Universidade Federal de Minas Gerais (UFMG)

Departamento de Ciência da Computação

Guilherme Mendes de Oliveira

guilhermemendes@ufmg.br

PREDIÇÃO DO NÚMERO DE COLÔNIAS DE ABELHAS

Belo Horizonte, MG – Brasil

2022

Guilherme Mendes de Oliveira

PREDIÇÃO DO NÚMERO DE COLÔNIAS DE ABELHAS

Trabalho prático de regressão da matéria de Mineração de Dados ministrada pelo professor Wagner Meira Jr.

Belo Horizonte, MG – Brasil

2022

1. Introdução

Este trabalho tem como objetivo a implementação de uma técnica de mineração de dados para regressão em uma base de dados. Para tal atividade foram utilizados os algoritmos ***Regressão Linear*** **e** ***SVM*****.**

Além disso, a escolha para base de dados foi feita a partir do Kaggle, um portal que disponibiliza diversas bases de dados para a prática de técnicas relacionadas à ciência de dados.

A base apresenta diversos atributos relacionados à produção de mel nos Estados Unidos nos anos de 1995 a 2021.

	Unnamed: 0	state	colonies_number	yield_per_colony	production	stocks	average_price	value_of_production	year
227	227	Indiana	8000	65	520000	286000	103.00	536000	2000
220	220	California	440000	70	30800000	11396000	58.00	17864000	2000
802	802	Hawaii	15000	93	140000	140000	229.00	3195000	2014
1014	1014	Missouri	10000	43	73000	73000	3.35	1441000	2019
273	273	Kentucky	3000	78	234000	94000	131.00	307000	2001
139	139	Idaho	120000	50	6000000	2220000	65.00	3900000	1998
97	97	Illinois	7000	69	483000	222000	127.00	613000	1997
145	145	Louisiana	41000	111	4551000	865000	59.00	2685000	1998
620	620	NewYork	47000	65	3055000	978000	183.00	5591000	2009
497	497	NewYork	60000	64	3840000	2458000	138.00	5299000	2006

Figura 1 – Amostra dos dados da base de produção de mel.

2. Motivação/ justificativa

A aplicação das técnicas de modelagem envolvidas no trabalho juntamente com a utilização de um algoritmo em uma base de dados que se aproxima de problemas reais encontrados no cotidiano é além de uma forma de avaliação uma oportunidade de desenvolver todas as competências técnicas necessárias para aplicar futuramente esse conhecimento no “mundo real” ou seja uma forma de ganhar experiência analítica dos resultados e competências técnicas de programação para exploração, limpeza dos dados , aplicação do algoritmo e extração de métricas para avaliação dos resultados.

3. Objetivo

A aplicação das técnicas de classificação nessa base de dados de produção de mel, busca avaliar qual dos dois algoritmos selecionados terá um melhor desempenho para predizer o número de colônias de abelhas.

Esse entendimento por exemplo poderia ser utilizado ainda que de forma genérica para uma

empresa que trabalha com mel ou um órgão que realiza um acompanhamento sobre produção agrícola como o SEBRAE realiza com diferentes setores da economia.

4. Metodologia

A metodologia utilizada foi inspirada no CRISP-DM, inspirada porque uma das etapas, mais especificamente a do entendimento do negócio, é realizada de forma simulada uma vez que que proposição de um problema só é possível após uma verificação preliminar da base de dados, ou ela é proposta e partir disso que os dados são buscados. Além disso, ela está direcionada a ser um problema que seja possível explorar e propor soluções com base em uma técnica específica, nesse caso a classificação.

Além disso, a etapa de *deployment* também não é pertinente uma vez que é uma prática e nesse caso não há um problema real a ser tratado, para as demais etapas todas foram possíveis e foram aplicadas para a conclusão do trabalho prático.

O fluxo de entendimento, preparação, modelagem e avaliação podem ser vistos aqui.

5. Desenvolvimento

5.1. Business Understanding

Essa etapa compreende o entendimento do problema, como foi propositivo a partir da base de dados e já direcionado para a aplicação de classificação esta etapa não se aplica.

5.2. Data Understanding

A base utilizada possui um total de 1115 linhas, 9 colunas e não possui amostras com valores nulos. Portanto, no entendimento da base de dados verificamos que a mesma está apta para a aplicação da técnica de mineração uma vez que possui uma quantidade considerável de registros, mas que será necessário avaliar e tratar as diferentes escalas de medição dos atributos.

Com o objetivo de entender se a distribuição de atributos pode ocasionar algum viés dados alguns valores foi realizado uma plotagem do histograma de cada atributo. A distribuição dos valores não indica viés uma vez mesmo existindo faixas dominantes, os mesmo não

representando majoritariamente a base de dados, não há grandes intervalos ausentes ou muito concentrados, a princípio as instâncias parecem bem heterogêneas.

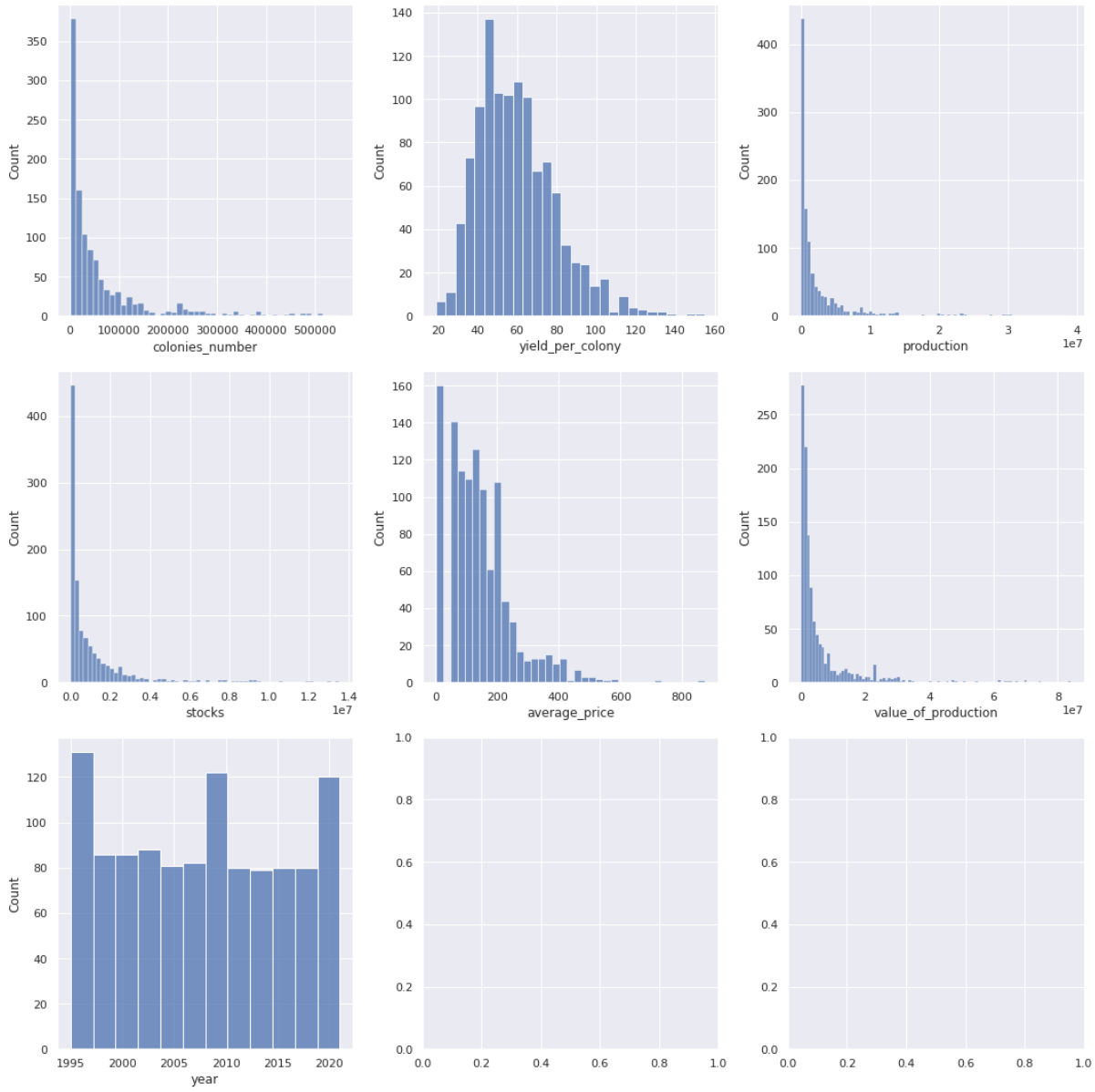

Figura 2 – Histograma dos atributos.

Com o objetivo de entender se há possibilidade de reduzir a dimensionalidade, a matriz de correlação foi plotada também, no entanto a correlação dos atributos uns com os outros mesmo sendo forte em alguns casos não indica seguramente que é possível suprimir algum atributo a ser passado para o modelo, para tal será implementada uma técnica de avaliação de acurácia com passo na iteração na quantidade e nos atributos que o algoritmo de treino irá receber chamada *RFE*.

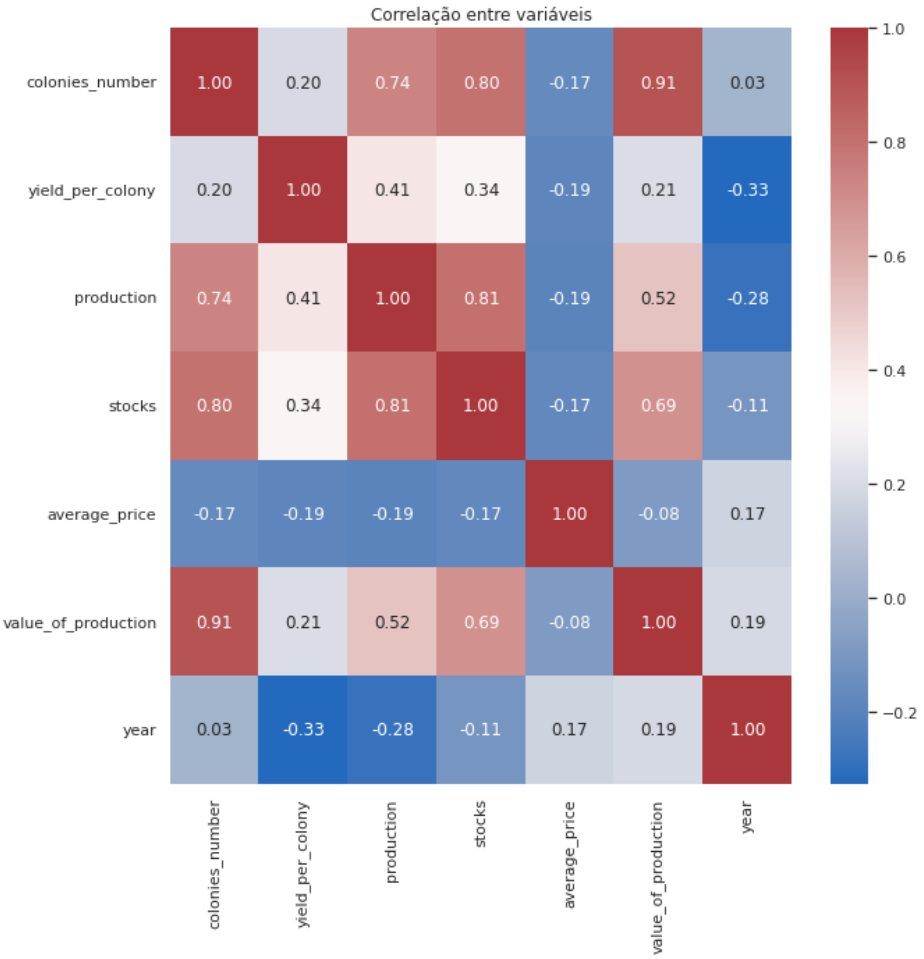

Figura 3 – Correlação entre os atributos da base de dados.

5.3. Data Preparation

Na etapa de preparação inicialmente não foi necessário remover instâncias ou atributos, apenas normalizar os dados e não tendo impacto em função da diferença de escala.

```
[ ]  #Normalizando os dados
     dfNorm = dfMel.copy()
     dfNorm['colonies_number'] = (dfNorm['colonies_number'] - dfNorm['colonies_number'].mean())/ dfNorm['colonies_number'].std()
     dfNorm['yield_per_colony'] = (dfNorm['yield_per_colony'] - dfNorm['yield_per_colony'].mean())/ dfNorm['yield_per_colony'].std()
     dfNorm['production'] = (dfNorm['production'] - dfNorm['production'].mean())/ dfNorm['production'].std()
     dfNorm['stocks'] = (dfNorm['stocks'] - dfNorm['stocks'].mean())/ dfNorm['stocks'].std()
     dfNorm['average_price'] = (dfNorm['average_price'] - dfNorm['average_price'].mean())/ dfNorm['average_price'].std()
     dfNorm['value_of_production'] = (dfNorm['value_of_production'] - dfNorm['value_of_production'].mean())/ dfNorm['value_of_production'].std()
```

Figura 4 – Normalização da base de dados.

5.4 Modelagem

Na etapa de modelagem da base para a aplicação dos algoritmos de regressão, inicialmente foi aplicada a técnica de *RFE*. A técnica consiste em iterar a quantidade de atributos que o modelo irá receber para o treino e para o teste e apurar a acurácia. Além disso, recursivamente o *RFE* irá medir a importância de cada atributo para o modelo e assim podemos avaliar com mais propriedade a remoção de atributos a fim de reduzir complexidade sem perder eficiência na predição.

A técnica foi utilizada para os dois algoritmos a serem comparados, a Regressão Linear e o SVM.

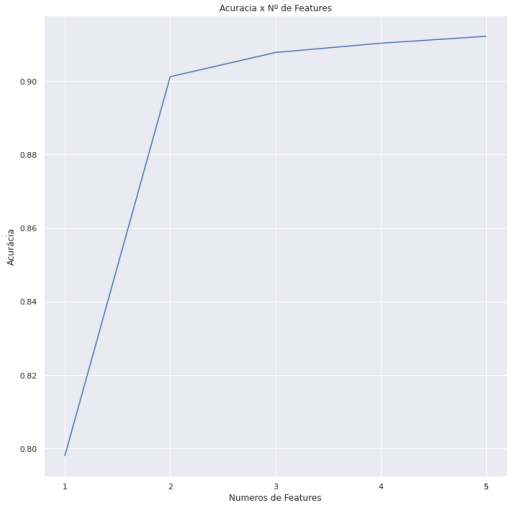

Figura 5 – Gráfico de acurácia x número de atributos - Regressão Linear.

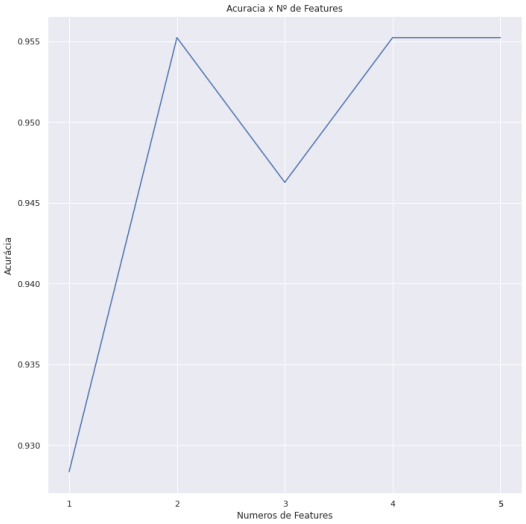

Figura 6 – Gráfico de acurácia x número de atributos - SVM.

Observamos que o melhor caso para a Regressão Linear são todos os atributos da base de dados com exceção da variável dependente que queremos predizer, o SVM temos um caso curioso, a acurácia é muito próxima utilizando 2,4 e 5 atributos, nesse caso por questão de complexidade na implementação escolhido o menor número de atributos como desempate.

Avaliando os atributos escolhidos pelo *RFE* observado que eles apresentam forte correlação com a variável dependente, o que era esperado, são a primeira e a terceira correlação mais forte.

A segunda correlação mais forte não foi selecionada, observando os valores do atributo em questão(*stocks*) a correlação é mais forte com um dos atributos que foi escolhido pelo *RFE* (*production*) do que com a variável dependente e talvez por isso o atributo não foi escolhido

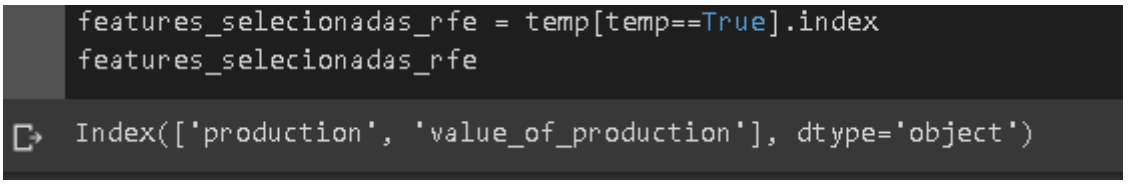

Figura 7 – Atributos selecionados pelo *RFE* - SVM.

6. Resultados experimentais e análise
6.1. Evaluation

Foi considerada a ideia de utilizar validação cruzada, no entanto não houve sucesso na implementação em tempo hábil para a entrega do trabalho. Avaliando as métricas mais rotineiras, acurácia que diz respeito ao quanto em média o modelo tem de desvio na predição em relação ao real e o R^2 score que diz respeito ao quanto da variância é explicada pelas variáveis independentes no modelo , computadas no teste, o algoritmo de Regressão Linear apresentou melhor desempenho que o SVM.

Figura 8 – RMSE - Regressão Linear.

```
rmseSVM = mean_squared_error(resSVM, y_testSVM)
rmseSVM
```

```
0.1827956989247312
```

Figura 9 – RMSE - SVM.

```
rscoreLin = r2_score(y_testRegLin, resRegLin)
rscoreLin
```

```
0.9342728616960904
```

Figura 10 – R² - Regressão Linear.

```
rscoreSVM = r2_score(y_testSVM, resSVM)
rscoreSVM
```

```
0.6889020070838253
```

Figura 11 – R² - SVM

7. Conclusões e perspectivas

O resultado do trabalho prático foi bem satisfatório, foi possível explorar algumas das técnicas de tratamento e preparação de dados, a implementação de dois algoritmos de regressão além de práticas de métricas de avaliação a forma de analisá-las e também entender e desenvolver mais a análise da relevância dos diversos atributos dentro de um modelo de predição ainda que a implementação esteja encapsulada na biblioteca utilizada.

8.Referências

Zaki, M and Meira, W. Data Mining and Machine Learning: Fundamental Concepts and Algorithms. Capítulo 19: Support Vector Machines;

Zaki, M and Meira, W. Data Mining and Machine Learning: Fundamental Concepts and Algorithms. Capítulo 23: Linear Regression;

Zaki, M and Meira, W. Data Mining and Machine Learning: Fundamental Concepts and Algorithms. Capítulo 27:Regression Evaluation;

<https://scikit-learn.org/stable/supervised_learning.html> . Acesso em: 12 de dezembro de 2022.

<https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html> Acesso em: 14 de dezembro de 2022.

<https://paulovasconcellos.com.br/como-selecionar-as-melhores-features-para-seu-modelo-de-machine-learning-2e9df83d062a>. Acesso em: 14 de dezembro de 2022.

<https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics> . Acesso em: 20 de dezembro de 2022.