PROJETO DE ESTATÍSTICA

Estatística e Probabilidade para Computação

Grupo:
- Gabriel Silva de Oliveira - gso
- João Marcos Alcântara Vanderley - jmav
- Kennedy Edmito Cunha Melo - kecm
- Rafael dos Reis de Labio- rrl3

Professor: Tsang Ing Ren
Centro de Informática - UFPE

O QUE SERÁ DISCUTIDO?

TÓPICOS DO PROJETO

1. Descrição do tema a ser abordado
2. Aplicação deste do tema
3. Experimentos para avaliação da técnica
4. Análise dos resultados para validar a técnica utilizada
5. Conclusão do projeto e considerações finais

DESCRIÇÃO DO TEMA

DATA SET "Wine Quality"

Utilizamos um dos vários Data sets disponíveis na plataforma do UCI Machine Learning, o nosso trata sobre Vinhos e suas carterísticas

Qual tópico é abordado?

O tópico apresentado e tratado será o uso do classificado do naive bayes para a previsão e classificação da qualidade de vinhos

Qual o problema é abordado?

O problema é conseguir desenvolver o nosso algoritmo para fazer com que nosso ele determine a qualidade de um Vinho

Como será a aplicação?

A partir do database, vamos classificar os vinhos por meio de suas caracterisiticas fisico-quimicas advindas de seus atributos, que são dados pelo data-set. E com isso, conseguiremos prever a qualidade do vinho a partir desses dados fornecidos e treinados

APLICAÇÃO DO TEMA

UTILIZAREMOS 12 ATRIBUTOS PARA IMPLEMENTAR OS CÓDIGOS

Atributos utilizados:

- Acidez fixa
- Volatilidade do ácido
- Ácido cítrico
- Açúcar residual
- Cloretos
- Dióxido de enxofre livre
- Dióxido de enxofre total
- Densidade
- PH
- Sulfatos
- Álcool
- Qualidade(entre 0-10)

NUMA APLICAÇÃO

Como será feito a implementação?

A implementação do naive bayes vai ser feita a partir do uso da biblioteca scikit learn em paralelo ao tratamento de dados, realizado com as outras bibs.

Quanto tempo vai durar a implementação?

0.004 segundos.

Todos os pontos do métodos são entendidos?

Sim, todas as funções ultilizadas,foram premeditadas a partir de um passo a passo para poder retirar a melhor eficiencia do algoritmo naive bayes.

NUMA APLICAÇÃO

Tecnica ultilizada:

Modelo de bayes

Base original

História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Desconhecida	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto
Desconhecida	Baixa	Nenhuma	>= 15.000 a <= 35.000	Moderado
Desconhecida	Baixa	Nenhuma	> 35.000	Alto
Desconhecida	Baixa	Nenhuma	> 35.000	Baixo
Desconhecida	Baixa	Adequada	> 35.000	Baixo
Ruim	Baixa	Nenhuma	< 15.000	Alto
Ruim	Baixa	Adequada	> 35.000	Moderado
Boa	Baixa	Nenhuma	> 35.000	Baixo
Boa	Alta	Adequada	> 35.000	Baixo
Boa	Alta	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	>= 15.000 a <= 35.000	Moderado
Boa	Alta	Nenhuma	> 35.0000	Baixo
Ruim	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto

Treinamento

Naive Bayes

Risco de crédito	História do crédito			Dívida		Garantias		Renda anual		
	Boa 5	Desconhecida 5	Ruim 4	Alta 7	Baixa 7	Nenhuma 11	Adequada 3	< 15000 3	>= 15000 <= 35000 4	> 35000 7
Alto 6/14	1/6	2/6	3/6	4/6	2/6	6/6	0	3/6	2/6	1/6
Moderado 3/14	1/3	1/3	1/3	1/3	2/3	2/3	1/3	0	2/3	1/3
Baixo 5/14	3/5	2/5	0	2/5	3/5	3/5	2/5	0	0	5/5

Teste (consulta)

Risco de crédito	História do crédito			Dívida		Garantias		Renda anual		
	Boa 5	Desconhecida 5	Ruim 4	Alta 7	Baixa 7	Nenhuma 11	Adequada 3	< 15 3	>= 15 <= 35 4	> 35 7
Alto 6/14	1/6	2/6	3/6	4/6	2/6	6/6	0	3/6	2/6	1/6
Moderado 3/14	1/3	1/3	1/3	1/3	2/3	2/3	1/3	0	2/3	1/3
Baixo 5/14	3/5	2/5	0	2/5	3/5	3/5	2/5	0	0	5/5

História = Boa
Dívida = Alta
Garantias = Nenhuma
Renda = > 35

Soma: 0,0079 + 0,0052 + 0,0514 = **0,0645**

P(Alto) = 6/14 * 1/6 * 4/6 * 6/6 * 1/6
P(Alto) = 0,0079
P(Alto) = 0,0079 / 0,0645 * 100 = **12,24%**

P(Moderado) = 3/14 * 1/3 * 1/3 * 2/3 * 1/3
P(Moderado) = 0,0052
P(Moderado) = 0,0052 / 0,0645 * 100 = **8,06%**

P(Baixo) = 5/14 * 3/5 * 2/5 * 3/5 * 5/5
P(Baixo) = 0,0514
P(Baixo) = 0,0514 / 0,0645 * 100 = **79,68%**

Risco de crédito	História do crédito			Dívida		Garantias		Renda anual		
	Boa 5	Desconhecida 5	Ruim 4	Alta 7	Baixa 7	Nenhuma 11	Adequada 3	< 15 3	>= 15 <= 35 4	> 35 7
Alto 6/14	1/6	2/6	3/6	4/6	2/6	6/6	0	3/6	2/6	1/6
Moderado 3/14	1/3	1/3	1/3	1/3	2/3	2/3	1/3	0	2/3	1/3
Baixo 5/14	3/5	2/5	0	2/5	3/5	3/5	2/5	0	0	5/5

História = Ruim
Dívida = Alta
Garantias = Adequada
Renda = < 15

Correção Laplaciana

P(Alto) = 6/14 * 3/6 * 4/6 * 0 * 3/6

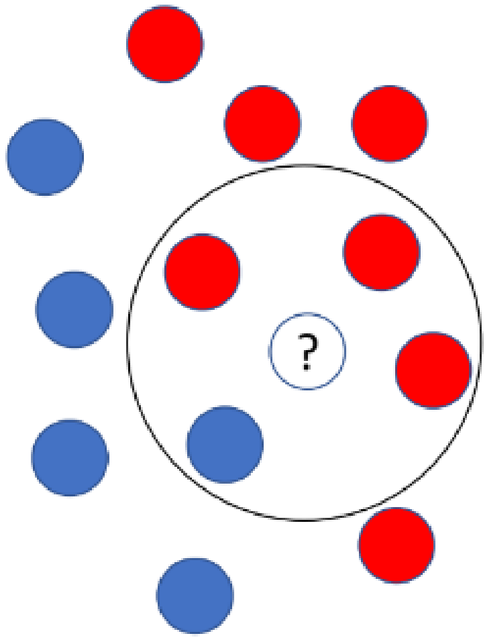
P(Moderado) = 3/14 * 1/3 * 1/3 * 1/3 * 0

P(Baixo) = 5/14 * 0 * 2/5 * 2/5 * 0

P(vermelho) = 7 / 12

P(azul) = 5 / 12

Probabilidades apriori



P’(vermelho) = 3 / 7

P’(azul) = 1 / 5

Probabilidades posteori

P’’(vermelho) = 7 / 12 * 3 / 7 = 21 / 84 = **0,25**

P’’(azul) = 5 / 12 * 1 / 5 = 5 / 60 = **0,08**

Vantagens x desvantagens

- Vantagens
  - Rápido
  - Simplicidade de interpretação
  - Trabalha com altas dimensões
  - Boas previsões em bases pequenas

- Desvantagem
  - Combinação de características (atributos independentes) – cada par de características são independentes – nem sempre é verdade

EXPERIMENTOS PARA AVALIAÇÃO TEÓRICA

Qual protocolo será utilizado nos experimentos?

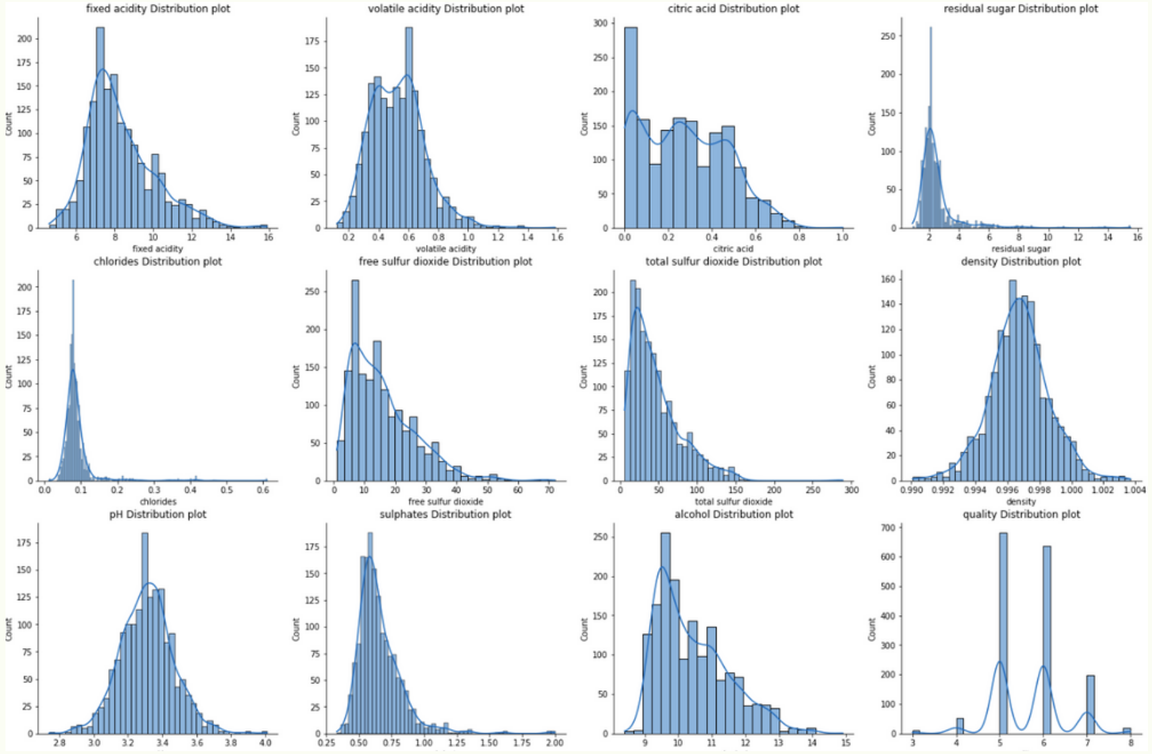
- Analise exploratória dos dados(graficos,tabelas)
- Tratamentos de dados (Checagem da presença de valores nulos, dentre outros pontos)
- Normalização dos dados
- Aplicação no algoritmo do Naive Bayes
- Test.

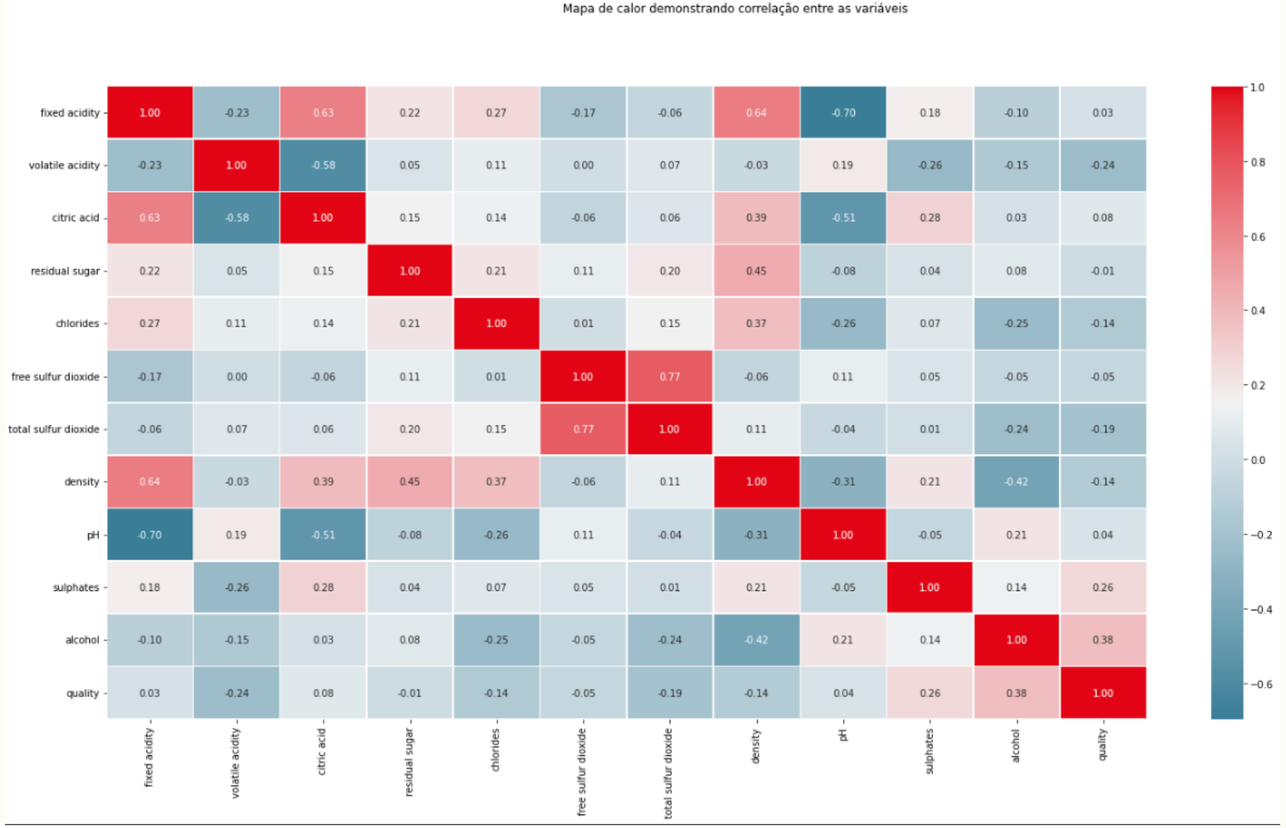
Divisão da base de treinamento e teste:

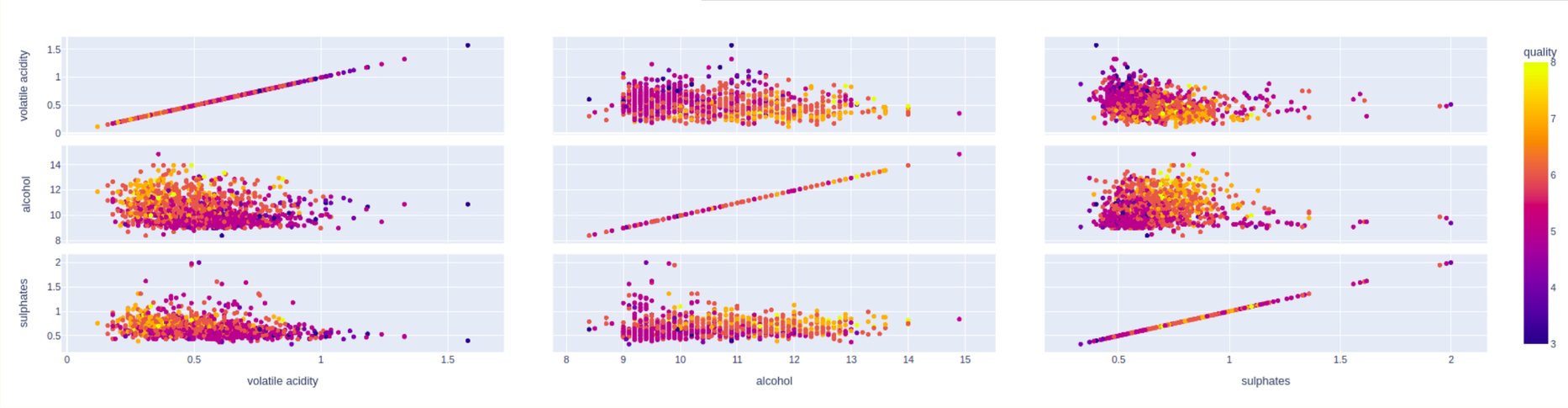
80% para treino e 20% para testes

Quais são os limites computacionais para o experimento?

- Dataset muito grande poderia não caber na memória ram e dificultaria o processamento dos dados.


fixed acidity Distribution plot
volatile acidity Distribution plot
citric acid Distribution plot
residual sugar Distribution plot
chlorides Distribution plot
free sulfur dioxide Distribution plot
total sulfur dioxide Distribution plot
density Distribution plot
pH Distribution plot
sulphates Distribution plot
alcohol Distribution plot
quality Distribution plot
Count


Mapa de calor demonstrando correlação entre as variáveis
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality


volatile acidity
alcohol
sulphates
quality

ANÁLISE DOS RESULTADOS VALIDANDO A TÉCNICA

Os experimentos comprovam a hipótese?

Sim. Conseguimos treinar o algoritmo para prever a qualidade do vinho.

Quais conclusões podem ser obtidas a partir destes experimentos? Dos dados? Das técnicas?

Que Naive Bayes foi efetivo para a classificação do dataset, é necessário o tratamento de dados para se ter um modelo com melhor desempenho

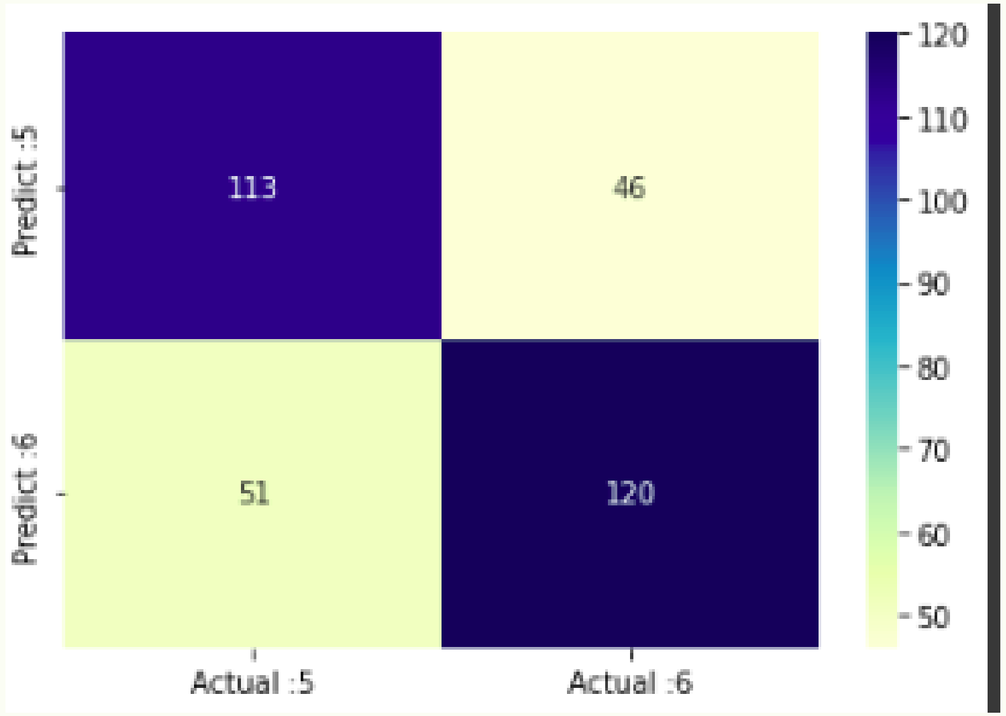
Com os resultados é possível fazer alguma previsão acerca do sistema avaliado?

Sim, é possível.

```
[ ] y_pred_train = gnb.predict(X_train);

[ ] print('Training-set accuracy score: {0:0.4f}'. format(accuracy_score(y_train, y_pred_train)))

    Training-set accuracy score: 0.6977
```




```
[ ] y_pred = gnb.predict(X_test)
```

Fazemos a acurácia para ver o rendimento do algorítmo com os casos testes

```
[ ] from sklearn.metrics import accuracy_score

    print('Model accuracy score: {0:0.4f}'. format(accuracy_score(y_test, y_pred)))

    Model accuracy score: 0.7061
```

CONCLUSÃO

- *Descrição do projeto*

Escolhemos o dataset *Wine Quality para trabalhar por conta da grande quantidade de atributos dentre outros pontos.*

- *Avaliação do projeto*

O projeto se mostrou de forma positiva,visto que, tornou possivel a aplicação de diversos tópicos estudados durante o curso de estatistica.

- *Aprendizados e resultados*

Tivemos resultados bons, porem, a acuracia foi satisfatória, ficou entre 70% de acerto.
Temos algumas hiposteses,como as classses ultilizadas para o treino, possuiam notas proximas, por conta disso, acreditamos que levou a classificação ter um desempenho menor.