Relatório Técnico (50%)

Pad: Padrão (o trabalho contém todos os itens requeridos): 10,0
Ling: Linguagem (o texto está bem escrito, correto gramaticalmente e as ideias bem expressas): 8,0
Mod:Modelo (o modelo está bem descrito e sua instanciação feita adequadamente): 10.0
Res: Resultados (os resultados foram devidamente descritos, inclusive utilizando gráficos): 10.0
Aval: Avaliação (os resultados foram adequadamente apresentados): 9,0

Implementação (50%)

Com: Comentários: (o código está devidamente documentado): 10,0
Cod: Codificação: (o código está bem escrito e roda adequadamente): 10,0
Result: Resultados: (os resultados foram produzidos adequadamente, inclusive utilizando gráficos): 10,0

Nota: 9,6

Relatório Técnico
sobre Naive Bayes

Thiago M. de Sousa *Luana G. B. Martins*
Ruan C. Rodrigues

Technical Report - RT-INF_000-19 - Relatório Técnico
July - 2019 - Julho

Instituto de Informática
Universidade Federal de Goiás
www.inf.ufg.br

Relatório Técnico sobre Naive Bayes

Thiago M. de Sousa
thiagomontelesofc@gmail.com

Luana G. B. Martins
luanagbmartins@gmail.com

Ruan C. Rodrigues
ruanchaves93@gmail.com

Abstract. *This report describes what were the decisions, solving process and results of the proposed problem in the field of Artificial Intelligence. The challenge of inducing a Naive Bayes that correlates with performance in the first two periods of the course with the final performance was introduced and allows to make predictions about the final performance of new students.*

Keywords: Technical Report, Naive Bayes.

Resumo. *Este relatório descreve quais foram as decisões, processo de resolução e resultados do problema proposto na matéria de Inteligência Artificial. Foi introduzido o desafio de induzir uma Naive Bayes que tenha uma correlação com o desempenho nos dois primeiros períodos do curso com o desempenho final e permite fazer previsões sobre o desempenho final de novos alunos.*

Palavras-Chave: Relatório Técnico, Naive Bayes.

1 Introdução

Este Relatório Técnico consiste na documentação de uma estratégia tomada para se obter um processo de tomada de decisões, buscando resolver o problema de se criar um modelo para obter previsões de desempenho final de novos alunos no Bacharelado em Ciência da Computação da Universidade Federal de Goiás, tendo em mãos apenas o histórico realizado nas matérias pertinentes aos dois primeiros períodos do curso.

O processo de decisão se deu por meio da utilização de uma Naive Bayes, as quais por sua vez tiram proveito de uma estrutura de dados com informações probabilísticos que levam o processo à sua conclusão através de uma classificação .

O problema consiste, em um primeiro momento, utilizar ~~de~~ dados de alunos que já concluíram o curso para definir parâmetros como período de ingresso, ano de conclusão e notas dos dois primeiros semestres para, em um segundo momento, criar um modelo que possa ser utilizado em uma Rede Bayesiana, com o objetivo final de gerar previsões de desempenho de novos estudantes do curso com base em seu desempenho nos dois primeiros períodos.

No restante deste documento estão definidas a forma abordada na base de dados (Seção 2), descrição geral da solução proposta contendo nela a descrição geral do modelo utilizado e dos dados selecionados para o modelo (Seção 3), dos resultados obtidos (Seção 4), das propostas para como utilizar os resultados obtidos (Seção 5), conclusões finais (Seção 6) e referências.

2 Descrição da base de dados

Os dados consistem em um arquivo no formato csv (Comma-separated values) que é representado por uma matriz de 22361 linhas por 66 colunas, onde existe em cada coluna um determinado atributo referente à relação de um aluno com as disciplinas que cursou durante os anos e períodos.

	id	ano_nascimento_discente	idade_conclusao_ensino_medio	idade_ingresso_universidade	idade_colacao_grau	uf_naturalidade_discente	c
0	1	1989	17.0	19	26.0	GO	
1	1	1989	17.0	19	26.0	GO	
2	1	1989	17.0	19	26.0	GO	
3	1	1989	17.0	19	26.0	GO	
4	1	1989	17.0	19	26.0	GO	

Os atributos nas colunas contém dados referente ao aluno e sua passagem no curso. Dados como ano de nascimento, idade de ingresso à universidade e nota do Enem são exemplos de dados relacionados ao aluno anteriormente ao ingresso na faculdade. Já atributos como quantidade de reprovações, média global e ano de conclusão estão ligados ao aluno após a entrada na universidade.

Devido às restrições da definição de aluno de bom desempenho fornecidas pelo problema, alunos que ingressaram após o primeiro semestre de 2015 foram desconsiderados na etapa de treinamento.

3 Descrição da solução

Com o objetivo de permitir fazer previsões sobre o desempenho final de novos alunos, foram considerados alguns aspectos como a criação de uma classificação de bom aluno e o treinamento de um modelo utilizando disciplinas dos dois primeiros períodos do curso. Para a arquitetura, foi proposto o uso de modelos Naive Bayes.

3.1 Descrição do modelo utilizado

Nessa seção será discutido os princípios do funcionamento de uma Rede Bayesiana baseado no Teorema de Bayes e como foi aplicado como solução do problema proposto.

3.1.1 Teorema de Bayes

se

Em estudos de teoria das probabilidades e estatística, o teorema de Bayes (também conhecido como regra de Bayes) foi criado pelo matemático inglês Thomas Bayes (1701 - 1761), dando a possibilidade de criar uma descrição da probabilidade de um determinado evento, inicialmente se baseando no conhecimento base relacionado a um evento, sendo esse conhecimento chamado de a priori. O teorema mostra como é possível criar alterações na probabilidade a priori obtendo novas evidências para obter uma nova probabilidade, que é comumente conhecida como probabilidade a posteriori.[2]

O Teorema de Bayes consiste em:

Por quê negativo?

$$P(\mathbf{Y}|\mathbf{X}) = \frac{P(\mathbf{X}|\mathbf{Y})P(\mathbf{Y})}{P(\mathbf{X})} \quad (-1)$$

- Onde **X** e **Y** são eventos
- P(**X**) e P(**Y**) são probabilidades a priori de **X** e **Y**.
- P(**X/Y**) é a probabilidade posteriori de **X** condicional a **Y**.
- P(**Y/X**) é a probabilidade a posteriori de **Y** condicional a **X**.

Para um conjunto de dados a expressão pode ser ampliada para $X = (X_1, X_2, X_3, ..., X_n)$, sendo Y uma variável de classe e X um vetor de recurso dependentes de tamanho N.

Quando queremos uma independência entre os recursos podemos então dividir as evidências entre os recursos. Assim obtendo:

$$P(\mathbf{Y}|\mathbf{x}_1, x_2, x_3, ..., x_n) = \frac{P(\mathbf{x}_1|\mathbf{Y})P(\mathbf{x}_2|\mathbf{Y})P(\mathbf{x}_3|\mathbf{Y})....P(\mathbf{x}_n|\mathbf{Y})P(\mathbf{Y})}{P(x_1,x_2,x_3,...,x_n)} \quad (-2)$$

3.1.2 Algoritmo Naive Bayes

O algoritmo Naive Bayes é um classificador probabilístico utilizado em aprendizagem de máquinas que foi baseado no Teorema de Bayes. O algoritmo assume que a presença de uma característica particular em uma classe não está relacionado com a presença de outro recurso. Por exemplo, uma determinada fruta pode ser classificado como uma maçã se ela é vermelha, redonda, e tiver cerca de 7 centímetros de diâmetro. Mesmo que esses atributos dependem uns dos outros ou até mesmo da existência de outras atributos não selecionados, todas essas propriedades contribuem de forma independente para que a probabilidade dessa fruta seja relacionado a uma maçã, por isso que é conhecido como Naive(Ingênuo).[3]

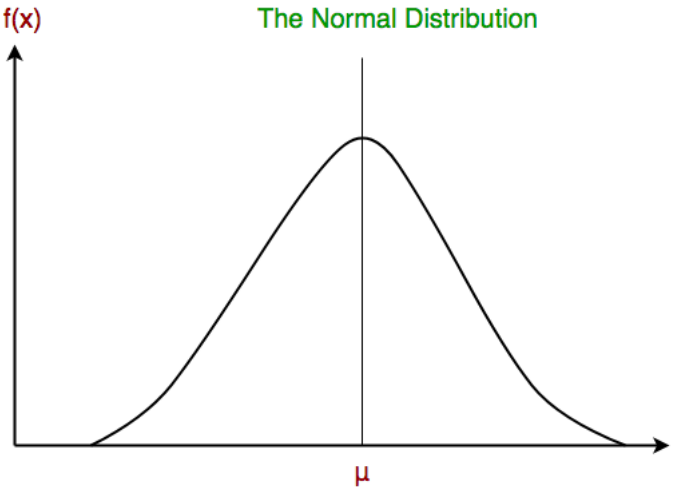
Utilizando-se do Teorema de Bayes, que nos permite calcular a probabilidade da ocorrência de uma classe c dado um conjunto de observações x_n, o algoritmo Naive Bayes visa escolher o valor da classe para o qual $P(x_1, x_2, ..., x_n|c)$ é o mais alto, ou seja:

$$Classifier(x_1, x_2, ..., x_n) = max_{C\in 0,1} P(x_1, x_2, ..., x_n|c) \quad (-3)$$

3.2 Abordagens para Naive Bayes

As suposições sobre distribuições de características são chamadas de modelo de evento do classificador Naive Bayes. Existem três tipos de modelo Naive Bayes para características discretas e contínuas.

- Gaussian Naive Bayes: Para valores contínuos associadas a cada atributo do evento são considerados em uma distribuição gaussiana, também conhecida como distribuição normal.[1]



Ao fazer o Plot dos atributos é formado uma curva em forma de sino que é simétrica em relação a média dos valores do recurso, conforme mostrado acima. A Probabilidade das características é assumida como Gaussiana.

- Multinomial Naive Bayes: O vetor de características representa as frequências com que certos eventos foram realizados por uma distribuição multinomial. Esse modelo é geralmente utilizado para classificação de documentos, com eventos sendo representados pela ocorrência de uma palavra no documento corrente.[2]

- Bernoulli Naive Bayes: No modelo Bernoulli, os recursos são booleanos e independentes que descrevem entradas, Assim como no modelo Multinomial, essa abordagem é popular em tarefas de classificação de documentos, onde as ocorrências de determinados termos são expressadas de forma binária, ou seja se o termo foi ou não citado no documento.[2]

3.3 Definição final do modelo

Afim de obter um modelo que seja capaz de trabalhar com o problema proposto, foi utilizado, inicialmente, dois classificadores, cada classificador para um conjunto de dados, de acordo com suas características. Para o conjunto de dados contínuo será utilizado o classificador Gaussian Naive Bayes, e para o conjunto de dados discreto foi utilizado o classificador Bernoulli Naive Bayes.

Para o classificador Gaussian Naive Bayes, como os dados possuem uma distribuição normal, sua fórmula da probabilidade condicional é dada por:

$$P(x_i|C = c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} exp\left(-\frac{(x_i - \mu_c)^2}{2\sigma_c^2}\right) \quad (-4)$$

Onde os parâmetros σ_c e μ_c são estimados durante o treinamento do modelo, e cada um representa, respectivamente, a variância do valor de x calculado para uma dada classe c.

No classificador de Bernoulli Naive Bayes, onde os atributos assumem valores binários, sua fórmula da probabilidade condicional é dada por:

$$P(x_i|C = c) = P(i|c)x_i + (1 - P(i|c))(1 - x_i) \quad (-5)$$

Onde, dado um conjunto de dados, cada atributo i corresponde aos valores x_i. A dimensão i de um vetor do conjunto de dados irá indicar se o aluno correspondente a este vetor passou ou não na matéria i em sua primeira tentativa. A probabilidade de cada x_i é independente da ocorrência de outros x_i em uma observação. E, a probabilidade de uma observação receber sua classe, é o produto da probabilidade dos valores de atributo sobre todos os atributos. A fórmula de Bernoulli Naive Bayes penaliza a não ocorrência de uma característica i que representa um indicador para a classe c.

Uma vez realizado as predições de cada um dos classificadores mencionados acima, essas predições serão utilizadas para avaliar a classificação final dos dados de observação utilizando por fim o classificador Bernoulli Naive Bayes. Cada um dos conjuntos de predições irá corresponder a um atributo para essa nova classificação.

3.4 Descrição dos dados selecionados

De acordo com a definição do problema, aluno de bom desempenho é aquele que se forma em até 4 anos e meio, ou se forma com média igual ou superior a 7, ou se forma com número de reprovações igual inferior a 5.

Temos, portanto, a seguinte representação lógica para esta classificação:

$$A \vee (B \wedge C) \vee (B \wedge D) \tag{-6}$$

A: O aluno se forma em 4 anos e meio.
B: O aluno se formou.
C: O aluno tem média igual ou superior a 7.
D: O aluno tem número de reprovações igual ou inferior a 5.

Seja "bom aluno"um termo para o aluno que possui bom desempenho. Considere um aluno qualquer a e uma matéria m. A este par (a, m) iremos atribuir as variáveis $X_{a,m}$ e $Y_{a,m}$.

Seja $T_{m,n}$ a quantidade de bons alunos que passaram na matéria m na n-ésima tentativa, e U_m a quantidade total de bons alunos que já cursaram a matéria m. Então:

$$P_{m,n} = \frac{T_{m,n}}{U_m} \tag{-7}$$

Seja $A_{a,m} = \{t_1, t_2, ..., t_n\}$ o conjunto de tentativas que o aluno a realizou para passar na matéria m.

Seja $M_{m,n}$ a média geral dos bons alunos que cursaram a matéria m por n vezes.

Seja N_{AP} a nota necessária para aprovação. No nosso dataset, essa nota é universalmente seis (6.0) em todas as matérias.

Seja $N_{a,n}$ a nota do aluno a na n-ésima tentativa de passar na matéria.

Seja $g_{a,m}$ uma função booleana que retorna 1 caso o aluno a tenha sido aprovado na matéria m, e 0 caso ele tenha reprovado na matéria m mesmo após todas as suas tentativas.

$$g_{a,m} = sgn(sgn(N_{a,|A_{a,m}|}) - N_{AP}) + 1) \tag{-8}$$

Então:

$$X_{a,m} = P_{m,|A_{a,m}|} * g_{a,m} \tag{-9}$$

Para o modelo Bernoulli Naive Bayes aqui apresentado, consideramos os valores binarizados de $X_{a,m}$:

$$b(X_{a,m}) = \lfloor X_{a,m} + 0.5 \rfloor \tag{-10}$$

E como $Y_{a,m}$, consideramos:

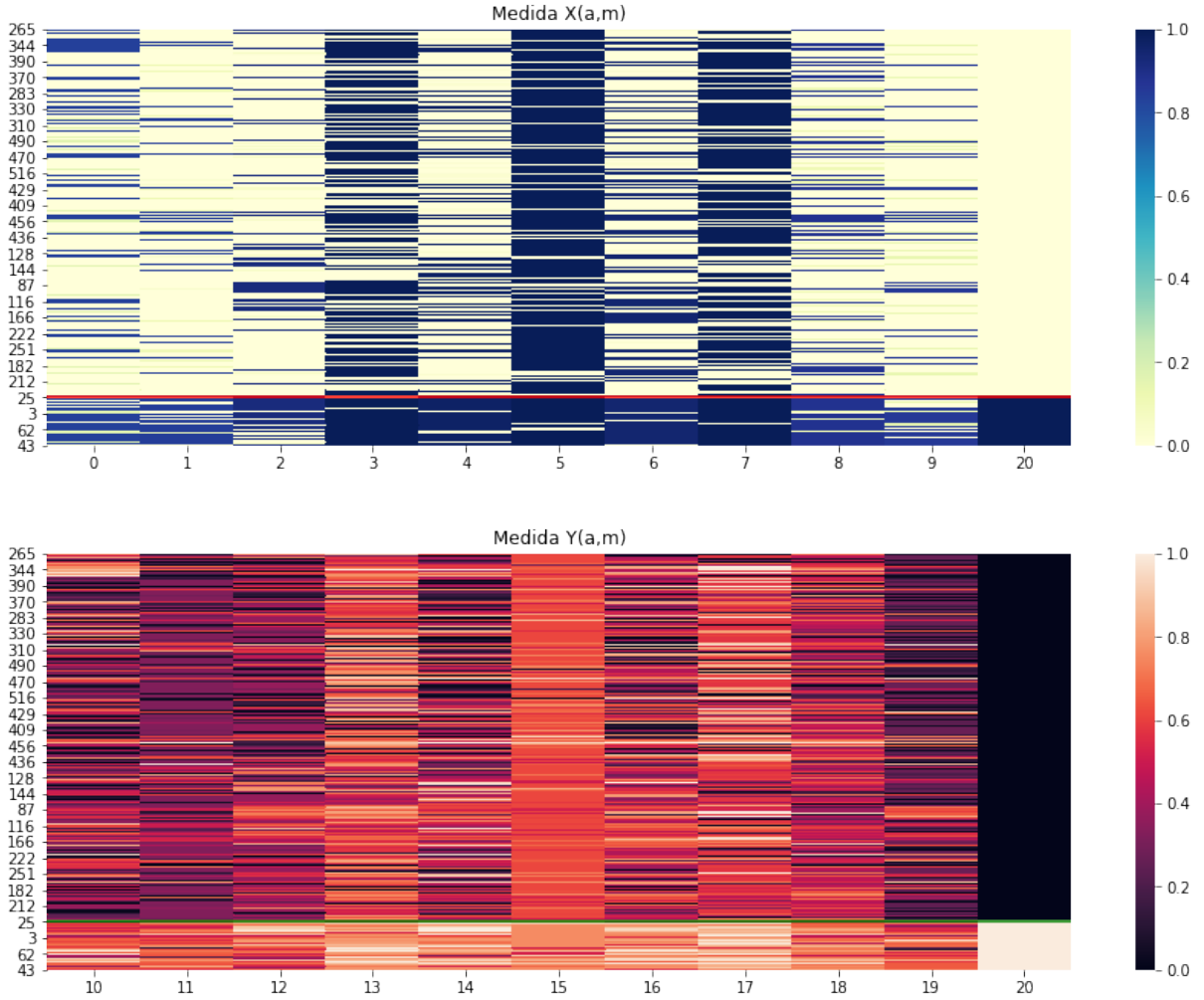
$$Y_{a,m} = \frac{\sum_{i=1}^{|A_{a,m}|} P_{m,i} * D_{a,m,i}}{\sum_{i=1}^{|A_{a,m}|} P_{m,i}} \tag{-11}$$

Sendo $D_{a,m,n}$ o desvio normalizado:

$$D_{a,m,n} = \frac{(N_{a,n} - M_{m,n}) + 10}{20} \tag{-12}$$

Explicar detalhadamente o significado destas figuras....

Após a inspeção visual dos dados através de *heatmaps*, percebemos que nosso modelo temático para os dados de cada disciplina foram capazes de separar bem os alunos de bom e mau desempenho.



As colunas nomeadas de 0 a 9 representam o atributo $X_{a,m}$ de cada matéria dos dois primeiros períodos. Já as colunas de 10 a 19 representam o atributo $Y_{a,m}$ de cada uma destas matérias. Atributos referentes à mesma matéria estão separados por dez unidades: sendo assim, as colunas 0 e 10, 1 e 11, 2 e 12, 3 e 13 e 5 e 15 são referentes à mesma matéria.

Somente a disciplina nas colunas 2 e 12 (Matemática Discreta) foi desconsiderada do *dataset*, por não estar mais sendo ofertada aos alunos que ingressaram no curso de Ciência da Computação após o primeiro semestre de 2017, em sua forma original ou em alguma forma equivalente.

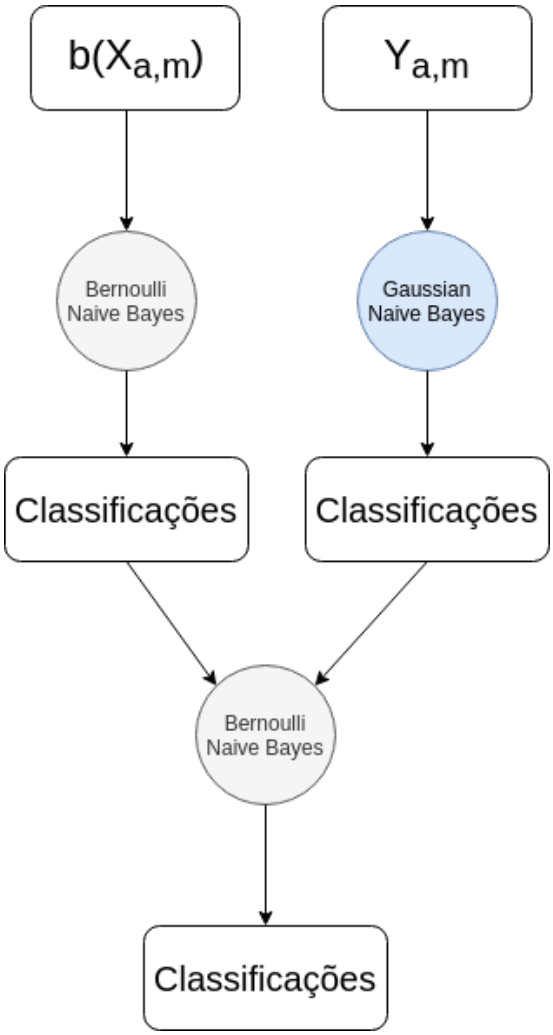
A coluna nomeada 20 é a classificação final, que denota se o aluno pode ser classificado como aluno de bom desempenho (valor 1) ou não (valor 0).

4 Resultados obtidos

Nessa seção serão apresentados os resultados obtidos após implementação do modelo e o seu treinamento com os dados obtidos na fase de tratamento de dados.

4.1 Modelo gerado

No final do processo foi gerada uma representação visual do modelo construído.



4.2 Justificativa

O modelo Naive Bayes se mostra adequado pois as variáveis consideradas são independentes entre si. Os valores de $b(X_{a,m})$ indicam se um aluno a passou na matéria m na sua primeira

tentativa, e os valores de $Y_{a,m}$ indicam em que medida a nota do aluno se desvia da média de notas dos bons alunos. Logo, por se tratarem de medidas relativas de desempenho, assumimos que não há interdependência entre as variáveis.

O desempenho de um aluno em uma determinada matéria está condicionado a fatores outros do que simplesmente o seu desempenho nas matérias restantes, mesmo que tais matérias estejam interrelacionadas conceitualmente de alguma forma.

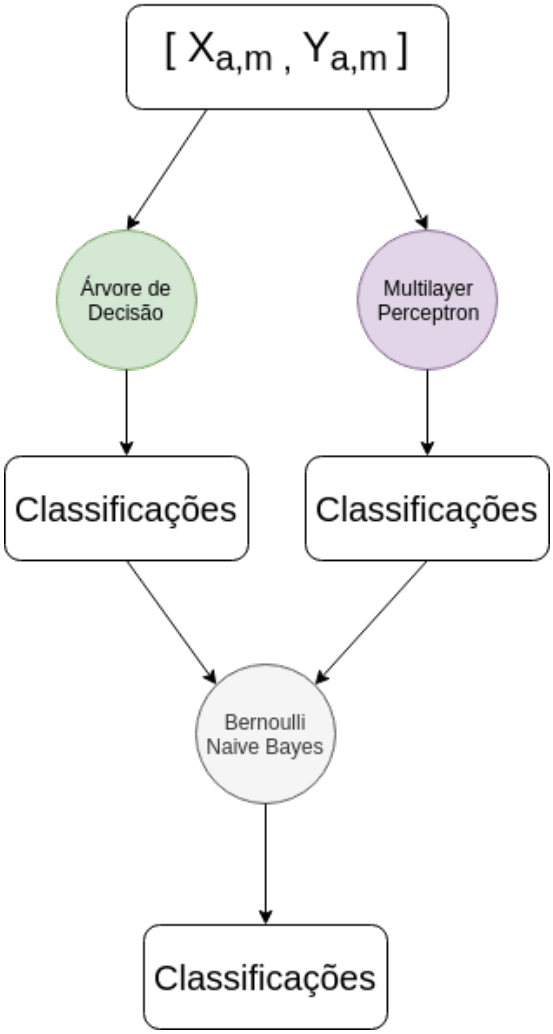
$b(X_{a,m})$ é uma medida de desempenho correlacionada a um valor fixo N_{AP}. Portanto, ela pode variar conforme os professores da matéria m sejam mais ou menos rigorosos, ou mais ou menos generosos em dar boas notas aos alunos; algo que, portanto, não pode ser deduzido como uma correlação entre matérias.

$Y_{a,m}$ é uma medida de desempenho correlacionada a um valor relativo determinado pelos bons alunos que cursaram a matéria m. Porém, a população de bons alunos que cursaram uma matéria m_1 não é a mesma população que cursou uma matéria m_2. Sendo assim, por se tratarem de populações distintas, não é possível extrapolar uma correlação entre os valores Y_{a,m_1} e Y_{a,m_2} para duas matérias quaisquer.

Dado que os valores de $Y_{a,m}$ para cada matéria seguem uma distribuição normal, um modelo gaussiano se mostrou adequado. E dado que existia uma grande disparidade entre os possíveis valores de $X_{a,m}$, percebeu-se que $b(X_{a,m})$ seria uma boa aproximação, a qual poderia ser tratada por um modelo que adota a distribuição de Bernoulli.

4.3 Acurácia

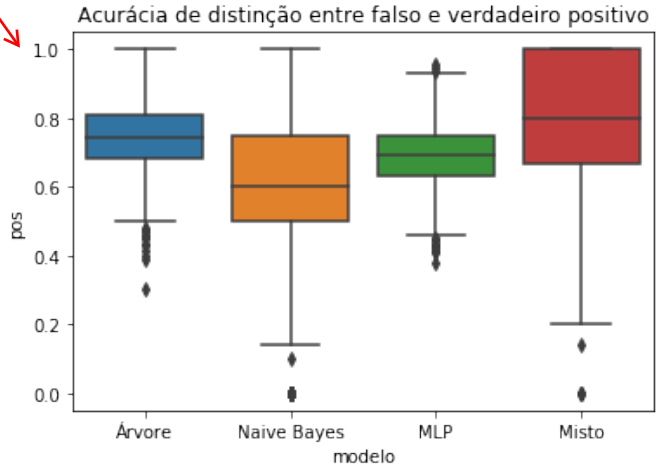
Além do modelo apresentado neste relatório, comparamos o modelo obtido com os modelos de relatórios anteriores, e também com o desempenho de um modelo misto, apresentado na figura a seguir:

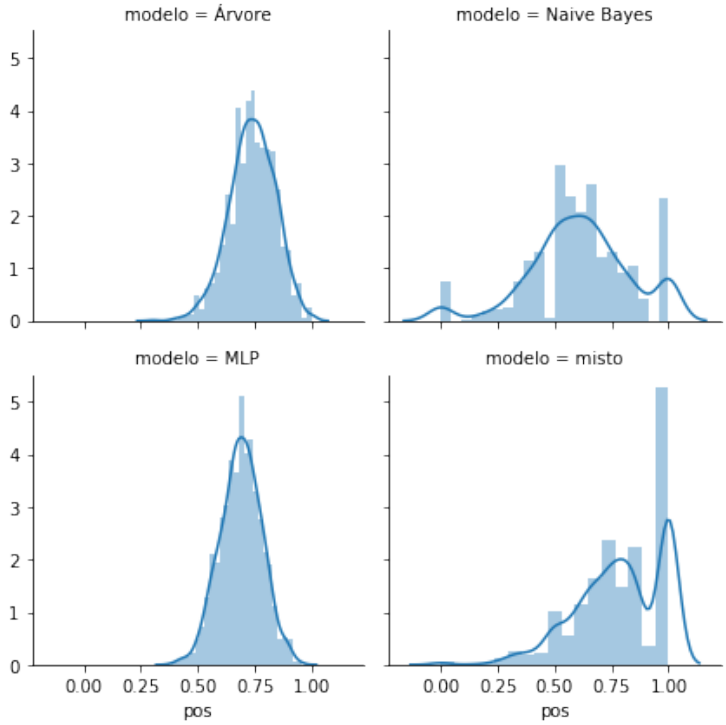


Neste modelo, a árvore de decisão e a rede neural MLP são treinados individualmente com as mesmas entradas de treinamento. Entretanto, a classificação emitida como saída por estes dois modelos não é tomada como a classificação final, mas como uma entrada que é fornecida ao Bernoulli Naive Bayes. Portanto, o Bernoulli Naive Bayes será treinado para prever, a partir das classificações emitidas pelos dois modelos anteriores, a classificação real de cada aluno.

As medidas de acurácia abaixo foram obtidas após 1400 testes para cada modelo.

Todas as figuras devem ser numeradas, rotuladas, referenciadas no texto e devidamente explicadas.





	acc		pos		neg	
modelo	*mean*	*std*	*mean*	*std*	*mean*	*std*
Naive Bayes	0.91	0.04	0.64	0.18	**0.96**	0.03
MLP	0.92	**0.02**	0.69	**0.09**	**0.96**	**0.02**
Árvore de Decisão	**0.93**	**0.02**	0.74	0.10	**0.96**	**0.02**
Misto	**0.93**	0.04	**0.79**	0.19	0.95	0.04

Onde estão esses gráficos???

- **Gráfico 1 (boxplot):** Representa a acurácia de cada modelo na distinção entre verdadeiro e falso positivo: Árvore de Decisão, o modelo Naive Bayes apresentado neste relatório, Multilayer Perceptron e o modelo misto apresentado no começo deste tópico.
- **Gráfico 2 (histograma):** Histogramas que correspondem às mesmas informações presentes no boxplot logo acima. Note que o modelo misto apresenta uma distribuição de

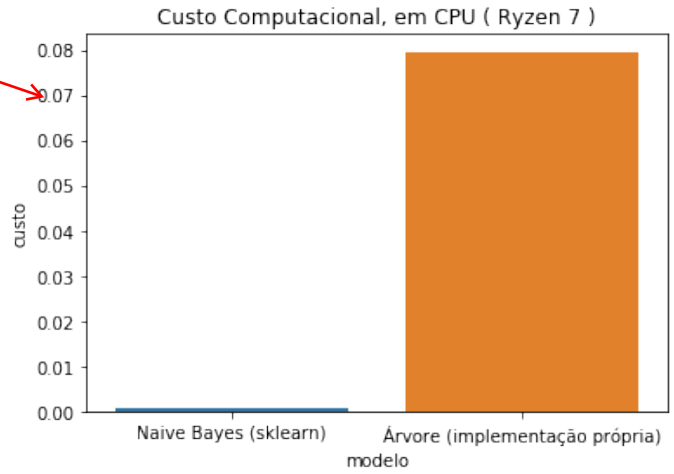
formato semelhante ao modelo Naive Bayes apresentado neste relatório, porém com uma maior tendência a valores mais altos de acurácia na distinção entre verdadeiro e falso positivo.

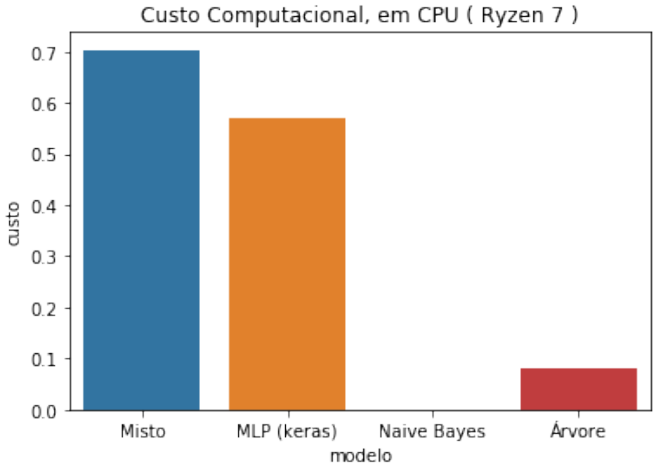
- **Tabela de acurácias: acc** indica a acurácia total; **pos** indica a acurácia de distinção entre verdadeiro e falso positivo (grandeza representada tanto no boxplot quanto no histograma); e **neg** representa a acurácia de distinção entre verdadeiro e falso negativo. Para cada medida de acurácia, é dado a média (*mean*) e o desvio-padrão (*std*) consideradas 1400 observações para cada modelo.

4.4 Custo computacional

O modelo apresentou um custo computacional bastante reduzido em comparação aos demais modelos já estudados.

O que significam esses valores?





O custo computacional foi medido como um fator diretamente proporcional a:

$$\frac{1}{t}$$

sendo t o tempo médio de uma execução do algoritmo utilizando todos os núcleos de um processador Ryzen 7.

uma especificação mais completa da máquina, senão não é possível ter uma idéia do custo.

5 Como a solução proposta pôde resolver o problema

Comparado aos modelos anteriormente examinados, o modelo Naive Bayes criado neste trabalho é capaz de distinguir entre verdadeiro e falso negativo com uma capacidade bastante semelhante aos demais modelos, embora exija um custo computacional muito menor do que todos os outros.

Além disso, o modelo misto gerado com o auxílio de Naive Bayes teve acurácia acima de todos os outros modelos individualmente na tarefa de distinguir entre falsos e verdadeiros positivos.

Sendo assim, os principais atrativos da solução proposta são o seu baixo custo computacional, o que possibilita grande escalabilidade em tarefas de detecção de desempenho de alunos em larga escala, com precisão satisfatória; a sua função de agir como *baseline* para modelos mais complexos e computacionalmente custosos para o mesmo problema; e a sua capacidade de facilitar a combinação de tais modelos para a elaboração de modelos mistos para a resolução do problema de distinguir entre alunos de bom e mau desempenho.

6 Conclusões

confuso!!!

Diante dos dados apurados, temos como conclusão que o algoritmo Naive Bayes apresenta um bom desempenho quando estamos obtendo o custo computacional como um peso a mais para decisão final da solução do problema proposto, também associado a isso ela apresenta uma boa acurácia quando utilizado com outros modelos, assim fazendo ela um auxiliar para outras soluções. Entretanto o Naive Bayes apresenta melhores desempenhos quando o problema apresenta uma classificação de texto ou algo do gênero, se a correlação entre os fatores forem extremamente importantes, o Naive Bayes pode apresentar falhas na predição de novos exemplos, com tudo isso posto, conclui-se que o modelo proposto obteve um bom resultado e correspondeu satisfatoriamente aos requisitos do problema que foi apresentado ao grupo.

Referências

[1] Gandhi, R. **Naive Bayes Classifier**. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c, último acesso em Agosto de 2019.

[2] Kumar, N. **Naive Bayes classifier**. https://www.geeksforgeeks.org/naive-bayes-classifiers,.

[3] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. **Scikit-learn: Machine learning in Python**. *Journal of Machine Learning Research*, 12:2825–2830, 2011.