Aprendizado de máquina aplicado em filtragem de spam em e-mails

Guilherme Henrique de Souza Nakahata

Universidade Estadual de Maringá
Aprendizagem de Máquina

01 de março de 2021

Overview

Introdução

- SPAM;
- Espaço nos servidores de e-mails;
- Responsável por 77 % de todo o tráfego do e-mail global;
- Saúde e namoro;
- Perdas financeiras.

Introdução

- Queda abaixo de 50 % desde 2003;
- Junho de 2015 - 49.7 %;
- Julho de 2015 - 46.4 %;
- BotNets.

Introdução

- Aumento a partir de 2015;
- Malware;
- Ransomware;
- Macros;
- Scripts maliciosos em Java Script.

Introdução

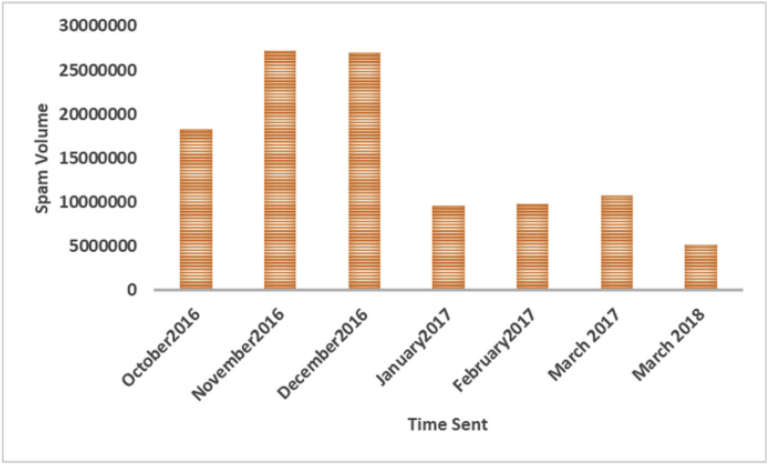

Figure: Volume de spams entre 2016 e 2018.

Introdução

- Provedores de e-mail;
- Combinações de diferentes técnicas (ML);
- Adaptação as variações de filtros;
- Regras pré-existentes;

Introdução

- Multinomial Naive Baies (MNB);
- Random Forests (RF);
- Support Vector Machine (SVM);
- Stacking;

Trabalhos Relacionados

- Diferentes técnicas;
- Probabilísticos;
- Arvores de Decisão;
- SVM;
- Redes Neurais Artificiais.

Trabalhos Relacionados

- Bo Yu e Zong-ben [5];
- Naive Bayes - 92%;
- SVM - 95.2%;
- Redes Neurais Artificiais - 85.3%;
- RVM - 96.1%.

Trabalhos Relacionados

- Karthika Renuka et al. [4];
- Multilayer perceptron (MLP);
- J48-classifier;
- Naive Bayes;
- Precision and Recall;
- Cross Validation.

Trabalhos Relacionados

- W.A Awad and S. M. Elseueofi [1];
- Naive Bayes;
- SVM;
- K-Nearest-Neighbours (KNN);
- Curva PR.

Trabalhos Relacionados

- Emmanuel Gbenga et al. [2];
- Revisão;
- Técnicas;
- Problemas não solucionados.

Trabalhos Relacionados

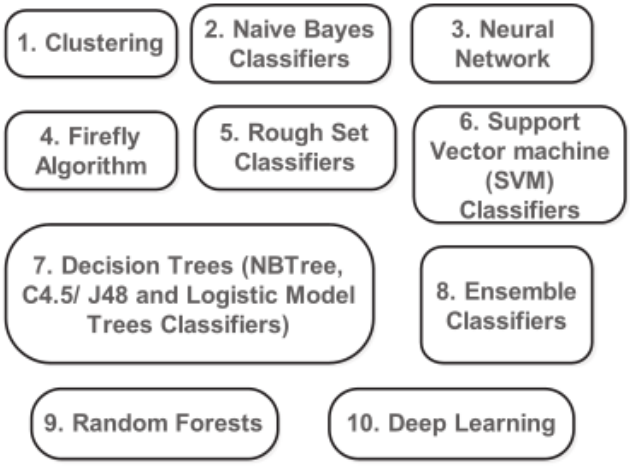

Figure: Principais técnicas utilizadas para classificação de spam [2].

Categorias de técnicas para filtros de spam

- Content Based Filtering Technique;
- Case Base Spam Filtering Method;
- Heuristic or Rule Based Spam Filtering Technique;
- Previous Likeness Based Spam Filtering Technique;
- Adaptive Spam Filtering Technique.

Support Vector Machine (SVM)

- Poderosa;
- Eficiente;
- Supervisionada;
- Kernel;
- Estado da Arte.

Multinomial Naive Bayes (MNB)

- Rápida convergência;
- Rápida classificação;
- Fácil implementação;
- Poucos dados de treinamento;
- Atributos discretos.

Random Forest (RF)

- Injeção de aleatoriedade;
- Árvores de decisões;
- Combinadas por votação;
- Aplicado a diferentes problemas.

Metodologia

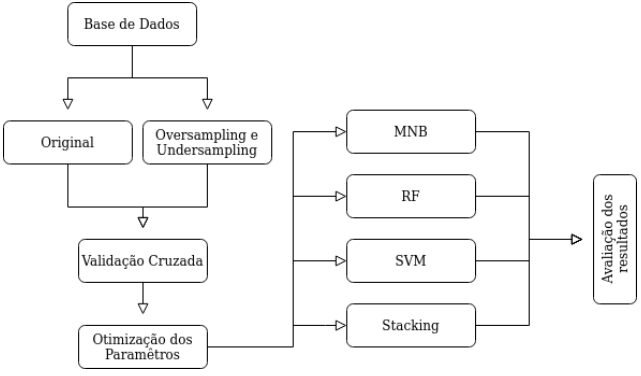

Figure: Etapas utilizadas.

Base de Dados

- Email Spam Classification Dataset CSV[1];
- 5172 e-mails;
- 1500 spam;
- 3672 não spam.

[1] https://www.kaggle.com/balaka18/email-spam-classification-dataset-csv

Base de Dados

- 5172 linhas;
- 3002 colunas;
- Palavras por e-mail;
- 1 spam;
- 0 não spam.

Email No.	the	to	ect	and	a	you	hou	label
Email 1	0	0	1	0	2	0	0	**0**
Email 2	8	13	24	6	102	1	27	**1**
Email 3	0	0	1	0	8	0	0	**0**
Email 4	0	5	22	0	51	2	10	**0**
Email 5	7	6	17	1	57	0	9	**1**

Figure: Exemplo base de dados.

Oversampling

- Synthetic Minority Oversampling Technique (SMOTE)[3];
- Metodo de interpolação;
- Segmento de linha;
- 50% SMOTE.

Undersampling

- Random UnderSampling (RUS) [3];
- Remoção aleatória;
- 80% RUS.

Implementação

- Python 3;
- Scikit-learn;
- Scipy;
- Pandas;
- Numpy.

Validação Cruzada

- Stratified K-folds;
- 10 folds;
- Random State: 10.

Otimização dos Paramêtros

- Randomized Search;
- Espaço de busca;
- Número de iteração: 500;
- Número de arvores: 50 - 400;
- Critério: Gini ou Entropy.

Otimização dos Paramêtros

- Grid Search;
- Posição do grid;
- Kernel: RBF ou Linear;
- Gamma: 1e-3 até 1e-4;
- C: 1, 10, 100, 1000.

Classificadores

- Multinomial Naive Bayes (MNB);
- Estimador de la place: 1;
- Random Forest (RF);
- Número de arvores: 339;
- Critério: Gini;
- Support Vector Machine (SVM);
- Kernel: RBF;
- Gamma: 0.0001;
- C: 100.

Classificadores

- Stacking;
- Nível 0 e Nível 1;
- Logistic Regression.

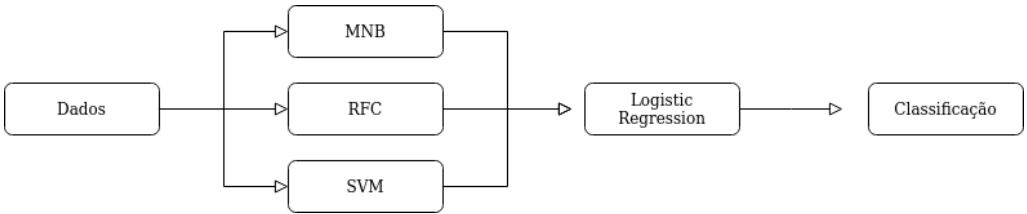

Figure: Arquitetura Stacking utilizada.

Resultados e discussões

Table: Resultados F-Score base de dados modificada.

Modificada	**MNB**	**RF**	**SVM**	**Stacking**
F score:	0.948197	0.968289	0.950859	0.979424

Table: Resultados F-Score base de dados original.

Original	**MNB**	**RF**	**SVM**	**Stacking**
F score:	0.943929	0.977185	0.959590	0.981052

Resultados e discussões

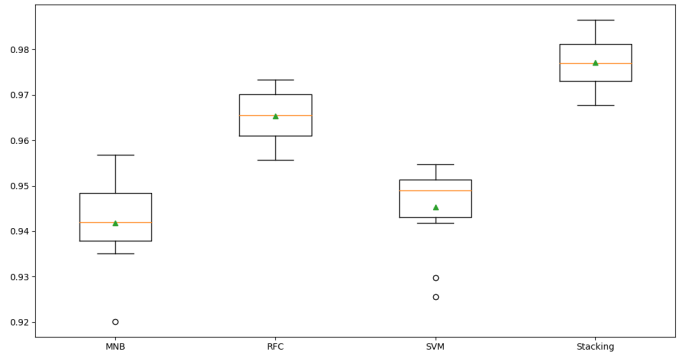

Figure: Boxplot base de dados modificada.

Resultados e discussões

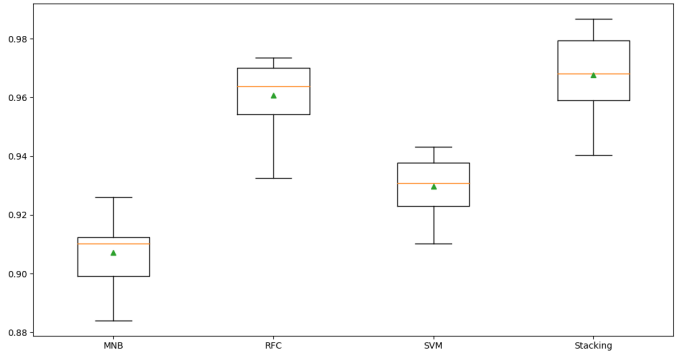

Figure: Boxplot base de dados original.

Resultados e discussões

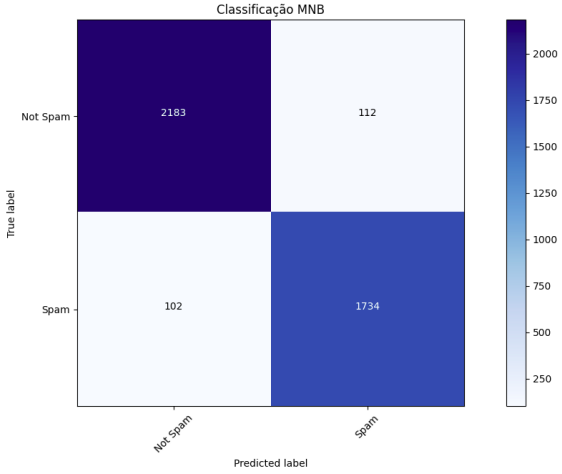

Figure: Matriz de confusão MNB base de dados modificada.

Resultados e discussões

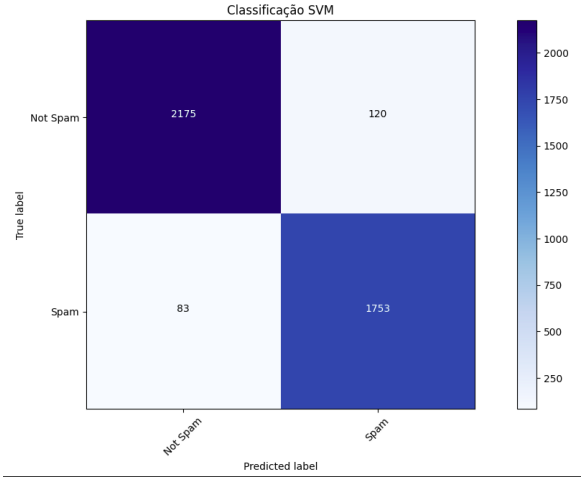

Figure: Matriz de confusão SVM base de dados modificada.

Resultados e discussões

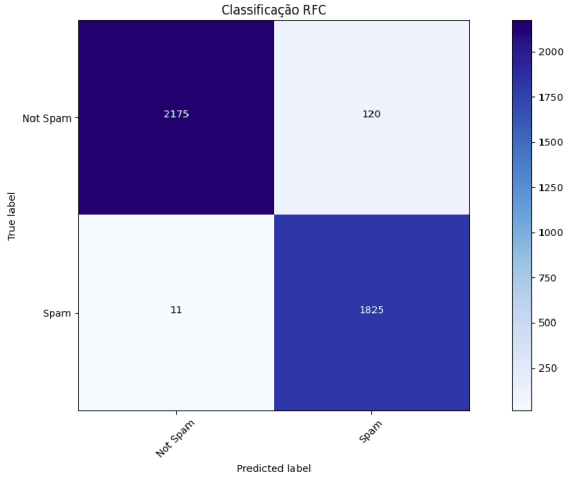

Figure: Matriz de confusão RF base de dados modificada.

Resultados e discussões

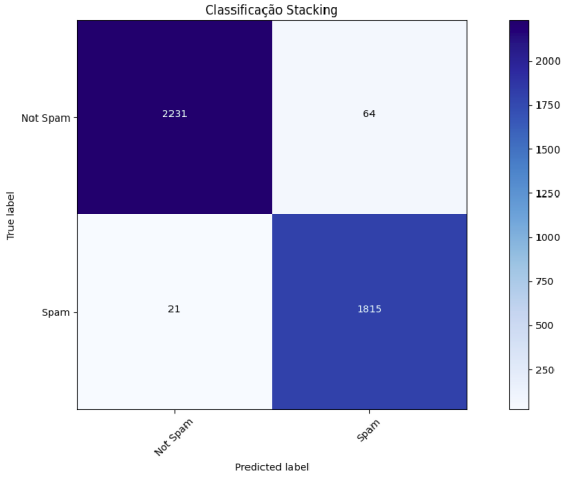

Figure: Matriz de confusão Stacking base de dados modificada.

Resultados e discussões

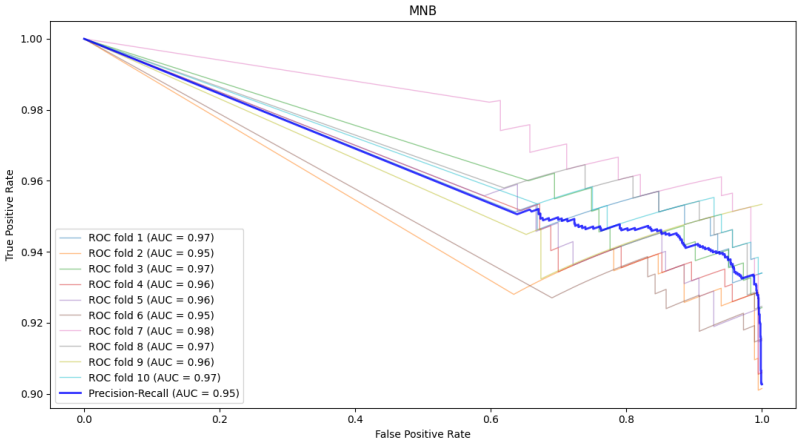

Figure: Curva PR para MNB base de dados modificada.

Resultados e discussões

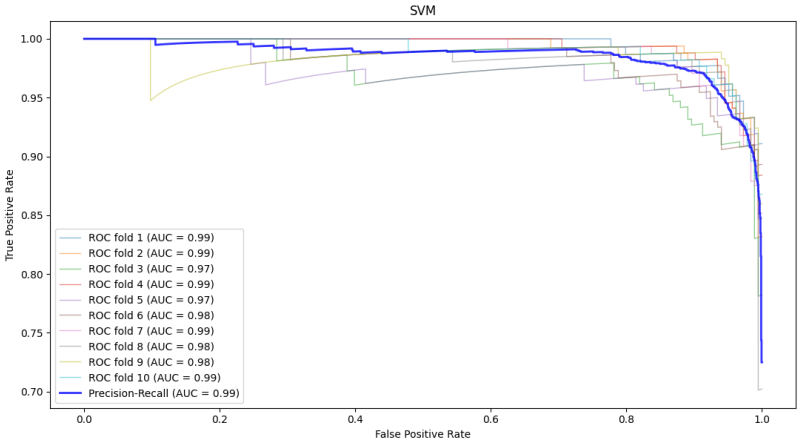

Figure: Curva PR para SVM base de dados modificada.

Resultados e discussões

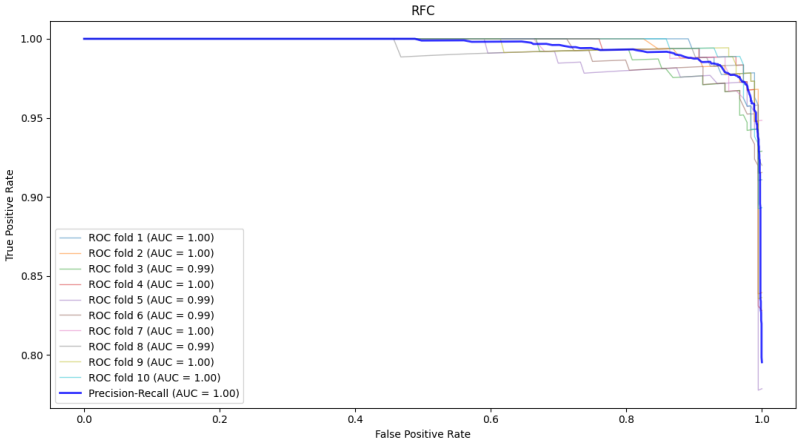

Figure: Curva PR para RF base de dados modificada.

Resultados e discussões

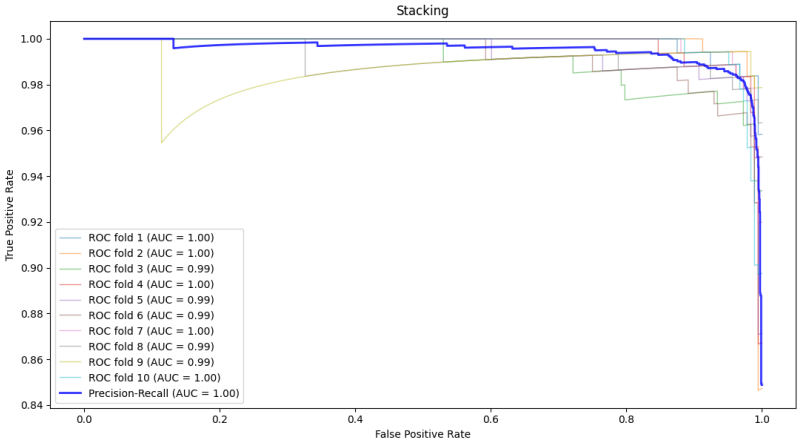

Figure: Curva PR para Stacking base de dados modificada.

Resultados e discussões

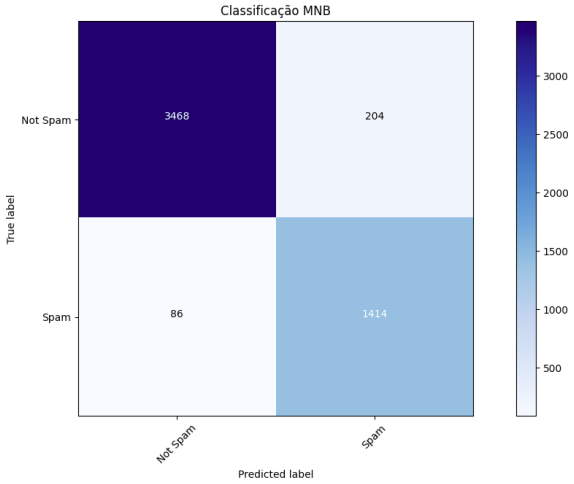

Figure: Matriz de confusão MNB base de dados original.

Resultados e discussões

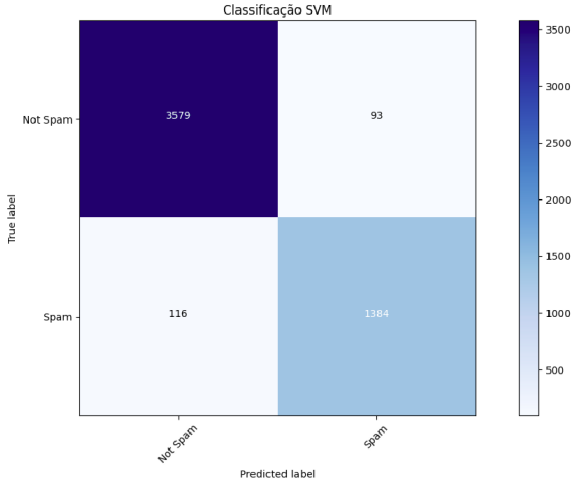

Figure: Matriz de confusão SVM base de dados original.

Resultados e discussões

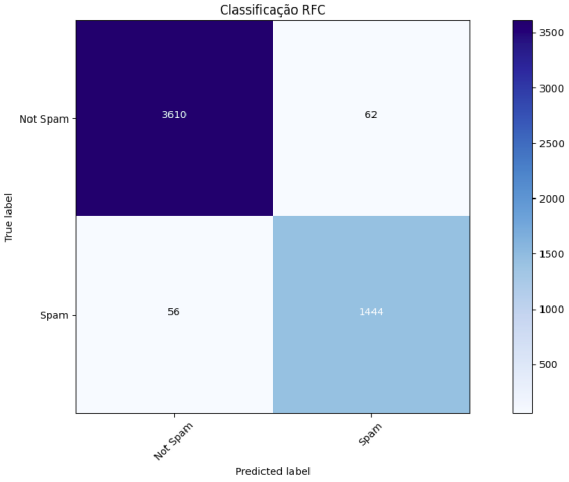

Figure: Matriz de confusão RF base de dados original.

Resultados e discussões

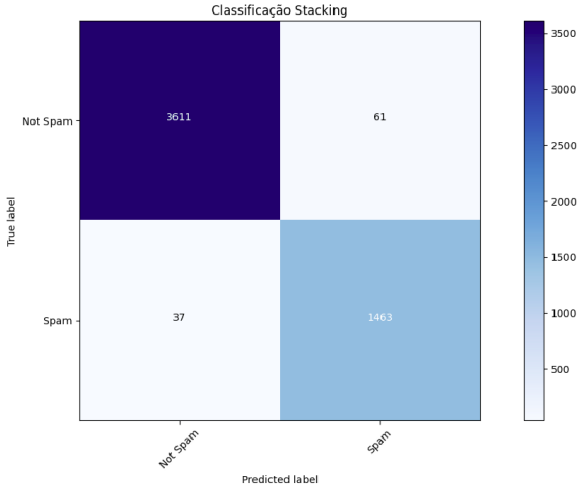

Figure: Matriz de confusão Stacking base de dados original.

Resultados e discussões

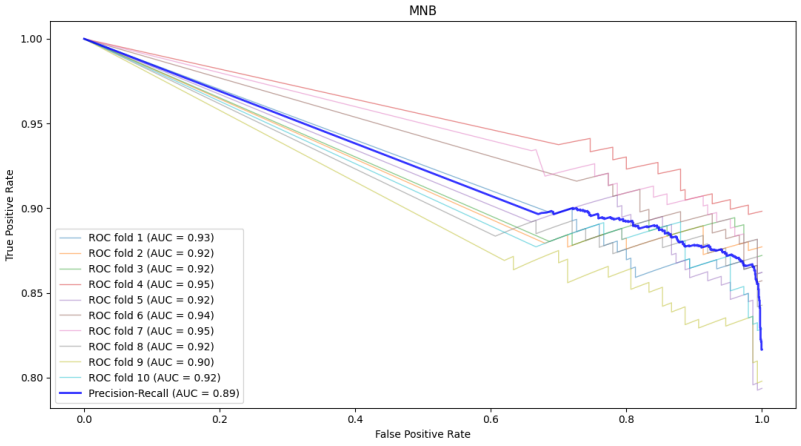

Figure: Curva PR para MNB base de dados original.

Resultados e discussões

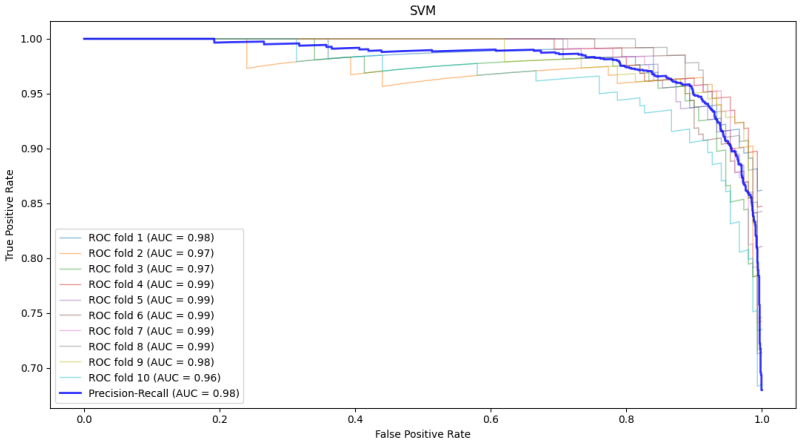

Figure: Curva PR para SVM base de dados original.

Resultados e discussões

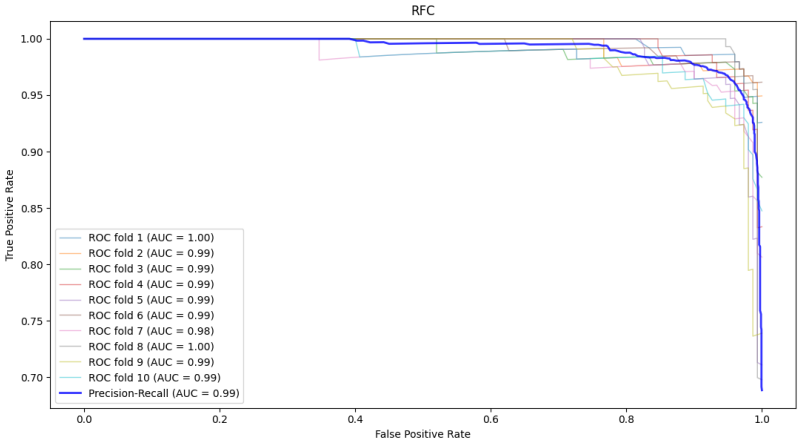

Figure: Curva PR para RF base de dados original.

Resultados e discussões

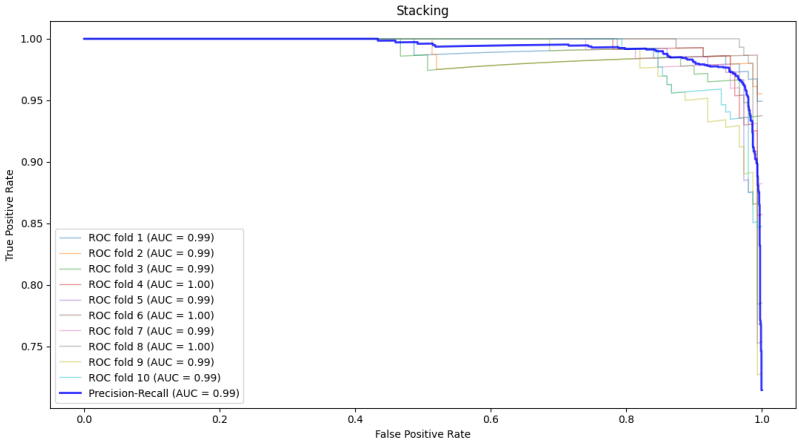

Figure: Curva PR para Stacking base de dados original.

Resultados e discussões

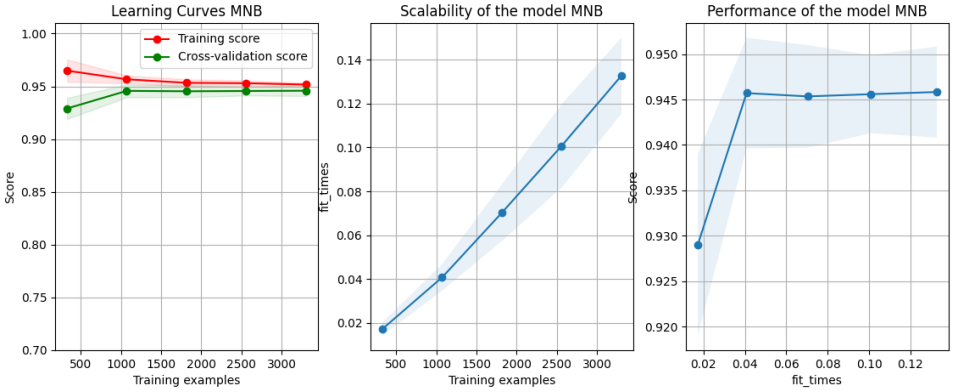

Figure: Curva de aprendizado para MNV base de dados modificada.

Resultados e discussões

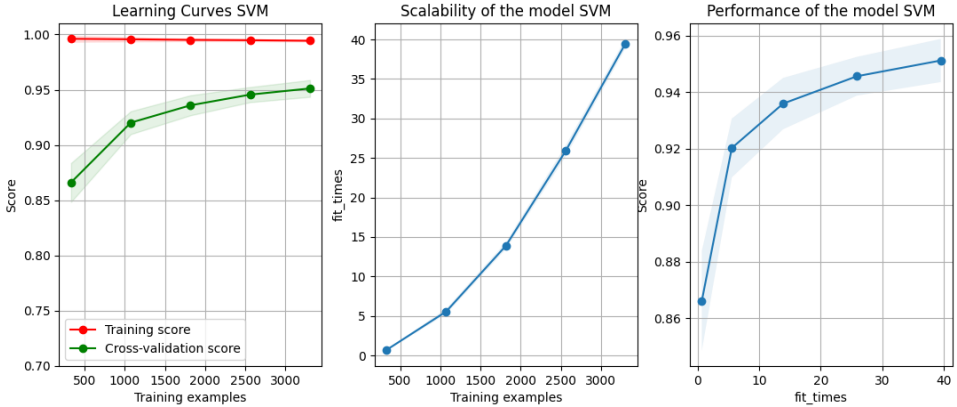

Figure: Curva de aprendizado para SVM base de dados modificada.

Resultados e discussões

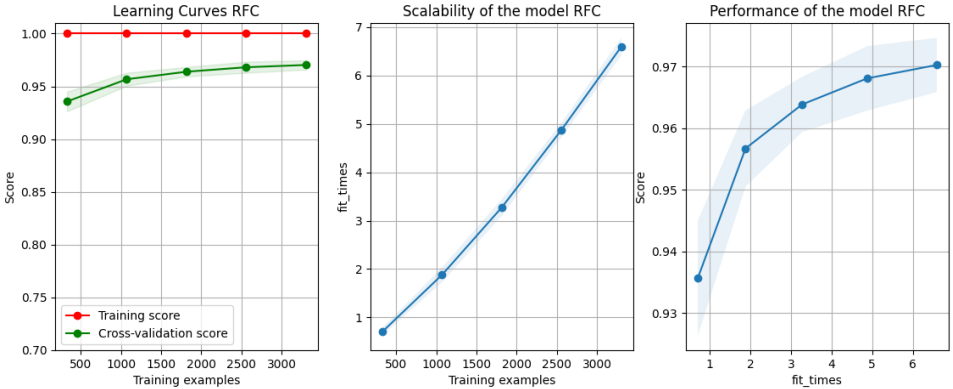

Figure: Curva de aprendizado para RF base de dados modificada.

Resultados e discussões

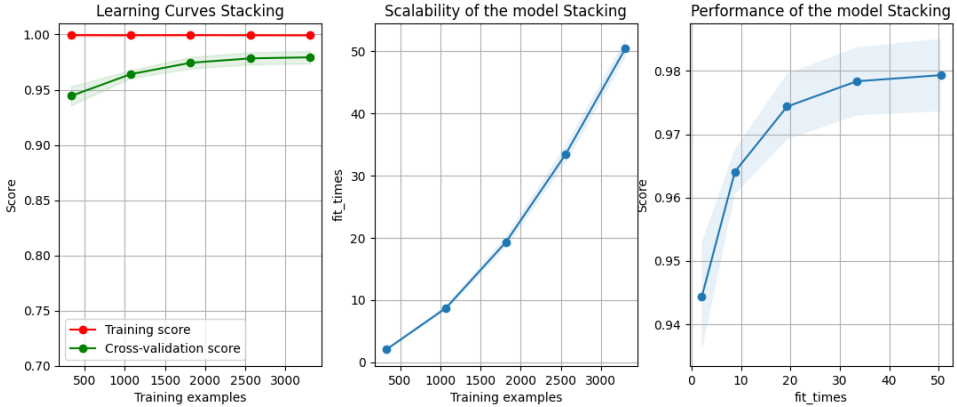

Figure: Curva de aprendizado para Stacking base de dados modificada.

Resultados e discussões

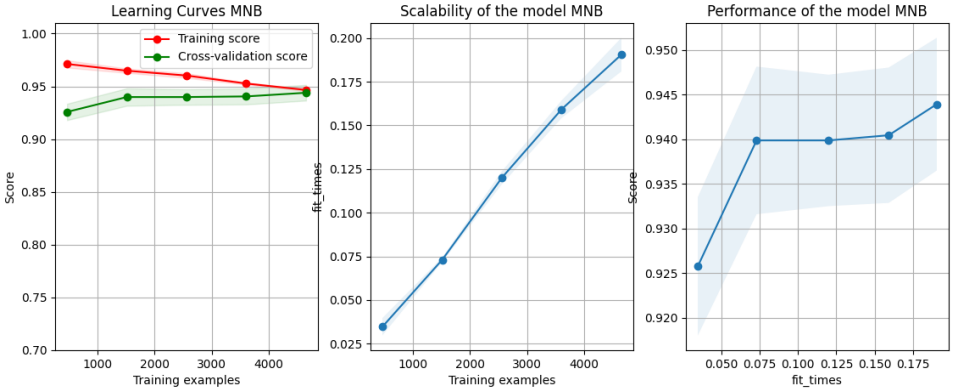

Figure: Curva de aprendizado para MNB base de dados original.

Resultados e discussões

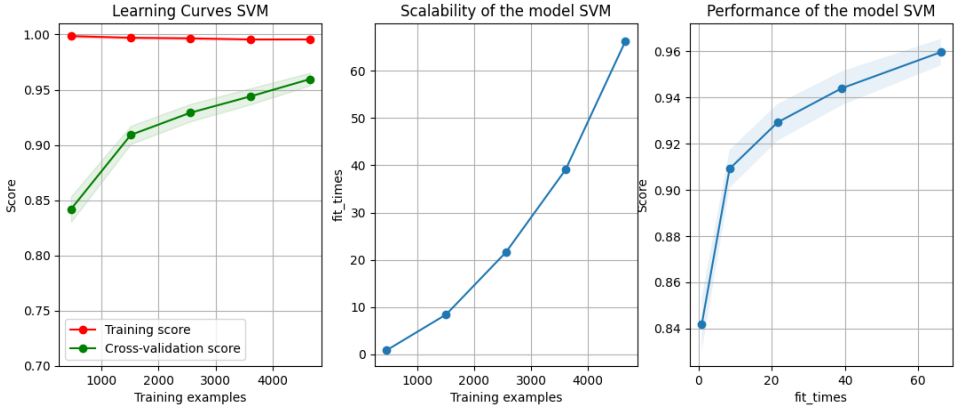

Figure: Curva de aprendizado para SVM base de dados original.

Resultados e discussões

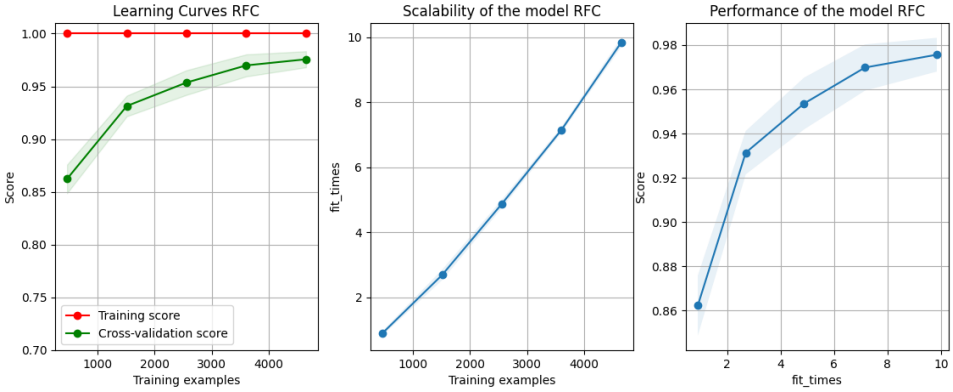

Figure: Curva de aprendizado para RF base de dados original.

Resultados e discussões

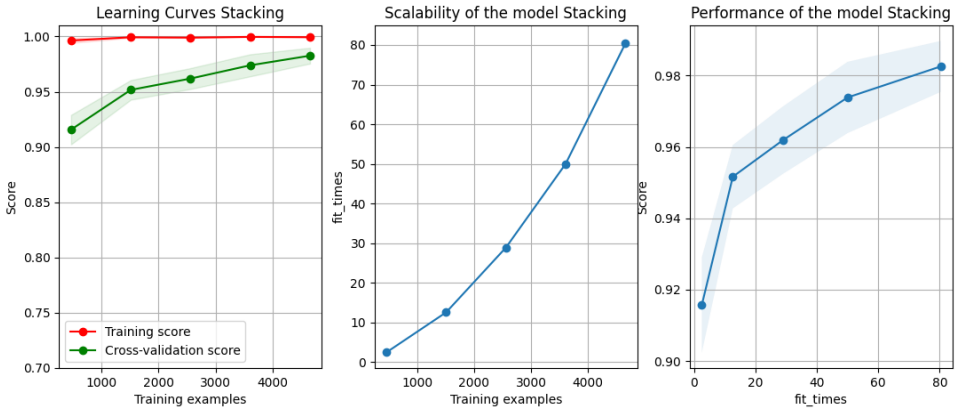

Figure: Curva de aprendizado para Stacking base de dados original.

Conclusão

- RF;
- Stacking;
- Grande custo computacional;
- Mais de 94 %;
- Undersampling e Oversampling;
- Base de dados com mais exemplos.

Trabalhos futuros

- Aumentar a base de dados;
- Incluir novas características;
- Meta-heurísticos;
- Reaproveitamento de partículas;
- Tempo de processamento e classificação.

Bibliografia I

W. Awad and S. ELseuofi.
Machine learning methods for spam e-mail classification.
International Journal of Computer Science & Information Technology (IJCSIT), 3(1):173–184, 2011.

E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa.
Machine learning for email spam filtering: review, approaches and open research problems.
Heliyon, 5(6):e01802, 2019.

P. Kaur and A. Gosain.
Comparing the Behavior of Oversampling and Undersampling Approach of Class Imbalance Learning by Combining Class Imbalance Problem with Noise, pages 23–30. 01 2018.

Bibliografia II

D. K. Renuka, T. Hamsapriya, M. R. Chakkaravarthi, and P. L. Surya.
Spam classification based on supervised learning using machine learning techniques.
In *2011 International Conference on Process Automation, Control and Computing*, pages 1–7, 2011.

B. Yu and Z. ben Xu.
A comparative study for content-based dynamic spam classification using four machine learning algorithms.
Knowledge-Based Systems, 21(4):355–362, 2008.

Obrigado!
Perguntas?

GuilhermeNakahata@gmail.com

https://github.com/GuilhermeNakahata/EmailSpamClassification