UNIVERSIDADE ESTADUAL DA PARAÍBA

CAMPUS I – CAMPINA GRANDE

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

CURSO DE GRADUAÇÃO EM ESTATÍSTICA

JOSÉ LUCAS COSTA DE OLIVEIRA

USO DE *MACHINE LEARNING* PARA PREDIÇÃO DE MORTE: UMA APLICAÇÃO A PACIENTES COM CÂNCER DE MAMA EM UMA CIDADE DA PARAÍBA

CAMPINA GRANDE – PB

2023

JOSÉ LUCAS COSTA DE OLIVEIRA

USO DE *MACHINE LEARNING* PARA PREDIÇÃO DE MORTE: UMA APLICAÇÃO A PACIENTES COM CÂNCER DE MAMA EM UMA CIDADE DA PARAÍBA

Trabalho de Conclusão de Curso (Artigo) apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Bacharel em Estatística.

Orientador: Prof. Dr. Tiago Almeida de Oliveira

CAMPINA GRANDE – PB

2023

O48u Oliveira, Jose Lucas Costa de.

Uso de *machine learning* para predição de morte [manuscrito] : uma aplicação a pacientes com câncer de mama em uma cidade da Paraíba / Jose Lucas Costa de Oliveira. - 2023.

32 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2023.

"Orientação : Prof. Dr. Tiago Almeida de Oliveira, Departamento de Matemática e Estatística - CCT. "

1. Inteligência artificial. 2. Câncer de mama. 3. Modelo de aprendizado de máquina. I. Título

21. ed. CDD 005.12

Elaborada por Talita M. A. Tavares - CRB - CRB 15/971 BC/UEPB

JOSÉ LUCAS COSTA DE OLIVEIRA

USO DE *MACHINE LEARNING* PARA PREDIÇÃO DE MORTE: UMA APLICAÇÃO A PACIENTES COM CÂNCER DE MAMA EM UMA CIDADE DA PARAÍBA

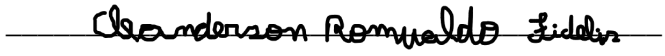
Trabalho de Conclusão de Curso (Artigo) apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Bacharel em Estatística.

Aprovada em: 28/06/2023.

BANCA EXAMINADORA

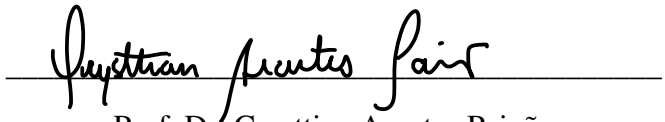
Prof. Dr. Tiago Almeida de Oliveira (Orientador)

Universidade Estadual da Paraíba (UEPB)



Prof. Me. Cleanderson Romualdo Fidelis

Universidade Estadual da Paraíba (UEPB)



Prof. Dr. Crysttian Arantes Paixão

Universidade Federal da Bahia (UFBA)

Dedico este trabalho as pessoas que confiaram no meu potencial e me deram apoio.

“Essencialmente, todos os modelos estão errados, mas alguns são úteis.”

(George Edward Pelham Box)

LISTAS DE ILUSTRAÇÕES

Figura 1 – Esquema de um projeto de Machine Learning 11
Figura 2 – Exemplos de Undersampling e Oversampling 13
Figura 3 – Tipos de Aprendizado de máquina 14
Figura 4 – Esquema de Treino e teste 15
Figura 5 – Validação cruzada 16
Figura 6 – Comparação entre CatBoost, XGBoost e LightGBM respectivamente 18
Figura 7 – Exemplo de curva ROC 21
Figura 8 – Exemplo de gráfico de Shapley value 22
Figura 9 – Matriz de confusão do modelo XGBoost 25
Figura 10 – Matriz de confusão do modelo LightGBM 26
Figura 11 – Matriz de confusão do modelo CatBoost 26
Figura 12 – Gráfico da área sob a curva ROC dos modelos CatBoost, XGBoost e LightGBM . 27
Figura 13 – Gráfico do SHAP value 27

LISTA DE TABELAS

Tabela 1 – Descrição do banco de dados 10
Tabela 2 – Exemplo de variável One-hot encolding 13
Tabela 3 – Hiperparâmetros do XGBoost, LightGBM e CatBoost 18
Tabela 4 – Exemplo de matriz de confusão 19
Tabela 5 – Informação dos dados 23
Tabela 6 – Descritiva das variáveis categóricas 23
Tabela 7 – Descritiva das variáveis numéricas 24
Tabela 8 – Resultado do oversampling 24
Tabela 9 – Avaliação dos modelos Catboost, LightGBM e XGBoost 25

SUMÁRIO

1	**INTRODUÇÃO**	09
2	**METODOLOGIA**	09
2.1	**Descrição do banco de dados**	09
2.2	**Linguagem Python**	10
2.3	**Pré-processamento dos dados**	11
2.4	**Limpeza dos dados**	11
2.4.1	***Dados faltantes***	11
2.5	**Tratamento dos dados**	12
2.5.1	***Padronização***	12
2.5.2	***Transformação de variáveis categóricas***	12
2.5.3	***Dados de classificação desbalanceados***	13
2.6	**Aprendizado de máquina**	14
2.6.1	**Treino e teste**	15
2.6.2	***Validação cruzada***	15
2.6.3	***Introduções ao modelo Gradient Boosting***	16
2.6.3.1	*Extreme Gradient Boosting*	16
2.6.3.2	*Light Gradient Boosting Machine*	17
2.6.3.3	*Categorical Boosting*	17
2.6.4	***Hiperparâmetros***	18
2.6.5	***Métricas de avaliação***	19
2.6.5.1	*Matriz de confusão*	19
2.6.5.2	Área sob a curva ROC	20
2.6.6	***Shapley value***	22
3	**RESULTADOS**	23
4	**DISCUSSÃO**	28
5	**CONCLUSÃO**	28
	REFERÊNCIAS.	29
	AGRADECIMENTOS.	32

USO DE MACHINE LEARNING PARA PREDIÇÃO DE MORTE: UMA APLICAÇÃO A PACIENTES COM CÂNCER DE MAMA EM UMA CIDADE DA PARAÍBA

José Lucas Costa de Oliveira *

RESUMO

Este estudo consiste em apresentar modelos de Machine Learning e seus resultados que foram utilizados para predição de morte de pacientes que foram diagnosticados com câncer de mama. Sendo assim inicialmente foi feita a coleta de uma amostra de 221 pacientes do gênero feminino do hospital fundação assistencial da paraíba (FAP). Com base nos dados obtidos, foi realizado um pré-processamento inicial que consiste em tratar algumas variáveis categóricas para colocar no padrão das dummies e assim possibilitar a partir do modelo de Machine Learning interpretar os dados categóricos. Por conta da quantidade de dados e por conta de dados faltantes que existem nos dados não foi possível realizar o preenchimento dos dados faltantes para não ocorrer o ajuste excessivo dos dados, porém foram utilizados modelos que funcionassem com esses dados, sendo os modelos Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM) e o Categorical Boosting (CatBoost) que são modelos que se adequam bem a dados faltantes por seguirem um modelo de árvore de decisão. Foi obtido alguns resultados relevantes a partir dos modelos utilizados que foram o do modelo LightGBM de acertar cerca de 85,00% das pacientes que não iriam morrer e cerca de 67,00% para as pacientes que morreriam com uma acurácia total do modelo de 82,08%. Também foi avaliado a curva ROC que teve sua área em torno de 00,71. Com isso para um modelo de aprendizado de máquina obteve-se resultados bastante significativos para o estudo.

Palavras-Chave: inteligência artificial; modelos; saúde.

ABSTRACT

This study consists of presenting Machine Learning models and their results that were used to predict the death of patients who were diagnosed with breast cancer. Initially, a sample of 221 female patients was collected from the hospital fundação assistencial da paraíba (FAP). Based on the data obtained, an initial pre-processing was performed, which consists in treating some categorical variables to put them in the pattern of dummies and thus enable the Machine Learning model to interpret categorical data. Because of the amount of data and because of missing data that exist in the data it was not possible to fill the missing data to avoid excessive adjustment of the data, but models that worked with these data were used, and the models Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM) and Categorical Boosting (CatBoost) are models that fit well to missing data by following a decision tree model. Some relevant results were obtained from the models used which were that the LightGBM model hit about 85,00% of the patients who would not die and about 67,00% for

* José Lucas Costa de Oliveira, Depto de Estatística, UEPB, Campina Grande, PB, jose.lucas.costa@aluno.uepb.edu.br

the patients who would die with a total accuracy of the model of 82.08%. It was also evaluated the ROC curve that had its area around 00,71. With this, for a machine learning model, it was possible to obtain very significant results for the study.

Keywords: artificial intelligence; models; Health.

1 INTRODUÇÃO

O câncer de mama é uma doença comum entre as mulheres, e sua incidência tem aumentado no mundo todo, mesmo em países desenvolvidos, o câncer de mama ainda é um dos maiores receios das mulheres, fatores de risco, como idade, histórico familiar e estilo de vida, estão associados a seu desenvolvimento. O procedimento de realizar a mamografia ajuda na detecção precoce e assim ter um tratamento com mais chances de cura que podem ser, cirurgia, radioterapia, quimioterapia e terapia hormonal (INCA, 2022).

Ocupando uma posição central entre as principais formas de câncer que afetam as mulheres, é compreensível que elas tenham preocupações em relação a essa doença, uma vez que seu diagnóstico e seus efeitos podem ter impactos diversos que abrangem aspectos psicológicos, relacionados às alterações corporais, à dor, à baixa autoestima, entre outros fatores, que podem ser agravados durante os tratamentos.

Cerca de 2,3 milhões de casos novos foram estimados para o ano de 2020 em todo mundo, o que representa cerca de 24,5% de todos os tipos de neoplasias diagnosticadas nas mulheres. No Brasil foram estimados 66,280 mil casos novos de câncer de mama em 2021, com um risco estimado de 61,61 casos a cada 100 mil mulheres, de acordo com o Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA, 2022).

Com base nisso, foi iniciado o estudo com uma amostra de 221 pacientes do sexo feminino, com a utilização da linguagem de programação Python que atualmente é uma das mais usadas em todo mundo e no mercado de trabalho.

Para processamento de dados, análises estatísticas e de inteligência artificial, onde na parte inicial foi feito um pré-processamento que consiste em tratar os dados com técnicas computacionais e estatísticas para obter dados que possam ser usados no aprendizado de máquina (machine learning), que é uma das áreas da inteligência artificial comumente usadas atualmente para prever padrões e resultados baseados nos dados.

Como segunda parte utilizou-se 3 dos melhores modelos de classificação que foram eles o *Extreme Gradient Boosting* (*XGBoost*), *Light Gradient Boosting Machine* (*LightGBM*) e o *Categorical Boosting* (*CatBoost*) que são modelos que utilizam como base o modelo de arvore de decisão para obter seus resultados **(JANSEN, 2020),** e para terceira parte ouve a demonstração dos resultados obtidos

2 METODOLOGIA

2.1 Descrição do banco de dados

Os dados utilizados foram coletados no Hospital Fundação Assistencial da Paraíba (FAP), com uma amostra de 221 pacientes do sexo feminino a partir de uma amostragem aleatória simples de cerca de 2000 pacientes, que teve base no estudo de *(SILVA, 2022)*, A pesquisa retrospectiva foi realizada com autorização do comitê de ética (Certificado de Apresentação para Apreciação Ética - CAAE) da Universidade Federal de Campina, número 97198518.9.0000.5182. Temos então na Tabela 1 a descrição mais detalhada do banco de dados.

Tabela 1 - Descrição do banco de dados

Variáveis	Descrição
Morte	Indicando se a paciente veio ou não a óbito sendo sim (1) para os que vieram a óbito e não (0) para aqueles que não vieram a óbito, essa será á variável a ser classificada pelos modelos utilizados.
Local	Local da mama em que foi identificado o câncer, sendo E para local esquerdo da mama (local_E), D para local direito da mama (local_D), D e E para local direito e esquerdo da mama (local_D e E).
Receptor estrogênio	São proteínas presentes nas células do tecido mamário que se ligam ao hormônio estrogênio, podendo ser classificados como positivo e negativo. No presente estudo, utilizou-se 1 para positivo e 0 para negativo (recptor_estrogenio).
Receptor progesterona	Assim como o receptor estrogênio a progesterona é ligado as células do tecido mamário, foi registrado sendo 1 para positivo e 0 para negativo (recptor_progesterona).
P53	A proteína capaz de interromper o processo de divisão celular. Quando essa proteína não é encontrada, não é possível eliminar uma célula com potencial para se transformar em tumor. Portanto, utiliza-se 1 para positivo e 0 para negativo (p53).
C – erb – b2	Tem grande influência no tamanho do tumor, sendo positivo para um aumento de tumor e negativo para um não aumento, sendo 1 para positivo e 0 para negativo (c – e r b- b2).
Subtipo molecular	O subtipo molecular inclui Luminal A (subtipo_molecular_1), Luminal B (subtipo_molecular_2), HER2 superexpressão (subtipo_molecular_3) e triplo negativo (subtipo_molecular_4).
Ki – 67	Ki-67 (ou MIB-1) é uma substância liberada durante a divisão celular, sendo um nível de agressividade Ki-67 elevado quanto mais as células se dividem. É definido como abaixo de 15% pouco agressivo (ki_67_1), entre 15% e 50% agressivo (ki_67_2) e acima de 50% extremamente agressivo (ki_67_3).
Terapia adjuvante	Um tratamento complementar que é realizado após a cirurgia para remover o tumor, com o objetivo de reduzir o risco da recorrência da doença e melhorar a chance de sobreviver ao câncer. Neste estudo, H representa a hormonioterapia, Q representa a quimioterapia e R representa a radioterapia, sendo as variáveis: (terapia_adjuvante_H, terapia_adjuvante_Q, terapia_adjuvante_QH, terapia_adjuvante_QRH, terapia_adjuvante_R, terapia_adjuvante_RH, terapia_adjuvante_RQ, terapia_adjuvante_RQH.) a letra indica se a paciente foi submetida a uma ou mais tipos de terapias adjuvantes.
Número de Radioterapia	Quantidade de Radioterapias que a paciente foi submetida (n_de_radio).
Número de Quimioterapia	Quantidade de Quimioterapia que a paciente foi submetida (n_de_quimio).
Número de Hormonioterapia	Quantidade de Hormonioterapia que a paciente foi submetida (n_de_hormonio).
Idade	Idade das pacientes (idade).
Tempo de estudo	Tempo (em dias) em que as pacientes ficaram em estudo (TempoMorte).

Fonte: Elaborado pelo autor, 2023.

2.1 Linguagem Python

A linguagem de programação Python (2022)[2] vem sendo comumente usada atualmente como uma das melhores linguagens para a análise de dados e criação de modelos de *machine learning.*

[2] Foi utilizado o software Python (2022). Versão 3.9.1. Criado por guido Van Rossum.
Link: https://python.org.

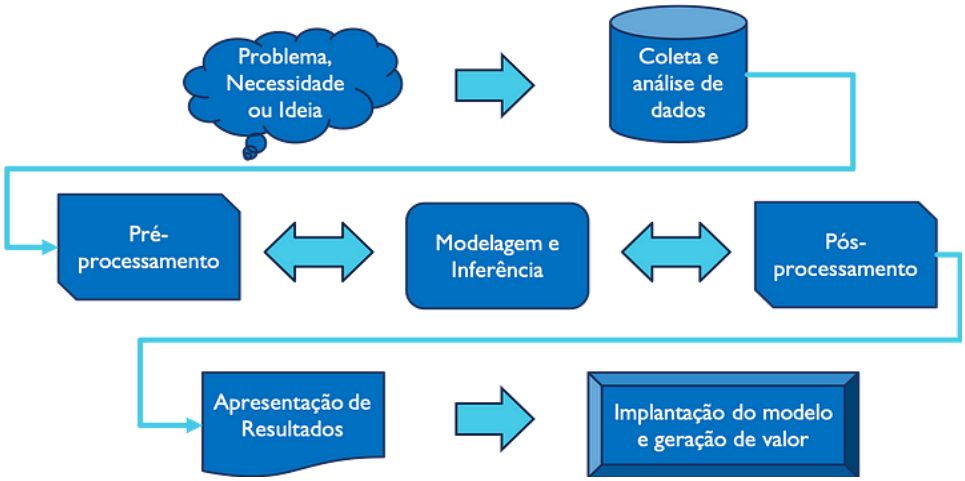
Foram utilizadas algumas bibliotecas como o Scikit-learn (Pedregosa et al., 2011) que conta com uma ampla gama de algoritmos e ferramentas para realizar tarefas de aprendizado de máquina, como classificação, regressão, *clustering* dentre outras, também contendo ferramentas de pré-processamento de dados e uma simplicidade em seu uso.

Outras bibliotecas utilizadas foram o Pandas que é utilizado para ciência de dados com o intuito de trabalhar com dados relacionais de maneira fácil, o *BayesSearchCV* (2020) que é utilizado para encontrar os melhores hiper parâmetros, *Numpy* (2021) para trabalhar com computação numérica, matrizes, vetores dentre outras funções que estão no seu escopo e as bibliotecas *Matplotlib* (2021) e *Seaborn* (2021) que são as principais do *Python* para criação de gráficos e visualização de dados de forma fácil e rápida.

2.3 Pré-processamento dos dados

São técnicas utilizadas para preparar os dados de aprendizado de máquina a terem um melhor desempenho, com principal foco em deixar os dados o mais perto do formato adequado para serem analisados, temos algumas etapas que normalmente são seguidas como exemplo: limpeza dos dados, transformação dos dados, redução de dimensionalidade, normalização ou padronização dos dados dentre outras etapas que dependem de como os dados estão estruturados *(MCKINNEY, 2017)*, como podemos ver a baixo na Figura 1, tem-se um esquema de como funciona as etapas de um projeto de aprendizado de máquina.

Figura 1 - Esquema de um projeto de *Machine Learning*



Fonte: *(ESCOVEDO, 2020)*

2.4 Limpeza dos dados

A limpeza dos dados consiste em realizar um processo para deixar os dados de forma que possam ser facilmente trabalhadas, sem valores faltantes, remoção de valores discrepantes e ouras inconsistências que possa haver no banco de dados.

2.4.1 Dados faltantes

Os dados faltantes podem ser um problema na hora de fazer uma implementação de algoritmos de *Machine Learning* e até mesmo para se trabalhar apenas com uma base para fazer

uma análise mais simples, alguns meios também são imputar a média, mediana, moda, regressão dentre outras formas de se obter bons dados para serem colocados no lugar dos dados faltantes *(ESCOVEDO, 2020)*.

No caso do aprendizado de máquina alguns algoritmos necessitam que não tenham dados ausentes funcionarem, porém foi utilizado modelos que não necessitam de tais implementações de dados pois eles trabalham bem com valores ausentes por usar um modelo de árvore de decisão, tornando assim o foco principal da pesquisa.

2.5 Tratamento dos dados

Na etapa de tratamento dos dados, tem-se como foco fazer algumas atividades como: remoção dos dados duplicados caso seja necessário, padronização ou normalização dos dados dependendo da situação que os dados e assim otimizando o aprendizado de máquina, também pode ser realizado a criação de variáveis em alguns casos entre outras formas de tratamento.

2.5.1 Padronização

Padronização dos dados têm como intuito principal transformar as variáveis na mesma ordem de grandeza assim se nos dados tiver variáveis com escalas diferentes, ajudam a evitar algum extremo ou até mesmo outliers nos algoritmos de *Machine Learning*. A padronização utilizada tem como resultado uma média igual a 0 e o desvio padrão igual 1 assim sendo obtido subtraído a média de cada valor dividido pelo desvio padrão *(BUSSAB, 2017)*.

2.5.2 Transformação de variáveis categóricas

No aprendizado de máquina, as variáveis categóricas precisam de uma transformação para ser atribuídas aos modelos para um bom ajuste, para isso algumas técnicas podem ser utilizadas, como:

- ***One-hot encoding***: É utilizada para transformar cada categoria em um vetor binário, onde cada vetor vai ter o número de elementos igual ao número de categorias únicas, será mostrado na Tabela 2.

- ***Label Encoding***: É utilizada para transformar dados categóricos em dados numéricos atribuindo um valor único para cada categoria, porém deve-se ter cuidado ao usar tal método pois pode gerar problemas de relação ordinal entre as categorias o que pode trazer viés aos resultados.

- ***Dummy encolding***: É utilizada para transformar as variáveis categóricas em novas variáveis binárias, ou seja, cria variáveis binárias para cada categoria na variável original, esta forma é bastante útil pois a maioria dos algoritmos não trabalham diretamente com variáveis categóricas.

Tabela 2 - Exemplo de variável One-hot encoding

Categorias	Café	Milho	Sal
Café	1	0	0
Milho	0	1	0
Sal	0	0	1

Fonte: Elaborado pelo autor, 2023.

2.5.3 Dados de classificação desbalanceados

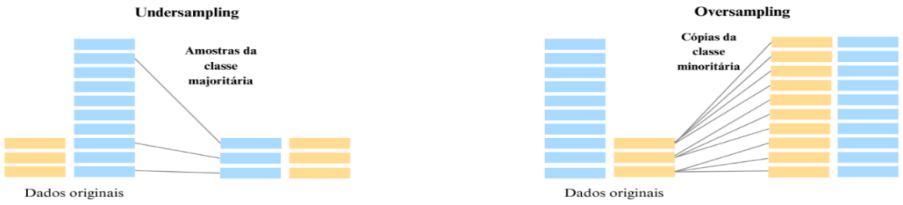
Os dados desbalanceados são algumas vezes um problema para os modelos de aprendizado de máquina até porque com os dados desbalanceados eles podem trazer resultados muito abaixo do esperado dependendo do algoritmo usado e da forma com que ele aprende.

Trabalhando com problemas reais muito dificilmente encontraremos dados que são balanceados, na maioria dos casos eles são desbalanceados o que pode trazer uma certa dificuldade, porém existem técnicas que são utilizadas para minimizar esse contratempo dos dados desbalanceados *(MURPHY, 2013)*, duas dessas técnicas são:

- ***Oversampling***: É uma das técnicas comumente utilizadas no aprendizado de máquina para dados desbalanceados. Essa técnica tem como foco aumentar o número de amostras da classe que têm menos valores, ela duplica dados a ponto de que a classe com menos dados fique equilibrado, algo que se deve-se ter cuidado com essa técnica, pois pode ocorrer o *overfitting* dos dados que é quando o modelo se ajusta bem aos dados de treinamento, mas não generaliza bem pra os dados de teste, fazendo com que o modelo não consiga generalizar e ser usado em aplicações, na Figura 2 pode-se observar o exemplo de *oversampling* e também será mostrado na Figura 9 alguns resultados.

- ***Undersampling***: É uma das técnicas que podem ser usadas para dados desbalanceados que faz com que tenha igualdade entre as classes fazendo com que selecione amostras da classe majoritária a ponto de igualar a quantidade de dados das classes, com isso pode-se obter classes iguais, porém é importante tomar cuidado com essa técnica pois como seleciona aleatoriamente pode ocorrer de não pegar os dados mais importantes nessa amostra o que vai levar a uma perca de informação, fazendo com que o modelo não se adeque bem *(GÉRON, 2019)*, na Figura 2 pode-se observar o exemplo de *undersampling*.

Figura 2 - Exemplos de *Undersampling* e *Oversampling*



Fonte: Elaborado pelo autor, 2023.

2.6 Aprendizado de máquina

O aprendizado de máquina é um campo muito amplo, que de diversas formas podem ser utilizados, no aprendizado de máquina utilizasse de programas computacionais para fazer com o ao ser inseridos dados de entrada a máquina possa identificar padrões, aprender sobre os dados e ainda ser capaz de tomar decisões com base nos dados *(ESCOVEDO, 2020)*, no campo da inteligência artificial o processo para o aprendizado de máquina envolve alguns passos e modelos que podem ser utilizados.

O aprendizado de máquina tem quatro tipos que são mais comumente usados que são eles: Aprendizado supervisionado, aprendizado não supervisionado, aprendizado semi-supervisionado, aprendizado por reforço.

- **Aprendizado supervisionado**: O aprendizado supervisionado é utilizado quando trabalhamos com algoritmos que aprendem com dados de treinamento, e que devem conter os dados de entrada. O modelo gera uma resposta que os dados de treinamento contêm as respostas que podem ser correspondentes com o gabarito, se houver divergência entre a resposta gerada pelo modelo e as respostas contidas nos dados de treinamento, o modelo deve ser reajustado para aprimorar sua capacidade de gerar respostas que sejam consistentes com o gabarito. *(BISHOP, 2006)*.

- **Aprendizado não supervisionado**: No aprendizado não supervisionado, não há um treinamento dos dados inseridos, ao contrário do aprendizado supervisionado que são chamados de dados não rotulados. O princípio desse aprendizado é o agrupamento e a redução de dimensionalidade, buscando padrões nos dados que podem servir como agrupamento ou simplificando o conjunto de dados, fazendo assim, ter uma redução na sua dimensionalidade *(BISHOP, 2006)*.

- **Aprendizado semi-supervisionado**: Nesse tipo de aprendizado, os dados são utilizados tanto de forma supervisionada, quanto de forma não supervisionada, utilizando dados não rotulados, isso permite que mais dados sejam utilizados para treinar o modelo resultando em um melhor desempenho *(ZHU, 2009)*.

- **Aprendizado por reforço**: No aprendizado por reforço, tem-se como principal foco uma inteligência que aprende com erros e acertos, isto é, ela recebe feedback a cada tentativa, fazendo com que ao longo do tempo melhore o desempenho com as tentativas para obter mais acertos *(SUTTON e BARTO, 2018)*, na Figura 3 pode-se observar os tipos de aprendizado de máquina e suas funcionalidades.

Figura 3 - Tipos de Aprendizado de máquina

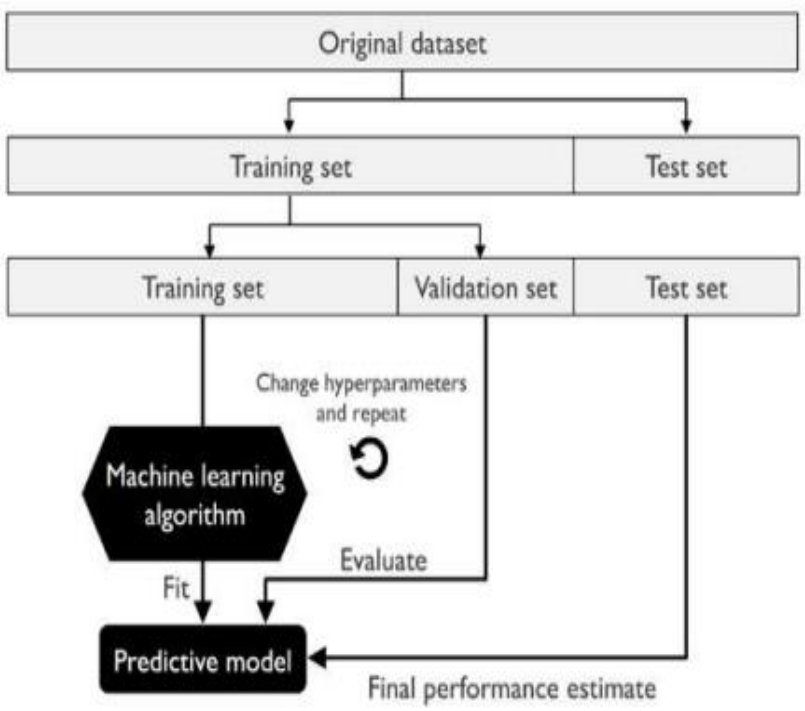
Fonte: (Alura, 2021)

2.6.1 Treino e teste

Uma das partes importantes sobre o aprendizado de máquina é definir o processo de treinamento e teste para que o modelo consiga generalizar e fazer previsões nas quais consigam bons resultados para novos dados. O treino e teste é uma etapa importante que divide os dados, de forma que tenha-se certa quantidade de dados para treino e para teste, alguns autores usam 70% dos dados para treino e 30% para testes, outros usam 80% para teste e 20% para treino *(MÜLLER e GUIDO, 2016)*. Essa divisão vai depender da quantidade de dados que estão em estudo, então deve-se escolher bem a quantidade para treino e teste. Na Figura 4 será mostrado o esquema de treino e teste e seu detalhamento.

Algumas observações importantes sobre o que podem trazer erro ao modelo é não dividir corretamente os dados, causando o *overffiting*, que é o ajuste excessivo dos dados, outro erro é usar os dados de teste no conjunto de treinamento fazendo com que o modelo não generalize bem e trazendo vazamento de dados para o modelo o que no meio da tecnologia chama-se (*data leakage*).

Figura 4 - Esquema de Treino e teste



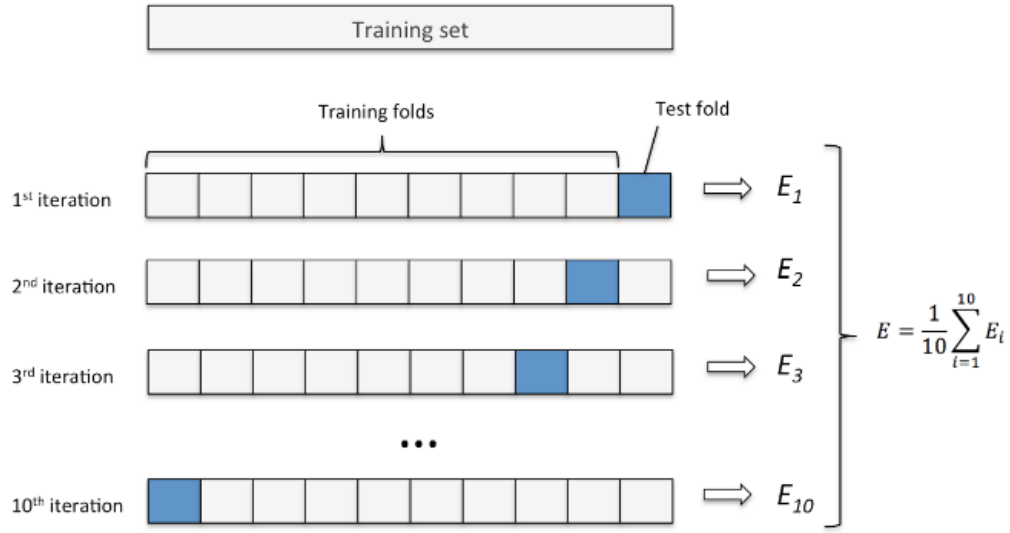
Fonte: *(RASCHKA e MIRJALILI, 2020)*

Na Figura 4 temos alguns termos do inglês que são eles: Original dataset, Training set, Test set, Validation set, Machine Learning, Fit, Predictive model, Change hyperparameters and repeat, evaluate e Final performance estimate que são na tradução para o português respectivamente eles: Conjunto de dados original, Conjunto de treino, Conjunto de teste, Conjunto de validação, Aprendizagem automática, Ajuste, Modelo preditivo, Alterar hiperparâmetros e repetir, avaliar e Estimativa de desempenho final

2.6.2 Validação cruzada

A validação cruzada também conhecida como *k-fold cross-validation* é uma técnica capaz de generalizar os dados, a técnica utiliza de uma divisão dos dados em treino e teste em que ele divide várias vezes em partições diferentes para obter a melhor avaliação do modelo. O *k-fold* divide os dados em k partes iguais, usando o k-1 para treinamento e a parte que sobra para teste assim repetindo o processo k vezes *(JANSEN, 2020)*, será mostrado na Figura 5 um exemplo de validação cruzada.

Figura 5 - Validação cruzada



Fonte: *(ROSAEN, 2016)*

Na Figura 5 temos alguns termos em inglês que são eles: Training set, training folds, test fold e iterations que no português são respectivamente: Conjunto de treino, dobras de treino, dobra de teste e iterações.

2.6.3 Introduções ao modelo Gradient Boosting

O Gradient Boosting é uma técnica do aprendizado de máquina que tem como base as árvores de decisão para melhorar o modelo a cada interação, isto é, ele faz novas árvores de decisão ajustando os resíduos do modelo anterior fazendo com que o minimize *(SILVA & SÁTIRO, 2022)*.

Ele pode ser usado em problemas de regressão e classificação *(WADE, 2020)*, sendo ele o modelo que fornece base para outros modelos mais robustos que são atualmente usados para grandes conjuntos de dados ele tem uma eficiência que é fundamental para algoritmos de aprendizado de máquina que são o *Extreme Gradient Boosting* (*XGboost*), *Light Gradient Boosting Machine* (*LightGBM*) e o *Categorical Boosting* (*CatBoost*) *(JANSEN, 2020)*.

2.6.3.1 Extreme Gradient Boosting

O modelo de aprendizado de máquina mais conhecido como XGBoost que foi criado por Tianqi Chen e Carlos Guestrin como projeto de pesquisa da universidade de Washington *(CHEN e GUESTRIN, 2016)*.

O XGBoost é um dos modelos de aprendizado de máquina que têm sido, atualmente, bastante utilizados no meio da inteligência artificial. Sua estrutura tem como base árvore de

decisão e com isso seguindo o modelo Gradient Boosting, busca diminuir os erros a cada nova árvore de decisão.

O modelo tem como umas das suas características a otimização que busca o ajuste dos pesos do modelo, que basicamente é quando o algoritmo busca os pesos ideais para minimizar a função de perda, de acordo com Jabeur, Mefteh-Wali e Viviani (2021), a fórmula de saída é calculada com a seguinte equação (1):

$$yhat_i^T = \sum_{k=1}^{T} f_k\,(X_i) = yhat_i^{T-1} + f_T(X_i) \quad (1)$$

Onde tem-se:

- $yhat_i^{T-1}$: é a arvore de decisão gerada.
- $f_T(X_i)$: é o modelo de árvore recém-criado.
- T: é o total de árvores no modelo.

Algumas vantagens de se usar o algoritmo XGBoost são a de trabalhar bem com dados faltantes, capacidade de lidar com dados não lineares e não homogêneos.

2.6.3.2 Light Gradient Boosting Machine

Também conhecido como LightGBM é um dos modelos de aprendizado de máquina que utilizam como base árvore de decisão e uma de suas peculiaridades é que suas árvores crescem verticalmente, enquanto a maioria dos algoritmos segue com árvores crescendo horizontalmente. Este algoritmo tem uma maior eficiência e velocidade de treinamento.

O modelo tem como característica manter a maior acurácia com o máximo de ganho de informações que é denominado de Gradient-based One-Side Sampling (GOSS) *(JABEUR, MEFTEH-WALI e VIVIANI, 2021)*. Segundo Jabeur, Mefteh-Wali e Viviani (2021), a função estimada integra várias T árvores de regressão definidas na equação (2):

$$Y_t = \sum_{h=1}^{T} f_t(X) \quad (2)$$

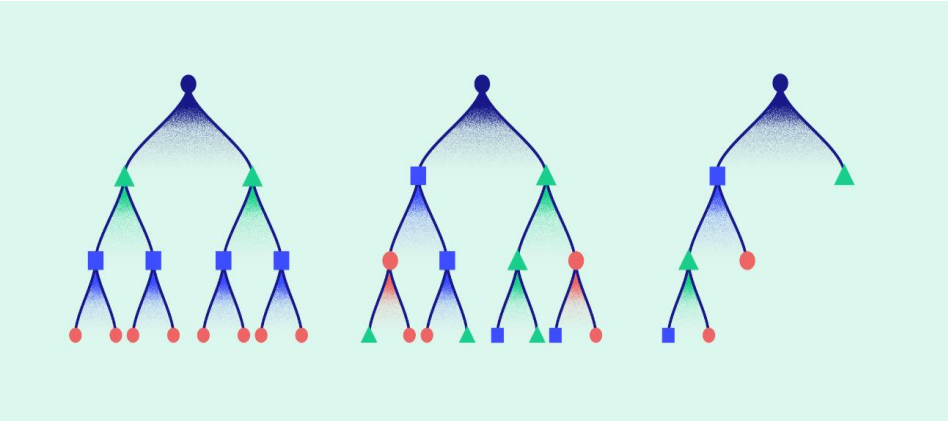
Onde $f_t(X)$: é a árvore de regressão.

2.6.3.3 Categorical Boosting

O modelo Categorical Boosting mais conhecido como CatBoost é um dos modelos que, assim como o XGBoost e LightGBM, tem como base árvores de decisão.

Ele tem a capacidade de trabalhar bem com dados categóricos sem a necessidade de transformar em *dummies*. Assim, como os outros dois modelos, ele consegue lidar com dados ausentes sem necessidade de fazer manipulações de pré-processamento *(JANSEN, 2020)*. Na Figura 6 pode-se observar.

Figura 6 - Comparação entre *CatBoost*, *XGBoost* e *LightGBM* respectivamente



Fonte: ALAL (2019)

Cada um desses algoritmos tem sua própria abordagem para selecionar as melhores divisões nos nós das árvores.

O Catboost utiliza um método chamado "árvores de decisão esquecidas". Antes do treinamento, as possíveis faixas de valores de cada característica são divididas em intervalos com base em valores de limiar. Esses pares característica-divisão são usados para selecionar a melhor divisão em cada nível da árvore, minimizando a perda de acordo com uma função de penalização (ALAL, 2019).

O XGBoost oferece diferentes métodos para selecionar as melhores divisões. Um deles é um algoritmo baseado em histograma, que agrupa características contínuas em compartimentos discretos e usa esses compartimentos para encontrar as divisões nos nós.

O LightGBM utiliza um método chamado "amostragem unilateral baseada no gradiente" (GOSS). Ele seleciona as divisões com base nos gradientes (ou seja, erros) das instâncias.

2.6.4 Hiperparâmetros

Os hiperparâmetros são parâmetros em que podem ser definidos para melhorar o modelo. Está é uma parte fundamental ao criar um modelo, pois pode-se obter os melhores resultados. Alguns desses hiperparâmetros são o número de árvores de decisão, a taxa de aprendizado em um algoritmo, máximo de folhas por árvore dentre vários outros *(RASCHKA e MIRJALILI, 2020)*, na Figura 3 será mostrado alguns dos hiper parâmetros dos modelos selecionados, fica a cargo do leitor a busca por cada um dos hiper parâmetros selecionados e suas funcionalidades.

Tabela 3 - Hiperparâmetros do XGBoost, LightGBM e CatBoost

Modelos	Hiperparâmetros
XGBoost	n_estimators, max_depth, learning_rate, subsample, colsample_bytree, gamma, lambda (reg_lambda), alpha (reg_alpha), min_child_weight, objective.
LightGBM	num_interations (n_estimators), learning_rate (eta), max_depth, num_leaves, min_data_in_leaf (min_child_samples), feature_fraction (colsample_bytree), bagging_fraction (subsample), bagging_freq (subsample_freq), lambda (reg_lambda), alpha (reg_alpha), min_gain_to_split (min_split_gain), max_bin.
CatBoost	interations, learning_rate (eta), depth, l2_leaf_reg, random_strength, border_count, bagging_temperature, min_data_in_leaf (min_child_samples), max_bin, grow_policy, cat_features.

Fonte: Elaborado pelo autor, 2023.

Encontrar os melhores hiperpârametros levar várias horas a depender da quantidade de dados. Para isso existem algumas técnicas que buscam maximizar o trabalho de forma a encontrar os melhores parâmetros em pouco tempo.

Uma delas, que será abordada, é a otimização bayesiana que consiste em encontrar os melhores hiperparâmetros com fundamentos do teorema de Bayes. A cada hiperparâmetros testado ele seleciona os melhores e faz com que a distribuição do parâmetro probabilidade do parâmetro seja atualizada usando o teorema de Bayes e obtendo os melhores hiperparâmetros com base nos resultados anteriores.

Não sendo necessária verificar todos os parâmetros possíveis levando uma grande quantidade de tempo *(GONZÁLEZ, 2020)*.

- **BayesSearchCV**: O BayesSearch é uma biblioteca para seleção de hiperparâmetros que utiliza de pesquisa bayesiana. Ele permite uma busca mais eficiente com otimização bayesiana que aplica uma abordagem baseada em modelos probabilísticos para inferir os melhores hiperparâmetros com base nas avaliações do modelo. Normalmente utilizado quando o espaço de hiperparâmetros é grande e necessita de uma demanda de tempo menor para encontrar os melhores hiperparâmetros diferente de outras bibliotecas que testam todas as combinações possíveis a custo de muito processamento de máquina e tempo como o GridSearchCV.

2.6.5 Métricas de avaliação

As métricas de avaliação do modelo são uma parte crucial do processo de aprendizado de máquina, elas permitem observar os resultados que os modelos obtiveram e classificam, prever ou agrupar os dados *(RASCHKA e MIRJALILI, 2020)*. Existem diversas métricas que podem ser utilizadas a depender do modelo e algumas delas são: acurácia, precisão, *recall*, área sobre a curva ROC e matriz de confusão.

2.6.5.1 Matriz de confusão

Temos como base principal para modelos de classificação binária a matriz de confusão. Com ela é possível de definir se o resultado pertence a uma das duas classes (1 e 0), a Matriz de confusão é formada por 4 elementos que são eles classificados como: verdadeiros positivos (VP); que é quando os resultados previstos foram corretamente classificados; verdadeiros negativos (VN); que é quando os resultados não previstos foram corretamente classificados; falso positivo (FP); que os resultados previstos foram incorretamente classificados e falso negativo (FN); que basicamente é quando os resultados não previstos foram incorretamente *(GÉRON, 2019)*. Na Tabela 4 pode-se ver um exemplo de como funciona a matriz de confusão.

Tabela 4 - Exemplo de matriz de confusão

	Fumante	Não fumante
Fumante	4 (VP)	3 (FP)
Não fumante	2 (FN)	1 (VN)

Fonte: Elaborado pelo autor, 2023.

Onde temos:

- Previu fumante 4 vezes corretamente.

- Previu não fumante 1 vez corretamente.
- Previu fumante 3 vezes incorretamente.
- Previu não fumante 2 vezes incorretamente.

A partir da matriz de confusão pode-se resultar em outras métricas, sendo a

Acurácia

A acurácia é umas da métricas utilizadas para avaliação de modelos, ela indica a proporção dos dados que foram classificados corretamente em relação ao total.

$$\frac{VP + VN}{VP + VN + FP + FN}$$

Precisão

É uma métrica que indica a proporção de dados que foram classificados corretamente em relação ao número total classificados como positivos.

$$\frac{VP}{VP + FP}$$

Recall

Também conhecida como sensibilidade essa métrica indica a proporção de dados que foram classificados corretamente em relação ao número total que realmente são positivos.

$$\frac{VP}{VP + FN}$$

F1-score

Essa métrica faz uma média harmônica entre precisão e recall dando um peso em ambas as medidas.

$$2x\frac{Precisão\ x\ Recall}{Precisão + Recall}$$

2.6.5.2 Área sob a curva ROC

A curva ROC (*Receiver Operating Characteristic*) mostra o quão bom o modelo classificou as classes, existem dois parâmetros utilizados que são eles a taxa de verdadeiro positivo e a taxa de falso positivo, com base nisso podemos calcular a área sob a curva ROC que é também conhecido como AUC (*Area Under the ROC Curve*). A AUC tem um valor entre 0 e 1 e quanto mais próximo de 1 melhor o modelo e quanto mais próximo de 0 pior *(MÜLLER e GUIDO, 2016)*.

No gráfico da curva ROC, o eixo y representa a taxa de verdadeiros positivos (TPR), enquanto o eixo x representa a taxa de falsos positivos (FPR), termos usados para criação da curva ROC:

$$TPR/sensibilidade = \frac{VP}{VP + FN}$$

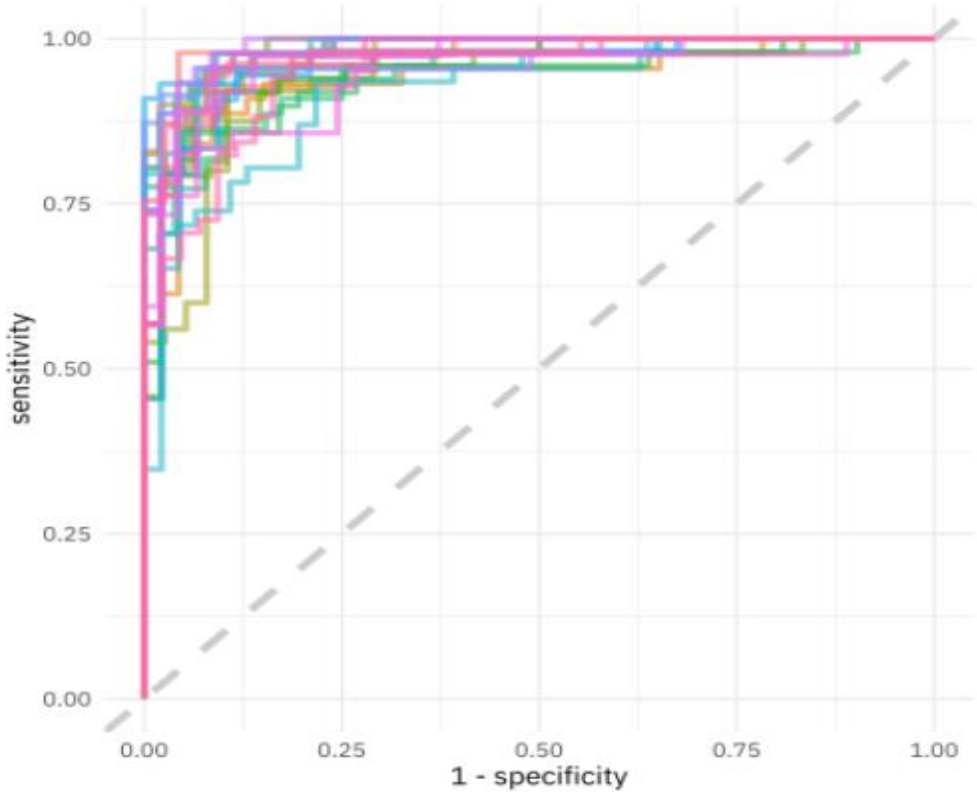
$$Especificidade = \frac{VN}{VN + FP}$$

$$FPR = 1 - Especificidade = \frac{FP}{VN + FP}$$

A sensibilidade e a especificidade têm uma relação inversa no aprendizado de máquina. Quando aumentamos a sensibilidade, a especificidade diminui e vice-versa. Isso significa que ao reduzirmos o limiar, a sensibilidade aumenta, mas a especificidade diminui. Por outro lado, ao aumentarmos o limiar, a especificidade aumenta, mas a sensibilidade diminui. É importante destacar que ao aumentar a sensibilidade, também aumentamos a taxa de falsos positivos (FPR).

Com base na curva ROC, pode-se avaliar a capacidade do modelo de equilibrar a taxa de verdadeiros positivos com a taxa de falsos positivos para diferentes pontos de corte. Assim, um modelo com um bom desempenho terá uma curva mais próxima ao canto superior esquerdo (BISHOP, 2006), Na Figura 7 podemos observar como funciona a curva ROC.

Figura 7 - Exemplo de curva ROC



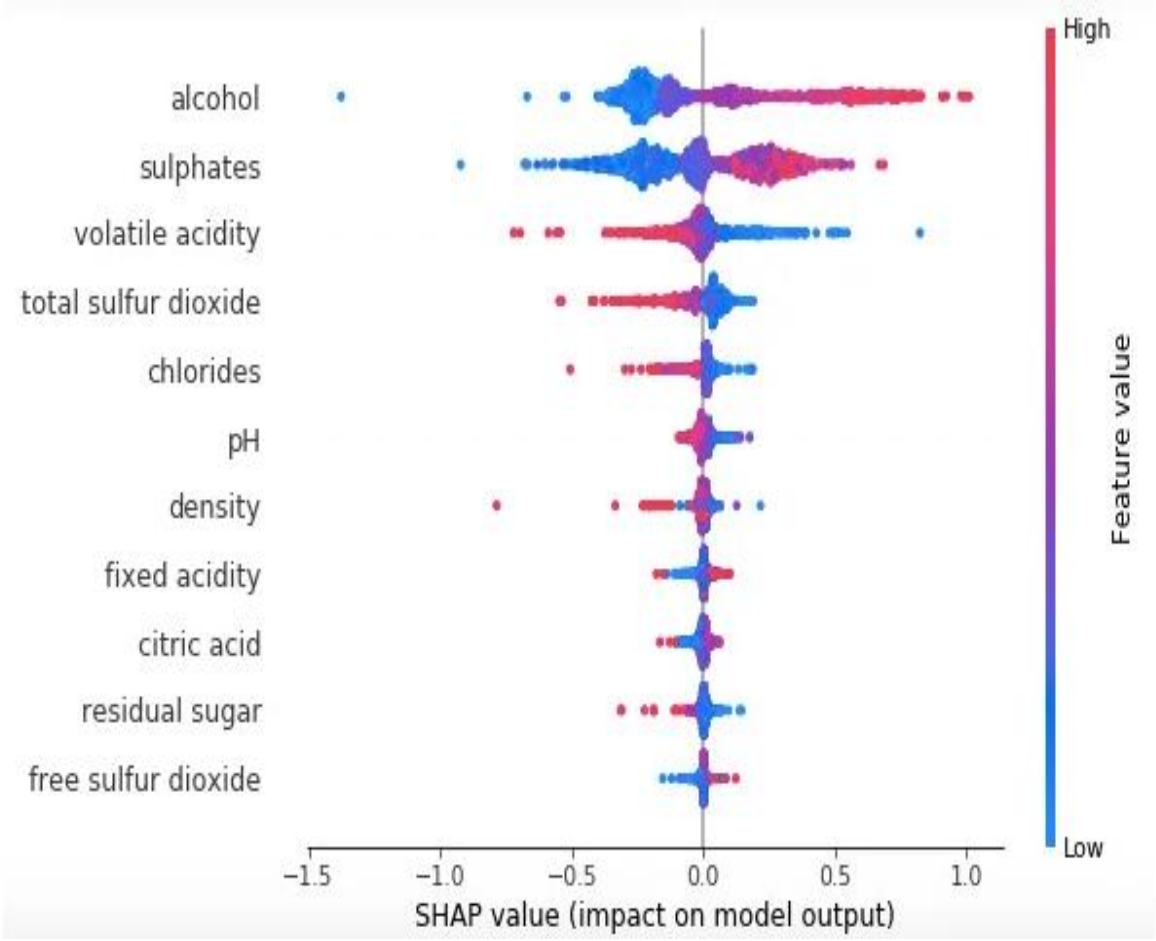
Fonte: *(SILGE, 2020)*

2.6.6 Shapley value

O Shapley value que foi desenvolvido por Lloyd Shapley, e inicialmente introduzido para encontrar a importância de cada jogador em um jogo cooperativo e medir sua contribuição (MOLNAR, 2019*)*. O exemplo de como obter valor de Shapley é calculando a contribuição marginal de cada jogador, quando este se junta a diferentes grupos de jogadores. A contribuição marginal é a diferença entre o valor total dos jogadores antes e depois da inclusão do jogador em questão. Logo para obter á contribuição correta de cada jogador, é feita uma média de contribuições marginais considerando todas as permutações possíveis da coalizão.

A partir do estudo de Shapley pode-se interpretar a contribuição de cada variável para previsão do modelo sendo usada atualmente como avaliador de variáveis no aprendizado de máquina *(KUO, 2019)*, Na Figura 8 temos um exemplo de Shapley value.

Figura 8 - Exemplo de gráfico de Shapley value



Fonte: (KUO, 2019)

Tem-se que na Figura 8 cada linha é uma variável e cada ponto no gráfico representa um exemplo do banco de dados. Quando tem-se um ponto vermelho indica um valor alto comparado com os demais valores da variável e quando o ponto for azul, tem-se um valor baixo comparado com os demais valores da variável. A posição do ponto no eixo horizontal indica basicamente o efeito, quando o ponto é mais à direita sua contribuição é positiva.

3 RESULTADOS

Tem-se como alguns resultados iniciais que são mostrados na Tabela 5 com informações sobre os dados do estudo que se pode observar a quantidade de variáveis, o número de observações dentre outras informações que são importantes para os resultados obtidos.

Tabela 5 - Informação dos dados

Número de variáveis	14
Número de observações	221
Células ausentes	1265
Células ausentes (%)	19,70%
Linhas duplicadas	14
Linhas duplicadas (%)	06,30%
Variáveis numéricas	5
Variáveis categóricas	9

Fonte: Elaborado pelo autor, 2023.

Com base na Tabela 5, tem-se uma porcentagem alta de dados ausentes, que fica em torno de 19,70% dos dados, isto é comumente adequado e seria necessário colocar valores como a média e moda para completar, porém, ao trabalhar com bons modelos que se utilizam bem com dados ausentes, se torna um meio de contornar a situação, o número de observações condiz com a quantidade de observações das pacientes (linhas do Excel). As células ausentes são as celular do Excel que estão ausentes. A porcentagem de dados ausentes é referente ao total do banco de dados que, no caso temos 19,70% de dados ausentes com base nisso foi trabalhado com 80,30% dos dados. A seguir temos as variáveis categóricas na Tabela 6 do seguinte estudo, na qual algumas contém muitos dados ausentes em algumas variáveis, com isso, temos resumos descritivos dos dados.

Tabela 6 - Descritiva das variáveis categóricas

(continua)

Variáveis	**Categorias**	**Quantidade**	**Porcentagem (%)**
Morte	Não	169	76,50%
	Sim	52	23,50%
	Ausentes	0	00,00%
Local	Esquerdo (E)	103	46,60%
	Direito (D)	105	47,50%
	Esquerdo e Direito (E e D)	7	03,20%
	Ausentes	6	02,70%
Receptor estrogênio	Negativo	22	10,00%
	Positivo	106	48,00%
	Ausentes	93	42,00%
Receptor progesterona	Negativo	41	18,55%
	Positivo	85	38,46%
	Ausentes	95	42,99%
P53	Negativo	74	33,49%
	Positivo	42	19,00%
	Ausentes	105	47,51%

Tabela 6 - Descritiva das variáveis categóricas

(conclusão)

Variáveis	Categorias	Quantidade	Porcentagem (%)
C – erb – b2	Negativo	48	21,72%
	Positivo	70	31,67%
	Ausentes	103	46,61%
Subtipo molecular	Luminal A	51	23,08%
	Luminal B	49	22,17%
	Her2 superexpressão	9	04,07%
	Triplo negativo	10	04,53%
	Ausentes	102	46,15%
Ki – 67	Menor que 15%	53	23,98%
	Entre 15% e 50%	41	18,55%
	Maior que 50%	15	06,79%
	Ausentes	112	50,68%
Terapia adjuvante	H	10	04,53%
	Q	7	03,17%
	QH	16	07,24%
	QRH	1	00,45%
	R	46	20,81%
	RH	39	17,65%
	RQ	29	13,12%
	RQH	70	31,67%
	Ausentes	3	01,36%

Fonte: Elaborado pelo autor, 2023.

Caso haja dúvida, mostrasse as descrições de todas as variáveis na Tabela 1 para apoio a Tabela 6.

Na Tabela 7, tem-se os resultados descritivos das variáveis numéricas que mostram a média de idade das pacientes com câncer de mama em torno de 59 anos e que a média de dias em que foram estudadas em torno de 1148 dias, mais ou menos 3 anos.

Tabela 7 - Descritiva das variáveis numéricas

Variáveis	Média	Desvio padrão	Mínimo	25%	50%	75%	Máximo
Tempo de estudo	1148 dias	1031 dias	10 dias	223 dias	951 dias	1848 dias	4809 dias
N° de radioterapia	25,65	14,58	0	25	25	30,25	90
N° de quimioterapia	08,91	12,67	0	0	4	14	67
N° de hormonioterapia	26,20	25,60	0	0	12	58	109
Idade	59 anos	13 anos	30 anos	49 anos	59 anos	68 anos	89 anos

Fonte: Elaborado pelo autor, 2023.

Inicialmente, foi utilizado um pré-processamento nos dados do qual nota-se que não há como fazer um adicionamento de valor para os dados ausentes. Haviam muitos dados que não tinham informações suficientes para tal procedimento, logo após foi realizada a padronização dos dados numéricos para estarem na mesma escala fazendo com que o algoritmo entenda de forma clara e sem dar mais pesos para valores maiores, assim ao fazer o processamento básico dos dados foi escolhido os modelos que mais se adequariam aos dados que foram os modelos *CatBoost*, *LightGBM* e *XGBoost*, dos quais as principais características para a escolha dos modelos foram que alguns dos modelos trabalham com dados faltantes e não necessariamente precisam de padronização dos dados mesmo já sendo realizado.

Tabela 8 - Resultado do *oversampling*

Média da acurácia sem smote	Média da acurácia com smote
81,96	79,46

Fonte: Elaborado pelo autor, 2023.

O resultado do *oversampling* para tentar contornar o balanceamento dos dados não obteve-se relevância para o modelo, causando um ajuste excessivo dos dados. Pode-se observar na Tabela 8 o resultado da acurácia sem o uso do *oversampling* funcionou melhor, a realização do *undersampling* não seria possível por conta da quantidade de dados.

Na Tabela 9 mostra-se os resultados obtidos dos modelos escolhidos no estudo e suas principais métricas de avaliação de modelos.

Tabela 9 - Avaliação dos modelos *Catboost*, *LightGBM* e *XGBoost*

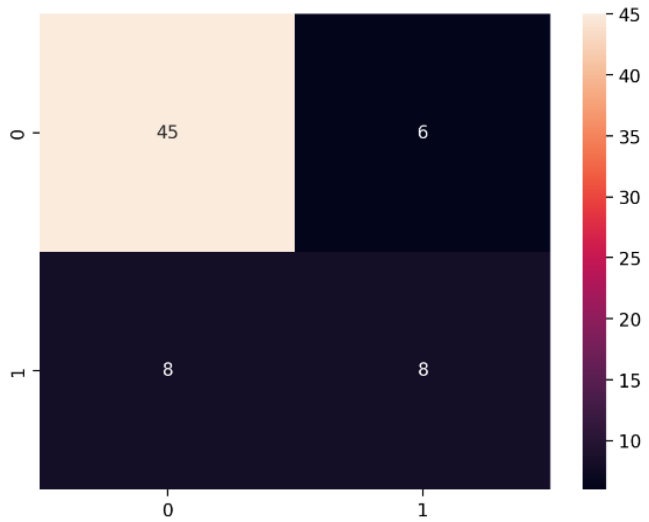
Modelos	**Acurácia**	**Precisão**	**Recall**	**F1-score**	**AUC**
CatBoost	00,78	00,68	00,64	00,65	00,64
LightGBM	00,82	00,76	00,71	00,73	00,71
XGBoost	00,79	00,71	00,69	00,70	00,69

Fonte: Elaborado pelo autor, 2023.

Com base na análise do desempenho do modelo na Tabela 9, observou-se que o *LightGBM* apresentou os melhores resultados, todas as métricas avaliadas foram ligeiramente superiores em comparação aos outros dois modelos. Esse fato é significativo, pois qualquer melhoria no desempenho do modelo é considerada relevante para atingir os objetivos do estudo, que, nesse caso, é a predição de óbito de pacientes com câncer de mama.

Ao analisar a matriz de confusão que organiza os resultados em uma matriz, onde as linhas representam as classes reais e as colunas representam as classes previstas pelo modelo, tem-se que as colunas que se encaixam na diagonal são os que o aprendizado de máquina conseguiu prever corretamente da classe 0 e 1 os resultados obtidos para os modelos foram:

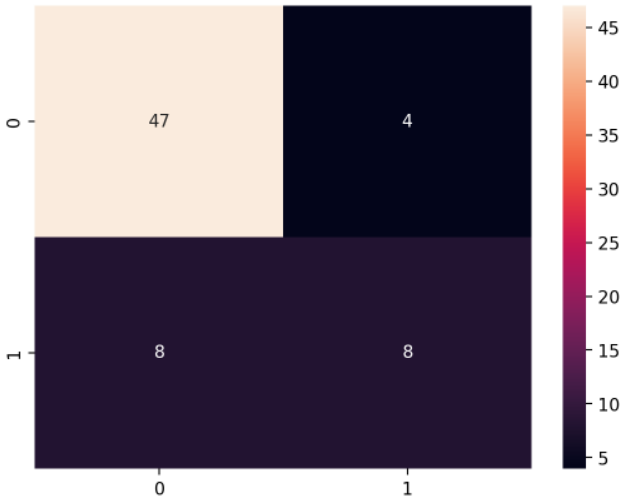
Figura 9 - Matriz de confusão do modelo *XGBoost*



Fonte: Elaborado pelo autor, 2023.

Onde tem-se que nos dados de teste na Figura 9, o modelo acertou 45 resultados de pacientes que não iriam vir a óbito de 51 possíveis que está em torno de 88,00%. Também acertou 8 de 16 possíveis, em torno de 50,00%. Com isso temos um modelo que identificou razoavelmente para pacientes que viriam a óbito com uma assertividade com bons resultados para pacientes que não iriam vir a óbito.

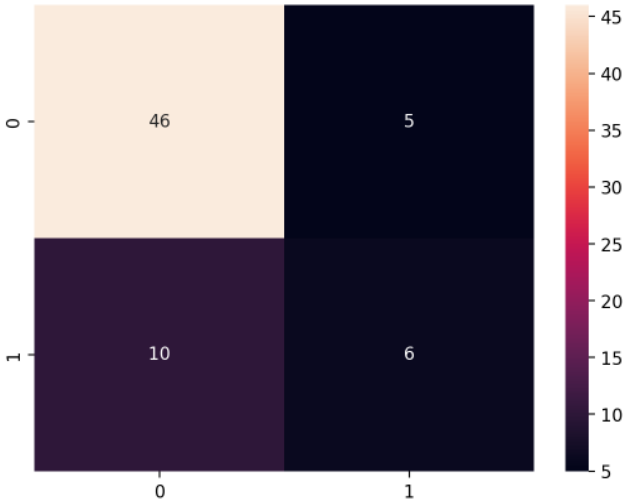
Figura 10 - Matriz de confusão do modelo *LightGBM*



Fonte: Elaborado pelo autor, 2023.

No modelo *LightGBM* da Figura 10, tem-se como resultados da matriz de confusão que a partir dos dados que foram para teste o modelo obteve 47 acertos de 51 possíveis, fica em torno de 92,00%. Obteve 8 acertos de 16 possíveis que dá em torno de 50,00%. Os resultados para pacientes que não vieram a óbito foram pouco melhores que o modelo *XGBoost*.

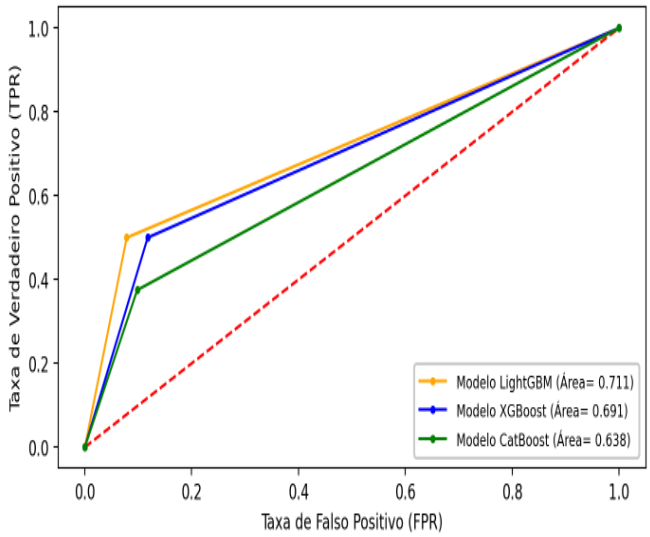
Figura 11 - Matriz de confusão do modelo *CatBoost*



Fonte: Elaborado pelo autor, 2023.

No modelo *CatBoost* (Figura 11) obteve-se resultados de assertividade para as pacientes que não viriam a óbito de 46 de um total de 51 possíveis ou seja 90,00%. E das pacientes que viriam a óbito ele identificou 6 de um total de 16 possíveis que está em torno de 38,00%.

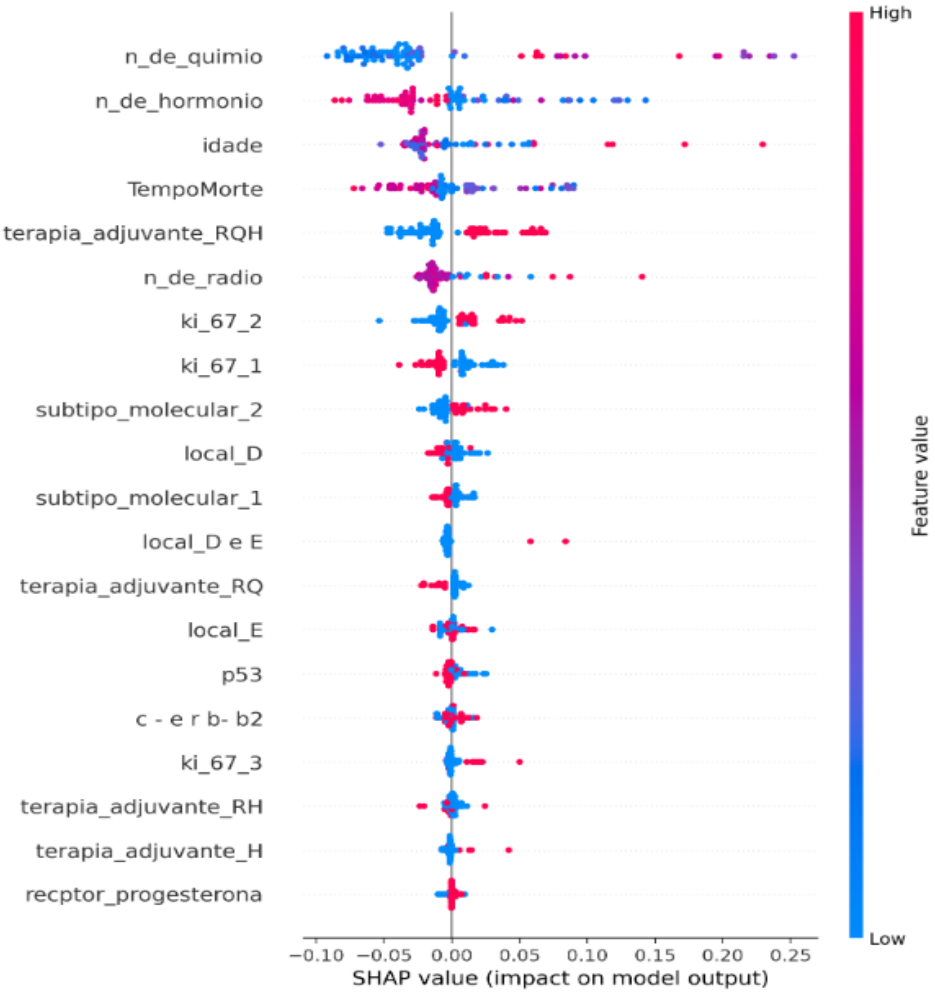
Figura 12 - Gráfico da área sob a curva ROC dos modelos *CatBoost*, *XGBoost* e *LightGBM*



Fonte: Elaborado pelo autor, 2023.

Tem-se na Figura 12 as curvas ROC dos respectivos modelos. Com isso o modelo *LightGBM* obteve resultados mais expressivos.

Figura 13 - Gráfico do SHAP value



Fonte: Elaborado pelo autor, 2023.

No gráfico ilustrado na Figura 13, constata-se que as variáveis de maior influência nos modelos são o número de quimioterapia e o número hormonioterapia, as quais têm o potencial de exercer tanto impacto positivo quanto negativo sobre o modelo em questão. A título de exemplo, a variável "n_de_quimio", quando apresenta um valor mais azul e ao lado esquerdo do eixo y, possui menor peso sobre o modelo, exercendo um efeito negativo, em virtude de sua posição no eixo horizontal esquerdo, destacado pela cor azul. Similarmente, a variável "n_de_hormonio", destacada em vermelho, também exerce um efeito negativo sobre o referido modelo, porém com uma um peso maior sobre o modelo por conta da sua coloração em vermelho, uma vez que se encontra no eixo horizontal esquerdo.

4 DISCUSSÃO

Com base nas informações, destacam-se pontos de relevância no estudo, uma das principais dificuldades enfrentadas foi a insuficiente quantidade de dados para permitir uma aprendizagem de máquina efetiva. No entanto, foi possível conduzir o estudo e obter resultados satisfatórios e notáveis.

Devido à escassez de dados, certos procedimentos padrões não puderam ser realizados, como a imputação de dados para auxiliar o modelo. Ao usar técnicas como o *oversampling* foi observado um ajuste excessivo dos dados (*overfitting*), o que inviabilizou condições adequadas para o modelo, sendo assim, para evitar o mínimo possível de erros, foi utilizado os modelos que tem configurações para suportar os dados ausentes e otimizar o modelo para diminuir o máximo possível de erros.

Além disso, é recomendado explorar alguns pontos para estudos futuros, um aspecto importante é a obtenção maior de dados, a fim de obter resultados ainda mais relevantes para a pesquisa, outra abordagem interessante seria a busca por outras variáveis significativas que possam auxiliar o modelo a compreender com maior precisão os padrões em análise.

5 CONCLUSÃO

Diante dos resultados obtidos, os modelos de aprendizado de máquina apresentados neste estudo, *XGBoost*, *LightGBM* e *CatBoost*, foram eficientes para a predição de morte em pacientes diagnosticados com câncer de mama. Mesmo com a presença de dados faltantes, os modelos foram capazes de realizar previsões com uma acurácia satisfatória, com destaque para o modelo *LigthGBM* que obteve uma acurácia de 82,00% e uma precisão de 76,00%.

Além disso, os resultados mostraram que o pré-processamento inicial foi essencial para que os modelos pudessem interpretar os dados categóricos, o que consequentemente realizou previsões mais precisas. A curva ROC que foi avaliada teve como principal modelo o *LightGBM*, com área em torno de 00,71 que indica uma qualidade eficiente para distinguir entre verdadeiros positivos e falsos positivos, mesmo com a quantidade de dados não sendo tão favorável e não deixando o modelo tão preciso quanto gostaria. Este estudo demonstrou a aplicação bem-sucedida de modelos de aprendizado de máquina para predição de morte em pacientes com câncer de mama, contribuindo assim para a melhoria da saúde pública e abrindo caminho para pesquisas futuras sobre estudos em pacientes com câncer de mama.

REFERÊNCIAS.

ALAL, N. **XGBoost, LightGBM or CatBoost – which boosting algorithm should I use?**,. 2019. <Urlhttps://www.riskified.com/resources/article/boosting-comparison/>. Acesso em: 7 mar. 2023. Citado na página 23.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. [S.l.]: Springer, 2006. Citado duas vezes nas página 18 e 28.

BRUNO, R. Alura. **Desmistificando Termos de Machine Learning:** Tipos de Aprendizado, 2021. Disponivel em: <https://www.alura.com.br/artigos/desmistificando-termos-machine-learning-tipos-aprendizado>. Acesso em: 02 mar. 2023. Citado na página 19.

BUSSAB, W. O. &. M. P. A. **Estatística básica**. 9. ed. São Paulo: Saraiva, 2017. Citado na página 15.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. **In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**, 2016. p. 785-794. Citado na página 22.

ESCOVEDO, T. &. K. A. S. **Introdução a Data Science:** Algoritmos de Machine Learning e métodos de análise. São Paulo: Casa do Código, 2020. Citado duas vezes nas páginas 14 e 17.

GÉRON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow:** Concepts, Tools, and Techniques to Build Intelligent Systems. [S.l.]: O'Reilly Media, 2019. Citado duas vezes nas páginas 17 e 25.

GONZÁLEZ, J. . D. Z. . &. D. **Bayesian optimization for machine learning**. [S.l.]: Morgan & Claypool Publishers, 2020. Citado na página 25.

INCA. Outubro Rosa. **gov.br**, 2022. Disponivel em: <https://www.gov.br/inca/pt-br/assuntos/campanhas/2022/outubro-rosa>. Acesso em: 25 jan. 2023. Citado na página 11.

JABEUR, S. B.; MEFTEH-WALI, S.; VIVIANI, J.-L. Forecasting gold price with the XGBoost algorithm and SHAP interaction values. **Annals of Operations Research**, 2021. p. 1-21. Citado na página 22.

JANSEN, S. **Machine Learning for Algorithmic Trading**. 2. ed. Birmingham: Packt Publishing, 2020. Citado duas vezes nas páginas 11, 20 e 22.

KUO, C. **Medium**, 2019. Disponivel em: <https://medium.com/dataman-in-ai/explain-your-model-with-the-shap-values-bc36aac4de3d>. Acesso em: 01 abr. 2023. Citado na página 29.

MATPLOTLIB. **Matplotlib: Python plotting**. Versão 3.4.2. 2021. Disponível em: https://matplotlib.org/. Acesso em: 05 mar. 2023. Citado na página 13.

MCKINNEY, W. A. A. T. **Python for Data Analysis:** Data Wrangling with Pandas, NumPy, and IPython. 2. ed. [S.l.]: O'Reilly Media, Inc., 2017. Citado na página 14.

MOLNAR, C. **Interpretable Machine Learning**. 1. ed. [S.l.]: Leanpub, 2019. Citado na página 29.

MÜLLER, A. C.; GUIDO, S. **Introduction to Machine Learning with Python:** A Guide for Data Scientists. [S.l.]: O'Reilly Media, Inc., 2016. Citado duas vezes nas páginas 19 e 27.

MURPHY, K. P. **Aprendizado de Máquina:** Uma Abordagem Estatística. [S.l.]: Bookman , 2013. Citado na página 16.

NUMPY. **NumPy: the fundamental package for scientific computing with Python**. Versão 1.21.0. 2021. Disponível em: https://numpy.org/. Acesso em: 05 mar. 2023. Citado na página 13.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). **Scikit-learn: Machine learning in Python**. Journal of Machine Learning Research, 12(Oct), 2825-2830. Citado uma vez na página 13

PYTHON SOFTWARE FOUNDATION. Python 3.10.0 documentation. **Python**, 2021. Disponivel em: <https://docs.python.org/3/>. Acesso em: 25 jan. 2022. Citado na página 13.

RASCHKA, S.; MIRJALILI, V. **Python Machine Learning:** Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2. 3. ed. [S.l.]: Packt Publishing, 2020. Citado três vezes nas páginas 20, 24 e 25.

ROSAEN, K. **ML Study**, 2016. Disponivel em: <http://karlrosaen.com/ml/learning-log/2016-06-20/>. Acesso em: 26 mar. 2023. Citado na página 21.

Scikit-optimize. (2023). **BayesSearchCV** (version 0.9.0). Disponível em: <https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html>. Acessado em: 3 de mar. de 2023. Citado na página 13.

SEABORN. S**eaborn: statistical data visualization**. Versão 0.11.2. 2021. Disponível em: https://seaborn.pydata.org/. Acesso em: 3 mar. 2023. Citado na página 13.

SILGE, J. **R NEWS**, 2020. Disponivel em: <https://r-craft.org/r-news/get-started-with-tidymodels-and-tidytuesday-palmer-penguins/>. Acesso em: 26 maio 2023. Citado na página 28.

SILVA, J. E.; SÁTIRO, R. M. O PODER PREDITIVO DOS MODELOS BOOSTING DE MACHINE LEARNING NO MERCADO BRASILEIRO DE AÇÕES. **Contabilometria**, v. 11, n. 1, 2022. Citado na página 21.

SILVA, S. J. D. **Predição dos tempos até a morte de mulheres com câncer de mama via random survival forest [manuscrito]**, 2022. p. 33. Citado na página 12.

SUTTON, R. S.; BARTO, A. G. **Reinforcement Learning:** An Introduction. [S.l.]: The MIT Press, 2018. Citado na página 18.

WADE, C. **Hands-on gradient boosting with xgboost and scikit-learn**. Birmingham: Packt Publishing Ltd., 2020. Citado na página 21.

ZHU, X. **Introduction to Semi-Supervised Learning**. [S.l.]: Morgan & Claypool Publishers, 2009. Citado na página 18.

AGRADECIMENTOS

Agradeço primeiramente a Deus por ter me dado saúde para continuar estudando, pois só ele sabe tudo que passei ao longo dos anos.

Aos meus pais Mizael Rocha de Oliveira e Francineide Costa de Oliveira que respeitaram minhas escolhas e se orgulharam de ter um filho universitário, principalmente a minha mãe que desde sempre me ajudou a me manter forte no que eu gostaria de ter para minha vida mesmo que indiretamente.

Aos meus colegas de turma que enfrentaram essa jornada comigo.

Aos meus colegas Samuel Souza e Giullber Valentin que foram fundamentais para não me deixar abater pelas dificuldades e me proporcionar conhecimento. Como principal agradecimento a Samuel que me abriu os olhos para além do curso e que eu poderia almejar ainda mais.

A todos os meus professores que sem eles eu não estaria onde estou em especial ao meu professor e orientador Dr. Tiago Almeida de Oliveira por ter me dado a oportunidade de seguir com seu projeto e me tornar o profissional que serei no futuro.

Aos meus amigos mais próximos por estarem comigo me ajudando em especial a Igor Henrique que tenho grande apresso.

Aos meus familiares que me deram apoio.

Agradeço a todos que me ajudaram de alguma forma a alcançar meus objetivos, seja diretamente ou indiretamente.

APOIO FINANCEIRO

Agradecimentos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) do Brasil pelo apoio fornecido ao nosso Projeto/Plano de Trabalho Uso de Machine Learning para predição de morte: uma aplicação a pacientes com câncer de mama em uma cidade da Paraíba sob processo 401821/2021-8 pela Chamada CNPq/MCTI/SEMPI Nº14/2021.