Identificação de tipos de vidro utilizando Naïve Bayes

1st Lucas de Souza
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
lsb4@cin.ufpe.br

1st João Guilherme
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
jgmsf@cin.ufpe.br

1st Mateus Elias
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
meap@cin.ufpe.br

***Abstract*—Visando aprofundar os nossos conhecimentos na área de Estatística e Probabilidade, assim como na análise exploratória de dados, decidimos implementar um algoritmo classificador a fim de por em prática toda a teoria que estudamos. Escolhemos como temática do nosso modelo classificador a identificação de tipos de vidros.**

***Index Terms*—indentificação de vidro, classificador probabilístico, naive bayes, teorema de bayes, análise exploratória de dados, classificador ingênuo de bayes.**

I. Introdução

A. Objetivos

O projeto tem como objetivo aprofundar os conhecimentos na área de Probabilidade e Estatística, mais precisamente em análise exploratória de dados e no Teorema de Bayes, com a implementação de um algoritmo classificador baseado nesse teorema. A partir disso, foi estudado e analisado o *dataset*, que contém informações sobre os tipos de vidro, a fim de escrever um programa em *Python* para calcular a probabilidade de uma amostra vidro ser de determinado tipo, utilizando o Classificador Ingênuo de Bayes.

B. Justificativa

Visto que o Algoritmo de Naive Bayes é um classificador probabilístico que utiliza o Teorema de Bayes, sendo um modelo simples, com fácil implementação e com bom funcionamento na maioria dos casos, decidimos nos basear nele, juntamente com algumas bibliotecas matemáticas de Machine Learning, para desenvolvermos o modelo.

II. Metodologia

Nesta seção, será descrita a metodologia usada para o desenvolvimento do projeto, detalhando o dataset escolhido e seus atributos e introduzindo o classificador probabilístico que será utilizado.

A. Dataset

A base de dados escolhida para a análise foi a de identificação de tipos de vidro, com dados fornecidos pelo Serviço de Ciência Forense dos Estados Unidos da América. [4].

Os atributos presentes na base de dados são:

1) Classes: Buscamos classificar os dados entre as classes *Janelas Flutuantes Processadas de Construção*, *Janelas Não Flutuantes Processadas de Construção*, *Janelas Flutuantes Processadas de Veículos*, *Janelas Não Flutuantes Processadas de Veículos*, *Contêineres*, *Talheres*, *Faróis*.
2) ID: Um número de 1 a 214.
3) RI: Índice de refração.
4) Na: Sódio (Unidade de medida: porcentagem em peso no óxido correspondente, assim como os items 5-11).
5) Mg: Magnésio.
6) Al: Alumínio.
7) Si: Silício.
8) K: Potássio.
9) Ca: Cálcio.
10) Ba: Bário.
11) Fe: Ferro.

B. Classificador Probabilístico

Nesta seção, iremos apresentar o classificador Ingênuo de Bayes, que será utilizado no desenvolvimento do projeto, e o teorema de Bayes, que é a base para o classificador.

1) Teorema de Bayes: O teorema de Bayes recebe esse nome por ter sido criado pelo pastor e matemático inglês Thomas Bayes (1702-1761), ele foi o primeiro a fornecer uma equação que permitia que novas evidências atualizassem a probabilidade de um evento a partir do conhecimento a priori (ou a crença inicial na ocorrência de um evento). O manuscrito de Bayes só foi publicado após a morte de Thomas, sendo editado significativamente por Richard Price antes disso. E hoje é usado para o cálculo da probabilidade de um evento dado que outro evento já ocorreu, o que é chamado de probabilidade condicional.

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)} \tag{1}$$

A equação (1) é a forma mais simples do teorema. P(A|B) representa a probabilidade do evento A ocorrer dado que o evento B já foi observado, consequentemente P(B) precisa ser diferente de zero.

Podemos classificar as probabilidades contidas no teorema da seguinte maneira:

- Probabilidades marginais: A probabilidade de um evento independentemente do restante. Ex.: P(A) e P(B).
- Probabilidade conjunta: A probabilidade de dois ou mais eventos ocorrerem simultaneamente. Ex.: P(A e B).
- Probabilidade condicionada: A probabilidade de um ou mais eventos dada a ocorrência de outro evento. Ex.: P(A|B), P(B|A).

Então, o teorema de Bayes trata de probabilidades condicionais, visto que, em (1) temos a probabilidade de A condicionada pelo evento B. Nele, P(A) e P(B) são as chamadas probabilidades a priori e P(A|B) e P(B|A) são as probabilidades a posteriori.

Outra forma de visualizar o teorema de Bayes é como a probabilidade conjunta de A e B, que pode ser simbolizada como P(A ∩ B).Na Fig. 1, temos a representação visual da igualdade.

$$P(B) \cdot P(A|B) = P(A \cap B) = P(A) \cdot P(B|A) \quad (2)$$

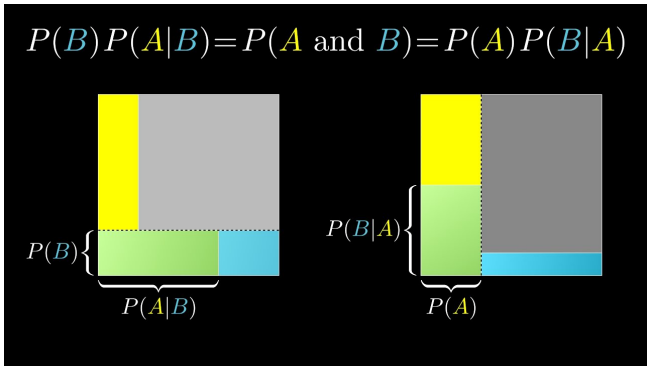


Fig. 1: Teorema de Bayes.

Podemos reescrever o teorema utilizando a igualdade em (2):

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

2) Classificador Ingênuo de Bayes: O classificador Ingênuo de Bayes ou Naïve Bayes é um popular classificador probabilístico usado frequentemente na área de aprendizagem de máquina (Machine Learning). Ele recebe o nome de ingênuo, pois desconsidera a correlação entre as variáveis, ou seja, trata cada uma de forma independente.

Uma das suas aplicações é a análise de texto de acordo com a frequência das palavras usadas, é comumente utilizado na classificação de e-mails como spam.

Por ser muito simples e rápido, possui um desempenho relativamente maior do que outros classificadores. Além disso, o Naive Bayes precisa de um pequeno número de dados de teste para concluir classificações com uma boa precisão.

No nosso modelo, existirão sete classes *Janelas Flutuantes Processadas de Construção*, *Janelas Não Flutuantes Processadas de Construção*, *Janelas Flutuantes Processadas de Veículos*, *Janelas Não Flutuantes Processadas de Veículos*, *Contêineres*, *Talheres*, *Faróis* e A será um vetor com as classes descritas na Seção II-A. Dessa forma, teremos:

$$P(C|A) = \frac{P(C)P(A|C)}{P(A)}$$

$$P(C|A) = \frac{P(C)\prod_{i=1}^{n} P(a_i|C)}{P(a_i, ..., a_n)}$$

Sendo C as classes *Janelas Flutuantes Processadas de Construção*, *Janelas Não Flutuantes Processadas de Construção*, *Janelas Flutuantes Processadas de Veículos*, *Janelas Não Flutuantes Processadas de Veículos*, *Contêineres*, *Talheres*, *Faróis* e a_i é o i-nésimo atributo do vetor A.

C. Aplicação

Para a criação do modelo, que será utilizado para a análise dos dados, será utilizada a linguagem Python no ambiente do Google Colaboratory.

A principal biblioteca a ser utilizada será a Pandas, que por sua vez é baseada em duas bibliotecas de Python: matplotlib e NumPy. Essa biblioteca é utilizada para a manipulação e análise de dados, utilizando matplotlib para a visualização gráfica e NumPy para as operações matemáticas. Outra biblioteca que pode ser usada é Scikit-Learn, uma biblioteca de Machine Learning que inclui vários algoritmos de classificação, regressão e agrupamento, foi projetada exatamente para interagir com bibliotecas numéricas e científicas como NumPy e SciPy.

III. Análise Exploratória de Dados

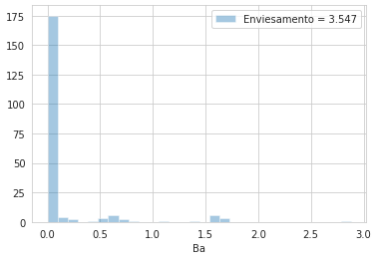
Após fazermos uma visualização inicial da nossa base de dados, nós visualizamos que a mesma consiste em 214 observações. Cada observação contendo 9 features relacionadas a uma determinada amostra de vidro e por fim, o respectivo tipo daquela amostra de vidro.

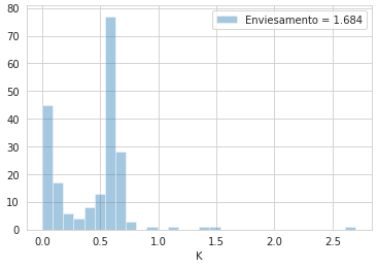
A. Estatística Descritiva

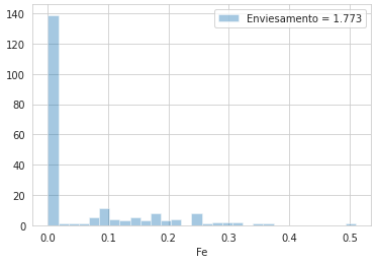
Observarmos que as features da nossa base de dados não estão na mesma escala. Por exemplo, Si tem a média de 72.65 enquanto Fe tem o valor médio de 0.057. É necessário que as features estejam na mesma escala para que algoritmos como o da regressão logística (método do gradiente) possam convergir suavemente. Olhando a distribuição dos tipos de vidro fica mais evidente ainda que a nossa base de dados é desbalanceada, pois as instâncias dos tipos 1 e 2 constituem mais de 67% dos tipos de vidro contidos em toda base de dados.

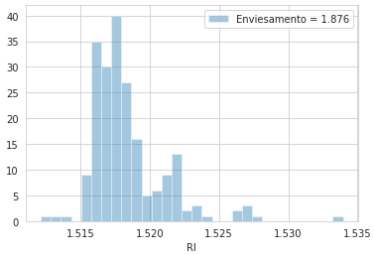
B. Visualização de Dados

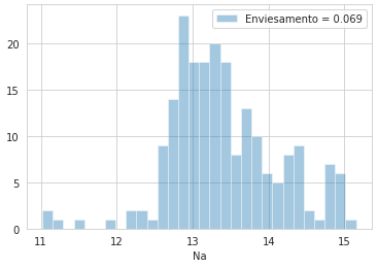
1) Gráficos Univariados: Nessa etapa iremos dar uma olhada na distribuição dos diferentes recursos desse conjunto de dados.

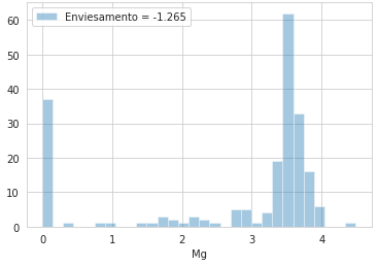


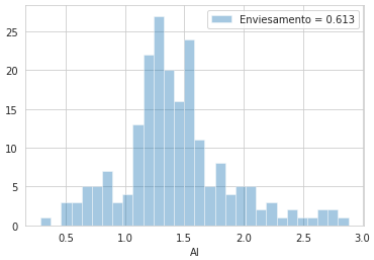


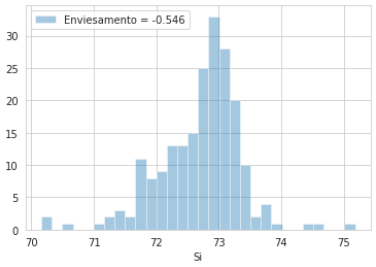


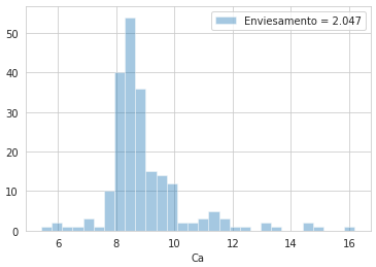






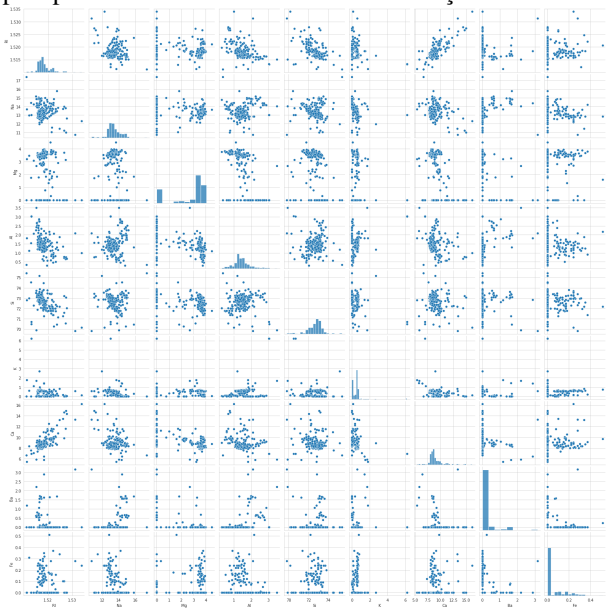






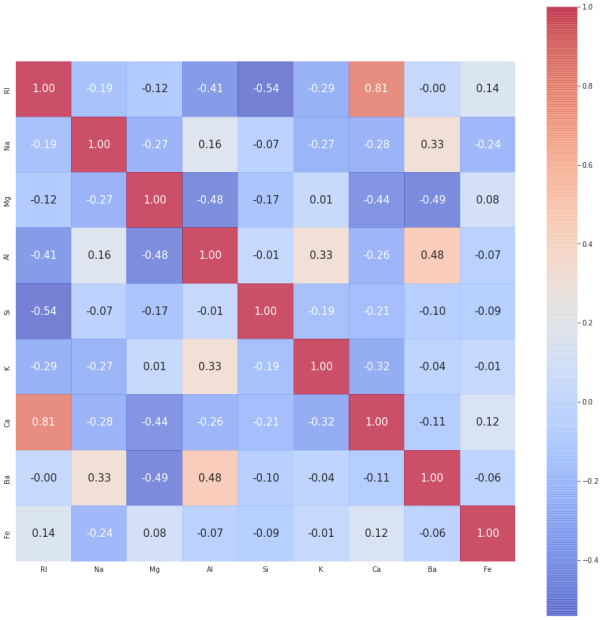
Como podemos ver, nenhuma das features é normalmente distribuída. As features Fe, Ba, Ca e K exibem os maiores coeficientes de assimetria. Além disso, a distribuição do Potássio (K) e do Bário (Ba) parece conter muitos outliers. Utilizando o método de Tukey, descobrimos que existem cerca de 14 observações com múltiplos outliers. Pelo fato disso poder prejudicar a eficiência de nossos algoritmos de aprendizagem iremos nos livrar deles nas próximas seções.

2) Gráficos Multivariados: Agora, vamos desenhar o pairplot para examinar visualmente a correlação entre as features.

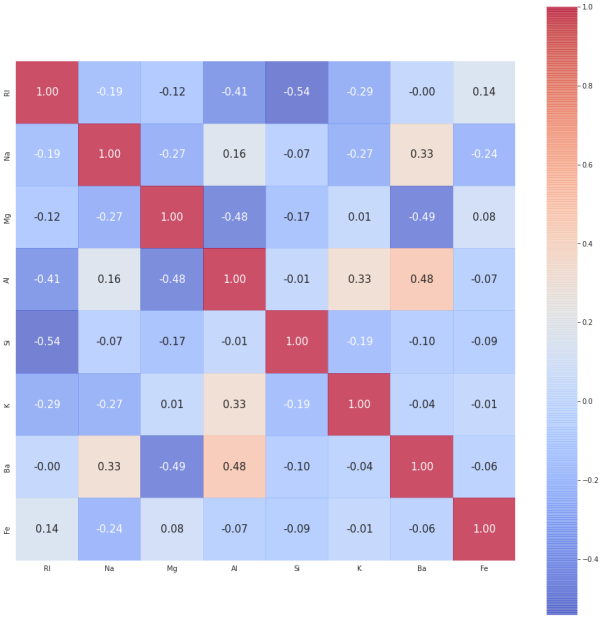


C. Tratamento das Correlações entre as Features

Como o classificador ingênuo de Bayes considera que as features são independentes entre si, é necessário observar se essa condição é respeitada entre as features da base utilizada. Por isso, iremos desenhar o heatmap das correlações, para analizá-las.



Olhando o heatmap, notamos que existe uma forte correlação positiva entre RI e Ca. Logo, para mantermos a condição de independência entre as features, iremos fazer a remoção de uma dessas duas features. A feature que escolhemos para remover foi o Ca.



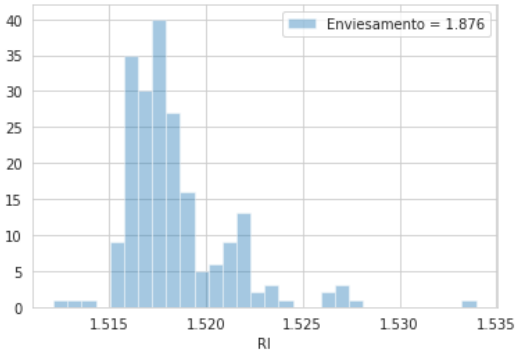
Após a remoção do Ca, podemos ver que o nosso novo heatmap cumpre com a relaçao de independência entre as features.

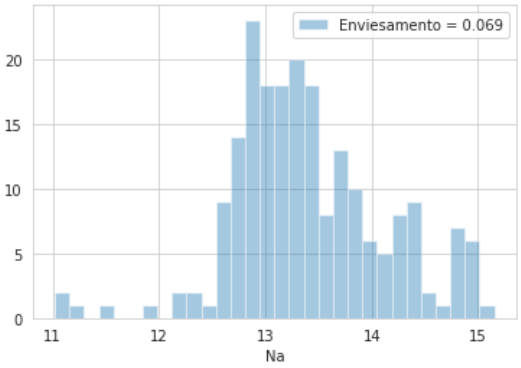
D. Limpeza de Dados

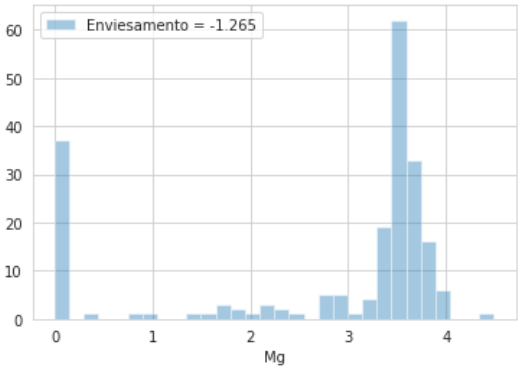
Ao checarmos a nossa base de dados, observamos que ela estava totalmente limpa, ou seja, não existem valores faltantes. Logo, não foi necessário fazer nenhuma limpeza de dados.

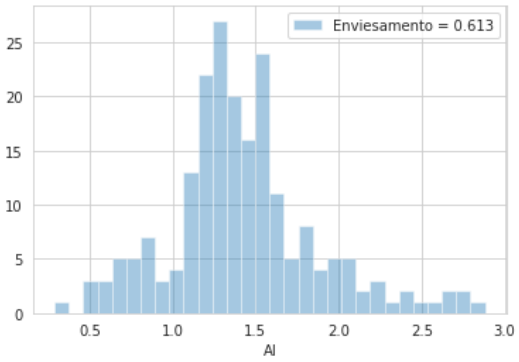
E. Localizando e Removendo os Outliers

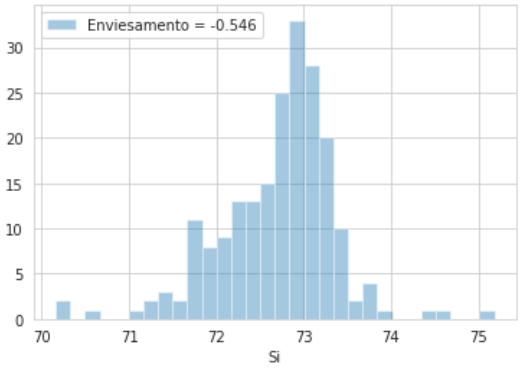
Por fim, como notamos na seção *Visualização de Dados* que existem muitos outliers, finalizamos fazendo a remoção deles. Com a remoção das observações com múltiplos outliers (mais de 2), nos restou um total de 206 observações para utilizarmos como base. Como fizemos a remoção da feature Ca, nosso trabalho foi facilitado, pois haviamos detectado anteriormente que existiam 14 observações na nossa base de dados com mais de 2 outilers, mas após a remoção dessa feature, nos restaram apenas 8 observações com múltiplos outliers.

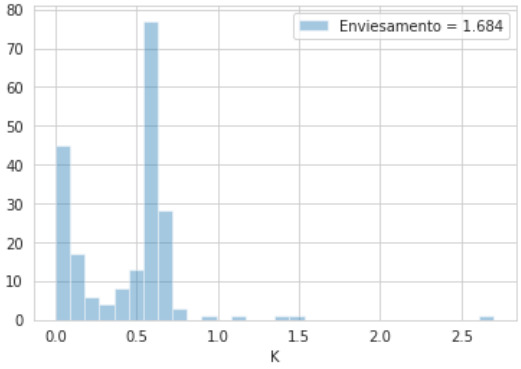


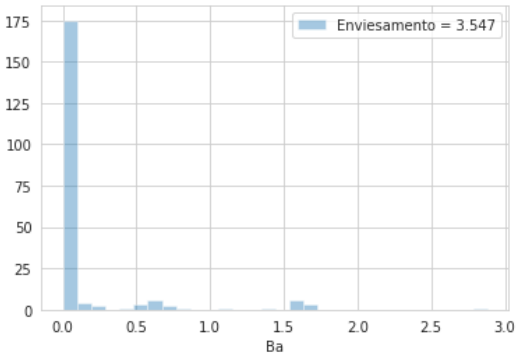


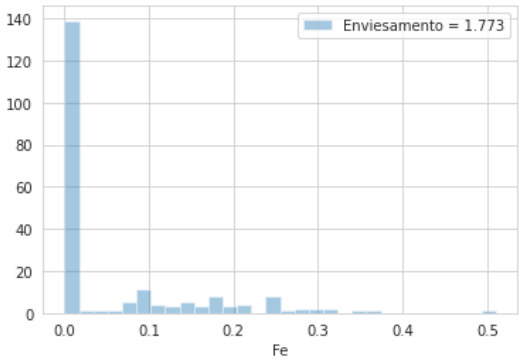












F. Transformação de dados

Vamos examinar se uma transformação Box-Cox pode contribuir para a normalização de alguns recursos, já que como utilizaremos do algoritmo de classificação gaussianNB, precisamos de nossa base normalizada para uma melhor acurácia do modelo. Deve-se enfatizar que todas as transformações devem ser feitas apenas no conjunto de treinamento para evitar espionagem de dados. Caso contrário, a estimativa do erro de teste será tendenciosa. Após a aplicação da transformação de Box-Cox, fizemos um teste e printamos gráficos onde se é mostrado que o enviesamento de todas as características foi reduzido.

IV. Treinamento do modelo

Com os dados tratados, é feito o treinamento e a medição de acurácia do classificador de Bayes, será utilizado o algoritmo classificador GaussianNB, já que todos elementos do nosso dataset se tratam de valores numéricos flutuantes em uma distribuição normal após a transformação de Box-cox.

Dividimos os dados de forma que 80% deles sejam usados para o treinamento e 20% para os testes.

V. Análise dos resultados

Como já foi especificado, dividimos nosso conjunto de dados em 80% para realizar o treinamento e 20% para a validação. Dessa forma, o nosso modelo, com o classificador gaussiano, conseguiu uma acurácia de 67%.

VI. Conclusões e discussões

Com base na análise da acurácia retornada pelo modelo proposto onde utilizamos o algoritmo gaussiano do classificador de naïve bayes, é possível concluirmos que sua acurácia de 67% é relativamente efetiva. Tendo em vista que o modelo tomado como referência [9] utilizou de abordagens mais complexas, e algoritmos mais sofisticados para adquirir uma acurácia próxima de 76% .

REFERENCES

[1] M. Paul, “Probabilidade: Aplicações à Estatística”. 2 Edição. livros Técnicos e Científicos Editora.

[2] https://en.wikipedia.org/wiki/Bayes’_theorem

[3] https://www.sciencedirect.com/topics/engineering/bayes-theorem

[4] https://www.3blue1brown.com/videos-blog/bayes-theorem-and-making-probability-intuitive

[5] Glass Identification [https://archive-beta.ics.uci.edu/ml/datasets/glass+identification].

[6] https://medium.com/turing-talks/turing-talks-16-modelo-de-predição-naive-bayes-6a3e744e7986

[7] https://towardsdatascience.com/how-i-was-using-naive-bayes-incorrectly-till-now-part-1-4ed2a7e2212b

[8] https://medium.com/analytics-vidhya/naive-bayes-for-mixed-typed-data-in-scikit-learn-fb6843e241f0

[9] https://www.kaggle.com/eliekawerk/glass-type-classification-with-machine-learning