8794	0.8281	0.8530	0.8552	0.9195
KNN	0.7593	0.8122	0.7849	0.8012	0.9195
Gradient Boosting	0.8646	0.8291	0.8465	0.8505	0.9200
Árvore de decisão	0.7794	0.7769	0.7782	0.7878	0.7892
Floresta aleatória	0.8504	0.8265	0.8383	0.8433	0.9097
Naive Bayes	0.5669	0.7898	0.6601	0.7211	0.8066

Conclusões

- O pré-processamento tem papel fundamental na qualidade final dos resultados;
- O modelo indicado para este conjunto de dados é o **Gradient Boosting**;
- A complexidade do modelo não necessariamente está relacionada à um melhor resultado;
- Modelo de classificação linear SVM se mostrou um ótimo classificador para esse conjunto de dados.

Conclusões

- Recomenda-se um estudo mais aprofundado na busca dos parâmetros para cada modelo (*Grid Search*);

- Moro *et al*. (2014) analisou o conjunto de dados *(*regressão logística, árvores de decisão, redes neurais e SVM) e obteve AUC = 0,8 com redes neurais.

Referências

- Geron, A. "Mãos à Obra: Aprendizado de Máquina com Scikit-Learn TensorFlow." (2019).
- Granik, Mykhailo, and Volodymyr Mesyura. "Fake news detection using naive Bayes classifier." *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*. IEEE, 2017.
- Zhang, Harry. (2004). The Optimality of Naive Bayes. Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004. 2.)
- UCI, Machine Learning Repository. Bank Marketing Data Set. Disponível em: <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>. Acesso em: 10 de ago. de 2020.
- S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014.
- S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011.
- Evsukoff, Alexandre. Inteligência computacional : Fundamentos e aplicações [recurso eletrônico] / Alexandre Evsukoff. - 1. ed. – Rio de Janeiro : E-papers, 2020.
- MCKINNEY, Wes. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.", 2012.
- Géron, A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, 2019.

Obrigada.