Campus Campinas Swift

USF

UNIVERSIDADE SÃO FRANCISCO

Banca do Trabalho de Conclusão de Curso da Engenharia de Computação

ANÁLISE COMPARATIVA DE MODELOS DE APRENDIZADO DE MÁQUINA NA DETECÇÃO DE DIABETES EM ESTÁGIO INICIAL

Boa noite!

Graduando

Lucas Gentil Carlos

Estudante do 10º semestre da Engenharia de Computação pela Universidade São Francisco e atualmente exercendo o cargo de Analista de Suporte JR no CPQD.

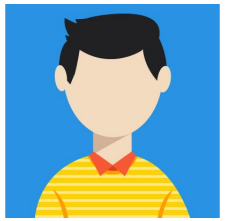
Graduando

Luan Ornelas de Souza

Estudante do 9° semestre de Engenharia da Computação pela Universidade São Francisco;

Faz estágio em ciência de dados no Instituto Eldorado;

Professor Orientador



José Guilherme Picolo

Introdução

OBJETIVOS

- Estudo comparativo de algoritmos de Machine Learning;
- Desenvolver um modelo para predição da diabetes em estágio inicial;

Introdução

JUSTIFICATIVAS

- Identificar antecipadamente os casos de Diabetes Mellitus através da analise comparativa de algoritmos;

Figura 1 - Informações sobre os portadores de Diabetes..
Fonte: (HOSPITAL ALEMÃO OSWALDO CRUZ, 2020).

Introdução

Diabetes Mellitus

- Atualmente 540 milhões de pessoas possuem a doença;
- Causas da Diabete;
- Principais sintomas;
- Identificação tardia da doença.

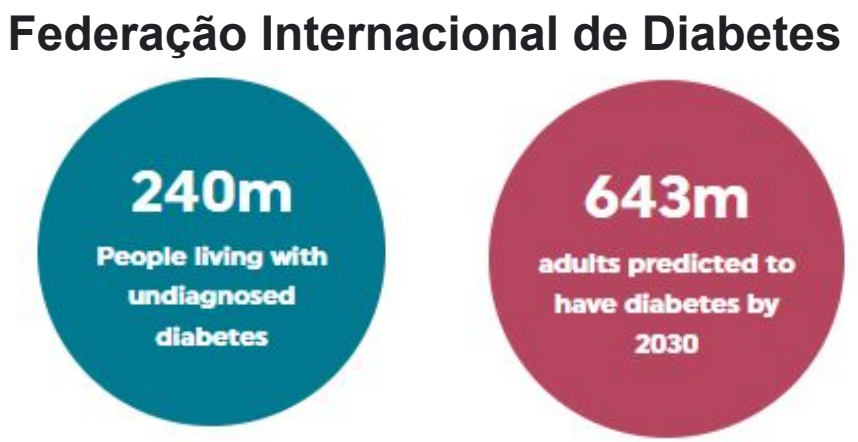


Figura 2 - Portadores de diabetes vivendo sem a identificação da doença.. Fonte: (INTERNATIONAL DIABETES FEDERATION, 2019).

Figura 3 - Previsão de adultos portadores de Diabetes em 2030.. Fonte: (INTERNATIONAL DIABETES FEDERATION, 2019).

Metodologia

- Base de dados;
- Linguagem e ambiente de desenvolvimento;
- Pré processamento dos dados;
- Algoritmos de Classificação;
- Validação cruzada;

Metodologia

BANCO DE DADOS

- Conjunto de dados público, disponível na UCI Machine Learning Repository - 520 Pacientes;

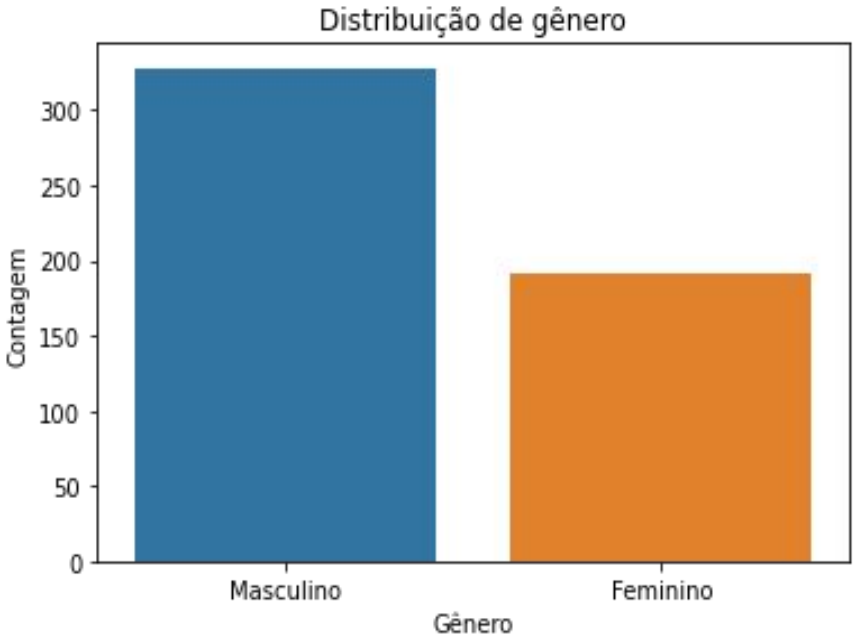


Figura 4 - Distribuição de pacientes por gênero. Fonte: Elaborada pelo autor.

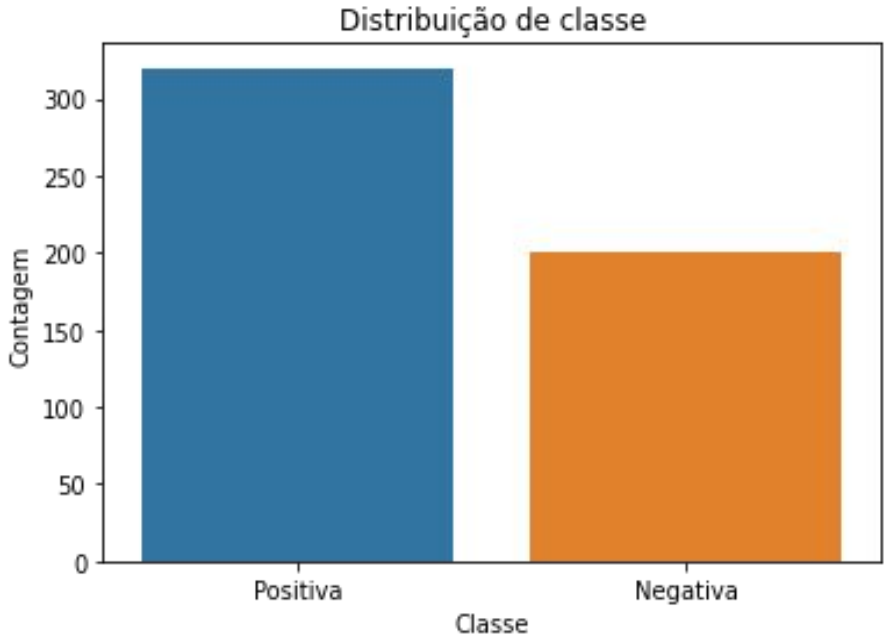


Figura 5 - Distribuição das classes. Fonte: Elaborada pelo autor.

Metodologia

BANCO DE DADOS

- Atributos do dataset

Nome da coluna	Descrição	Range de valores
Idade	Idade do paciente	16 - 90 anos
Gênero	Gênero biológico do paciente	Masculino ou Feminino
Poliúria	Necessidade excessiva de urinar	Sim ou Não
Polidipsia	Sede excessiva	Sim ou Não
Perda repentina de peso	-	Sim ou Não
Fraqueza	Sensação de cansaço constante	Sim ou Não
Polifagia	Fome excessiva	Sim ou Não
Candidíase genital	Infecção por fungos na genital	Sim ou Não
Visão borrada	-	Sim ou Não
Coceira	-	Sim ou Não
Irritabilidade	-	Sim ou Não
Cicatrização retardada	-	Sim ou Não
Paresia parcial	Grau leve a moderado de fraqueza muscular	Sim ou Não
Rigidez muscular	-	Sim ou Não
Alopecia	Queda de cabelos ou de pelos	Sim ou Não
Obesidade	Excesso de gordura corporal	Sim ou Não
Classe	Classificação positiva ou negativa para diabetes	Positiva ou Negativa

Tabela 1: Descrição do conjunto de dados. Fonte: Elaborada pelo autor.

Metodologia

LINGUAGEM E AMBIENTE DE DESENVOLVIMENTO

- Linguagem - Python versão 3.9;
- Ambiente de desenvolvimento - Jupyter Notebook;

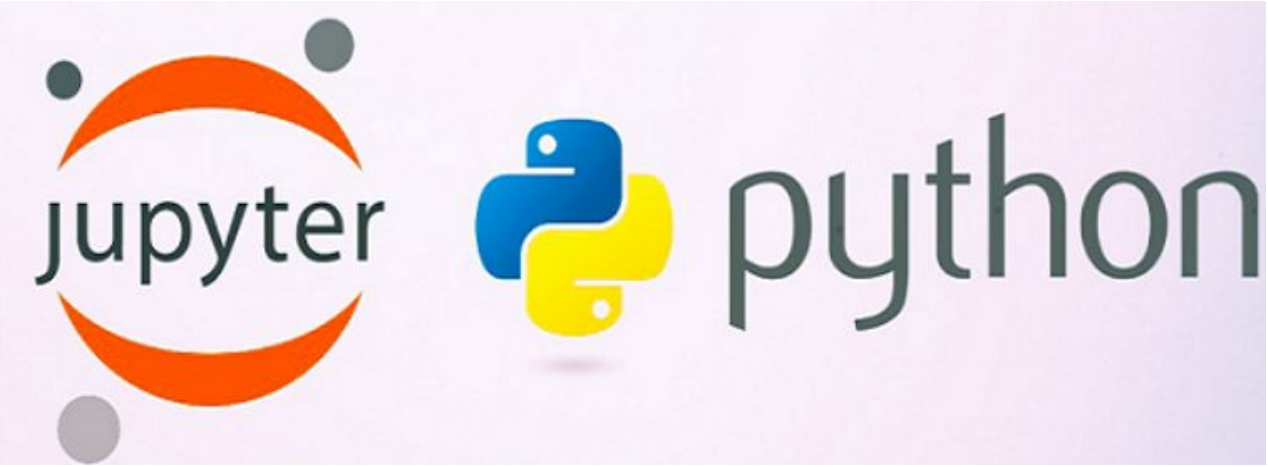


Figura 6 - Jupyter & Python. Fonte: (MAULANA, 2020).

Metodologia

Pré-processamento dos dados

- Conversão dos valores categóricos para numéricos - Para 0 ou 1;
- Normalização dos dados numéricos - Entre 0 e 1;
- Balanceamento dos dados presentes no dataset;

Metodologia

Pré processamento dos dados

- Conversão dos valores categóricos para numéricos - 0 ou 1;

Nome da coluna	Descrição	Range de valores
Idade	Idade do paciente	16 - 90 anos
Gênero	Gênero biológico do paciente	Masculino ou Feminino
Poliúria	Necessidade excessiva de urinar	Sim ou Não
Polidipsia	Sede excessiva	Sim ou Não
Perda repentina de peso	-	Sim ou Não
Fraqueza	Sensação de cansaço constante	Sim ou Não
Polifagia	Fome excessiva	Sim ou Não
Candidíase genital	Infecção por fungos na genital	Sim ou Não
Visão borrada	-	Sim ou Não
Coceira	-	Sim ou Não
Irritabilidade	-	Sim ou Não
Cicatrização retardada	-	Sim ou Não
Paresia parcial	Grau leve a moderado de fraqueza muscular	Sim ou Não
Rigidez muscular	-	Sim ou Não
Alopecia	Queda de cabelos ou de pelos	Sim ou Não
Obesidade	Excesso de gordura corporal	Sim ou Não
Classe	Classificação positiva ou negativa para diabetes	Positiva ou Negativa

Tabela 1: Descrição do conjunto de dados. Fonte: Elaborada pelo autor.

Metodologia

Pré processamento dos dados

- Normalização dos dados numéricos - Entre 0 e 1;

Nome da coluna	Descrição	Range de valores
Idade	Idade do paciente	16 - 90 anos

Tabela 2: Demonstração do range de valores do atributo idade. Fonte: Elaborada pelo autor.

Metodologia

Pré processamento dos dados

- Balanceamento dos dados presentes no dataset;

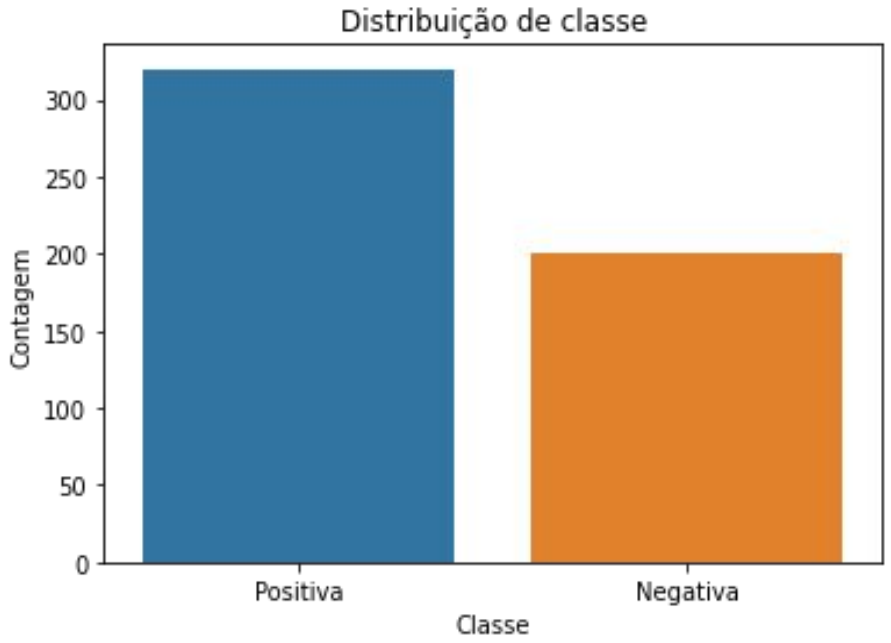


Figura 5 - Distribuição das classes. Fonte: Elaborada pelo autor.

Metodologia

Algoritmos de Machine Learning

- Inteligência artificial;
- Machine Learning;
  - O que é?
  - Quais foram utilizados?

Algoritmos de Machine Learning

Inteligência Artificial

"Capacidade de um sistema para interpretar corretamente dados externos, aprender a partir desses dados e utilizar essas aprendizagens para atingir objetivos e tarefas especificos através de adaptação flexivel." (Kaplan e Haenlein, 2018)

Figura 7 - Ilustração do campo de Inteligência Artificial. Fonte: (MORAIS, 2023).

Machine Learning - O que é?

Subconjunto de IA que utiliza métodos estátisticos permitindo que máquinas melhorem tarefas através do aprendizado e experiência.

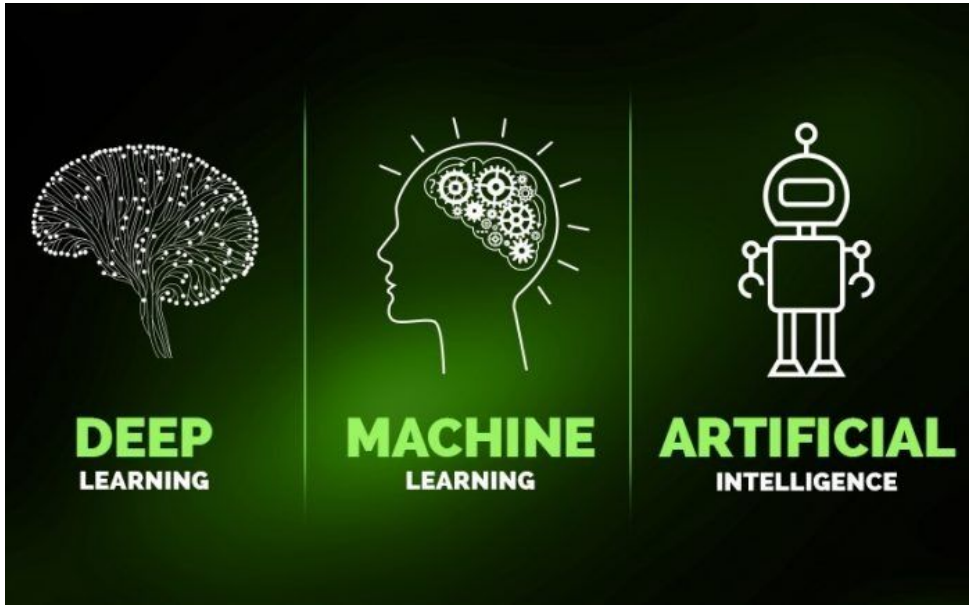


Figura 8 - Sub-áreas da Inteligência Artificial Fonte: (HELDER, 2018).

Machine Learning - Quais algoritmos utilizados?

- Random Forest
- XGBoost
- Support Vector Machine
- Naive Bayes

Algoritmos de Machine Learning

Random Forest

- É um método de aprendizado em conjunto (ensemble), ou seja, combina várias estruturas aleatórias de Decision Tree (DT);
- Cada DT realiza a classificação;
- Por fim, uma votação é feita e a classe mais votada é o resultado do algoritmo;

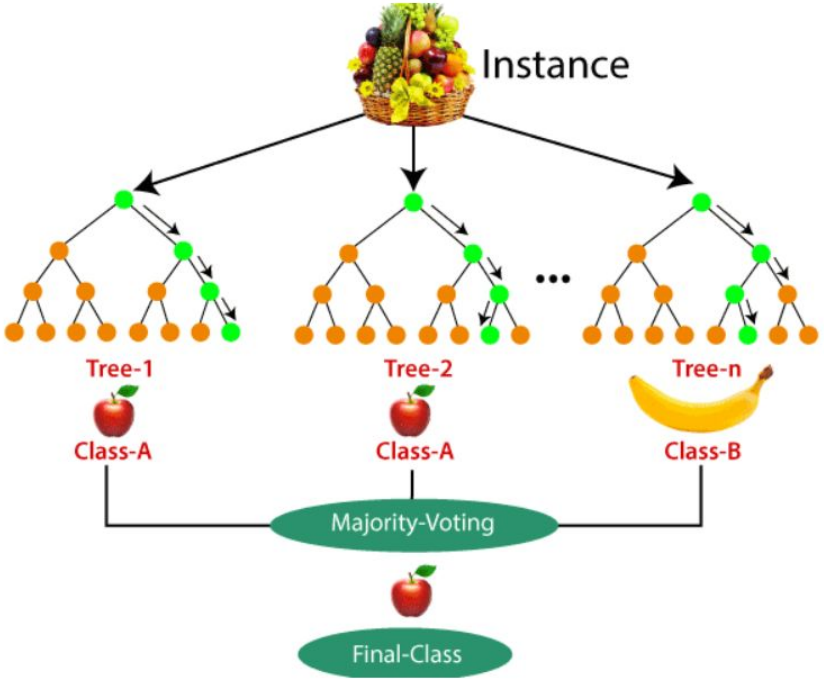


Figura 9 - Exemplo de um algoritmo RF. Fonte: (BOSE, 2023).

XGBoost

- XGBoost também é um método ensemble;
- Se baseia em Decision Tree com Gradient Boosting;
- Cada DT aprende nos exemplos que seu predecessor errou, minimizando a perda;

Algoritmos de Machine Learning

Support Vector Machine - SVM

- SVM é um algoritmo desenvolvido para realizar a criação de um hiperplano responsável por realizar a separação das classes;
- Vetores de suporte;
- Margens de separação;
- Função da reta para determinar o hiperplano central e paralelos.

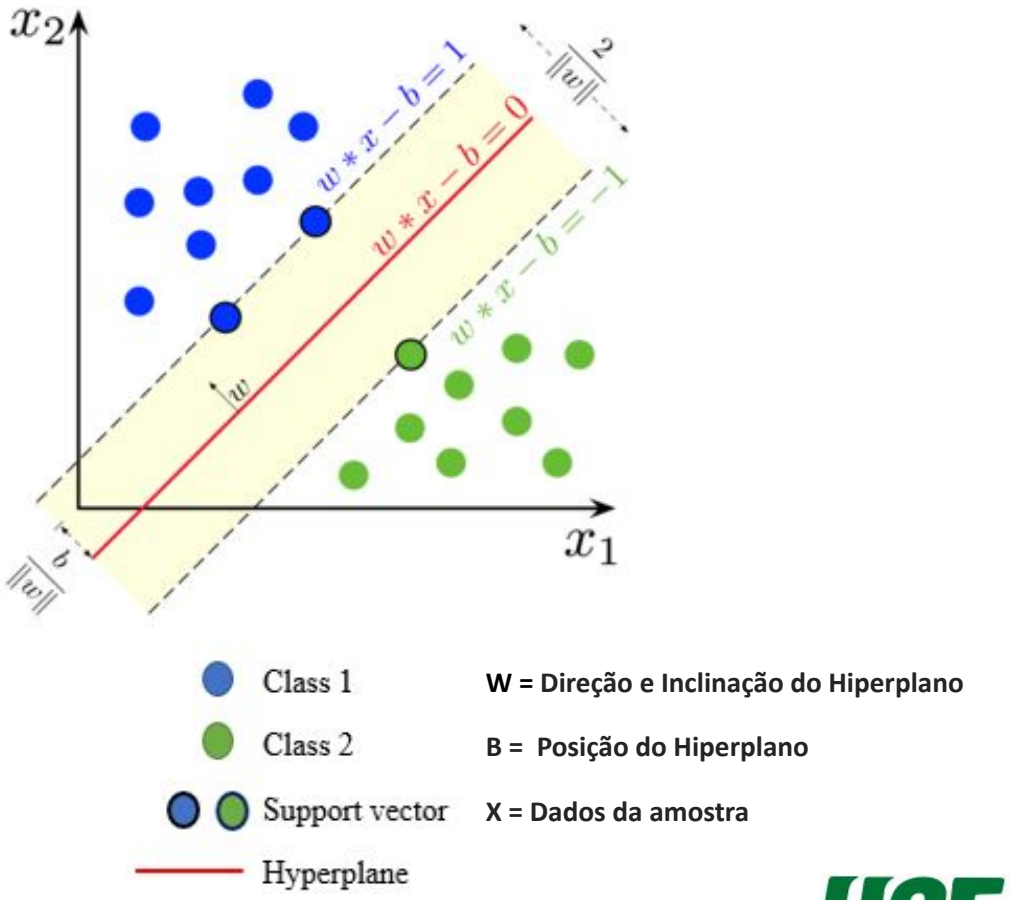


Figura 10 - Exemplo de um hiperplano traçado utilizando o algoritmo SVM.. Fonte: (ELSAYED et al., 2019).

Algoritmos de Machine Learning

Naive Bayes

- Algoritmo fundamentado no teorema de Bayes;
- Análise de probabilidade estatistica;
- *Atributos independentes entre si;*
- *Utilizado em conjuntos de dados grandes devido a simplicidade.*

Teorema de Bayes

$$P(A \mid B) = P(A) \times \frac{P(B|A)}{P(B)}$$

P = Probabilidade a ser calculada
A = Evento a ocorrer
B = Evento ocorrido.

USF
UNIVERSIDADE SÃO FRANCISCO

Metodologia

Validação Cruzada

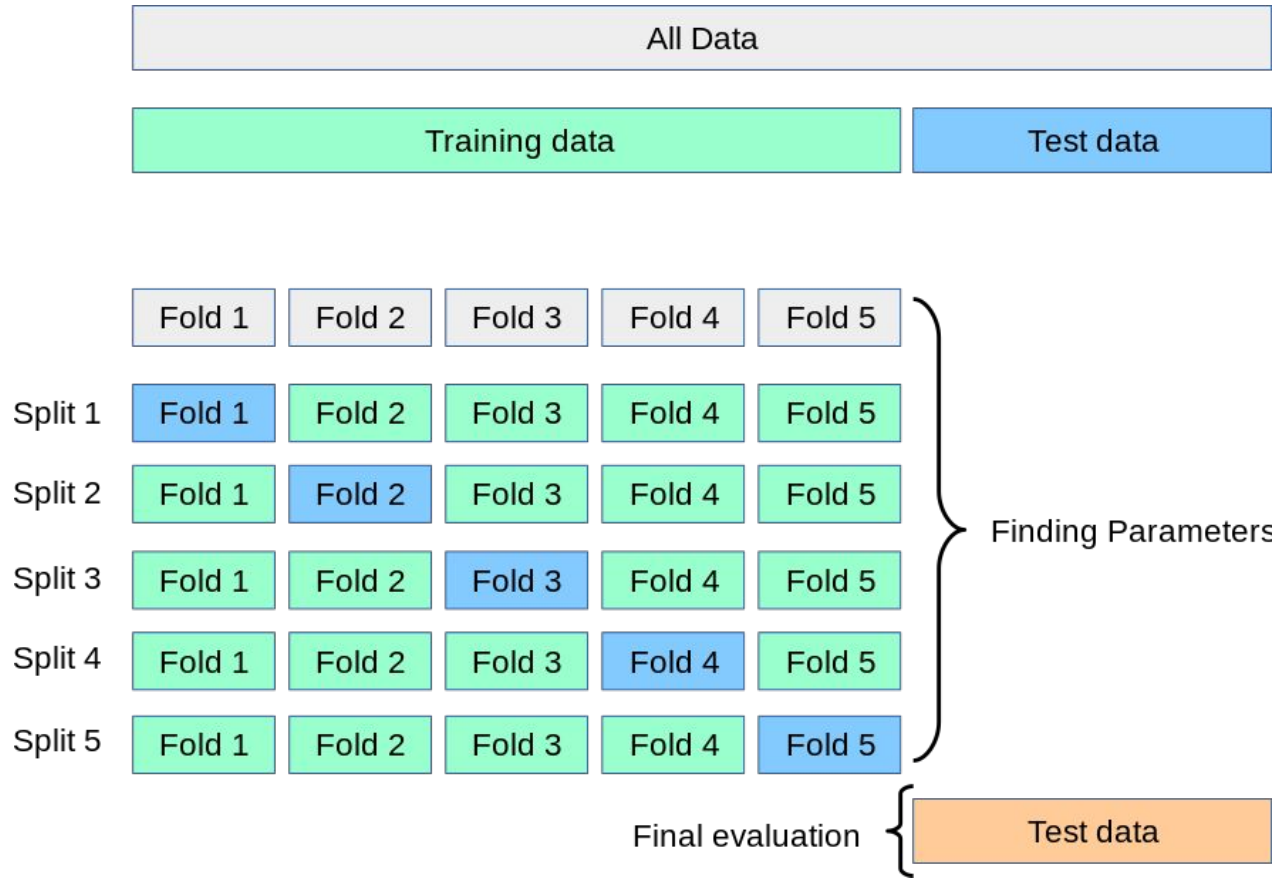


Figura 11 - Exemplificação da Validação Cruzada..
Fonte: (SCIKIT LEARN, 2023).

Resultados e Discussões

- Apresentar:
  - Métricas de avaliação;
  - Análise dos resultados.

Resultados e Discussões

Métricas de avaliação

- Tradicionalmente elaborada a matriz de confusão para classificação dos dados;
- Distribuidos entre VP, VN, FP e FN;
- Diagonal Principal e Secundária;
- Posteriormente são realizados os cálculos para análise de desempenho do algoritmo.

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 12 - Exemplo de uma matriz de confusão utilizada para classificação. Fonte: (RODRIGUES, 2019).

Resultados e Discussões

Métricas de avaliação

- Métricas utilizadas - Precisão, Sensibilidade, Acuracia, F1-Score, Acuracia Balanceada e Área sob a curva ROC;
- Valores variam entre 0 e 1 ou 0% a 100%;
- Mais próximos de 1 são considerados os melhores resultados.

$$Precision = \frac{VP}{VP + FP} \qquad Recall = \frac{VP}{VP + FN} \qquad Accuracy = \frac{VP + VN}{VP + FP + VN + FN} \qquad F1\ Score = \frac{2 \times VP}{2 \times VP + FP + FN}$$

$$Balanced\ Accuracy = \frac{Recall + Specificity}{2}$$

Resultados e Discussões

Área sob a curva ROC

- VP (Sensibilidade) por FP (Especificidade);
- Probabilidade de ocorrência de um dado positivo ser maior do que um dado negativo;
- Valor 1 o valor ideal para classificação e o valor de 0.5 como uma classificação aleatória.

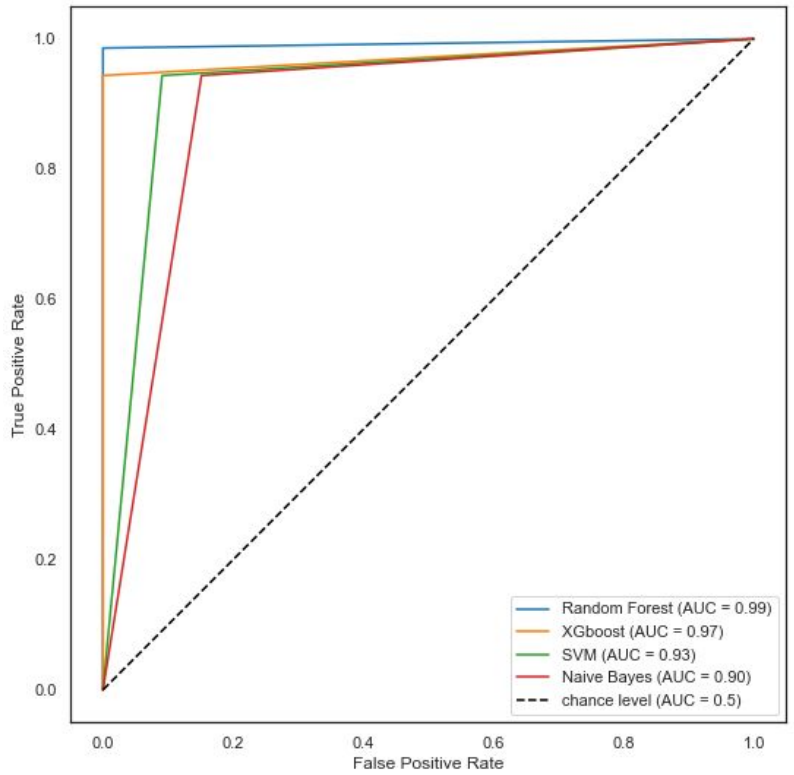


Figura 13 - Característica de Operação do Receptor (ROC).

Fonte: Elaborada pelo autor.

Resultados e Discussões

Análise dos resultados

Model	Accuracy	Balanced Accuracy	F1_Score	Precision	Recall
Random Forest Classifier	0.99	0.993	0.993	1	0.986
XGBoost Classifier	0.962	0.972	0.971	1	0.944
SVM Linear	0.933	0.926	0.95	0.957	0.944
Naive Bayes Classifier	0.913	0.896	0.937	0.931	0.944

Tabela 3 - Resultados dos classificadores. Fonte: Elaborada pelo autor.

- Random Forest foi o mais eficaz em todas as métricas;
- Naive Bayes foi o pior algoritmo para esse experimento;

Conclusão

- Melhor algoritmo;
- Pontos de melhorias;
- Próximos passos.

Referências Bibliográficas

- CAMPBELL, C.; YIMING, Y. Learning with Support Vector Machines. Morgan & Claypool Publishers, 2011. 100 p. ;
- CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, USA, p.785–794. Aug. 2016. DOI: https://doi.org/10.1145/2939672.2939785
- INTERNATIONAL DIABETES FEDERATION. Diabetes Atlas. Bélgica: International Diabetes Federation, 2006, 3rd. ed. Disponível em: https://diabetesatlas.org/upload/resources/previous/files/3/Diabetes-Atlas-3rd-edition.pdf. Acesso em: 15 set. 2023.;
- KANNAN, H. et. al. Bayesian Reasoning and Gaussian Processes for Machine Learning Applications. Chapman and Hall/CRC, 2022. 133 p.
- LIU, Y.; WANG, Y.; ZHANG, J. New Machine Learning Algorithm: Random Forest. Information Computing and Applications (ICICA), Springer, Berlin, v.7473, p.246-252. 2012, DOI: https://doi.org/10.1007/978-3-642-34062-8_32.
- RUSSEL, S.; NORVIG, P. Artificial Intelligence: A Modern Approach. 3. ed., Inglaterra: Pearson, 2009. 1152 p.
- RODRIGUES, V. Métricas de Avaliação: acurácia, precisão, recall… quais as diferenças?. Medium., Abr. 2019. Disponível em: https://vitorborbarodrigues.medium.com/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-recall-quais-as-diferen%C3%A7as-c8f05e0a513c. Acesso em: 29 out. 2023.

USF
UNIVERSIDADE SÃO FRANCISCO

Referências Bibliográficas

- ELSAYED, H. et. al. Integrating Modern Classifiers for Improved Building Extraction from Aerial Imagery and LiDAR Data. American Journal of Geographic Information System 213-220. Oct, 2019. Disponível em: https://www.researchgate.net/publication/336810849_Integrating_Modern_Classifiers_for_Improved_Building_Extraction_from_Aerial_Imagery_and_LiDAR_Data. Acesso em: 29 out. 2023
- INTERNATIONAL DIABETES FEDERATION. Facts & figures. In: Facts & figures, 2023. Disponível em: https://idf.org/about-diabetes/diabetes-facts-figures/. Acesso em: 15 set. 2023.
- BOSE, S. Random Forest Algorithm. Inside AIML, 2021. Disponível em: https://insideaiml.com/blog/Random-Forest-Algorithm-1029. Acesso em: 13 nov. 2023.
- HOSPITAL ALEMÃO OSWALDO CRUZ. Tratamento do Diabetes. In: Tudo Sobre Diabetes, 2020. Disponível em: https://centrodeobesidadeediabetes.org.br/tudo-sobre-diabetes/tratamento-do-diabetes/. Acesso em: 20 nov. 2023.
- MAULANA, A. How to Install Python and Jupyter Notebook on Windows 10 64 bit. Medium. Abr. 2020. Disponível em: https://medium.com/@akbar.maulana2298/how-to-install-python-and-jupyter-notebook-on-windows-10-64-bit-9933aa3c0ae4. Acesso em: 20 nov. 2023.
- MORAIS, F. Branding com Inteligência Artificial. Medium. Nov. 2023. Disponível em: https://felipemorais2309.medium.com/branding-com-intelig%C3%AAncia-artificial-f6eea9c75dc6. Acesso em: 20 nov. 2023.

Referências Bibliográficas

- HELDER. Entenda o que é Deep Learning e Como Funciona. Cultura Analítica. Set. 2018. Disponível em: https://culturaanalitica.com.br/deep-learning-oquee-como-funciona/. Acesso em: 20 nov. 2023.
- SCIKIT LEARN. Cross Validation: evaluating estimator performance. In: Cross Validation, 2023. Disponível em: https://scikit-learn.org/stable/modules/cross_validation.html. Acesso em: 15 set. 2023.

UNIVERSIDADE SÃO FRANCISCO

Campinas – SP
2023