BayesGraphics: Ferramenta para Auxílio na Correlação de Grande Quantidade de Variáveis Utilizando Naive Bayes

Matheus F. M. Lima
Universidade Federal do Pará - UFPA
Instituto de Tecnologia / Faculdade de Engenharia da Computação e Telecomunicações
Belém, Pará
7matheus.maia@gmail.com

Marcos C. R. Seruffo
Universidade Federal do Pará - UFPA
Instituto de Tecnologia / Faculdade de Engenharia da Computação e Telecomunicações
Belém, Pará
seruffo@ufpa.br

***Resumo* —** A utilização de correlação de dados é uma importante forma para entendimento de problemas e uma maneira de encontrar possíveis soluções. Assim, compreender a correlação entre variáveis e o grau de dependência entre as mesmas para determinado problema, é um passo fundamental, não só para entendimento do cenário geral, como também para identificar quais fatores têm mais e menos impacto sob o domínio pesquisado. Neste aspecto, pesquisadores adotam técnicas de Machine Learning (ML), sendo o algoritmo Naive Bayes (NB) aplicado para mais diversas soluções, no qual, trata-se de um algoritmo simples e versátil de classificação. Entretanto, quando o problema analisado apresenta uma grande quantidade de variáveis, o processo de correlação torna-se trabalhoso e demorado, visto que a inferência entre os pares é dispendiosa. Com o intuito de otimizar o processo de extração de conhecimento a partir da correlação de dados utilizando Redes Bayesianas (RB), este trabalho propõe a ferramenta BayesGraphics, que organiza de forma gráfica o percentual de correlação entre todas as variáveis de uma RB que utilize o algoritmo de aprendizagem NB. A coleta dos valores de correlação, visualização da relevância das variáveis e ordenação são feitas de forma simplificada e automatizada, tornando o aplicativo facilmente manipulável, com o intuito de que possa ser aplicado de maneira interdisciplinar. A ferramenta foi validada a partir da aplicação em dois cenários e os resultados mostram a facilidade de manipulação e rapidez de obtenção da correlação das diversas variáveis envolvidas.

Palavras Chave— Análise de Correlação; Naive Bayes; Java; Aprendizado de Máquina.

I. INTRODUÇÃO

Ao se estudar duas ou mais variáveis, é interessante conhecer se as mesmas têm alguma relação entre si, isto é, se valores altos ou baixos de uma das variáveis implicam em valores altos ou baixos da outra variável [1]. A título de exemplo, em um prontuário medico, as variáveis utilizadas para cada paciente, podem ser: idade, sexo, sintomas e resultados de exames, no qual, determinadas variáveis podem ser relacionadas a presença de uma doença, ou não. A este fator, dá-se o nome de correlação.

Ao se estudar correlação entre variáveis um dos principais objetivos é identificar o grau de relacionamento entre as variáveis analisadas. Conforme descreve em [2], "O estabelecimento da existência de uma correlação entre duas variáveis pode constituir o objetivo precípuo de uma pesquisa (...). Mas também representar apenas um passo, ou estágio, de uma pesquisa com outros objetivos, como, por exemplo, quando empregamos medidas de correlação para comprovar a confiabilidade de nossas observações".

Uma das técnicas computacionais aplicadas na verificação de correlação entre múltiplas variáveis é o algoritmo NB. Por se tratar de um algoritmo simples e versátil, que se adéqua aos diversos cenários diferentes. Através deste algoritmo pode ser criado uma RB, que são modelos gráficos para raciocínio, onde representam variáveis e as conexões diretas entre elas.

O classificador NB é uma RB com uma estrutura fixa onde cada variável de entrada é independente das outras variáveis, dada a variável alvo. As conexões saem da variável alvo em direção a todas as variáveis entrada [3].

O uso de NB para correlação entre variáveis tem sido muito utilizado na literatura, em trabalhos de diferentes áreas de conhecimento. Por exemplo, no trabalho de [4] para aprimorar o método de detecção de *malware* em celulares *Android*, utiliza-se NB para fazer a correlação entre características retiradas do celular quando o vírus está em execução. Por sua vez, em [5] é estudado a correlação entre alguns fatores de risco que promove o câncer de mama usando NB e outros algoritmos, com o intuito de comparar entre os algoritmos, qual gera melhor resultado. No estudo de [6] é trabalhado em conjuntos de dados médicos, relacionados a doenças hepáticas, onde é correlacionado as variáveis envolvidas usando algumas técnicas de ML, uma delas é o NB, para auxílio de predição dessas doenças.

Porém, quando o problema analisado apresenta uma grande quantidade de variáveis para serem correlacionadas, o processo de extração de informações da relação entre estas

variáveis em RB torna-se um processo custoso e extenso, desta forma, desestimulando pesquisadores no estudo de ML e análise de correlações.

Neste trabalho é proposto, uma ferramenta de auxílio a pesquisadores que apliquem algoritmos de ML para análise de correlações. Foi desenvolvido uma ferramenta chamada de *BayesGraphics*, que realiza operações relacionadas a ML, sendo a principal motivação para o uso dessa aplicação, as informações de correlação entre os atributos de entrada e o atributo alvo de um conjunto de dados selecionado, organizados em uma lista, do maior atributo para o menor, segundo seu valor de relevância para o atributo alvo. O *BayesGraphics* tem como seu diferencial dos softwares existentes, o foco na visualização das correlações de forma clara e objetiva.

Este artigo está organizado da seguinte forma: a seção 2 apresenta a metodologia adotada. A seção 3 apresenta e discute os resultados encontrados. Finalmente, a seção 4 articula algumas conclusões gerais.

II. METODOLOGIA

A metodologia adotada para desenvolvimento deste artigo é apresenta na Fig. 1, que contém o fluxograma que explica as etapas adotadas:

i. Levantamento Bibliográfico: esta subseção apresenta uma breve descrição sobre conceitos, programas e trabalhos relacionados a NB além de abordar algumas definições sobre a linguagem de programação Java[1], com ênfase na biblioteca JavaFX [2] e a *Application Programming Interface* (API) do Weka[3];
ii. Prototipação: nesta subseção é proposto soluções para os problemas supracitados, através de protótipos de forma que possibilitam adequar às expectativas e alinhar aos objetivos;
iii. Resultados: será abordado na seção 3, que apresenta e discute os resultados encontrados na metodologia, seção 2.

Fig. 1 – Metodologia.

A. Levantemanto Bibliográfico

Nesta seção é fundamentado a correlação entre variáveis e discutida o contexto geral de Inteligência Artificial (IA), focado no algoritmo NB, e por fim, são apresentadas as tecnologias e ferramentas utilizadas, sendo estas: Java[1], JavaFX[2] e API do Weka[3].

a) Correlação

A área de análise de correlação tem como objetivo fornecer um número, para indicar como duas variáveis variam conjuntamente. Mede a intensidade e a direção da relação linear ou não-linear entre duas variáveis. Afirmando isto, Bussab [7] diz que, correlação é qualquer relação dentro de uma ampla classe de relações estatísticas que envolvam dependência entre duas variáveis. Tendo como exemplo, associações entre a taxa de emprego e a criminalidade, entre verba investida em propaganda e retorno nas vendas, entre outras.

Um exemplo prático de análise de correlação é apresentado no trabalho de [8], no qual, foi analisado os dados de transformadores no período de um ano, com o intuito de analisar a correlação entre temperatura ambiente, carga e ponto quente (do inglês, hotspot). No qual, como resultados tiveram a confirmação da grande influência do hostspot com a temperatura ambiente e baixa relação do mesmo com a carga.

Desta forma, fica claro a importância da análise de correlação, com a possibilidade de agir de forma preventiva e identificar quais fatores são impactantes.

b) Redes Bayesianas

IA é uma área de estudo que pode ser definida como um ramo de pesquisa que se ocupa em desenvolver mecanismos e dispositivos tecnológicos que possam simular o raciocínio humano, ou seja, a inteligência que é característica dos seres humanos [9].

Uma das principais técnicas de IA é o ML, que cria por si própria, a partir de experiências passadas, uma função, capaz de resolver o problema que deseja tratar [9]. Para isso, existem diferentes algoritmos de ML, por exemplo, árvores de decisão, clustering, algoritmos genéticos e naive bayes [6]. No qual, não existe o melhor dentre eles, todos são eficientes para determinados cenários, entretanto, neste esse artigo, será destacado o algoritmo probabilístico supervisionado NB.

Esse que tem como características ser supervisionado, significa que o conjunto de dados a ser analisado deve ser previamente validado e confiável [10]. Além disso, ele desconsidera completamente a correlação entre as variáveis, daí o motivo de receber "naive" (ingênuo) no nome[9]. O fato de que o algoritmo assume a independência entre as variáveis o torna muito rápido em seus cálculos, mesmo utilizando grandes quantidades de dados, o que, em situações práticas, faz com que seja muito utilizado em trabalhos de estudo de caso em diversas área de conhecimento.

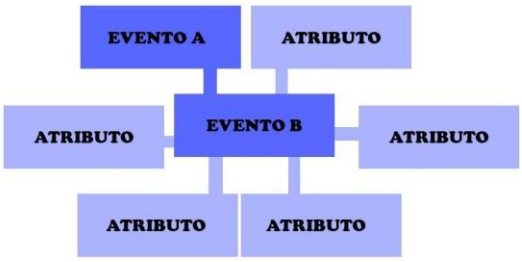

Fig. 2 – Rede Bayesiana.

[1] Disponível em: https://www.java.com/pt_BR/
[2] Disponível em: https://www.oracle.com/technetwork/pt/java/javafx/overview/index.html
[3] Disponível em: https://www.cs.waikato.ac.nz/~ml/weka/index.html

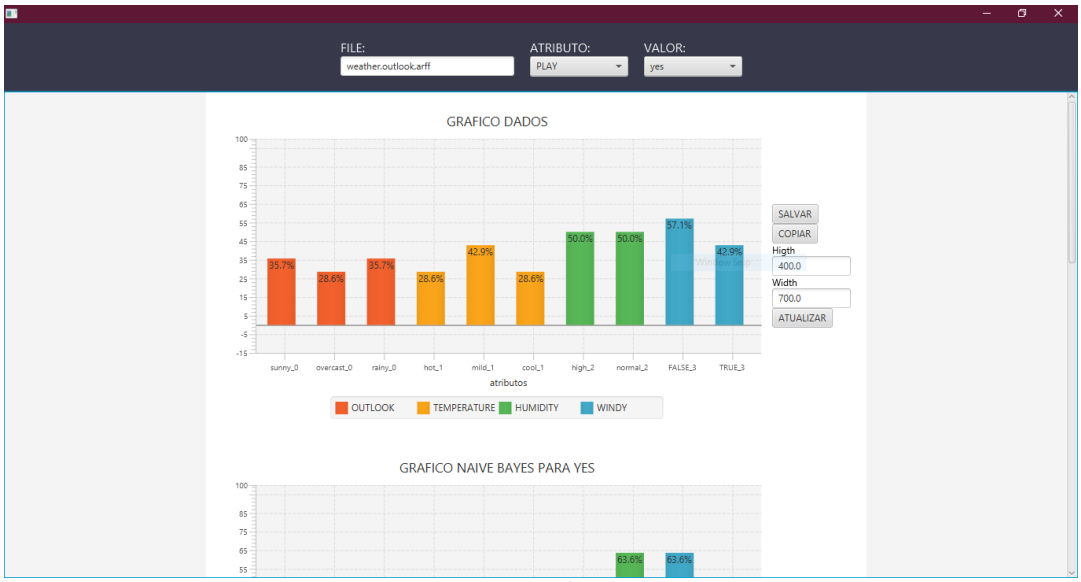

Fig. 3 – Protótipo 1.

O NB tem seus códigos baseado no Teorema de Bayes, no qual, sendo dois eventos, A e B, calcula a probabilidade do evento A acontecer, P(A), dado o fato em que o evento B ocorreu, P(A|B). Da mesma forma, para P(B|A) calcula a probabilidade do evento B acontecer, P(B), dado o fato em que o evento A ocorreu [11], como mostra a Equação 1

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{Equação (1)}$$

Dessa forma, em um conjunto de dados com *n* variáveis, cada variável de entrada torna-se um evento, como mostra na Fig. 2, no qual, um destes é destacado para ser alvo do estudo. Utilizando o algoritmo NB, é relacionado o atributo alvo com cada um dos atributos restantes. Como resultado, será gerada uma rede de atributos que estão correlacionadas, chamada de RB.

Com a RB montada é possível fazer inferências, que consiste na avaliação de hipóteses pela máxima verossimilhança, consideradas as evidências e as hipóteses de interesse[11]. No qual, pode ser imposto o valor (ou estado) de um determinado atributo (ou variável), e como resposta da RB, será mostrado a probabilidade de cada estado dos atributos restantes ocorrer dado o estado inferido.

Desta forma, por exemplo, é possível sabe qual a probabilidade de uma pessoa está com dengue, dado o fato de ela está com os sintomas de febre alta, dor nos olhos e manchas vermelhas no corpo. Por outro lado, é possível determinar qual a probabilidade de uma pessoa ter manchas vermelhas no corpo dado que está com dengue[12].

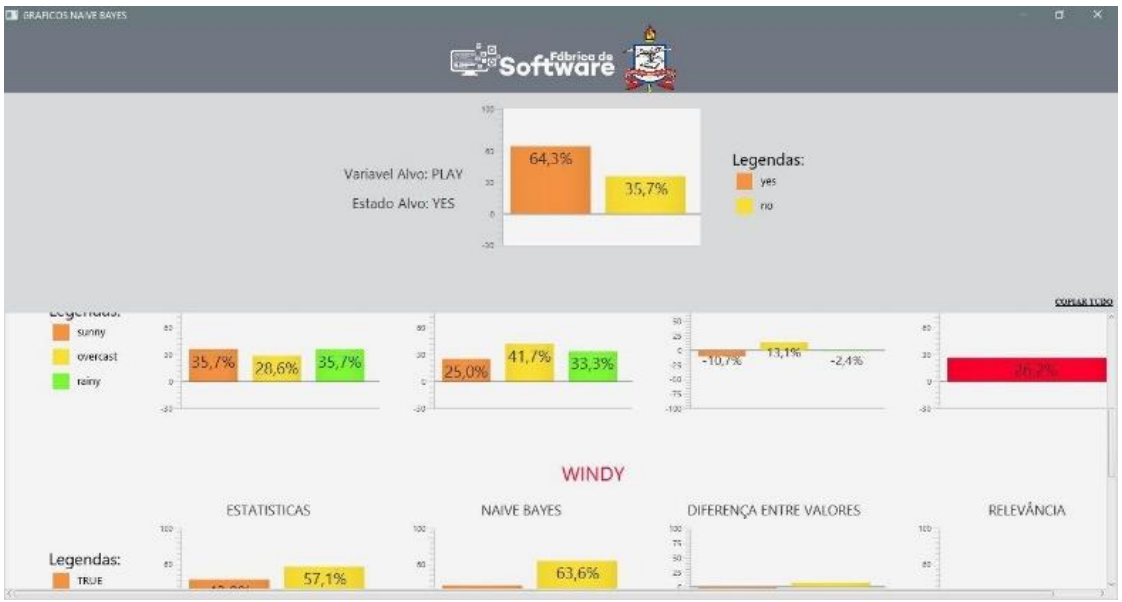

Fig. 4 – Protótipo 2.

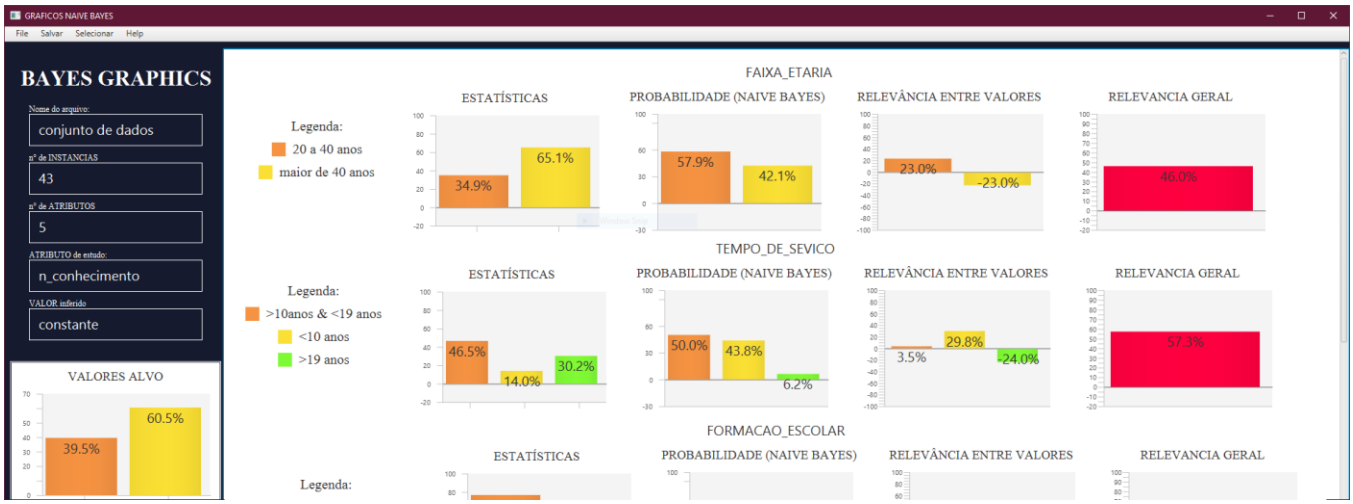

Fig. 5 – Protótipo 3.

Para esse trabalho, utilizou-se o software de código aberto WEKA[3], que contém um conjunto de algoritmos de ML, incluindo NB, onde pode ser acessado a sua API, que possibilita diversos criadores de software a desenvolverem produtos associados aos seus serviços.

B. Prototipação

Com o intuito de desenvolver uma ferramenta de auxílio a pesquisadores interessados no estudo de ML para correlação entre variáveis. Foi desenvolvida uma aplicação que lista os valores de correlação entre os atributos de entrada e o atributo alvo, como explicado na seção A.

Com base na análise empírica de softwares de ML, como GeNIe [4] e Weka[3], foram separadas as principais informações e operações relacionado ao processo de ML com NB. Algumas são: nome do arquivo, número de variáveis, número de instâncias e lista com os nomes de todos os atributos. Além disso, a lista com probabilidades NB dos valores para cada atributo, que se modificam de acordo com a inferência feita, como explicado na subseção A. Essas informações serão utilizadas em todos os protótipos.

Para os cálculos referentes ao algoritmo NB, foi consumida a API disponibilizada pelo programa Weka[3], no qual, após a coleta do arquivo com os dados, foi utilizado a API para criar a RB e salvar em variáveis. Com essas variáveis foi possível criar gráficos através da biblioteca JavaFX[2].

Esta etapa do projeto foi resumida em 3 partes, chamadas de protótipo 1, 2 e 3, que serão apresentadas através de seus resultados obtidos, de acordo com erros e acertos.

a) Protótipo 1

Para o protótipo 1, foi elaborada uma *interface* que após o processo de seleção do conjunto de dados, escolha da variável alvo e seu valor, será mostrado as correlações, como visto na Fig. 3, através de gráficos de barras verticais, separando os atributos por cor.

As informações estão organizadas em 4 grupos de informações, divididos em linhas diferentes, que são:

i. **Dados estatísticos**, informações estatísticas retiradas do conjunto de dados de entrada;
ii. **Inferência NB**, probabilidades retiradas da RB criada;
iii. **Relevância dos Valores**, diferença entre os valores de cada atributo para estatísticas e probabilidades;
iv. **Relevância Geral**, soma absoluta da diferença de cada valor do atributo.

Porém, com o uso pratico, essa organização das informações apresentou algumas dificuldades para o usuário. Sendo estas, letras pequenas, excesso de informações nos gráficos e dificuldade de salvar os resultados gerados. Além de apresentar vários erros e problemas, um dos principais ocorria quando no conjunto de dados inserido possuía muitas variáveis.

b) Protótipo 2

Para o protótipo 2, apresentando na Fig. 4, foi adotada uma organização onde os grupos de informações sejam divididos em colunas, ao invés de linhas, e cada atributo separado em linhas, sendo dessa forma, organizados segundo seu grau de relevância, de ordem crescente. Dessa forma, resolvendo o problema de escalabilidade existente no protótipo anterior.

Com a nova organização foi possível separar as informações em diversos gráficos, desta forma, foram mais aceitas entre pesquisadores que fizeram os testes iniciais, sendo considerados os gráficos mais legíveis e úteis. Além disso, foi disponibilizado para o usuário a possibilidade de capturar qualquer parte dos resultados.

Porém, ainda apresentando pequenos erros e uma interface pouco amigável e a atraente, deu-se a necessidade da criação de um novo protótipo.

[4] https://www.bayesfusion.com

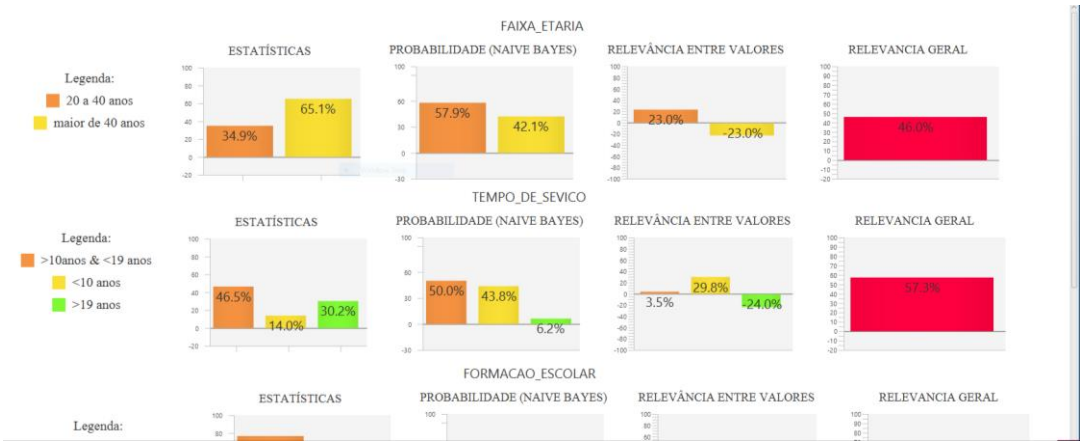

Fig. 8 – Lista de Atributos

a) Protótipo 3

Para o protótipo final, foram realizadas algumas alterações, de modo que permaneça a representação dos atributos em lista e as informações divididas em colunas. Com o foco na correção dos erros e na criação de uma interface mais atraente, como visto na Fig 5. Além disso, nesta etapa foi criada novas funcionalidades para a aplicação, por exemplo, a possibilidade de salvar os gráficos e as linhas de gráficos no formato .jpeg, e a lista completa nos formatos .docx e .pdf.

III. RESULTADOS

Esta etapa será apresentada através do trabalho do SILVA, Mateus Borges, que analisar o efeito de uma intervenção educacional sobre o Conhecimento, Atitude e Prática (CAP) de profissionais de saúde do município de Curuçá em relação à doença de raiva, no qual, foi oferecida uma capacitação à equipe de 105 profissionais de saúde do município, na qual foram abordados aspectos sobre a o agente etiológico, formas de transmissão do vírus e medidas profiláticas através de uma palestra e oficina.

A análise deu-se através de dois questionários, um antes e após a capacitação, esses que foram classificados como "aumentou" ou "constante". Com essas informações foi gerado um conjunto de dados com o perfil dos participantes junto com a sua classificação respectiva. Sendo as variáveis:

- faixa_etaria;
- formacao_escolar;
- tempo_de_sevico;
- ultima_capacitacao;
- n_conhecimento;

No qual, as variáveis de entrada são seus dados pessoais e a variável alvo é o *n_conhecimento*, que é classificação atribuída de seu resultado do questionário aplicado. Uma parte deste conjunto de dados está sendo apresentada na tabela 1, apenas 5 instancias (ou exemplos) das 105 instancias existentes.

Para entender a correlação entre as variáveis de entrada e com a variável alvo, com o intuito de analisar os efeitos da intervenção a partir da classificação atribuída às respostas dos participantes e de seus perfis respectivos, foi utilizado a aplicação proposta por este artigo, o BayesGraphics, no qual será apresentado o processo para a extração de informações sobre a correlação dessas as variáveis e em paralelo apresentar os resultados obtido por este artigo.

Na inicialização do programa, apresentada na Fig. 6, é necessário a inserção do arquivo do conjunto de dados, escolha do atributo alvo e seu valor. Neste trabalho foi utilizado para "n_conhecimento" como variável alvo e com seu valor "constante" para ser inferido.

Fig. 6 – Tela de Início.

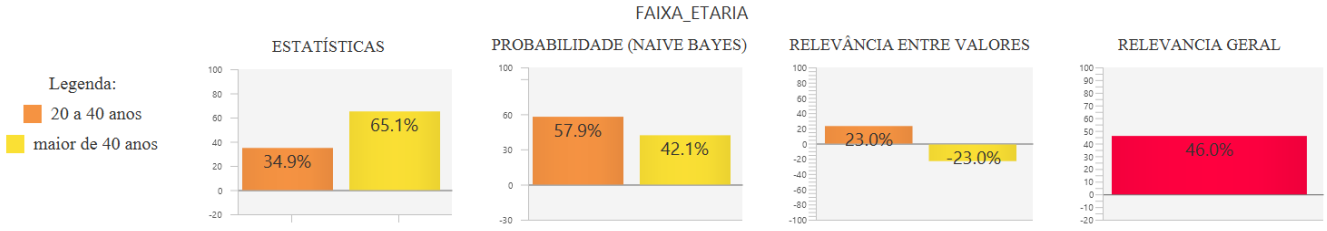

Fig. 10 – Linha / Atributo.

n_conhecimento	faixa_etaria	formacao_escolar	tempo_de_sevico	ultima_capacitacao
constante	"maior de 40 anos"	"ensino médio completo"	">10anos & <19 anos"	"há mais de 5 anos"
aumentou	"maior de 40 anos"	"ensino médio completo"	">19 anos"	"1 ano há 2 anos"
aumentou	"maior de 40 anos"	"ensino médio completo"	">19 anos"	" há 3 a 4 anos"
aumentou	"maior de 40 anos"	"ensino médio completo"	">10anos & <19 anos"	" há 3 a 4 anos"
constante	"maior de 40 anos"	"ensino médio completo"	">10anos & <19 anos"	" há 3 a 4 anos"

Tabela 1 – Conjunto de dados

Como mostrado Fig. 5, após inserir o conjunto de dados, atributo e estado alvo, o programa abre uma tela com os resultados gerados. Sendo destacado na parte esquerda, Fig. 7, as informações sobre a RB montada.

Na parte central da aplicação, Fig. 8, apresenta uma lista vertical com as variáveis, ordenada segundo o seu grau de relevância para a variável alvo. Neste caso estudado, nota-se que *tempo_de_sevico* e *faixa_etaria* estão no topo da lista. Sendo então considerados as variáveis mais relevantes para a inferência *constante* (em *n_conhecimento)*.

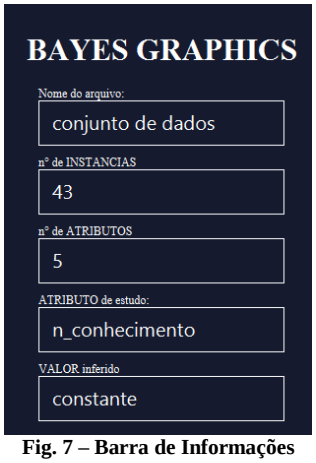

Fig. 7 – Barra de Informações

Cada variável possui uma lista horizontal de gráficos, Fig. 10, representando 4 grandes grupos de informações, gráfico de estatísticas, gráfico de probabilidade NB, relevância entre valores e a relevância geral, além da legenda. Para o caso estudado, ao se analisa a lista da variável *faixa_etaria* nota-se que seus estados, "20 a 40 anos" e "maior 40 anos ", estão sendo mostrados na parte da legenda, assim como os restantes das informações através de gráficos.

Com a intenção de facilitar na produção de materiais acadêmicos o programa deixa a disposição, capturar todas as partes da tela de resultados, Fig. 5, que o especialista desejar. Podendo ser somente um gráfico único, Fig. 9, uma linha de gráficos de um atributo, Fig. 10, ou todos as linhas com seus atributos como na Fig. 11, e além de disponibilizar a legenda de maneira separada para se adequar em qualquer forma de representação, como mostrado na Fig. 12.

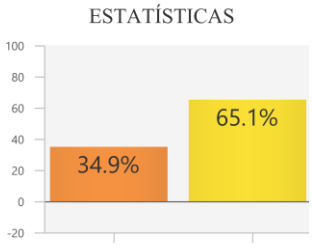

Fig. 9 – Gráfico Único.

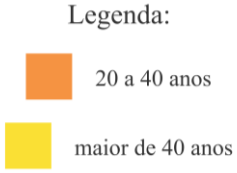

Fig. 12 – Legenda Geral

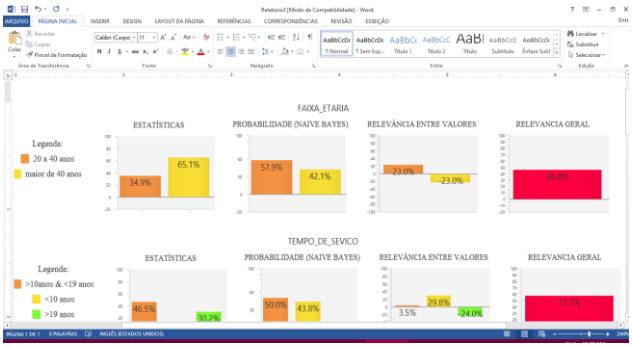

Fig. 11 – Documento com a Lista dos Atributos.

IV. Conclusão

Neste trabalho, foi proposto uma ferramenta para análise de correlação entre variáveis, utilizando o algoritmo NB. Como resultado, teve o desenvolvimento de uma aplicação feita em Java[1], consumindo a biblioteca JavaFX[2] e a API disponibilizada pelo software de código aberto WEKA[3], para a criação de uma interface que cria uma lista de forma ordenada, do maior para o menor grau de relevância, além de simplificar o processo do ML. Além do que, a aplicação criada foi validada em dois trabalhos acadêmicos, que somaram para o desenvolvimento, através da avaliação dos protótipos. Desta forma, possibilitaram adequar às expectativas e alinhar aos objetivos.

Referências

[1] Andrew Bruce, Peter Bruce. (2019) "Estatística Prática Para Cientistas de Dados: 50 Conceitos Essenciais" in Alta Books.

[2] SIEGEL, Sidney. Estatística não-paramétrica: para as ciências do comportamento. São Paulo: McGraw-Hill do Brasil, 1975. 350 p.

[3] OLIVEIRA, André Rodrigues. Comparação de algoritmos de aprendizagem de máquina para construção de modelos preditivos de diabetes não diagnosticado. Porto Alegre: UFRGS, Instituto de Informática, p. 15 – 70. 2016.

[4] Min Tan, Min Yu, Yongjian Wang, Song Li and Chao Liu, "Android malware detection combining feature

correlation and Bayes classification model - IEEE Conference Publication", Ieeexplore.ieee.org, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8230195.

[5] Y. Khourdifi and M. Bahaj, "Feature Selection with Fast Correlation-Based Filter for Breast Cancer Prediction and Classification Using Machine Learning Algorithms - IEEE Conference Publication", Ieeexplore.ieee.org, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8618688. [Accessed: 30- Aug- 2019].

[6] L. Alice Auxilia, "Accuracy Prediction Using Machine Learning Techniques for Indian Patient Liver Disease - IEEE Conference Publication", Ieeexplore.ieee.org, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8553682. [Accessed: 30- Aug- 2019].

[7] Bussab, Wilton de O.; Morettin, Pedro A. (2010). Estatística Básica 6ª ed. [S.l.]: Saraiva. p. 73. 540 páginas

[8] Michał Kunicki, "Correlation analysis of hotspot temperatures in power transformer – a case study", Ieeexplore.ieee.org, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8778179.

[9] Mitchell, T. (2017). “Machine learning” in McGraw-Hill Science/Engineering/Math, p.1-19.

[10] Katti Faceli, Ana Carolina Lorena, João Gama, André C. P. L. F. de Carvalho (2011) “Inteligência Artificial. Uma Abordagem de Aprendizado de Máquina” Editora: LTC p.70-82

[11] NETO, Joaquim. Inferência Bayesiana. Juiz de Fora: Departamento de Estatística – ICE, p. 22. Disponível em: < http://www.ufjf.br/joaquim_neto/files/2011/11/MBP-Inferência-Bayesiana.pdf >. Acesso em: 5/10/2019.

[12] Shameem Fathima, Nisar Hundewale, " Comparison of classification techniques-SVM and naives bayes to predict the Arboviral disease-Dengue", Ieeexplore.ieee.org, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/6112426.

.