UNICARIOCA
CENTRO UNIVERSITÁRIO CARIOCA
CURSO DE ENGENHARIA DA COMPUTAÇÃO

LEONARDO DOS SANTOS LOURES

ANÁLISE ESTATÍSTICA COMO PARTE DA MELHORIA DO PROCESSO LOGÍSTICO DE IMPRESSÃO DE PROVAS INSTITUCIONAIS

Rio de Janeiro
2019

LEONARDO DOS SANTOS LOURES

ANÁLISE ESTATÍSTICA COMO PARTE DA MELHORIA DO PROCESSO LOGÍSTICO DE IMPRESSÃO DE PROVAS INSTITUCIONAIS

Trabalho de Conclusão de Curso apresentado ao Centro Universitário Carioca, como requisito parcial para btenção do grau de Bacharel em Engenharia da Computação.

Orientador: Prof. Antonio Felipe Podgorski Bezerra - M.Sc.,

Rio de Janeiro

2019

Loures, Leonardo dos Santos.
Análise estatística como parte da melhoria do processo logístico de impressão de provas institucionais. / Leonardo dos Santos Loures. – Rio de Janeiro, 2019.

47f.

Orientador: Antonio Felipe Podgorski Bezerra.

Trabalho de Conclusão de Curso (Graduação Superior em Engenharia de Computação) – Centro Universitário Carioca, Rio de Janeiro, 2019.

1. Sustentabilidade. 2. Análise de dados. 3. Aprendizado de máquina. I. Bezerra, Antonio Felipe Podgorski, prof. orient. II. Título.

CDD 621.39

LEONARDO DOS SANTOS LOURES

ANÁLISE ESTATÍSTICA COMO PARTE DA MELHORIA DO PROCESSO LOGÍSTICO DE IMPRESSÃO DE PROVAS INSTITUCIONAIS

Trabalho de Conclusão de Curso apresentado ao Centro Universitário Carioca, como requisito parcial para btenção do grau de Bacharel em Engenharia da Computação.

Aprovação em ____/____/ 2019

Banca Examinadora

_____

Prof. Antonio Felipe Podgorski Bezerra - Orientador

Centro Universitário Carioca

_____

Prof. Neury Nunes Cardoso - Coordenador

Centro Universitário Carioca

_____

Prof. Manuel Martins Filho - Convidado

Centro Universitário Carioca

Dedico este trabalho à minha noiva Isabel Cristina Gomes da Silva, por todo seu amor e carinho nessa etapa tão importante da minha vida.

Obrigado, meu amor!

AGRADECIMENTOS

A todos os professores da Unicarioca por seus ensinamentos em aulas que contribuiu para a minha formação.

Ao meu professor e orientador Antonio Podgorski, por aceitar conduzir o meu trabalho e me ajudou desde o início, compartilhando seu conhecimento com paciência.

Ao meu amigo Alexander James (in memorian), que sempre me dizia para fazer a faculdade, pois mudaria a minha vida. Ele estava certo.

Ao meu irmao Frederico Santos Silva que sempre vibra com as minhas conquistas, se preocupa e torce por mim.

À minha noiva Isabel Gomes, que sempre teve orgulho de mim e me apoiou durante toda a minha formação com muita motivação, otimismo e amor.

À minha minha mãe, uma grande mulher, guerreira e de coração puro. A quem devo tudo em minha vida e a quem me tornei.

Leonardo dos Santos Loures

“Entrega o teu caminho ao Senhor, confia nele, e o mais Ele fará.”

Salmos 37:5

RESUMO

Em época de crises com grandes e rápidas mudanças, num mundo cada vez mais competitivo, é comum as empresas olharem para o ambiente interno e buscarem respostas para a melhoria de recursos e diminuição dos custos para melhorar o desempenho econômico, porém outros fatores devem ser considerados como a responsabilidade social e compromisso com o meio ambiente. Um dos meios mais importantes para se alcançar os objetivos para otimização dos seus processos e obter maior conhecimento sobre seus ambientes interno, é a cultura de tomada de decisões baseado em dados. Técnicas e métodos para melhorar o armazenamento em quantidade e manutenção da qualidade, torna os dados um dos bens mais valiosos de uma instituição. Devido a impressão das provas finais e nem todos os alunos comparecerem para a avaliação, há sobras de folhas de papel gerando excedente de material. Algumas análises iniciais são feitas para avaliar o impacto econômico através dos dados históricos. Posteriormente são aplicadas técnicas estatísticas e de inteligência computacional em busca de padrões para otimizar a impressão das provas finais e reduzir o desperdício deste processo acadêmico. O ponto de partida para o processo de decisões baseada em dados é a coleta, preparação, limpeza para análise exploratória dos dados através de visualizações de tabelas e gráficos estatísticos. A extração, seleção e criação de atributos são etapas para a modelagem de técnicas de inteligência artificial com a utilização de algorítmos de aprendizagem de máquina. Na fase exploratória, muitas perguntas foram respondidas com propostas de soluções, muitas outras podem ser feitas e servem como ferramentas de criação de modelos preditivos para otimizar o processo. Foi possível observar a possibilidade de economia de pelo menos 20 porcento nas impressões de provas finais, onde alunos não compareceriam e com os modelos de aprendizagem de máquina os resultados podem ser ainda melhores com o mapeamento dos perfis de alunos por turmas. Os potenciais das técnicas aplicadas são crescentes a medida que mais dados sejam disponibilizados. Os dados disponíveis para este estudo foram de 46 turmas de três anos de apenas um professor, no total de 2239 observações, o que é muito pouco para resultados ótimos. Apesar de material disponível insuficiente, o estudo trouxe à luz uma

demonstração da possibilidade de compreender melhor os processos da instituição e otimiza-los, responder perguntas que não foram pensadas antes para diversos outros domínios em que os dados estejam disponíveis.

Palavras-chave: Sustentabilidade; Outsourcing de impressão; Análise de dados; Aprendizado de máquina;

ABSTRACT

In times of major and rapidly changing crises in an increasingly competitive world, companies often look to the internal environment and seek answers to resource improvements and cost reductions to improve economic performance, but other factors must be considered. as social responsibility and commitment to the environment. One of the most important means of achieving the goals of optimizing your processes and gaining insight into your internal environments is the data-driven decision-making culture. Techniques and methods for improving quantity storage and maintaining quality make data one of the most valuable assets of an institution. Due to the printing of the final exams and not all students attend for the evaluation, there are leftovers of paper generating excess material. Some initial analyzes are made to assess economic impact through historical data. Subsequently, statistical and computational intelligence techniques are applied in search of standards to optimize the printing of finals and reduce the waste of this academic process. The starting point for data-based decision making is the collection, preparation, cleansing for exploratory data analysis through table views and statistical graphs. Extracting, selecting, and creating attributes are steps for modeling artificial intelligence techniques using machine learning algorithms. In the exploratory phase, many questions have been answered with proposed solutions, many more can be asked and serve as predictive modeling tools to optimize the process. At least 20 per cent savings on final exam impressions where students would not attend and machine learning models could be further improved by mapping student profiles by class. The potentials of the applied techniques are increasing as more data becomes available. The data available for this study were 40 one-year, one-teacher classes out of 2259 observations, which is too little for optimal results. Despite insufficient available material, the study brings to light a demonstration of the possibility of better understanding and optimizing the institution's processes, answering questions that were not thought of before for several other areas where data is available.

Keywords: Sustainability; Printing outsourcing; Data analysis; Machine learning;

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Cada vez mais as organizações buscam ser sustentáveis, é comum encontrar em emails organizacionais frases como "Antes de imprimir pense em seu compromisso com o Meio Ambiente" e "Responsabilidade ambiental: utilize o verso de qualquer folha para imprimir o e-mail. Proteja o meio ambiente", mas além de envios de notas de rodape, a atenção e cuidados em seus processos internos são determinantes para o sucesso de suas premissas econômicas, sociais e ambientais. O descarte de excedente de papeis impressos não utilizados e não reaproveitados, gera desperdício de folhas, tinta de impressora, dentre outros recursos, no entanto abre oportunidades para compreender melhor sua própria infraestrutura e criar soluções que otimizam a impressão e reduzem o desperdício.

Em um cenário academico é comum que o aluno seja avaliado através de provas teóricas sendo essas comumente impressas em um sistema de três avaliações, calculado a média das duas maiores notas excluindo a menor. Em uma instituição privada foi observado que na última avaliação (AV3) se torna opcional quanto o aluno alcança a média necessária para aprovação nas duas últimas, então muitos alunos não comparecem por este motivo, porém outros o fazem a AV3 para tentar melhorar a média final e ainda, há os que não comparecem por desistência, dentre outras razões. Por não ter a informação dos que farão a avaliação final e quem não a farão, por garantia, a impressão da prova é contada para todos os alunos sem levar em conta a presença desses. Uma solução seria a impressão de provas por demanda no momento da avaliação, porém o custo operacional se tornaria inviável. O resultado é o descarte de excedente de papel das provas não utilizadas e não reaproveitadas, um cenário ideal para propor soluções e reduzir o desperdício deste processo acadêmico.

Este trabalho está organizada da seguinte forma: no capítulo 2, serão apresentados os conceitos básicos relacionados à sustentabilidade, modelos de impressão e métodos estatísticos. No capítulo 3, a metodologia do fluxo de impressão e coleta de dados para análise exploratória será conceitualizada. No capítulo 4, será descrito como a metodologia foi implementada. No capítulo 5, serão apresentados os

resultados obtidos. No capítulo 6, serão realizadas as considerações finais e apresentadas sugestões de trabalhos futuros. O capítulo 7, contém as referências bibliográficas utilizadas como base neste trabalho.

2 CONCEITUALIZAÇÃO BÁSICA

2.1 SUSTENTABILIDADE ORGANIZACIONAL

Sustentabiliade pode ser descrito como equilíbrio e respeito aos limites do ecossistema em que está inserido, uma característica ou condição de um processo ou sistema que permite a sua permanência. A capacidade de o ser humano interagir com o mundo, preservando o meio ambiente para não comprometer os recursos naturais das gerações futuras.

De acordo com (MUNCK, 2008), proposições relacionadas à sustentabilidade adquiriram destaque pela incorporação da promessa de evolução da sociedade rumo a um mundo mais harmonioso, no qual o meio natural e as conquistas culturais procuram ser preservados para as gerações futuras. Sendo assim, para as novas gerações surgirem, a sustentabilidade garante cuidados do meio ambiente.

É necessário ações para que haja harmonia entre suprir as necessidades da geração atual e preservação do meio ambiente para subsistência das gerações futuras, portanto sustentabilidade é melhor tratado como desenvolvimento sustentável. Nesse entendimento, além das questões de preservação ambiental, ser sustentável passa a incluir a justiça social juntamente com os interesses econômicos que formam a base de três pilares que compõem o triple bottom line (CARVALHO, 2015), (MUNCK, 2008).

Segundo (CARVALHO, 2015), esses pilares se correlacionam na forma de ecoeficiência, nas práticas de comércio justas e de inserção social e na justiça ambiental, correspondente aos resultados de uma organização e tratado como sustentabilidade organizacional. Uma organização sustentável gera lucro enquanto protege o meio ambiente e melhora a vida das pessoas com as quais mantém relações. (CARVALHO, 2015) continua, que este não é um assunto exclusivo para as grandes corporações, mas a toda cadeia de negócios, passando pelos grandes e pequenos empreendimentos. A sustentabilidade não poderá ser alcançada sem o apoio das empresas, pois elas representam o setor produtivo da economia.

Em (MUNCK, 2008) a gestão sustentável é uma boa oportunidade a considerar e é proposto a discussão de um modelo de gestão que leve em conta fatores, que a

disciplina econômica denomina como externalidades, positivas e negativas. Como exemplos de externalidades negativas envolvem custos oriundos do descarte de produtos, custos por serviços ambientais de recuperação e limpeza, etc.

Para (CARVALHO, 2015), as empresas poderiam tratar a sustentabilidade como oportunidade de negócios, fonte de inovação, aumento de rendimentos, diminuição de custos ou investimentos de longo prazo na competitividade futura. Ser sustentável é importante para dar lucro, pois os consumidores acabam dando atenção maior por empresas que seguem com ações sustentáveis.

2.2 MODELOS DE IMPRESSÃO

Outsourcing de impressão é a terceirização dos equipamentos e gerenciamento de cópias e impressão com o objetivo de minimizar o trabalho, as operações por meio do gerenciamento de tarifas por página com a ajuda de software, ficar livre da depreciação dos equipamentos e dos investimentos de insumos e pagar somente as páginas copiadas impressas (MENEZES, 2014), (FILHO, 2014). O modelo de outsourcing está compreendido na contratação de serviços que tradicionalmente são desenvolvidos por recursos próprios e passam a ser contratados de um fornecedor (MENEZES, 2014).

De acordo com (FILHO, 2014), o outsourcing de impressão tem sido visto cada vez mais como meio de redução de custos de manutenção, suprimentos e infraestrutura, inerentes a um parque de impressão. A atual situação econômica obriga as organizações a repensarem as infraestruturas e os processos. Para uma contratação bem sucedida é necessário considerar os aspectos qualitativos, além dos custos, continua (MENEZES, 2014).

Segundo (FILHO, 2014), a instituição poderá ter uma economia de, aproximadamente, 30% com impressões, porém o outsourcing não é uma solução para todas as organizações. (FILHO, 2014), (MENEZES, 2014) destaca as vantagens e as desvantagens de uma contratação de serviços de outsourcing de impressão, a seguir.

Vantagens com o outsourcing de impressão:

- Redução nos gastos de impressão e cópia;
- Eliminação do estoque e logística consumível;
- Atualização tecnológica sem investimentos;
- Rastreabilidade de utilização das impressoras por Centro de Ensino;
- Nova cultura de trabalho na área de impressão, evitando desperdícios e perdas;
- Gerenciamento dos equipamentos em rede;
- Pró-atividade de atendimento aos usuários;
- Substituição de equipamentos somente impressão por multifuncionais;
- Serviço especializado - qualidade e alta disponibilidade;
- Substituição de impressoras jato de tinta por laser e;
- Aprimoramento no ambiente de trabalho a fim de aumentar sua produtividade.

Desvantagens com o outsourcing de impressão:

- Maior necessidade de controle da execução do contrato;
- Perda da confidencialidade;
- Má qualidade dos serviços prestados e diminuição do nível de satisfação;
- Dependência excessiva em relação à empresa contratada;

- Elevados custos de um eventual regresso a situação anterior;
- Desmotivação do pessoal gerada pela instabilidade do processo;
- Surgimento de melhores alternativas para o serviço contratado;
- Eventual ocorrência de custos ocultos.

O efeito da administração de uma parte da infraestrutura da organização por uma empresa externa, é buscar economia financeira, ganho de agilidade, flexibilidade e eficiência, alcançando novas tecnologias e elevando o nível de qualidade de processos internos, por outro lado, abre-se espaço para que um terceiro passe a integrar sua estrutura e administre alguns de seus processos, o que gera naturalmente uma

dependência mútua de ambas as empresas nessa relação comercial (MENEZES, 2014).

A comparação do custo da administração própria com os custos de uma terceirização auxiliará a obtenção do ponto de equilíbrio, isto é, a situação na qual os custos totais do outsourcing de impressão se igualam aos custos totais de administração própria. Dessa forma, o ponto de equilíbrio representa o corte entre a desvantagem de se terceirizar o serviço e a vantagem de contratar uma empresa especializada (MENEZES, 2014).

Para a decisão do modelo de gestão de impressão seja própria ou de Outsoursing é necessário fazer análises pela empresa ou instituição e não faz parte deste estudo. Outsourcing é muitas vezes preferível e adotada em instituições públicas e privadas e entre vantagens e desvantagens declaradas em outras fontes de pesquisa, neste estudo será considerado somente a vantagem de Redução nos gastos de impressão e cópia e será considerado estudos de ambos os modelos calculados por (MENEZES, 2014) e (ROSSI, 2010).

Algumas pesquisas feitas com relação à diminuição da utilização de papel para impressão, em substituição por tecnologias de Tecnologia da Informação (TI), (MORAES, 2011) observou que investimentos em TI pelas empresas não reduziu o consumo de papel e, ao contrário houve aumento no consumo de outros tipos de papel, como o papel de impressão e embalagem.

2.3 ANÁLISE ESTATÍSTICA DE DADOS

Os números constituem a única linguagem universal (Nayhanael West). Entretando os números podem ser manipulados, pois por mais que os números constituam uma verdade universal, se forem manipulados ainda assim os números não valem muita coisa. Claro que olhar para os números é a melhor forma de se compreender um problema e também buscar uma solução. Estatística é a ciência que nos permite aprender a partir dos dados e utilizar seus recursos para extrair

informações relevantes para análise estatística, compreenção da situação atual e tomada de decisões.

A Estatística oferece uma série de ferramentas para coletar dados com técnicas de amostragem, organizar os dados com técnicas de tabulação, calcular frequência, colocar os dados de maneira organizada para poder realizar processos de análise, apresentar os dados através de gráficos estatísticos ou de visualizações que resumem ou simplificam aquilo que temos nos dados e interpretar os dados, pois "Os números constituiem a única verdade universal", porém é necessário compreender o que os dados querem dizer, através do processo de interpretação com as ferramentas que a estatística oferece e a partir dessa interpretação podemos fazer inferências, coletar amostras, organizar esses dados, descrevê-los e apresentá-los e depois interpretar, para que possa fazer inferências sobre uma população ou sobre um fenômeno qualquer.

É muito importante definir qual o tipo de dado com o qual estou trabalhando para que eu possa escolher a técnica adequada de análise estatística. Os dados podem se qualitativos (nominais ou ordinais) e quantitativos (discretos ou contínuos). Importante tambem saber qual o tipo de estudo estatístico ao qual está trabalhando, observação ou experimentação. Em um estudo de observação, os dados e as características específicas são recolhidos e observados, entretanto, não há iniciativa de modificar os estudos que estão sendo realizados. Em um estudo de experimentação, cada indivíduo é aleatoriamente atribuído a um grupo de tratamento, em seguida, os dados e as características específicas são observados e coletados.

Os estudo experimentais ajudam a proteger contra potenciais viéses desconhecidos que interferem o resultado da análise estatística, já os estudo observacionais não oferece esse nível de proteção contra fatores de confusão, por exemplo, na hora de observar os dados. Cada estudo será mais aderente ao objetivo de acordo com a necessidade da análise.

A Análise de Dados é o meio através do qual utilizamos a estatística para apresentar e demonstrar os resultados dos dados que foram avaliados. Ou seja, é uma forma de avaliar os dados utilizando raciocínio lógico e estatístico de cada componente

fornecido no conjunto de dados. A estatística não tem sido usada apenas por técnicos, mas também por gestores de todos os níveis. Para onde se olha, se vê Estatística sendo aplicada, desde o planejamento corporativo atpe decisões simples do dia a dia.

Uma das primeiras etapas em ciência de dados é a coleta dos dados e começar uma análise exploratória, então aplicar estatística descritiva para obter um valor representativo dos dados, como a média, por exemplo, uma medida para avaliar a dispersão ou a variância dos dados e uma medida para avaliar a forma da distribuição dos dados.

Estatística descritiva é um conjunto de métodos estatística utilizados para descrever as principais características dos dados. Cada método é concebido para proporcionar insight distinto da informação disponível. Os dados são analisados para coletar informações. Além de compreender como aplicar os métodos de estatística descritiva, é essencial perceber o que casa método revela sobre os dados.

Com base nessas informações como os dados como estão organizados, tomar as decisões seguintes sobre que ferramentas usar para tratar os dados, para fazer a limpeza, para fazer o processo de transformação, para então preparar, normalizar ou padronizar os dados para a modelagem preditiva. Os dados em si são apenas material bruto, mas quando analizados, descritos e interpretados é que realmente consegue-se coletar informações e depois se tornarão em conhecimento para então ajudar na tomada de decisões.

Em geral os métodos estatísticos descritivos são de natureza gráfica ou numérica. O principal propósito de métodos gráficos é organizar e apresentar os dados de forma gerencial e ágil. É importante primeiro coletar informações sobre as características dos dados.

Algumas das principais ferramentas gráficas utilizadas para visualização de dados são o histograma, cujo objetivo principal é mostrar a distribuição de frequência de uma variável (análise univariada) e com isso analisar se os dados seguem ou não o formato de uma distribuição normal. O Gráfico de Dispersão que é a representação de duas variáveis (análise bivariada) e como elas se correlacionam. A correlação permite determinar quão fortemente os pares de variáveis estão relacionados. A correlação

permite analizar duas variáveise então extrair como informação a força da relação entre elas. Dentre outras ferramentas como gráfico de Barras, gráfico de Linha e gráfico de pizza.

2.4 MACHINE LEARNING

Machine Learning (ML) ou aprendizado de máquina é um subcampo da Inteligência Artificial (IA) que permite dar aos computadores a habilidade de aprender sem que sejam explicitamente programados para isso. Embora esteja atualmente nos holofotes, IA vem sendo pesquisada desde a década de 50, com algumas décadas de trabalho, de pesquisa, de descobertas em diversas técnicas e uma dessas técnicas é o aprendizado de máquina que tem como objetivo básico, treinar um algoritmo com um conjunto de dados (pré-processados, limpos e tratados) é executado em um computador. Durante o treinamento o algoritmo vai aprendendo o modelo matemático que representa os dados. Uma vez esse modelo aprendido, novos dados são oferecidos e ele será capaz de fazer previsões.

Aprendizado pode ser descrito como a capacidade de se adaptar, modificar e melhorar seu coportamento e suas respostas, sendo portanto uma das propriedades mais importantes dos seres ditos inteligentes, sejam eles humanos ou não, dentro de cinco características: adaptação, correção, otimização, representação e interação.

Adaptação é a mudança de comportamento de forma a evoluir, ou seja, melhorar segundo algum critério. Um sistema biológico ou artificial que não seja capaz de evoluir ou de mudar seu comportamento diante de novas situações que lhe são propostas, é um sistema sem inteligência.

Correção dos erros cometidos no passado de modo a não repeti-los no futuro. Assim como a adaptação, um sistema deve modificar o seu comportamento caso o atual não satisfaça algum tipo de exigência, ou seja, o aprendizado deve permitir a correção de erros.

Otimização é a melhora da performance do sistema como um todo. O aprendizado pode implicar em uma mudança de comportamento que busque economia

de energia gasta para realizar uma tarefa, a reduçao do tempo gasto de uma tarefa e assim por diante.

A representação do conhecimento adiquirido, o sistema deve ser capaz de armazenar uma massa muito grande de conhecimentos e isso requer uma forma de representar esses conhecimentos que permita o sistema (biológico ou artificial) explorá-los de maneira conveniente.

Interagir com o meio através do contato com o mundo que nos cerca e com o qual se pode trocar ou realizar experiências de forma a adquirir novos conhecimentos.

Há diferentes técnicas de aprendizagem e os três tipos principais são a aprendizagem supervisionada, aprendizagem não supervisionada e aprendizagem por reforço. Aprendizagem Supervisionada é o termo usado sempre que o algoritmo é "treinado" sobre um conjunto de dados pré-definido. A figura 2.1 ilustra o fluxo do processo de aprendizado supervisionado, onde os algoritmos fazem previsões com base em um conjunto de exemplos, na tentativa de reproduzir nas máquinas observando a forma como o ser humano aprende.

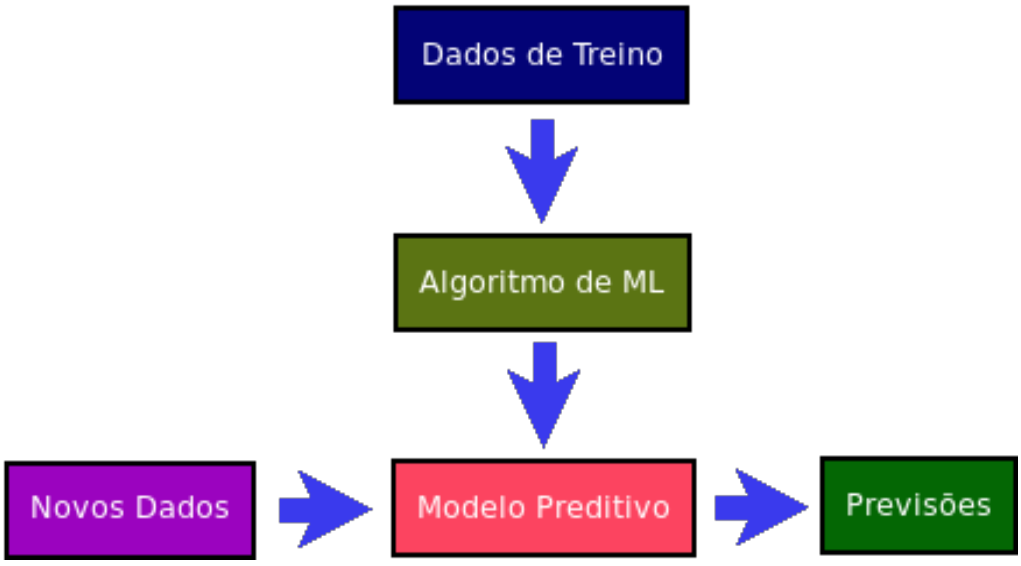

FIG. 2.1: Processo de Aprendizagem Supervisionada

No processo de aprendizagem supervisionada utiliza dados históricos, alimenta o algoritmo, faz com que o algoritmo passe por esses dados diversas vezes para aprender as diferenças, correlações, etc e no final cria o modelo que é uma função matemática.

Podemos representar a realidade e toda a sua complexidade através de funções matemáticas e o objetivo de ML é encontrar uma função matemática aproximada que melhor explica a relação entre os dados. Dentro da Aprendizagem Supervisionada há duas subcategorias importantes, a Classificação (prever uma categoria) e a Regressão (prever um valor numérico). Classificação é o processo de identificar a qual conjunto de categorias uma nova observação pertence, com base em um conjunto de dados de treino contendo observações (ou instâncias) cuja associação é conhecida.

Apesar de décadas de pesquisas, ML e IA começaram a receber mais atenção ao longo dos últimos anos devido ao Big Data, que é caracterizado pelo grande conjunto de dados definido por alto volume, alta variedade, alta velocidade de geração de dados e com veracidades e a evolução da capacidade computacional, pois mesmo com muitos dados, sem capacidade computacional, não tería avançado tanto quanto nos últimos anos. Principalemte com a computação paralela em GPUs (Unidade de Processamento Gráfico). A computação paralela permitiu, também dar um salto em termos de capacidade computacional o que aliado ao Big Data e aos algoritmos de ML formaram o que chamou-se de "Tempestade Perfeita".

Observando tudo em volta, neste momento, está gerando dados. Smartphones, computadores, geladeiras, automóveis, etc. Esse volume de dados tem crescido de maneira exponencial e embora já existisse ML há algumas décadas, nunca se teve tantos dados como nos dias de hoje e esse foi o grande fator para extrair cada vez mais dos algoritmos de aprendizado de máquina e desenvolver aplicações de IA.

Os dados são o novo petróleo, pois são o ativo cada dez mais precioso para qualquer empresa atenta e aos benefícios que a análise de dados pode trazer para qualquer compania. Quanto mais dados tiver, melhor será o modelo no final (na maioria dos casos).

3 METODOLOGIA

Este capítulo apresenta a metodologia proposta neste trabalho para a coleta e tratamento de dados para uma análise exploratória e o conhecimento sobre o fluxo de impressão.

Esta metodologia tem por objetivo a construção de modelos de machine learning a partir de métodos estatísticos e encontrar padrões para otimizar e reduzir o desperdício de impressão de provas em um processo acadêmico.

3.1 FLUXO DE IMPRESSÃO GENÉRICO

A figura 3.1 apresenta um diagrama com um fluxo de impressão genérico desde o início da preparação das provas, passando pela impressão que tem como origem as infraestruturas própria e/ou terceirizada, até a utilização e descarte das folhas.

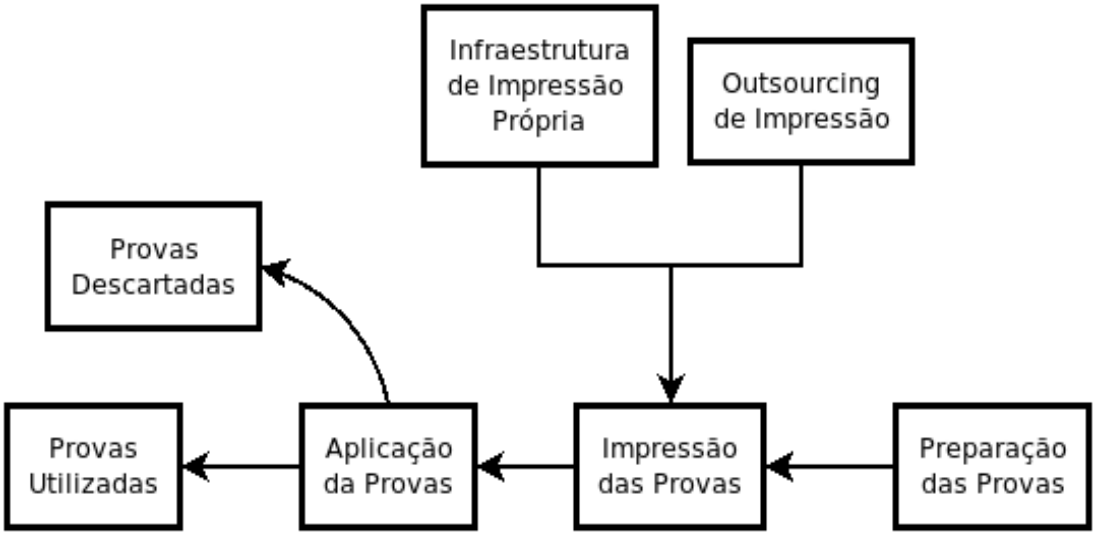

FIG. 3.1: Fluxo de Impressão Genérico

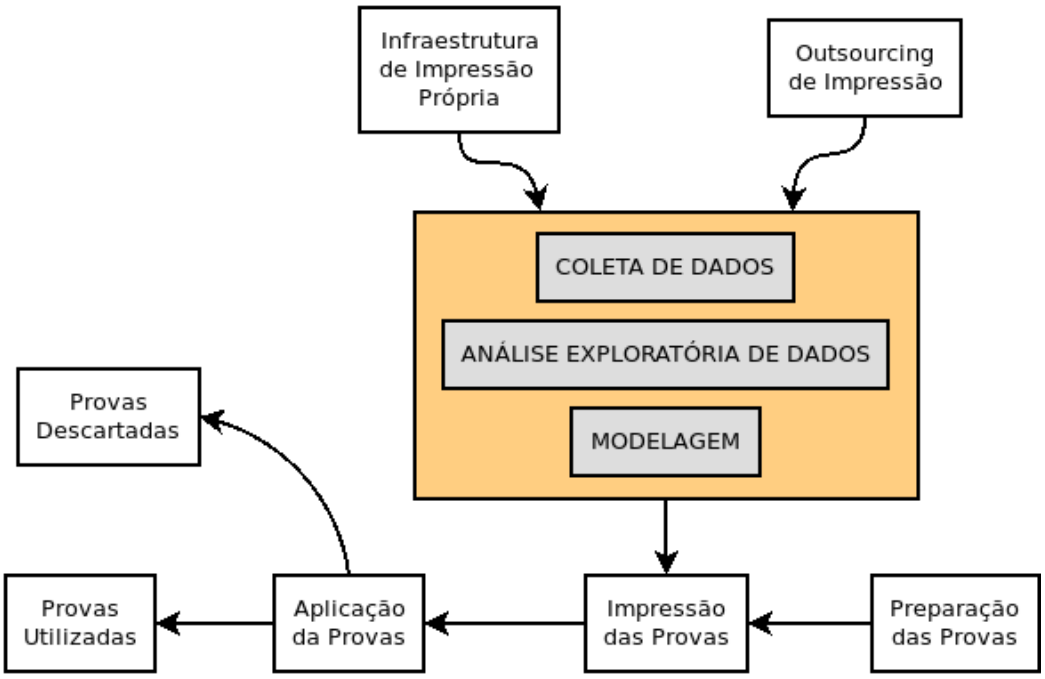

FIG. 3.2: Fluxo de Impressão Proposto

Este estudo propõe contribuir com tomadas de decisões para otimização na cadeia de processos de impressão de provas, através de coleta de dados, análises exploratória de dados e modelagem preditiva, demostrado na figura 3.2, com a utilização do que a instituição tem de mais valioso que são os seus dados históricos e incentivar a cultura de decisões a partir de dados.

3.1.1 COLETA DE DADOS

A partir da definição do objetivo que direcionará o foco da solução do problema a ser resolvido, a etapa imediatamente posterior é a coleta de dados que deve ser executada com critério e planejamento para garantir a qualidade em quantidade. Os dados sempre foram importantes e utilizados para traçar metas, definir estratégias ou qualquer outra ação que envolva uma organização, seja pública ou privada e atualmente dados são gerados continuamente em grande quantidade e diversidade.

Com o constante crescimento e complexidade dos dados, surge abordagens do conceito do Big Data que considera o Volume, a Velocidade, a Variedade, a Veracidade e o Valor dos dados (NETTO, 2016). Portanto, o investimento em para transformar dados em informações e em conhecimento é determinante para o sucesso de uma empresa. Quanto mais dados, melhores resultados serão apresentados.

3.1.2 ANÁLISE EXPLORATÓRIA DE DADOS

A disponibilidade dos dados, somente, não é o suficiente para a qualidade da informação, tampouco revela a solução para o problema de negócio de forma clara. Além de coletar os dados segundos os critérios que garantem as suas características, é necessário aplicar a metodologia adequada para interpretar e compreender os dados. Análise exploratória é uma das partes fundamentais que não deve ser negligenciada e o processo de preparação e exploração de dados pode ser subdividida em cinco etapas:

- Identificação das variáveis: são definidas as variáveis de entrada ou preditoras (inputs) e as variáveis de saída ou variável alvo(output), os tipos de dados e categorias das variáveis. As variáveis são as colunas em um dataframe, de uma planilha ou uma tabela em um banco de dados. É comum que as variáveis estejam mal distribuídas e uma mesma variável pode estar espalhada por mais de uma coluna e os dados em formatos diferentes.

- Tratamento de Valores Missing: são valores que não estão presentes, que estão faltando em determinadas linhas e colunas. Os dados dificilmente virão limpos e organizados e valores poderão estar faltando em determinadas variáveis. Os valores Missing podem levar a resultados totalmente distorcidos e podem ser um problema na fonte de dados ou um problema durante o processo de extração e coleta dos dados. Para o tratamento de valores missing existem diversas técnicas, pode

remover a linha onde o valor missing aparece ou pares de linhas dependendo da relação entre as variáveis. Cada método terá vantagens e desvantagens.

- Tratamento de Outliers: são valores extremos que fogem muito da média dos valores, uma observação que está muito distante do padrão observado naquele conjunto de dados e existem basicamente os tipos de outliers unicariado e multivariado. Pode ser causado por entrada errada de dados, erros de experimento, pode ser intencional ou uma observação natural de acordo com o conjunto de dados. Seja qual for a causa, os outliers terão impacto na análise e pode simplesmente ser removidos, tratados separadamente ou transformados.

- Transformação de Variáveis: uma única variável pode ter valores missing, outliers, valores incorretos que precisam de limpeza, entre outras possibilidades. Nesta fase pode mudar a escala de uma variável, transformar relacionamentos não lineares complexos e relacionamentos lineares simples, mudar a simetria da distribuição de dados entre outras opções.

- Criação de Variáveis: Depois que os dados estiverem coletados e devidamente explorados, pode haver necessidade de gerar relacionamentos ou definir melhor determinada análise. Nesse momento pode ser necessário criar variáveis, seja transformando as existentes e então gerando uma nova ou a partir da trasformação de outras variáveis ou ainda coletando mais dados. É possível ainda aplicar funções às variáveis existentes e o resultado dessa função gerar uma nova variável.

É possível explorar os dados sem que haja um problema inicialmente definido afim de buscar padrões ocultos e responder perguntas que ainda não foram feitas, pois nem todos os problemas estão claros num primeiro momento. Apesar de que se deve ter um mínimo de noção do que se está procurando.

3.1.3 MODELAGEM

Passado a fase crítica de transformação para que os dados estejam limpos e organizados, é a etapa de definir o modelo que ajude a explicar o relacionamento entre as variáveis e fazer as previsões. Essa etapa requer uma série de escolhas, afinal existem mais de 60 algoritmos de ML e a primeira é escolher qual o algorítmo utilizar de acordo com a necessidade.

Na sequência os dados devem ser subdivididos em dados de treino e dados de teste e então realizar um trabalho iterativo. Depois de treinar, avaliar e testar o modelo, verificar se a precisão está adequada e se não estiver, retorna e modifica os dados, altera os parâmetros do algoritmo, treina novamente, testa e avalia e esse ciclo se repete algumas vezes até encontrar o modelo adequado.

O modelo encontrado poderá ser automatizado de modo que ao receber novos conjuntos de dados seu processo possa ser executado de uma única vez, preferencialmente sem a interferência humana. É nesta fase que é pensado em como transformar o processo de análise em um produto ou serviço que poderá ser publicado através de um dashboard usado pelos gestores para tomada de decisões, uma aplicação web ou uma aplicação para smartphone de acordo com a forma que se quer automatizar o processo.

4 ABORDAGEM METODOLÓGICA

Este capítulo apresenta a implementação da metodologia para a coleta e análise exploratória dos dados e construção de modelos preditivos. Para a coleta dos dados é necessário que o professor disponibilize os dados histórico de suas turmas, pois dados fictícios não gerariam relevância nos resultados. A quantidade de dados solicitada inicialmente foi de 40 turmas de um professor.

O ambiente escolhido para a implementação foi o Anaconda, uma distribuição Python que vem com os principais pacotes necessários para trabalhar com ciência de dados, facilitando o procedimento de instalação de vários pacotes manualmente. A linguagem Python versão 3 foi escolhida para a implementação da metodologia, uma linguagem de programação de uso geral, fácil de aprender, amplamente utilizada e poderosa.

4.1 TECNOLOGIAS UTILIZADAS

Segue a descrição das tecnologias utilizadas para implementar essa abordagem.

4.1.1 PYTHON

Python é uma linguagem de programação criada para produzir código bom e fácil de manter de maneira rápida. Entre as características da linguagem que ressaltam esses objetivos sao o baixo uso de caracteres especiais, o que torna a linguagem muito parecida com pseudo-código executável, o uso de identação para marcar blocos, quase nenhum uso de palavras-chave voltadas para a compilação, coletor de lixo para gerenciar automaticamente o uso da memória, etc.

4.1.2 ANACONDA

Anaconda é uma plataforma de ciência de dados para Python. Possibilita a instalação de diferentes versões da linguagem Python com a criação de ambientes de desenvolvimento específicos.

4.1.3 PANDAS

Pandas é uma biblioteca de código aberto, que fornece estruturas de dados de alto desempenho e fáceis de usar e ferramentas de análise de dados para a linguagem de programação Python.

4.1.4 SCIKIT-LEARN

Scikit-Learn é uma biblioteca de aprendizado de máquina de código aberto para a linguagem de programação Python. Ela inclui vários algoritmos de classificação, regressão e agrupamento incluindo máquinas de vetores de suporte, florestas aleatórias, gradient boosting, k-means e DBSCAN, e é projetada para interagir com as bibliotecas Python numéricas e científicas NumPy e SciPy.

4.1.5 MATPLOTLIB

Matplotlib é uma biblioteca Python de plotagem 2d, que auxilia a biblioteca matemática NumPy. Pode ser usada em scripts Python, no shell Python e IPython, em servidores de aplicação web e outras ferramentas de interface gráfica.

4.1.6 GITHUB

Github é um serviço online de hospedagem de repositórios Git (como são chamados os projetos que utilizam Git). Com ele podemos manter todos os projetos sincronizados entre os membros do time.

4.2 CENÁRIO CONSTRUÍDO

O Cenário em questão é composto por informações de 46 turmas de resultados recuperadas de um professor dos anos de 2016 a 2019. O dados foram disponibilizados por meio de um arquivo em formato xlsx, organizados em uma turma para cada aba da planinha, posteriormente tratados para manter a privacidade dos dados pessoais dos alunos, professores e da instituição e gerado novos arquivos genéricos em formato csv para as pesquisas.

4.3 SELEÇÃO E TRATAMENTO DAS VARIÁVEIS

Os dados receberam tratamento dos campos nulos para zeros, os nomes dos alunos foram excluídos do dataset bem como os nomes dos professores, as matrículas alteradas com uma operação de criptografia hash, que manterá a unicidade preservando a privacidade.

O Anaconda Python possui uma ferramenta já instalada chamada Jupyter Notebook que carrega o ambiente Python via browser local, pelo endereço *http://localhost:8888/*, após a execução do comando jupyter notebook no prompt de comando. A ferramenta é apresentada em forma de células para execução dos códigos Python e tem recursos de cores e identação padrões da linguagem. A leitura e carga do arquivo em formato *xlsx* fornecido pelo professor, foi carregado em memória, através dos recursos das bibliotecas da linguagem, e os dados tratados e organizados em um novo arquivo em extensão do tipo *csv*.

No arquivo novo, pronto para a exploração e visualização, foi feita a primeira pergunta e respondida na figura 4.1: quantos alunos fizeram a avaliação final e quantos não fizeram, por turma, em proporção percentual e destaque para a turma com maior presença.

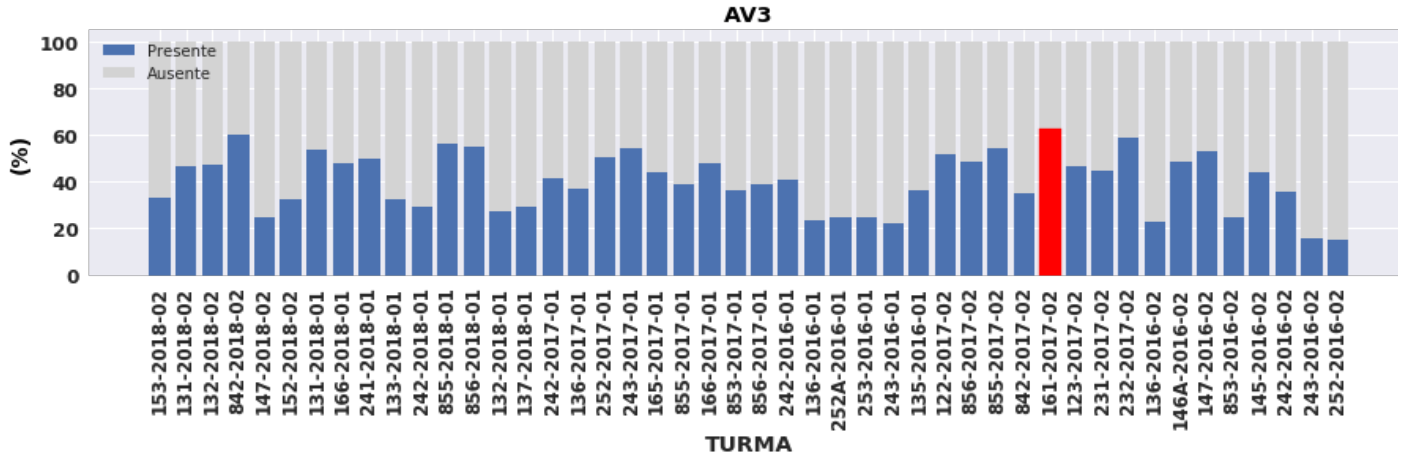

FIG. 4.1: Presença na AV3 por Turma em proporção percentual

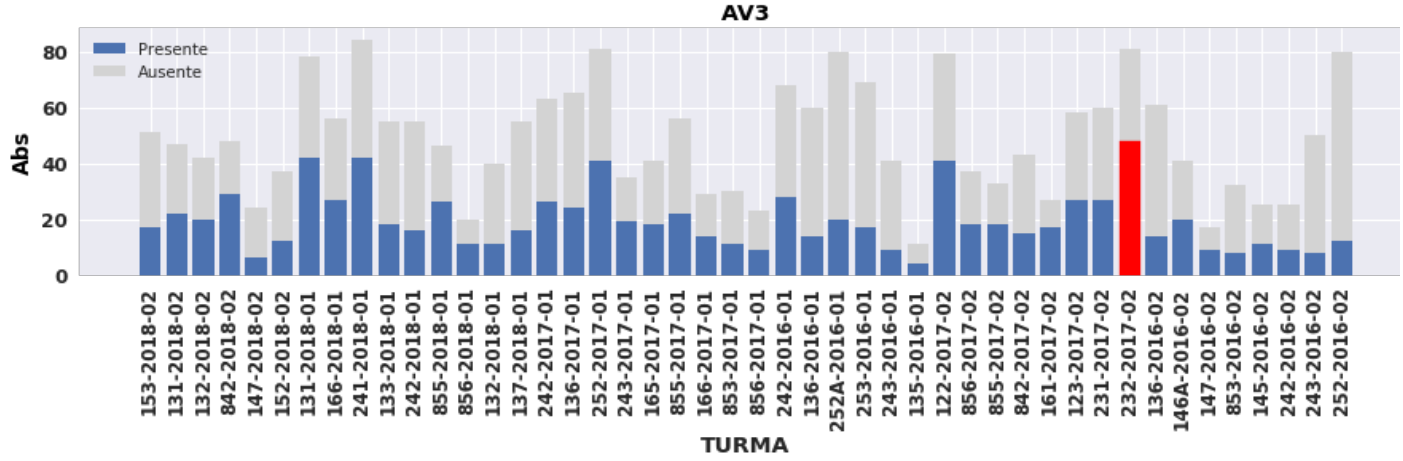

FIG. 4.2: Presença na AV3 por Turma em valores absolutos

A quantidade de alunos presentes e ausentes por turma podem ser observadas, também em forma de valores absolutos, conforme o gráfico na figura 4.2.

A função desenvolvida para a visualização da presença de alunos na avaliação final é generalizavel e tem como parâmetros, além da opção de visualização de valores absolutos e proporção percentual, uma lista de variáveis a ser observada. A figura 4.3 demonsta por disciplina e por período.

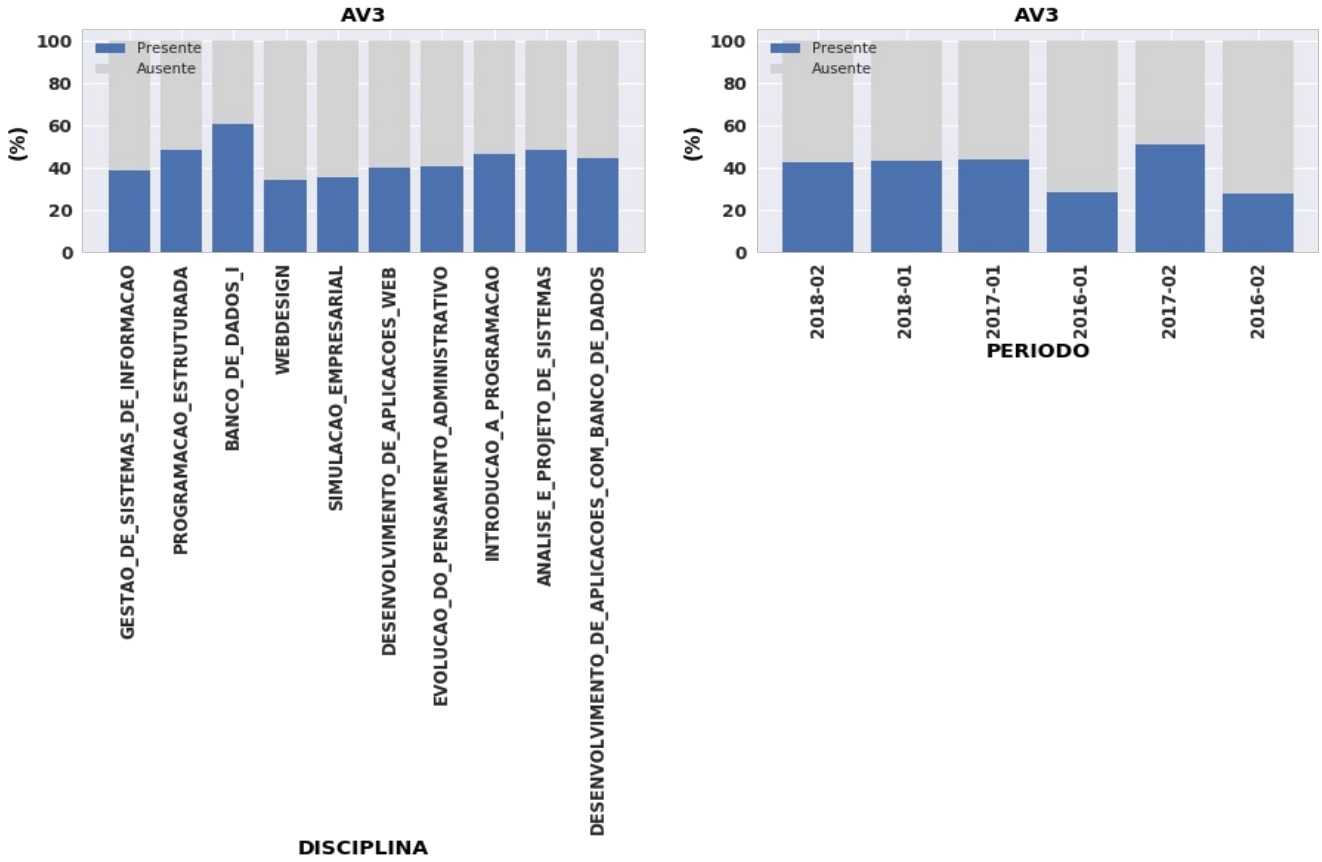

FIG. 4.3: Presença na AV3 por Disciplina e Período em proporção percentual

A figura 4.4 demonstra a presença por curso e turno.

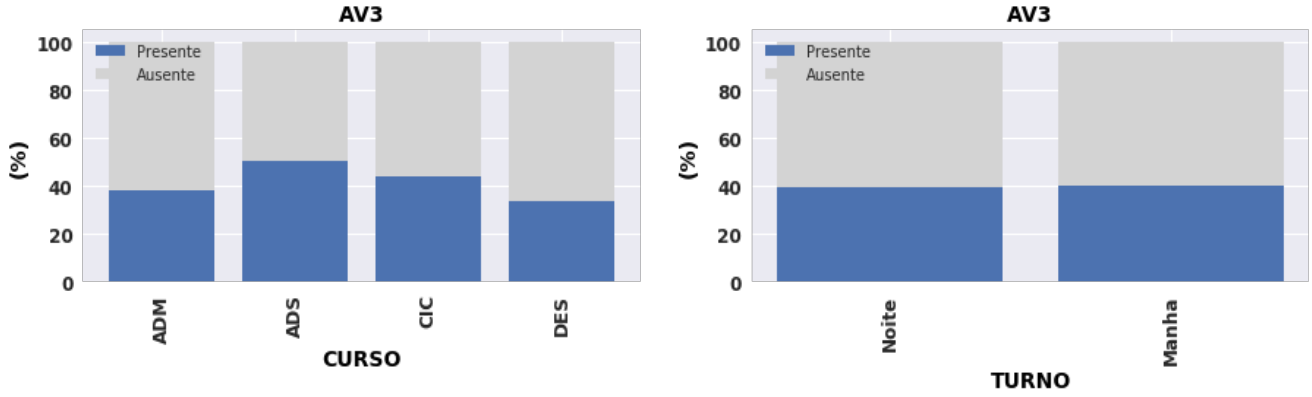

FIG. 4.4: Presença na AV3 por Disciplina e Período em proporção percentual

O gráfico ilustrado na figura 4.5 representa a influência do tamanho da turma na presença dos alunos na AV3 e verificou-se que não tem qualquer relação.

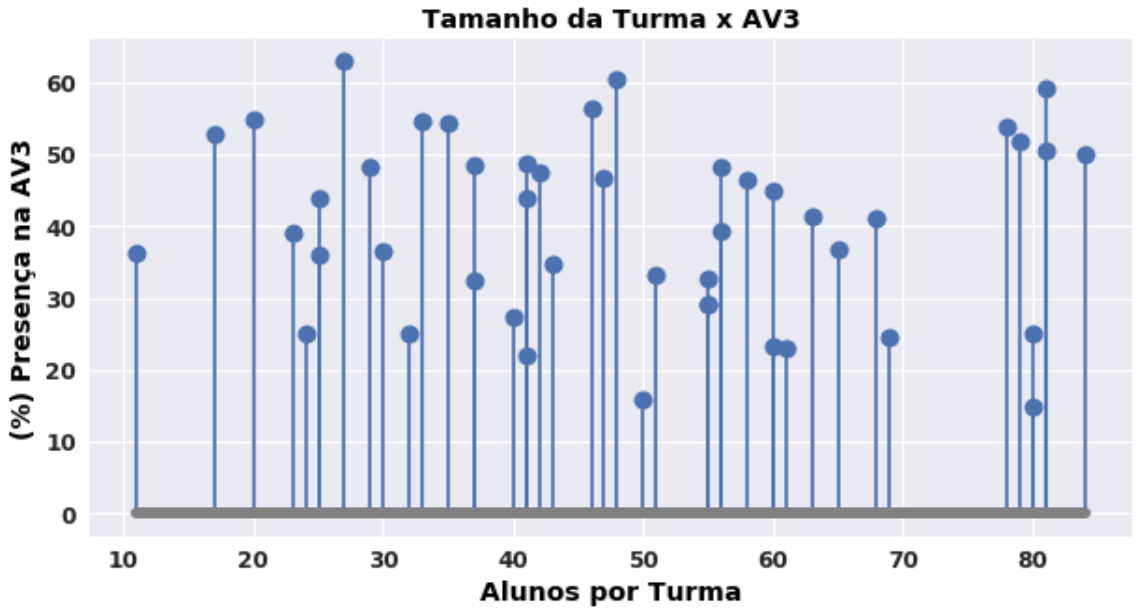

FIG. 4.5: Influência do Tamanho da Turma na Presença na AV3

É interessante observar quantos alunos comparecem para a AV3, mesmo tendo alcançado a média e somente para aumentar a nota da média final. A representação é de menos de cinco porcento com os dados do estudo, conforme ilustrado na figura 4.6 no gráfico de pizza.

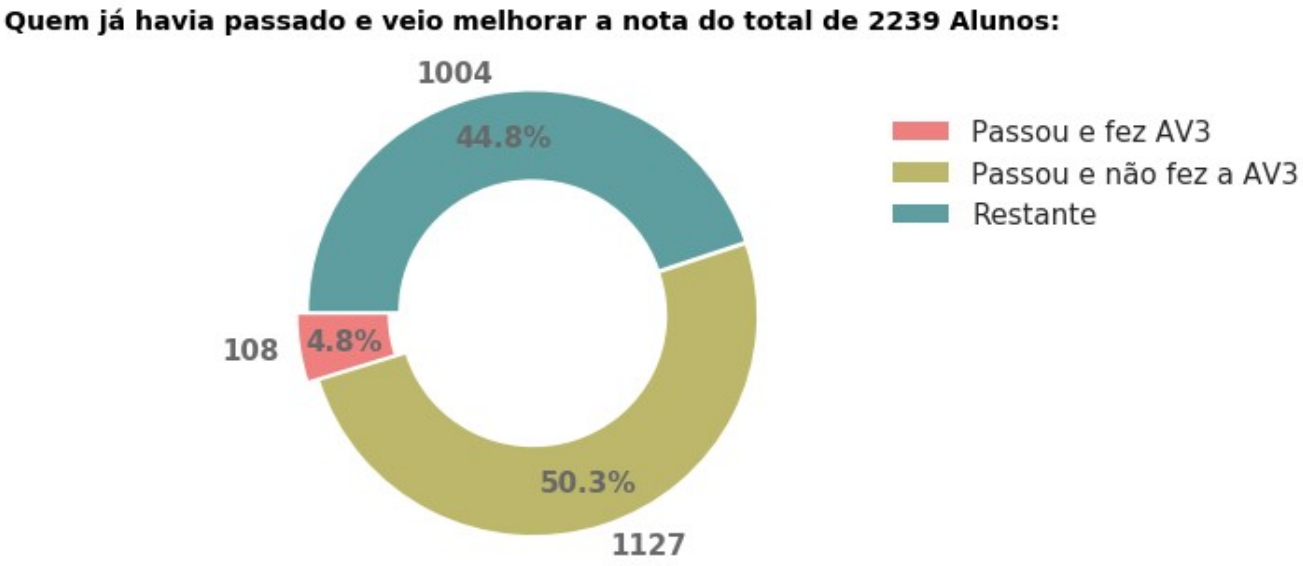

FIG. 4.6: Presença na AV3 para melhorar a Média Final

As próximas observações tem como objetivo analisar a quantidade de faltas dos alunos durante o período e qual a influência nas presenças para as avaliações. A figura 4.7 ilustra um gráfico do tipo histograma dos que fizeram somente uma das avaliações e na figura 4.8 demonstra as faltas geral e compara com as faltas de quem não fez a AV3.

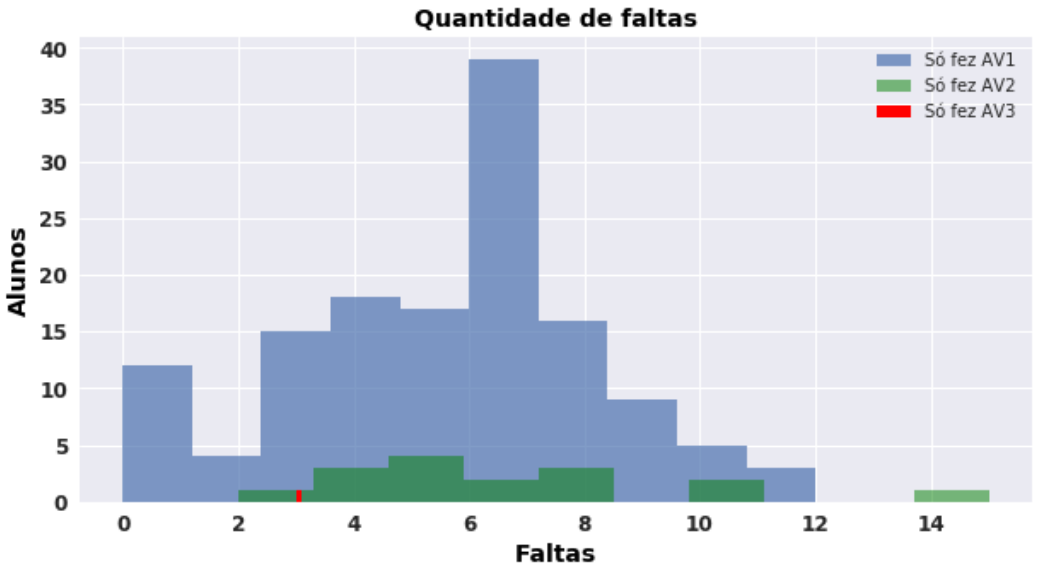

FIG. 4.7: Número de faltas por Avaliação

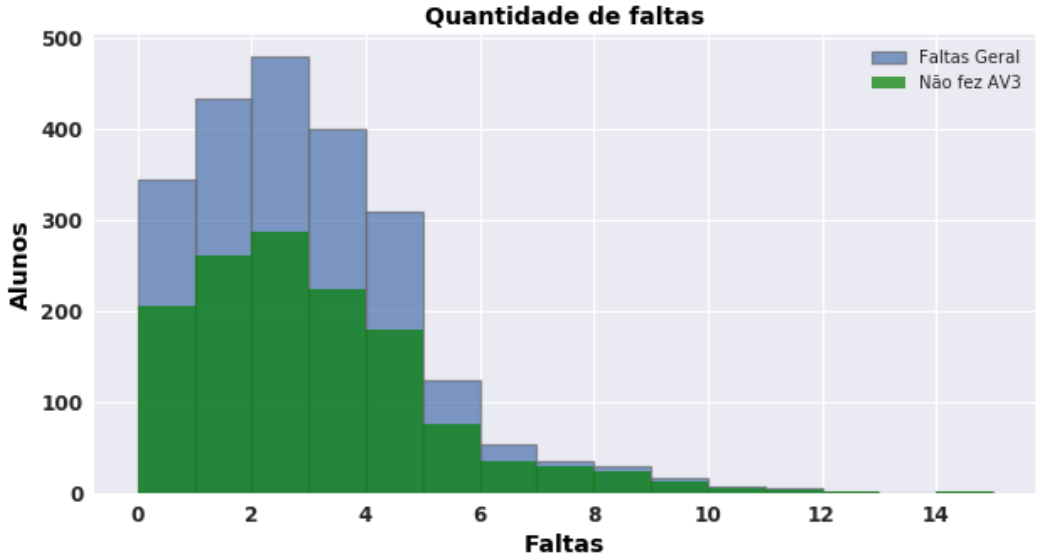

FIG. 4.8: Número de faltas Geral e na AV3

A tabela 4.1 demonstra as dez turmas ordenadas por maior presença na AV3, em proporção percentual e valores absolutos de ausência e presença por turma e no geral. Em destaque, a turma 161 com maior adesão à AV3, com 37,04% de provas descartadas, em comparação com o descarte geral de 60,12%. É possível perceber pela tabela que, pelo total de alunos por turma não há relação entre o tamanho da turma com a presença na AV3, observado na figura 4.5.

		Geral	TURMA										Média
			161 2017-02	842 2018-03	232 2017-03	855 2018-02	856 2018-02	855 2017-03	243 2017-02	131 2018-02	147 2016-03	122 2017-03	
Ausente	(%)	60,12	37,04	39,58	40,74	43,48	45	45,45	45,71	46,15	47,06	48,1	43,831
	ABS	1346	10	19	33	20	9	15	16	36	8	38	20,4
Presente	(%)	39,88	62,96	60,42	59,26	56,52	55	54,55	54,29	53,85	52,94	51,9	56,169
	ABS	893	17	29	48	26	11	18	19	42	9	41	26
TOTAL		2239	27	48	81	46	20	33	35	78	17	79	46,4

TAB. 4.1: As dez turmas com maior presença em proporção percentual na AV3

4.4 MODELOS DE MACHINE LEARNING

Com o conjunto de dados tratados, limpos, organizados e bem observados com as análises feitas até aqui, a etapa de definir o modelo preditivo está pronta para ser iniciada. Foram criadas algumas variáveis para média parcial (MEDIAParcial), que é o resultado após as primeiras avaliações e a variável de saída (AV3Bin) que será a avaliação final trocando os campos das notas por valores um (1) e os campos vazios por valores zero (0), representando uma classe binária. Serão descartados do tadaset os valores zeros para todas as avaliações, pois entende-se que quem não compareceu a nenhuma avaliação fica fora da análise.

A classe de saída é a variável que o algoritmo mapeará as variáveis preditoras no modelo de aprendizagem supervisionado e o balanceamento entre a quantidade de cada valor impacta no resultado final no modelo preditivo.

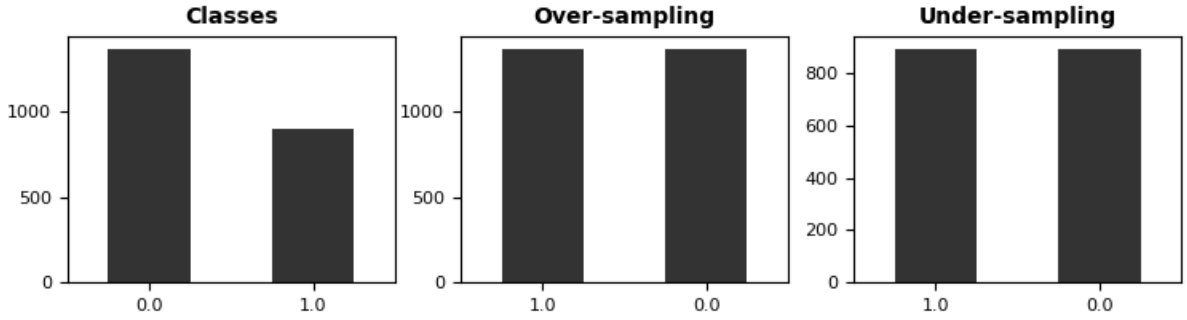

FIG. 4.9: Balanceamento de Classes

Ao avaliar os dados e detectar desbalanceamento na classe, algumas técnicas podem ser aplicadas e os que vem sido muito utilizados por serem mais simples e terem gerados sólidos resultados, são chamados de over-sampling e under-sampling. A figura 4.9 ilustra o formato da classe no primeiro gráfico e o resultado após a aplicação das técnicas de over-sampling e under-sampling no gráficos gráficos em sequência.

Para a escolha do modelo de ML dentre mais de 60 algoritmos, conforme descrito no capítulo 3, a estratégia adotada neste estudo foi a de criar diversos modelos diferentes, compará-los e selecionar os melhores. Os dados foram separados em dados de teste e dados de treino e aplicados para cada modelo, alterando os parâmetros de cada algoritmo de forma automatizada.

Accuracy		Cross Validation		Repeated Cross Validation	
Rsr_	: 89.97	RSr_	: 87.29	SVM	: 87.21
SVM	: 89.38	SVM	: 87.12	RSr_	: 87.09
EXT_	: 89.38	GB	: 87.12	EXT_	: 87.03
RNN2	: 87.61	EXT_	: 87.03	XGB	: 86.88
GB	: 87.46	XGB	: 86.99	GB	: 86.65
XGB	: 87.32	RNN	: 85.26	RNN2	: 86.13
RNN	: 86.87	RNN2	: 85.26	RNN	: 85.35
RF2	: 85.69	GB2	: 84.37	RF	: 84.50
RF	: 84.96	RF2	: 84.29	RF2	: 84.31
EXT2	: 84.51	EXT	: 84.15	GB2	: 84.10
EXT	: 84.22	RF	: 84.11	EXT	: 84.05
GB2	: 84.22	EXT2	: 83.75	EXT2	: 84.03
CART	: 83.63	CART	: 83.09	CART	: 82.83
CART2	: 83.63	CART2	: 83.00	CART2	: 82.78
NB	: 71.09	NB	: 75.08	NB	: 75.19
LR	: 63.72	LR	: 64.94	LR	: 64.92

TAB. 4.2: Lista dos Algoritmos após treinamento ordenado por melhor acurácia

A escolha do algoritmo utilizando este método exige maior recurso computacional, portanto mais tempo para ser executado. Então deve-se avaliar a melhor escolha (*Trade-off*) entre o melhor modelo dentro do tempo viável e recursos disponíveis. A tabela 4.2 mostra os modelos treinados em ordem de melhores resultados após aplicação de técnicas de amostragem de *cross-validation*. Na figura 4.9 os resultados são apresentados em gráficos boxplots com mais detalhes sobre cada modelo.

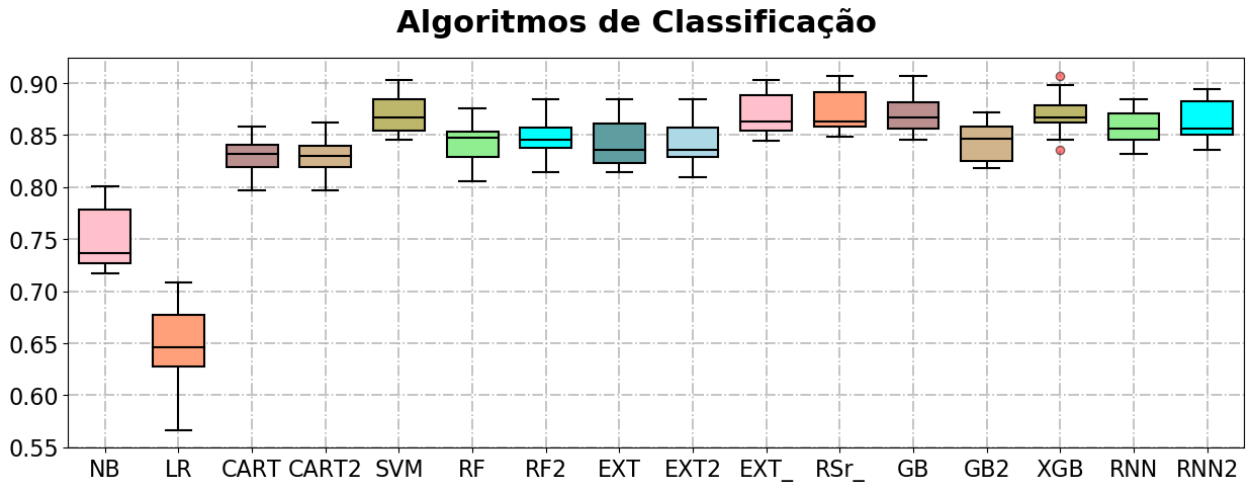

FIG. 4.10: Algoritmos treinados comparados em boxplot

A precisão do modelo é definida no início, junto com a fase da definição do problema. O nível de acertos aceitável é a partir de 70%, em geral e dificilmente chegaria a 100%, pois todo modelo tem um nível de erros esperado da função matemática genérica mapeada de acordo com os dados que o alimenta. Se um modelo alcançar precisão total na fase de testes após o treinamento, haveria forte evidência de sofrer com overfiting, quando o conjunto de dados recebidos são memorizados, mas haveria muitos erros ao prever para novas entradas.

Com o melhor modelo selecionado, um novo treinamento é feito, dessa vez com todos os dados disponíveis e salvo em um arquivo para posterior recarga. O modelo obteve uma acurária de 87,9 % e aplicado aos dados de uma turma de 83 alunos como exemplo, obteve o resultado de 59 que fará a prova e o restante, 24 anulos, não farão a prova. Equivalente a 71.08% para os presentes e 28.92% para os ausentes para essa turma.

A avaliação do modelo pode ser visualizada através de uma tabela que mostra as frequências de classificação para cada classe, chamada Matriz de Confusão (Confusion Matrix). A figura 4.10 representa a Confusion Matrix de forma gráfica e mostra na diagonal principal a classe prevista corretamente como presente para a prova (verdadeiro positivo) e prevista incorretamente para presença (falso positivo). Na diagonal secundária são as previsões para quem não estarão presente para a prova (verdadeiro negativo) e previsões incorretas para ausência na prova, mas que estaria presente (falso negativo).

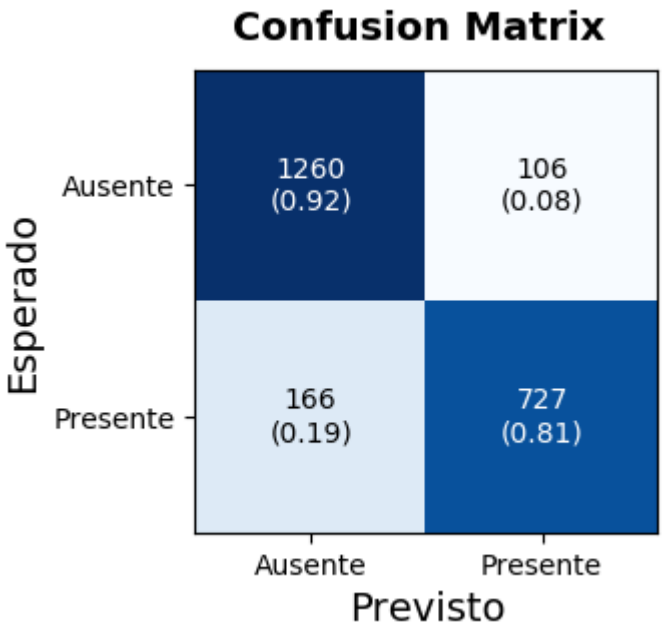

FIG. 4.11: Matriz de Confusão (Confusion Matrix)

5 ANÁLISE DE RESULTADOS

Os resultados das análises exploratória de dados consistiu na visualização da quantidade de alunos presentes por disciplina, período, curso, turno e, principalmente, por turma que é a dimensão principal das decisões de impressões de provas. A relação entre o tamanho da turma com a presença dos alunos na prova final foi verificada e não encontrada. Alunos que obtiveram aprovação, mas fizeram a AV3 mesmo assim para aumentar a média foi possível visualizar e foi de menos de 5%. Sobre a quantidade de faltas no geral segue a mesma distribuição das faltas dos alunos que não fizeram a AV3, indicando que as faltas não influenciam quem fará ou não a final.

As observações das referências para impressão de provas em valores de proporções percentuais são mais representativas do que os valores absolutos, pois como visto, o tamanho da turma não influencia na presença para as avaliações e a turma com maios presença na AV3 pode ser uma turma pequena em comparação com as outras, o que de fato foi o caso.

A turma 161 do segunto semestre de 2017 foi observada com a maior presença de 62.96%, conforme demonstrado na figura 4.1, e poderia ser a base de referência para a impressão das provas. Como a turma observada com a menor presença foi de 37.04 %, com uma margem de erro sugerida de 5 a 10 porcento, as impressões de provas para a AV3 poderiam ter uma redução de 100% atual para 70 a 75 porcento para todas as turmas futuras. Uma economia de 25 a 30 porcento do total de impressões de provas finais.

A presença total de todas as turmas é de 39.88 % e ausência 60.12 %, em destaque na figura 4.1, muito acima do observado na turma descrita como referência, demonstrando uma decisão conservadora e mesmo assim um ganho de economia. Com as técnicas de ML com um modelo de 87,9% de precisão, após o treinamento com os dados disponíveis, com uma margem de erro de 15% sugerida, as previsões seriam feitas para cada turma e obter maior redução de impressões ajustadas para cada turma. Com a mostra de uma turma de 83 alunos, o modelo previu que 59 (71.08%) farão e 24 (28.92%) não farão a prova final, a impressão de provas para esta turma

seria de 80 a 85 porcento. O estudo foi feito com os dados de 46 turmas de três anos de um professor.

Para uma melhor visão sobre o significado do impacto com o excedente, cada prova tem em média, de 5 a 7 páginas, uma para folha de rosto, uma de rascunho, uma para respostas e mais duas a quatro para o corpo da prova contendo as questões, podendo conter mais folhas. Com 1346 provas descartadas em 2239 total, equivalente a 6730 folhas considerando 5 folhas por prova. Com a análise estatística proposta, de forma conservadora, seriam impressas 1680 provas, uma economia de 25% do total de 2239 ou ainda, 2795 folhas deixariam de ser descartadas (559 provas). Com as técnicas de ML a economia poderia ser o dobro, pois a diferença entre a turma com maior presença, tomada como referência para a solução estatística, e o total de presença nos dados são consideraveis e haveriam ganhos significativos com análises personalisadas de mapeamento dos dados e busca de padrões com aprendizagem de máquina.

6 CONCLUSÕES

O conjunto de habilidades nas áreas da matemática, estatística, programação de computadores e conhecimento de negócios são necessários para ciência de dados, cujo principal objetivo é extrair e interpretar os dados de forma eficaz e apresentá-los em uma linguagem simples e não técnica para os usuários finais e tomadores de decisão. Neste trabalho foi apresentado técnicas de análise preditiva, modelagem, engenharia de dados e visualização, com o objetivo de demonstrar de maneira simplificada a importância do tratamento e exploração dos dados de uma instituição para conhecer e otimizar seus processos internos.

O responsável pelo sucesso dos resultados em ciência de dados, juntamente com as habilidades e técnicas nas principais áreas da matemática, estatística e ciência da computação, foi o grande volume de dados disponível para serem processados, portanto a quantidade e qualidade de dados é imprescindível. A quantidade de dados utilizados para o estudo, apesar de insuficientes para resultados conclusivos e produtivos, permitiu a compreenção do processo com a visualização e análise dos dados e motivação para buscar questões sobre outros domínios além da presença de alunos em provas finais por turma.

A principal contribuição deste trabalho foi o desenvolvimento e disponibilidade de funções de programação e construção de datasets, mediante a coleta de dados que foram tratados, organizados e limpos com privacidade preservada. As funções também são utilizadas para análise estatística e visualização de dados em forma de tabelas e gráficos e modelos preditivos com utilização de algoritmos de aprendizado de máquina.

Os datasets disponíveis em formato csv e dados reais, porém genéricos sem exposição de dados pessoais, servem como base de estudos futuros. A programação foi escrita de forma organizada, seguindo padrões de engenharia de software, podendo ser aproveitado para desenvolvimento de aplicações (web ou mobile) que automatiza o processo deste a coleta, tratamento e visualização até a contrução e aplicação do modelo preditivo.

7 REFERÊNCIAS BIBLIOGRÁFICAS

ANACONDA. **Anaconda Distribution**. Disponível em: https://www.anaconda.com/distribution/. Acesso em: 10 nov. 2019.

CARVALHO, ANA, STEFANO, SILVIO e MUNCK, LUCIANO. **Competências voltadas à sustentabilidade organizacional: um estudo de caso em uma indústria exportadora**. Em *Gestão & Regionalidade* - Vol. 31 - N o 91 - jan-abr/2015.

DOMINGOS, PEDRO. **O Algoritmo Mestre: Como a Busca Pelo Algoritmo de Machine Learning Definitivo Recriará Nosso Mundo**. Em *Novatec; Edição: 1,* 2017.

FILHO, JAIR, NUNES, PAULA, ALVES, ROBERTO e PACHECO, ANDRESSA. **Eficiência na Administração Pública: o modelo do outsourcing de impressão na Universidade Federal de Santa Catarina**. Em *Universitas Gestão e TI*, Brasília, v.4, n.2, p. 15-21, jul./dez. 2014.

GIT. **Link de disponibilização dos códigos e datasets**. Disponível em: https://github.com/luvres/TCC. Acesso em: 10 nov. 2019.

GITHUB. **The world's leading software development platform**. Disponível em: https://github.com/. Acesso em: 10 nov. 2019.

HUFF, DARRELL. **Como Mentir com Estatística (Português)**. Capa dura – por Darrell Huff (Autor), Irving Geis (Ilustrador), Bruno Casotti (Tradutor), 18 jun 2019.

MATPLOTLIB. **Plotting Library for the Python Programming Language**. Disponível em: https://matplotlib.org/. Acesso em: 10 nov. 2019.

MENEZES, DIEGO. **Análise dos custos de outsourcing de serviços de impressão**. Em *IX SAEPRO - Simpósio de Engenharia de Produção*, Universidade Federal de Viçosa, 20, 21 3 22, Novembro 2014.

MORAES, SALATI, HERMÍNIO, GUSTAVO, CAPPELLOZZA, ALEXANDRE, MEIRELLES, FERNANDO. **Será o fim do papel? os avanços tecnológicos e seus possíveis impactos no consumo de papel**. Em *Revista Eletrônica de Negócios Internacionais* (Internext), vol. 6, núm. 2, julio-diciembre, 2011, pp. 48-65.

MORETTIN, PEDRO e BUSSAB, WILTON. **Estatística básica (Português)** Capa Comum – 8 jul 2017. Em *Saraivauni*; Edição: 9, 2017

MUNCK, LUCIANO, DIAS, BÁRBARA e SOUZA, RAFAEL. **Sustentabilidade organizacional: uma análise a partir da institucionalização de práticas ecoeficientes**. Em *Revista Brasileira de Estratégia*, Curitiba, v. 1, n. 3, p. 285-295, set./ dez. 2008.

NETTO, ADRIANA, MORO EVANDRO e FERREIRA, FERNANDA. **Os 5 V's do Big Data**. Em *Big Data e suas Influências sobre a Estratégia das Empresas*, ufrj, 2016. Disponível em: https://www.gta.ufrj.br/grad/15_1/bigdata/vs.html

NORVING, PETER e RUSSELL, STUART. **Inteligência artificial**. Em *Elsevier Editora Ltda*, 2013.

O'NEIL, CATTHY. **Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy**. Em *Broadway Books*, 2019.

PANDAS. **Python Data Analysis Library**. Disponível em: https://pandas.pydata.org/. Acesso em: 10 nov. 2019.

PYTHON. **Python Software Foundation**. Disponível em: https://www.python.org/. Acesso em: 10 nov. 2019.

ROSSI, ALEXANDRE, SANTOS, ANDRÉ, BELI, DANILO e FONSECA, EDUARDO. **Economia de tinta de impressão utilizando um novo sistemade cotas nas unidades da unicamp**. Em *Revista Ciências do Ambiente On-Line*, Dezembro, 2010 Volume 6, Número 2.

SCIKIT-LEARN. **Machine Learning in Python**. Disponível em: https://scikit-learn.org/stable/. Acesso em: 10 nov. 2019.

WHEELAN, CHARLES. **Estatística: O que é, para que serve, como funciona**. Em *Zahar; Edição: 1ª*, 2016.