UNICESUMAR - CENTRO UNIVERSITÁRIO DE MARINGÁ
CENTRO DE CIÊNCIAS EXATAS TECNOLÓGICAS E AGRÁRIAS
CURSO DE GRADUAÇÃO EM ENGENHARIA DE SOFTWARE

O USO DO APRENDIZADO DE MÁQUINA NA INSTITUIÇÃO UNICESUMAR PARA DESCOBERTA DE NOTA NA DISCIPLINA ESCOLA DE TI

JOÃO PAULO FRANÇA DE CARVALHO

MARINGÁ – PR
2022

João Paulo França de Carvalho

O USO DO APRENDIZADO DE MÁQUINA NA INSTITUIÇÃO UNICESUMAR PARA DESCOBERTA DE NOTA NA DISCIPLINA ESCOLA DE TI

Artigo apresentado ao Curso de Graduação em Engenharia de Software da UNICESUMAR – Centro Universitário de Maringá como requisito parcial para a obtenção do título de Bacharel em Engenharia de Software, sob a orientação do Prof. Msc. Aparecido Vilela Junior.

MARINGÁ – PR

2022

FOLHA DE APROVAÇÃO

JOÃO PAULO FRANÇA DE CARVALHO

O USO DE APRENDIZADO DE MÁQUINA NA INSTITUIÇÃO UNICESUMAR PARA DESCOBERTA DE NOTA NA DISCIPLINA ESCOLA DE TI

Artigo apresentado ao Curso de Graduação em Engenharia de Software da UNICESUMAR – Centro Universitário de Maringá como requisito parcial para a obtenção do título de Bacharel(a) em Engenharia de Software, sob a orientação do Prof. Msc. Aparecido Vilela Junior.

Aprovado em: ____ de _____ de _____.

BANCA EXAMINADORA

_____

Nome do professor – (Titulação, nome e Instituição)

_____

Nome do professor - (Titulação, nome e Instituição)

_____

Nome do professor - (Titulação, nome e Instituição)

O USO DE APRENDIZADO DE MÁQUINA NA INSTITUIÇÃO UNICESUMAR PARA DESCOBERTA DE NOTA NA DISCIPLINA ESCOLA DE TI

João Paulo França de Carvalho

RESUMO

O presente trabalho tem como objetivo apresentar o estudo sobre a história do Aprendizado de Máquina (AM), os principais tipos e a aplicabilidade de um modelo de AM, GaussianNB, baseado no Teorema de Bayes. Dessa forma, analisando, filtrando, compreendendo, aplicando e testando em uma base de dados dos anos de 2013, 2014, 2015, 2016 e 2019, dos cursos de Tecnologia da Informação (TI) da Universidade Unicesumar, na disciplina de Escola de TI. Buscando assim um retorno de precisão maior que 80%, referente à aprovação dos alunos com média igual ou superior a 6.0 pontos, utilizando a linguagem Python com jupyter notebook e suas devidas bibliotecas relacionadas à ciência de dados, Pandas, matplotlib, numpy e sklearn.

Palavras-chave: Análise de dados. Inteligência Artificial. Tecnologia da Informação.

THE USE OF MACHINE LEARNING IN THE UNICESUMAR INSTITUTION FOR GRADE DISCIPLINE IN SCHOOL OF

ABSTRACT

The present work aims to present the study on the history of Machine Learning (ML), the main types and the applicability of an ML model, GaussianNB, based on Bayes' theorem. In this way, analyzing, filtering, understanding, applying and testing in a database of the years 2013, 2014, 2015, 2016 and 2019, of Information Technology (IT) courses at Unicesumar University, in the IT School discipline. Thus seeking a return of accuracy greater than 80%, referring to the approval of students with an average equal to or greater than 6.0 points, using the Python language with jupyter notebook and its appropriate libraries related to data science, Pandas, matplotlib, numpy and sklearn.

Keywords: Artificial intelligence. Data analysis. Information Technology.

1 INTRODUÇÃO

A importância da inteligência artificial (IA) e do aprendizado de máquina (ML) estão sendo pesquisados e desenvolvidos em escala global, atraindo a atenção de instituições de ensino. A inteligência artificial de hoje imitará e, em alguns casos, substituirá completamente os humanos em atividades que agora são realizadas por eles (CRUZ, 2020). Várias empresas de TI, incluindo Amazon, Facebook, Microsoft e Google, incorporaram IA e aprendizado de máquina. Enquanto isso, poucas pessoas estão cientes de que a inteligência artificial e o aprendizado de máquina entraram no campo da educação e da instrução. O crescimento das escolas foi acompanhado pelo desenvolvimento da tecnologia, com livros didáticos online e softwares práticos que são utilizados na área contábil em nível universitário. Dessa forma, é possível citar o fundador da Microsoft, Bill Gates, que é um dos defensores da inteligência artificial e do Aprendizado de Máquina na educação.

O ensino presencial da Universidade Unicesumar está há mais de uma década disponibilizando a disciplina Escola de TI nos cursos de engenharia de software e análise de desenvolvimento de sistemas, e um dos maiores problemas é o baixo rendimento dos estudantes. Nesse sentido as técnicas de aprendizagem de máquinas também conhecido como “Machine Learning”, tem o intuito de auxiliar encontrando padrões nos estudantes que têm um baixo rendimento na matéria. Utilizando de técnicas o grau de aceitabilidade é superior as 75% (KAPPEL; MARCO, 2020), impulsionando os alunos que apresentam baixos rendimentos na matéria de Escola de TI da Universidade Unicesumar.

Este trabalho tem como objetivo auxiliar os alunos com rendimento insatisfatório na disciplina de Escola de TI, utilizando sua documentação, realizada na plataforma do Redmine, para isso será aplicado a linguagem Python com bibliotecas open-source voltadas para a análise de dados e criação de modelos de Aprendizado de Máquina.

2 DESENVOLVIMENTO

2.1 FUNDAMENTAÇÃO TEÓRICA

2.1.1 MACHINE LEARNING HISTÓRIA

O “Machine Learning” vem trazendo um extraordinário avanço para a Inteligência artificial (IA) (LUDERMIR, 2021), mas o que significa Machine Learning (ML)? Sua tradução diz respeito a “Aprendizado de Máquina”, uma técnica da IA que foi criada em meados da década de 50 pelo cientista da computação Arthur Lee Samuel, objetivo do aprendizado de máquina é proporcionar um sistema se aprimore com base em exemplos (MITCHELL, 1997), por conseguinte para que se consiga uma melhora, é preciso de uma grande massa de dados como exemplos. O ML cria hipóteses com base nos dados aprendidos automaticamente (LUDERMIR, 2021), isso significa que quanto maior o número de dados, mais assertivo será seu programa, portanto Machine Learning é orientado para dados.

O Aprendizado de Máquina é capaz de analisar grandes volumes de dados e extrair padrões, sendo baseado na observação. O ML está presente no nosso dia a dia como por exemplo no nosso mecanismo de pesquisa na web, a pesquisa produz alguns termos, onde com eles o ML realiza o processo para trazer o que mais se identifica com o que foi pesquisado.

2.1.2 TIPOS DE MACHINE LEARNING

Para que se consiga chegar ao resultado esperado o Machine Learning tem inúmeros tipos de sistema (GÉRON, 2019), dentre eles podemos citar os aprendizados de máquinas com supervisão de humanos, que são divididos em aprendizado supervisionado, o aprendizado não supervisionado e o aprendizado por reforço.

O aprendizado supervisionado obtém informações com base na amostragem, tendo suas entradas e saídas esperadas (SILVA, 2021). Um exemplo seria apresentar uma base de dados e dizer que uma bicicleta tem seus padrões, como pedais, corrente e guidão, com isso a pessoa entenderá que qualquer objeto com pedais, corrente e guidão é uma bicicleta, porém Steven Choi nos alerta para o seguinte risco “O aprendizado a partir de um conjunto de dados pode induzir um viés — as máquinas podem repetir preconceitos humanos”(FREITAS, 2019) ele diz também que já aconteceu de de uma máquina dizer, por exemplo, que homens negros eram gorilas porque o banco de dados analisado não tinha diversidade étnica. Importante citar que grandes empresas como Amazon e Walmart estão utilizando o aprendizado de máquina com a técnica de aprendizado supervisionado para o controle de reposição de estoque.

Já o aprendizado não supervisionado não é muito utilizado por empresas, tendo em vista que para a máquina aprender sozinha um conceito, é necessário muito tempo (CHOI).

Ele cita também que para uma máquina aprender por si só a separar o que é garrafa, por exemplo, é um processo demorado, pois é um conceito nunca visto antes.

Por fim, o aprendizado por reforço é quando um sistema realiza inúmeras tentativas, passando por erros até chegar na melhor resposta (SILVA, 2021). Esse tipo de aprendizado também é utilizado na robótica, e pode ser relacionado com o processo que passamos até aprendermos a andar, afinal, primeiro engatinhamos, tentamos levantar, caímos várias vezes, até, por fim, alcançarmos nosso equilíbrio (FREITAS, 2019).

2.1.3 LINGUAGEM PYTHON

Utilizar a linguagem Python para se trabalhar com Aprendizado de Máquina é uma das melhores opções, pois Python tem uma maior capacidade de manipular grandes volumes de dados (FORMIGONI, 2021). A linguagem citada, teve seu nascimento em 1991 por um programador Holandês chamado Guido Van Rossum, suas características são: orientadas a objetos, funcional, interpretada por script, imperativa e de tipagem dinâmica e forte. Uma linguagem interpretada é aquela que precisa de um programa interpretador para ser executada, outra observação relevante sobre a linguagem Python é que seus códigos precisam ser identados, diferente de outras linguagens em que a indentação é opcional, realizada apenas para facilitar a compreensão dos códigos. Sua popularização se deu por meados do ano 2005 (MCKINNEY, 2019), McKinney diz também que Python desenvolveu uma comunidade grande e ativa de processamento científico e análise de dados.

2.1.4 DISCIPLINA ESCOLA DE TI

A Escola de TI é uma disciplina presente nos cursos da área da computação da Universidade Unicesumar, tendo como ementa a criação de um trabalho técnico utilizando os conhecimentos ensinados ao longo do curso (UNICESUMAR, 2022), ela tem como objetivo realizar a criação de um projeto de ponta a ponta, para que o aluno tenha visão profissional e vivência em equipe (UNICESUMAR, 2022). A realização de um projeto de ponta a ponta consiste em 26 competências descritas na ementa e vão desde ler textos técnicos na língua inglesa até conhecer os limites da computação.

A Escola de TI consiste também em um regulamento onde nele é apresentado as diretrizes que devem ser seguidas. No processo de criação do software, é sugerida a utilização do processo Kanban, realizando atribuições de atividades aos integrantes com no máximo duas semanas (REGULAMENTO, 2022). Kanban é uma das metodologias ágeis e para Sommerville os métodos ágeis consistem em produzir softwares úteis rapidamente (SOMMERVILLE, 2011), ele diz também que o software não é criado de uma única unidade, mas sim algo que recebe melhorias contínuas (SOMMERVILLE, 2011). O Kanban foi criado em 1953, pelo engenheiro japonês Taiichi Ohno, e não era utilizado em softwares, mas em uma empresa para produção em massa de veículos, a Toyota. Esse processo consistia em um quadro com 3 colunas To do, Doing e done, e ao longo do tempo ele foi sofrendo melhorias e adaptações e hoje a estrutura do Kanban para o desenvolvimento de software é a de backlog, design, desenvolvimento, teste, deploy e pronto (REIS, 2021).

Para realização da Escola de TI os estudantes são divididos em times de até 6 pessoas, que precisam realizar a elicitação de requisitos do seu projeto, para que eles possam criar suas tarefas no quadro Kanban, e assim dividi-las em Sprint. Sprint Roberto Gil Espinha (2020) afirma que é um período de tempo limitado a um mês ou menos, no qual uma versão incremental e usável de um produto é desenvolvida, no regulamento pede que a Sprint tenha um prazo máximo de 15 dias (REGULAMENTO, 2022). Após separar as tarefas que serão desenvolvidas, cada integrante da equipe começa a resolver as que foram delegadas a si. No regulamento tem um tópico de não conformidades (NC), isso ocorre quando não é cumprida alguma regra tanto do regulamento quanto da equipe, deixando a responsabilidade de auditar a um integrante da equipe. As não conformidades ocorrem também quando alguma tarefa não é entregue dentro do prazo proposto inicialmente as Sprints. A Escola de TI é desenvolvida dessa forma, realizando interdisciplinaridade e trazendo um aprendizado significativo, aplicando muitas vezes esse processo em uma rotina profissional (PPC. Eng. Software). Por conta disso ela tem tamanha importância e é realizada ao final do curso.

2.2 METODOLOGIA

O desenvolvimento de Aprendizado de Máquina é formado por 5 pilares e eles são, o Business Problem, que significa a definição do problema a ser resolvido, Preparação de Dados que fica responsável por mais de 80% do processo, Seleção do Algoritmo, definindo qual algoritmo de aprendizagem melhor se aplica ao processo, Treinamento do modelo e teste de

avaliação do modelo. Todo esse processo foi realizado no Jupyter Notebook, que é uma interface gráfica muito consolidada, fundada em meados do ano de 2014, que desde então vem sendo uma das principais utilizadas por cientistas de dados.

O primeiro passo, seguindo os 5 pilares já citados, se trata da identificação de um problema que pode ser resolvido com aprendizado de máquinas, que é o de descoberta de alunos com o rendimento abaixo de 6 pontos na Escola de TI, média essa escolhida com base na nota mínima necessária para aprovação na matéria. Com base nisso foi realizada a preparação dos dados, com toda base de dados já unificada em um único arquivo CSV, utilizando o Python com a biblioteca do Pandas, abrimos o arquivo CSV para leitura e utilizando a função “shape” já obtivemos a informação de que o arquivo continha 11673 linhas e 29 colunas, porém não eram linhas de alunos, mas sim de tarefas como pode ser visto na tabela 1.

Tabela 1 – Tabela de tarefas.

#	Project	Tracker	Status	Priority	Subject	Assignee	Updated	Parent task	Author	...	Created	Closed
1903.0	escoladeti2013time03	Task	Novo	Normal	Correção das Interfaces		11/21/2013 11:05 PM	1823.0		...	11/21/2013 11:05 PM	NaN
1880.0	escoladeti2013time02	Task	Executando	Normal	6 - Elaborar a Interface Gráfica - Front-End		11/21/2013 10:55 PM	1832.0		...	11/07/2013 08:47 PM	NaN
1879.0	escoladeti2013time02	Task	Executando	Normal	4 - Elaborar o Diagrama de Sequencia da Estória		11/07/2013 08:14 PM	1832.0		...	11/07/2013 08:14 PM	NaN
1878.0	escoladeti2013time01	Task	Novo	Normal	Interface - Desenvolvimento da Homepage		11/21/2013 08:54 PM	1223.0		...	11/07/2013 08:03 PM	NaN
1874.0	escoladeti2013time02	Task	Executando	Normal	9 - Elaborar as Classes de Negócio		11/06/2013 10:31 PM	1832.0		...	11/06/2013 10:31 PM	NaN

Fonte: Autoria própria.

Na tabela 1 como pode ser observado, há 2 colunas com dados cobertos, onde estão os nomes dos alunos, esses dados sensíveis não são necessários para a execução do algoritmo de Aprendizado de Máquina. O que também pode ser visto na tabela é que ela não contém a coluna de notas dos alunos, pois estão em um arquivo separado, e que para isso precisa primeiramente agrupar os dados por alunos, sendo escolhido a coluna de “Author”, após realizar o agrupamento com a função “groupby”, juntamente com o “count” para que os dados possam ser quantificados, e apenas após tudo isso Machine Learning faz o seu papel. Depois da realização desse processo a tabela diminuiu ficando com seu “shape” de 219 linhas e 29

colunas, cada linha sendo representada por 1 estudante e suas colunas apresentando o quanto foi executado de cada tarefa.

Tabela 2 – Tabela de alunos e quantidades de atividades.

Author	#	Project	Tracker	Status	Priority	Subject	Assignee	Updated	Parent task	...	Created	Closed
	13	13	13	13	13	13	12	13	0	...	13	13
	1	1	1	1	1	1	1	1	1	...	1	1
	27	81	81	81	81	81	60	81	3	...	81	57
	38	38	38	38	38	38	6	38	1	...	38	7
	0	76	76	76	76	76	12	76	2	...	76	14

Fonte: Autoria própria.

Como mostrado na tabela 2, os dados estão quantificados e transformados em números, e é preciso incluir a coluna de médias e 218 registros, processo que manualmente não seria nada fácil, portanto foi utilizado o Python ao nosso favor, para isso foi criado uma função que procura o Author correspondente e inclui sua média.

Quadro 1 – Função de correção do nome Author

```
for linha in tabela_dados['Author']:
    for nota in tabela_notas['Author']:
        contem = 0
        for palavra_nota in nota.split():
            if palavra_nota in linha.upper():
                contem = contem +1
        if contem >= 2:
            tabela_dados.replace(to_replace=linha, value=nota, inplace = True)

tabela_dados.to_csv('arquivos/dados_corrigidos.csv',  index = False)
```

Fonte: Autoria própria.

O trecho de código apresentado no quadro 1 resolve o problema gerado pelo fato da média estar em um CSV diferente dos demais e os dados da coluna de Author estavam com os nomes escritos de formas diferentes, para utilização da função Pandas responsável por incluir colunas de um CSV em outro. Fazia-se necessário corrigir esse problema do nome, e uma forma de exemplificar essa dificuldade é: em uma linha de código o nome está identificado como “João Paulo França de Carvalho” e em outro CSV está apresentado como “João

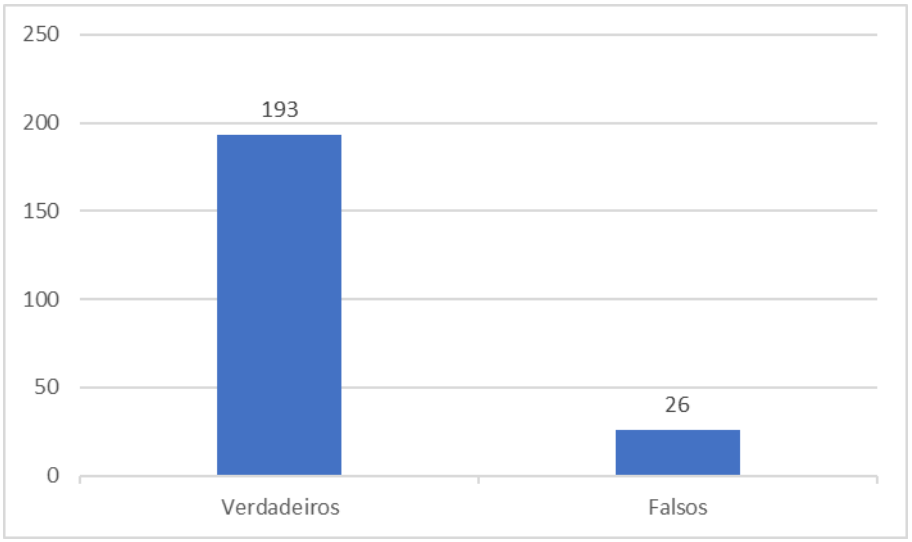
Carvalho", é a mesma pessoa só que o nome está escrito de maneiras diferentes e para conseguir deixar os nomes iguais nos 2 CSV foi realizado esse código em que passam todos os nomes da tabela de notas, e os separam em palavras, validando se cada trecho contém na tabela de dados e se 2 ou mais nomes estiverem contidos, realiza-se assim a troca do nome da tabela de dados para o mesmo nome da tabela de notas. Após realizar esse processo foi utilizado a função "merge" do Pandas, que é responsável por unir 2 CSV com base em uma coluna correspondente. Por fim, precisávamos resolver um último tratamento de dados na base, os valores nulos, para a realização do Aprendizado de Máquina os valores nulos é um problema no processo, pois eles induzem o aprendizado ao erro e para resolver esse problema utilizamos uma das maneiras mais simples, que é realizar a mediana para os valores nulos, utilizando a função "median" do próprio Pandas e em determinada coluna para o cálculo dela e em segunda a função "fillna" também do Pandas para realização da substituição dos valores nulos. Após essas etapas os dados estão totalmente prontos para a realização do processo.

Para facilitar a visualização e análise dos dados foi utilizado a biblioteca "matplotlib", que é responsável por criar gráficos. A primeira análise feita foi a de correções, afinal é importante saber quais colunas têm forte correção com outra para ajudar na escolha dos atributos, que nada mais é do que as colunas que são levadas em consideração para execução do treinamento do ML. O gráfico 01 evidenciado no apêndice mostra os níveis de correções de uma coluna, onde o que está em amarelo tem forte correção e quando está azul tem baixa correção.

A análise do gráfico 01 é complexa, porque totalizam 30 quantidades de colunas. Dessa forma, foi importante a visualização do gráfico, pois pode ser visto que a coluna "tipoNaoConfirmidade" não continha nenhum registro e por isso ela não foi utilizada como atributo, nesse sentido, outras colunas também não foram utilizadas no modelo de aprendizado de máquinas de "Assignee" e "Author", que possuem dados sensíveis não necessários para o modelo.

Outro ajuste que houve necessidade de ser feito nos dados foi a conversão da coluna "media", onde todos os valores maiores ou iguais a 6 receberam 1, e os menores que 6 receberam zero, após isso, realizando a separação dos valores foi possível visualizar no gráfico 2 que a maioria dos casos foram verdadeiros, totalizando 88,13% de casos verdadeiros e 11,87% de casos falsos, ou seja, alunos que não atingiram a média para aprovação.

Gráfico 2 – Números de casos verdadeiros e falsos

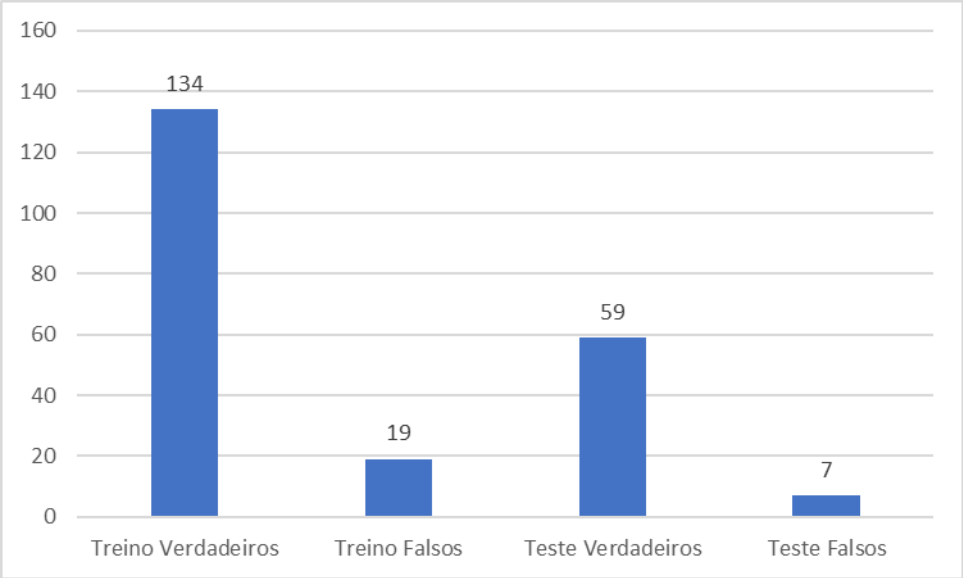


Fonte: Autoria própria.

Para o Aprendizado de Máquina essa distribuição dos dados não é o ideal, pois deveria ser balanceado para evitar problemas na construção do modelo de ML, mas em virtude do baixo volume de dados não foi possível a realização.

Para a realização do processo do Machine Learning, precisamos dividir nossos dados em dados de treino e dados de testes, dividindo a maior parte para treino onde será responsável criar o modelo de teste e uma outra parte para a realização da validação do modelo criado. Para a realização do processo foi utilizado outra biblioteca, a “sklearn” que fica responsável por nos ajudar a dividir os dados de treino e de teste. Para execução do processo precisamos identificar as variáveis preditoras, que são as variáveis que têm o poder de influenciar na variável que queremos prever, chamada de alvo. As variáveis preditoras totalizam 26, sendo assim, as únicas que não foram utilizadas são as que possuem dados sensíveis e a coluna que não continha registro, dessa forma é preciso também identificar a variável alvo que no caso é a “media”. Por fim, a divisão dos dados foi realizada em 30% testes e 70% treino.

Gráfico 3 – Divisão dos dados em treino e teste



Fonte: Autoria própria.

Como pode ser visto o gráfico ficou dividido de maneira parecida, quando se trata dos valores de verdadeiros e falsos, ficando assim os de treino com 87,58% de verdadeiros e 12,42% de falsos e os do teste ficando com 89,39% de verdadeiros e 10,61% de falsos.

Por fim, será criado e treinado o modelo escolhido de Aprendizado de Máquina, GaussianNB, baseado no “Teorema de Bayes”, teorema que é usado para cálculo de probabilidade de um evento com base em outro evento que já ocorreu com uma fórmula matemática (COUTINHO, 2020).

Concluindo a criação do modelo de Aprendizado de Máquina, é realizada a validação da exatidão do modelo, apresentando a precisão calculada, utilizando a função “accuracy_score” da biblioteca “sklearn” que apresentou um valor maior que 80% de precisão. Levando em consideração o baixo volume de dados esse valor é muito bom, e para surpresa quando verificado os dados para testes foi visto que quase alcançou 90%, cena muito difícil de acontecer com dados de treino, pois esses dados nunca tinham sido apresentados ao modelo.

2.3 COLETA DE DADOS

Os dados obtidos para a realização do aprendizado de máquina foram por meio do software Redmine, criado para o controle e gerenciamento de projeto, permitindo a criação de tarefas, sprints, apontamento de horas e toda documentação necessária. Com base nisso foram

obtidos dados de 5 anos de Escola de TI, 2013 a 2016 e 2019, portanto não foram utilizados os dados de 2017 e 2018 devido a falta da média dos anos, dado responsável pela previsão.

Primeiro passo da coleta de dados foi a unificação, onde separamos os arquivos em pastas com seus respectivos anos, depois foi possível observar que em 2013 havia 7 equipes com um total de 57 alunos, já em 2014 havia a mesma quantidade de equipes, porém com um total de 60 alunos, no ano 2015 houve 6 equipes, mas com um total de 23 alunos, em 2016 havia 8 equipes e um total de 54 alunos e por fim em 2019, foram 5 equipes e um total de 24 alunos, totalizando nos 5 anos, 33 equipes e 218 alunos.

Após a organização dos dados obtidos pelo Redmine foi preciso organizar os dados responsáveis pelas notas dos alunos, para isso foram escritos arquivos em CSV, no mesmo formato dos arquivos obtidos pelo Redmine. Arquivos CSV são arquivos muito comuns na ciência de dados, eles são simplesmente arquivos de texto com seus dados separados por vírgulas, formando assim tabelas, e são muito utilizados na ciência de dados por conta da sua leveza e facilidade de acesso, hoje por exemplo o Excel importa e exporta arquivos CSV.

Por fim, realizamos a unificação desses dados, e para isso foi utilizado a linguagem Python com a biblioteca Pandas. Bibliotecas para programação significa um conjunto de recursos, funções e procedimentos e o Python conta com mais de 137 mil delas (CATUNDA, 2022), cada uma com sua finalidade específica. Pandas é voltada para o trabalho com dados, sendo uma biblioteca open-source, que são softwares de código aberto, possibilitando o direito de qualquer pessoa modificar, estudar e distribuí-los totalmente de graça.

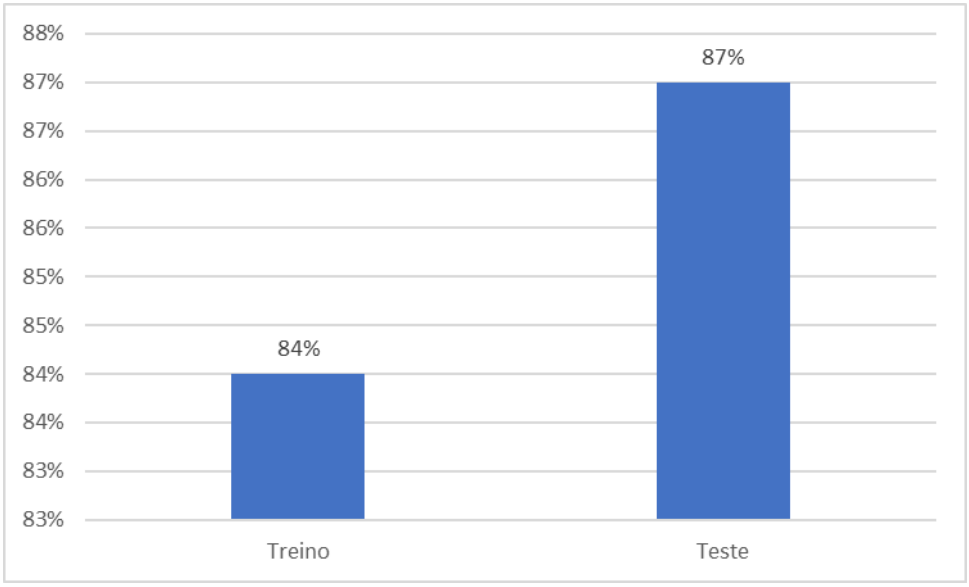
3 RESULTADOS E DISCUSSÃO

Um dos grandes aprendizados obtidos neste trabalho, está fundamentado no fato de que a maior parte do tempo de um cientista de dados (CD) é gasto realizando a organização dos dados. Barbosa afirma que 80% do tempo útil dos CD é utilizado em tarefas para achar, limpar e organizar as informações (BARBOSA, 2018). Foi compreendido também que Python é uma das melhores linguagens de programação para os cientistas de dados utilizarem, não só por conta da sua facilidade de aprender, mas também por conta da infinidade de bibliotecas voltadas para o trabalho de cientistas de dados.

Os resultados obtidos com o Aprendizado de Máquina foram excelentes, em virtude do fato de que o volume de dados era baixo e a base de dados não estava balanceada, no entanto, ao adquirir uma precisão maior que 80% ficou evidente que a escolha do modelo

baseado em Teorema de Bayes foi a melhor a ser aplicada, devido ao contexto obtido. Lima afirma que comumente os modelos acima de 85% já são considerados eficientes (LIMA,2021).

Gráfico 4 – precisão do modelo de aprendizado de máquina



Fonte: Autoria própria.

Como visto no gráfico 4, os dados de testes não costumam ter uma precisão maior que os de treino, entretanto, no modelo acima isso se mostrou diferente, em virtude dos maiores números de atributos que interferem na variável alvo, apresentada ao modelo.

3 CONCLUSÃO

Após a realização do trabalho podemos concluir que o Aprendizado de Máquina está cada vez mais presente no nosso dia a dia e que surgem inúmeras tecnologias para auxiliar nesse processo, tornando assim, gradativamente mais assertivo e simples de ser executado. Dado ao momento mais tecnológico que vivemos, o Machine Learning pode ser realizado diariamente tendo em algo simples, quanto no mais complexo, com seu propósito de trazer números mais precisos e nos ajudar em descobertas, como também na otimização de tempo e para maximizar algo que e deseja. Quando aplicado em empresa, ela terá maiores chances de lucros e rendimentos. O ML é utilizado também na bolsa de valores para prever algumas regras de possíveis volatilidades.

Com este trabalho foi possível entender a importância dos dados e como podemos realizar diversas coisas com eles, além de conhecer mais sobre a linguagem Python, uma linguagem de fácil compreensão e muito poderosa, que contém uma comunidade muito ativa e forte, estimulando sua evolução.

Ao final desse trabalho podemos contribuir na realização e desenvolvimento para encontrar meios que favorecem melhorias de média dos alunos da Escola de TI. Os resultados obtidos provam que se pode realizar melhorias na forma de coleta e tratamento dos dados para alcançar percentuais maiores de assertividade no processo de Aprendizado de Máquina.

REFERÊNCIAS

BARBOSA, SURIA**. O que um cientista de dados faz de acordo com 35 profissionais**. Na prática. Disponível: <https://www.napratica.org.br/o-que-um-cientista-de-dados-faz/>. Acesso em: 21 out. 2022.

CASTRO, R. V.; ALMEIDA, L. S. **Ser Estudante no Ensino Superior: O caso dos estudantes do 1° ano,** 2016. Universidade do Minho.

CATUNDA, HEITOR**. Bibliotecas do Python:conheça as melhores finalidade!**. Hashtag treinamento. Disponível: <https://www.hashtagtreinamentos.com/bibliotecas-mais-importantes-do-python/>. Acesso em: 01 out. 2022.

COUTINHO, THIAGO**. Teorema de Bayes: saiba o que é e aprenda a utilizar**. voitto. Disponível: < https://www.voitto.com.br/blog/artigo/teorema-de-bayes/>. Acesso em: 30 set. 2022.

CRUZ, M. T. S. **Impactos da Inteligência Artificial na gestão de Pessoas**. São Paulo: TIKI Books. 2020.

ESPINHA, R. G**. Você realmente sabe o que é SPRINT? Veja definição e aprenda como fazer na sua empresa**. artia. Disponível: < https://artia.com/blog/sprint/>. Acesso em: 21 out. 2022.

FORMIGONI, P. A. F. **Python na análise de dados: estudo de caso com dados de acidentes aéreos no brasil,** 2021.Trabalho de Conclusão de Curso. (Graduação em engenharia de produção) - Universidade Federal Fluminense, Niterói, RJ.

FREITAS, TAINÁ**. Os três tipos de aprendizado de machine learning, um ramo da inteligência artificial**. .startSe. Disponível: <https://www.startse.com/noticia/nova-economia/machine-learning-inteligencia-artificial-aprendizado//>. Acesso em: 16 out. 2022.

GÉRON, AURÉLIEN. **Mãos à Obra: Aprendizado de Máquina com Scikit-learn e TensorFlow**. Rio de Janeiro, Alta Books Editora. 2019.

KAPPEL, M. A. A. **Aplicação de técnicas de aprendizado de máquina para predição de risco de evasão escolar em instituições públicas de ensino superior do Brasil.** 2020. Revista Brasileira de Informática na Educação.

LIMA, A. M. **Como avaliar se um modelo de machine learning está indo bem ou não?.** Linkedin. Disponível: < https://www.linkedin.com/pulse/como-avaliar-se-um-modelo-de-machine-learning-est%C3%A1-ou-mendon%C3%A7a-lima/?originalSubdomain=pt/>. Acesso em: 5 nov. 2022.

LUDERMIR, T. B. **Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências,** 2021. Estudos Avançados 35.

MCKINNEY, WES. **Python para análise de dados: Tratamento de dados com Pandas, NumPy e IPython**. Novatec Editora. 2019.

SANTOS, GUILHERME**. Kanban: O Que é, Como Funciona e Dicas**. Automação Industrial. Disponível: < https://www.automacaoindustrial.info/kanban />. Acesso em: 14 out. 2022.

SARTORI, JUNIOR**. Kanban no desenvolvimento de software: origem e conceitos**. EZ.devs. Disponível: < https://ezdevs.com.br/kanban-desenvolvimento-de-software />. Acesso em: 14 out. 2022.

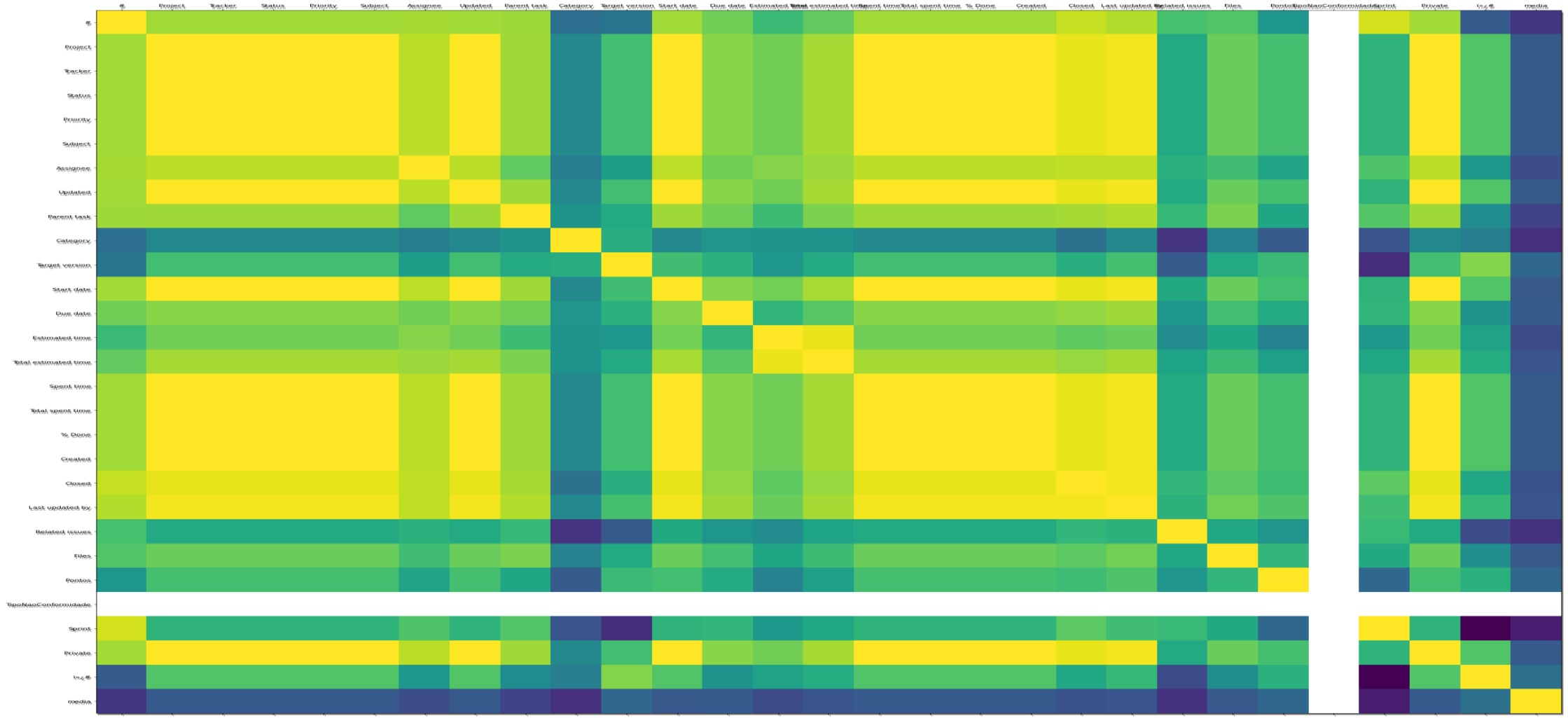
SOUZA, E. G**. Implementando Regressão Linear Simples em Python**. Medium. Disponível: <https://medium.com/data-hackers/implementando-regress%C3%A3o-linear-simples-em-python-91df53b920a8/>. Acesso em: 30 set. 2022.

SUMMERVILLE, LAN. **Engenharia de Software**. São Paulo: Pearson Edicaion do Brasil. 9° edição 2011.

VIANA, W. M. O. **Comparativo de alguns modelos de Machine Learning utilizando dados de domínio público e a linguagem python,** 2021.Trabalho de Conclusão de Curso. (Graduação em engenheiro eletricista) - Universidade Estadual Paulista Júlio de Mesquita Filho Faculdade de Engenharia, Ilha Solteira, SP.

APÊNDICE

GRÁFICO 1 – Correlações dos atributos



Fonte: Autoria própria.