UNIVERSIDADE FEDERAL DO ACRE

CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS – CCET

BACHARELADO EM SISTEMAS DE INFORMAÇÃO

MATEUS DE SOUZA LOPES

TRABALHO DE MINERAÇÃO DE DADOS

RIO BRANCO

2023

MATEUS DE SOUZA LOPES

TRABALHO DE MINERAÇÃO DE DADOS

ALGORITMOS DE MINERAÇÃO DE DADOS

SISTEMAS DE APOIO À DECISÃO

Trabalho de mineração de dados do Curso de Bacharelado em Sistemas de Informação da Universidade Federal do Acre, a fim de obter nota referente a N2.

Orientador: Prof. Manoel Limeira

RIO BRANCO

2023

1. MINERAÇÃO DE DADOS E O SOFTWARE WEKA

A mineração de dados é um processo que envolve a descoberta de padrões úteis, informações relevantes e conhecimento previamente desconhecido a partir de grandes conjuntos de dados. Essa disciplina multidisciplinar combina elementos da estatística, aprendizado de máquina, inteligência artificial, banco de dados e visualização de dados para extrair informação.

Para obter os resultados e aplicar os algoritmos foi utilizado o software Weka. O Weka desempenha um papel fundamental na implementação de algoritmos de mineração de dados, ela é uma suíte de software de código aberto que oferece uma ampla variedade de ferramentas e algoritmos para análise de dados e mineração de conhecimento.

2. CLASSIFICAÇÃO

2.1 A BASE DE DADOS

A base de dados selecionada para realizar a classificação aborda o tema de e-mail spam. Uma base de dados usada para classificar e-mails como spam é fundamental para a implementação de filtros de spam em sistemas de correio eletrônico. Essa base de dados é comumente utilizada em conjunto com algoritmos de aprendizado de máquina para treinar modelos capazes de distinguir entre e-mails legítimos (não spam) e e-mails indesejados (spam).

O conjunto de dados escolhido tem 5172 linhas que correspondem aos e-mails e 3002 atributos onde o primeiro atributo corresponde ao identificador do e-mail e o último corresponde a classificação do mesmo como spam ou não, os demais são palavras.

Quadro 1 – Dicionário de Dados dos campos presentes na base utilizada

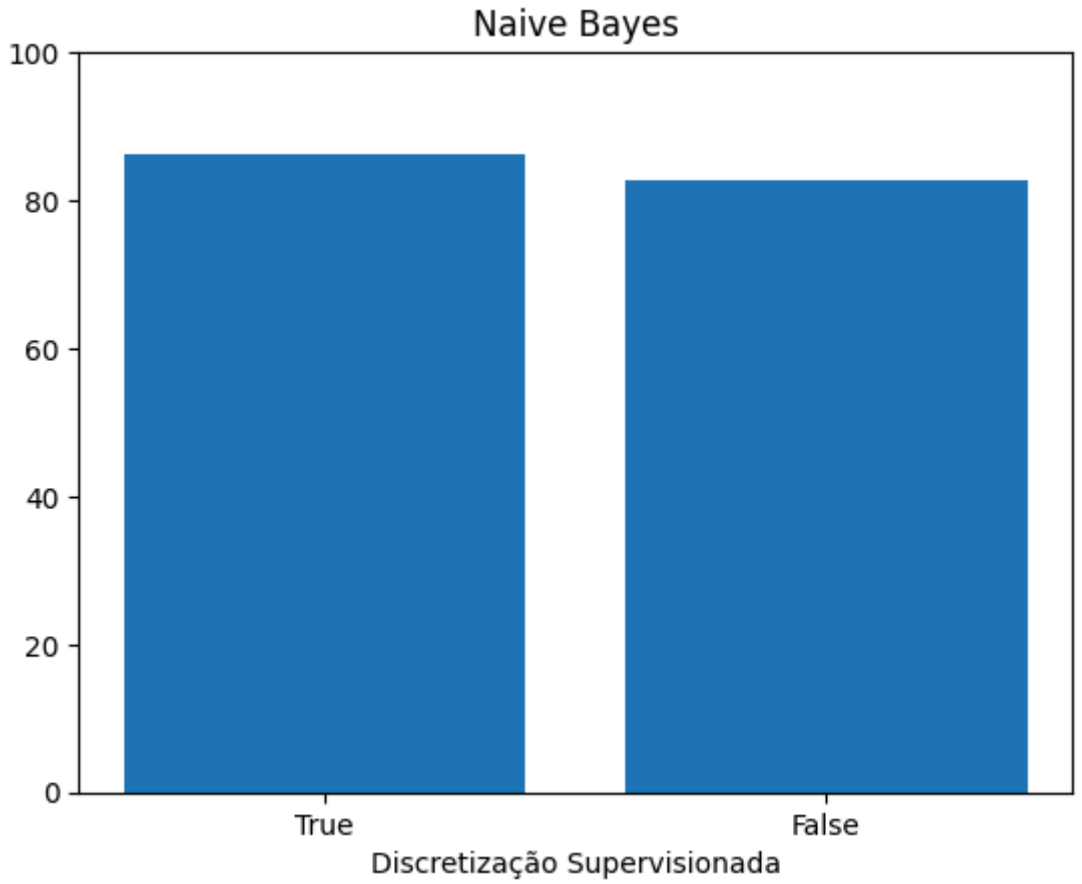
ATRIBUTO	DESCRIÇÃO
Email No.	Número correspondente ao identificador do e-mail.
Palavras	“Palavras-chave”, onde o valor do atributo é a quantidade daquela palavra presente no e-mail.
Prediction	Diz se um e-mail é spam ou não, (valor de 0 para não e 1 para sim).

2.2 EXECUÇÃO DOS ALGORITMOS

Para utilizar a base de dados de e-mails, a fim de realizar os algoritmos de classificação foi necessário transformar o atributo Prediction em um atributo nominal, utilizando o filtro NumericToNominal presente no Weka, já que o mesmo é numérico onde 0 e 1 correspondem a não spam e spam, respectivamente.

Foram utilizados os algoritmos de classificação Naive Bayes, Ibk (classifica novos pontos de dados com base na classe da maioria dos k-vizinhos mais próximos no espaço de características.), J48 (constrói uma árvore de decisão recursiva, dividindo os dados com base nos atributos mais informativos), SMO e Random Forest (consiste em um conjunto de árvores de decisão, onde cada árvore é treinada em uma amostra aleatória dos dados), onde cada algoritmo foi analisado verificando a acurácia e variando os parâmetros. O objetivo da classificação é conseguir identificar se um e-mail é spam ou não.

Figura 1 – Resultados obtidos com o algoritmo Naive Bayes



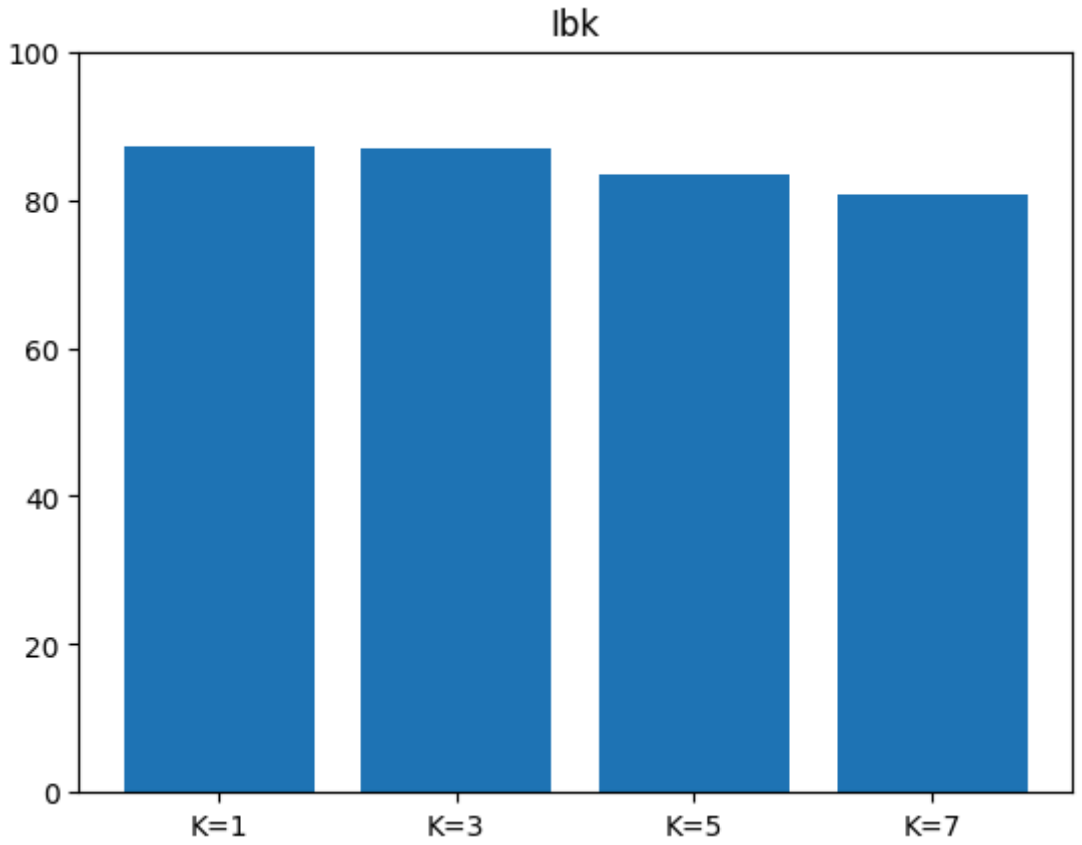
Fonte: Resultados obtidos no WEKA.

Com o algoritmo Naive Bayes foram obtidos bons resultados, alternamos a Discretização Supervisionada entre True e False e obtivemos os seguintes resultados:

Tabela 1 – Resultados Naive Bayes

DS	Acurácia
True	86,1206% (1514 instâncias)
False	82,7645% (1455 instâncias)

Figura 2 – Resultados obtidos com o algoritmo Ibk



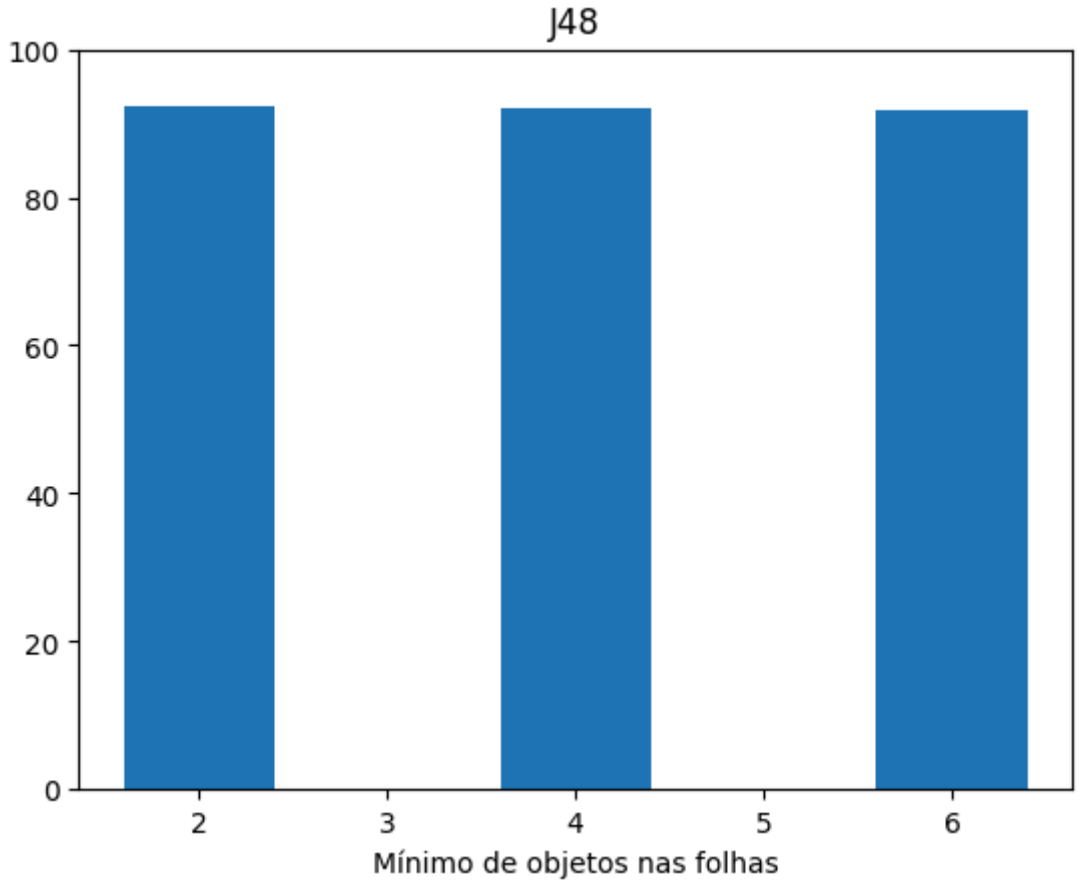
Fonte: Resultados obtidos no WEKA.

Com o algoritmo Ibk foram obtidos resultados um pouco melhores em relação ao Naive Bayes, o parâmetro alterado foi o número de vizinhos K = (1, 3, 5 e 7). Obtivemos os seguintes resultados:

Tabela 2 – Resultados Ibk

K	Acurácia
1	87,1445% (1532 instâncias)
3	86,917% (1528 instâncias)
5	83,4471% (1467 instâncias)
7	80,7167% (1419 instâncias)

Figura 3 – Resultados obtidos com o algoritmo J48



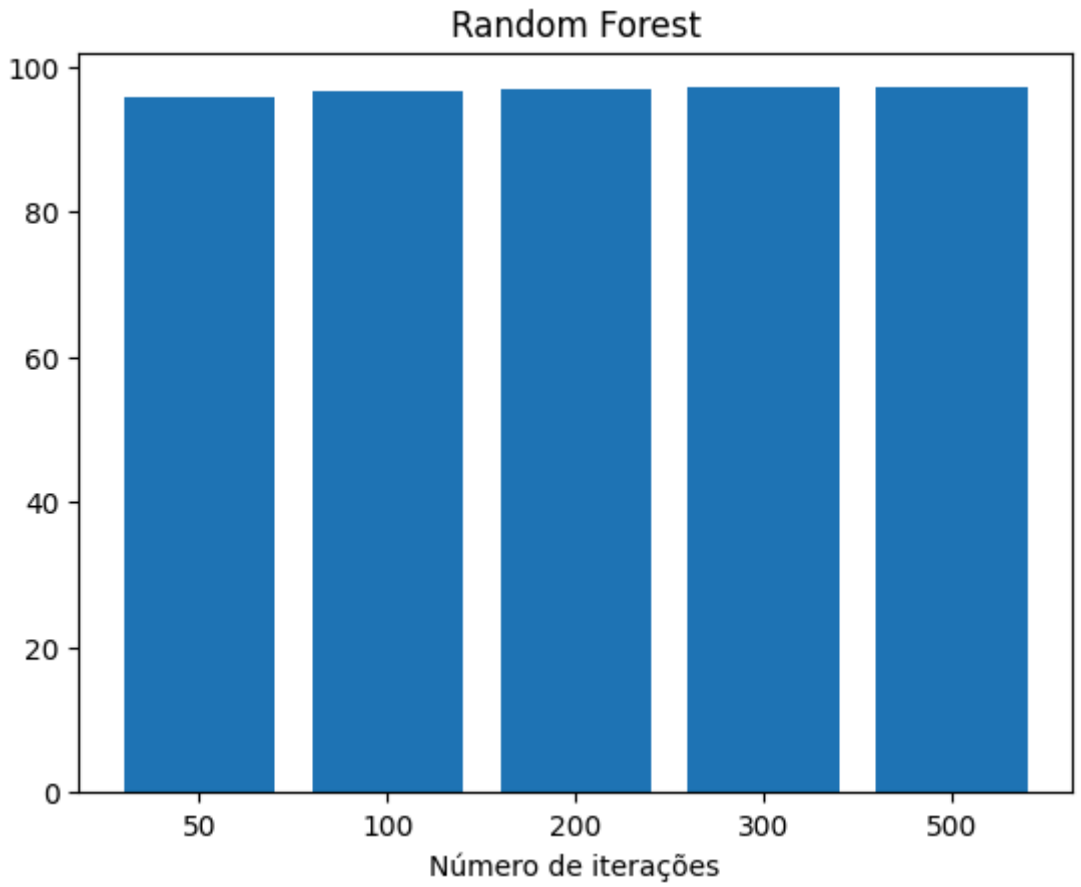
Fonte: Resultados obtidos no WEKA.

Com o algoritmo J48 alterando o número mínimo de elementos nas folhas (2, 4 e 6), aumentando esse parâmetro verificamos que a precisão variou pouco. O algoritmo apresentou resultados muito bons, melhor que os dois anteriores (Naive Bayes e Ibk). O algoritmo apresentou os seguintes resultados:

Tabela 3 – Resultados J48

MinNumObj	Acurácia
2	92,4346% (1625 instâncias)
4	91,9795% (1617 instâncias)
6	91,8089% (1614 instâncias)

Figura 4 – Resultados obtidos com o algoritmo Random Forest



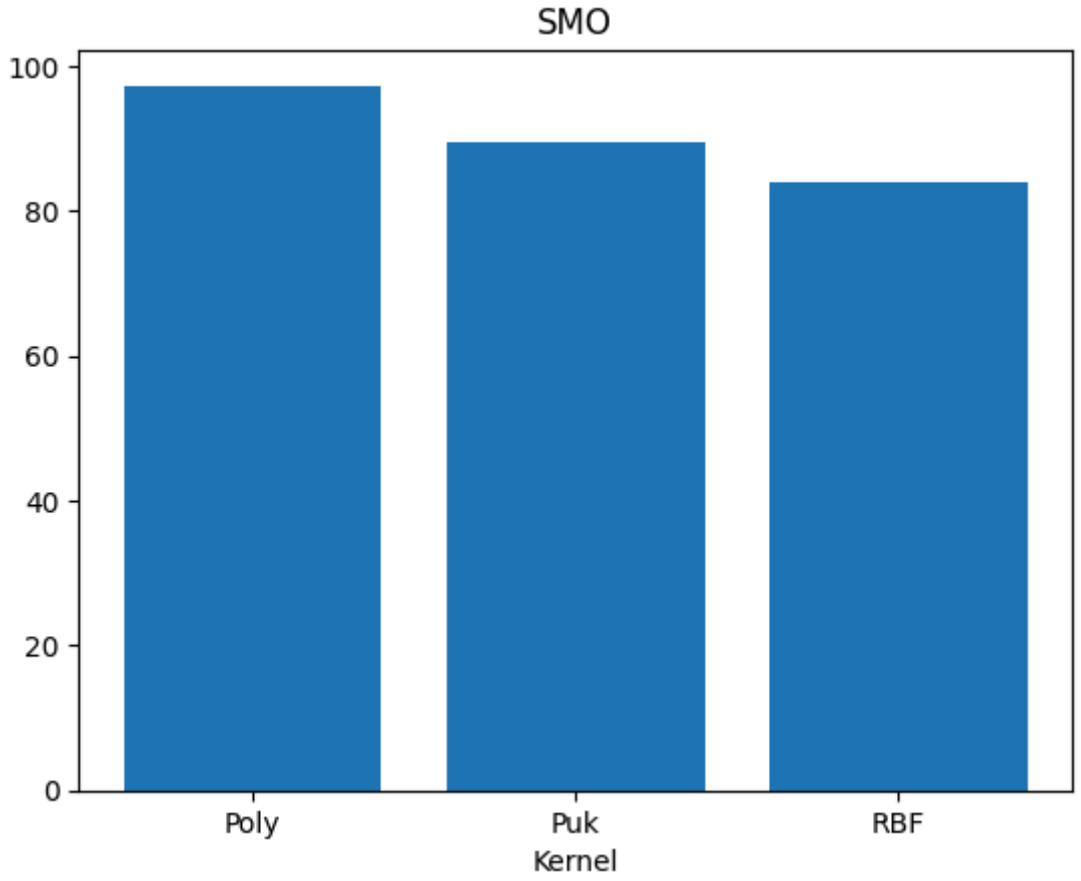
Fonte: Resultados obtidos no WEKA.

No algoritmo Random Forest variamos o parâmetro do número de iterações, que gerou resultados bem parecidos, utilizamos os valores (50, 100, 200, 300, 500). Foram obtidos os seguintes resultados:

Tabela 4 – Resultados Random Forest

Iterações	Acurácia
50	95,7338% (1683 instâncias)
100	96,4733% (1696 instâncias)
200	96,8714% (1703 instâncias)
300	97,0421% (1706 instâncias)
500	97,099% (1707 instâncias)

Figura 5 – Resultados obtidos com o algoritmo SMO



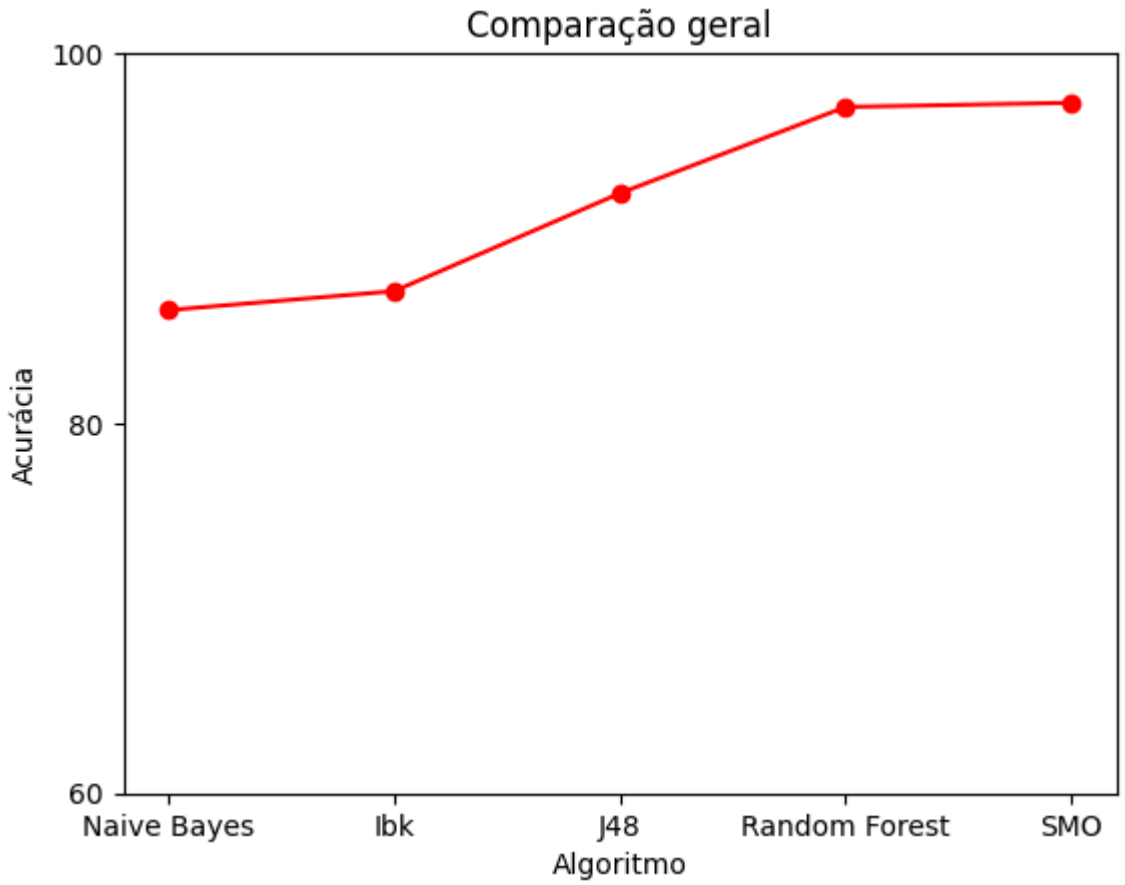
Fonte: Resultados obtidos no WEKA.

No algoritmo SMO alteramos o kernel entre Poly, Puk e RBF, dentre todos os algoritmos de classificação, o SMO com o kernel Poly obteve os melhores resultados. Resultados:

Tabela 5 – Resultados SMO

Kernel	Acurácia
Poly	97,3265% (1711 instâncias)
Puk	89,5336% (1574 instâncias)
RBF	84,0159% (1477 instâncias)

Figura 6 – Resultados obtidos com os algoritmos



Através dos resultados obtidos, foi visto que o melhor algoritmo para classificar um e-mail como spam ou não, foi o algoritmo SMO (Poly Kernel) ele alcançou uma acurácia de 97,3265%, próximo ao resultado desejado que são os 100%. Realizando a seleção de atributos o valor de 97,3265% do SMO (Poly Kernel) caiu para 90,7966%.

2.3 SELEÇÃO DE ATRIBUTOS

Após a execução dos algoritmos foram realizados procedimentos de seleção de atributos a fim de melhorar a acurácia. Inicialmente foi encontrado um problema, que era o alto número de atributos presente na base, com isso foi necessário usar o filtro do Weka “Attribute Selection”. Esse filtro executa os algoritmos de seleção de atributos e remove os descartados.

Usando o algoritmo Cfs, obtivemos resultados piores, isso pode ser explicado devido ao fato de o número de atributos diminuir de 3002 para 128, acaba que o algoritmo retirou muitos atributos, e com certeza ali haviam atributos importantes.

Não foi possível executar os outros algoritmos pois a ferramenta não excluía os atributos, isso pode ter acontecido devido à alta quantidade de atributos.

Tabela 6 – Resultados seleção de atributos

Algoritmo	Acurácia
SMO (Poly)	97,3265%
SMO (Poly) com seleção	90,7966%

3. Regressão

3.1 A BASE DE DADOS

Já a base de dados escolhida para a regressão foi uma de vendas em supermercados, oferecendo insights sobre o comportamento de compra dos consumidores e tendências de produtos. A análise desses dados permite que os supermercados e profissionais de marketing otimizem operações, personalizem ofertas para clientes, melhorem a eficiência do estoque e tomem decisões estratégicas informadas. Além disso, a análise desses dados pode revelar padrões de compra, sazonalidades e preferências dos clientes, contribuindo para o sucesso e a competitividade do supermercado no mercado. A base selecionada conta com 1001 linhas e 17 atributos.

Quadro 2 – Dicionário de Dados dos campos presentes na base utilizada

ATRIBUTO	DESCRIÇÃO
Invoice ID	Identificador único para Nota Fiscal ou compra.
Branch	A filial ou localização de onde ocorreu a transação.
City	Localização onde a filial está localizada.
Customer Type	Indica se o consumidor é novo ou habitual.
Gender	Gênero do consumidor.
Product Line	Categoria ou tipo do produto comprado.
Unit price	Preço da unidade.
Quantity	O número de unidades do produto

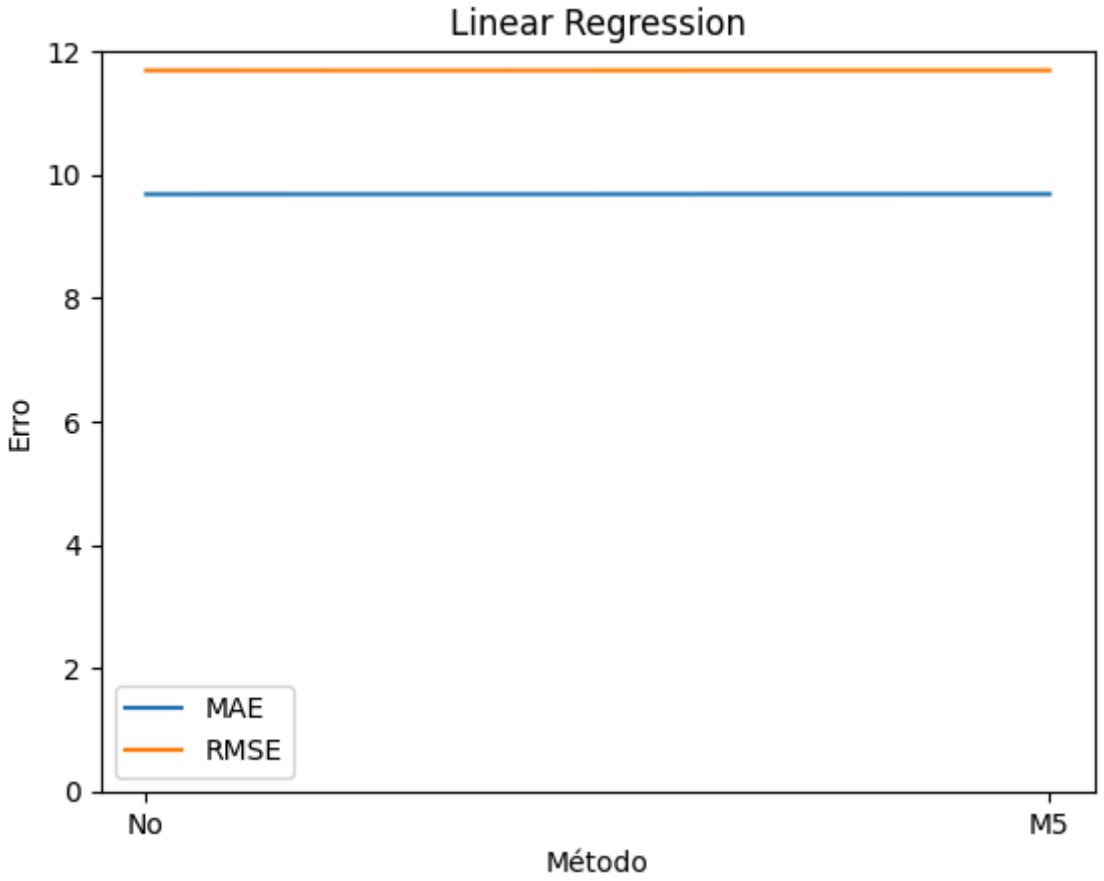
	comprado.
Tax 5%	A quantia da taxa de 5% aplicada a transação.
Total	Custo total da transação.
Date	A data que a transação aconteceu.
Time	A hora que a transação aconteceu.
Payment	O método de pagamento utilizado.
COGS	O custo de produção ou de compra do produto vendido.
Gross Margin Percentage	Margem de lucro da transação em porcentagem.
Gross Income	O lucro total ganho na transação.
Rating	Avaliação da satisfação do consumidor ou feedback da transação.

3.2 EXECUÇÃO DOS ALGORITMOS DE REGRESSÃO

Diversos algoritmos de regressão foram aplicados para prever o lucro total das transações. Os algoritmos utilizados incluem Linear Regression, IBk (k-Nearest Neighbors), SMOReg (Sequential Minimal Optimization for Regression), M5P (M5' Model Tree), e Random Forest. Cada algoritmo tem suas próprias características e abordagens, visando capturar padrões nos dados que levam a previsões precisas do lucro total. O desempenho dos algoritmos será avaliado usando métricas apropriadas para determinar qual deles oferece a melhor capacidade de prever o lucro total com base nas transações do supermercado.

O "Mean Absolute Error" (Erro Médio Absoluto ou MAE) e o "Root Mean Squared Error" (Erro Quadrático Médio ou RMSE) são duas métricas comumente utilizadas para avaliar a precisão de modelos de previsão ou regressão. Enquanto o MAE fornece uma média das diferenças absolutas entre previsões e valores reais, o RMSE fornece uma média das diferenças quadráticas, sendo mais sensível a erros grandes. A escolha entre essas métricas depende da natureza específica do problema e da importância de se penalizar ou não os erros grandes de forma mais acentuada. Portanto, quanto menor o MAE e o RMSE, melhor o resultado.

Figura 7 – Resultados obtidos com o algoritmo Linear Regression

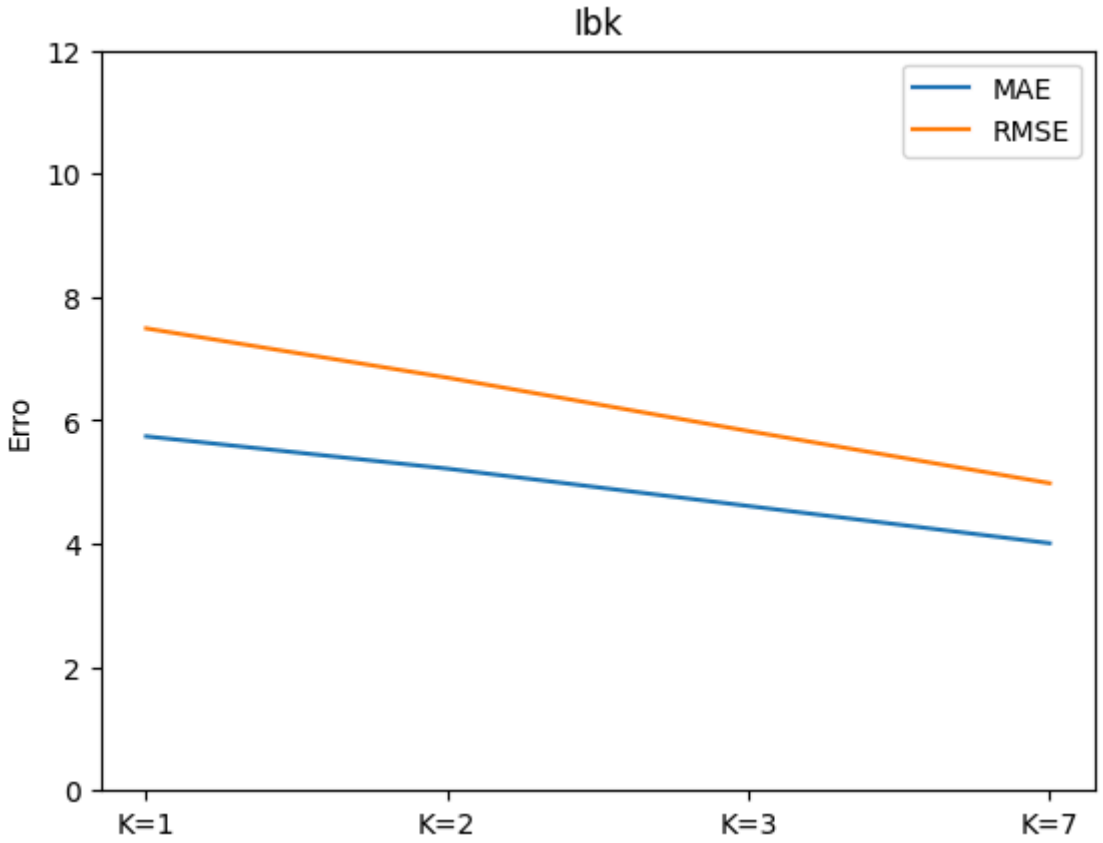


Fonte: Resultados obtidos no WEKA.

Tabela 7 – Resultados Linear Regression

Método	MAE	RMSE
No	9,6826	11,69
M5	9,6848	11,6913

Figura 8 – Resultados obtidos com o algoritmo Ibk

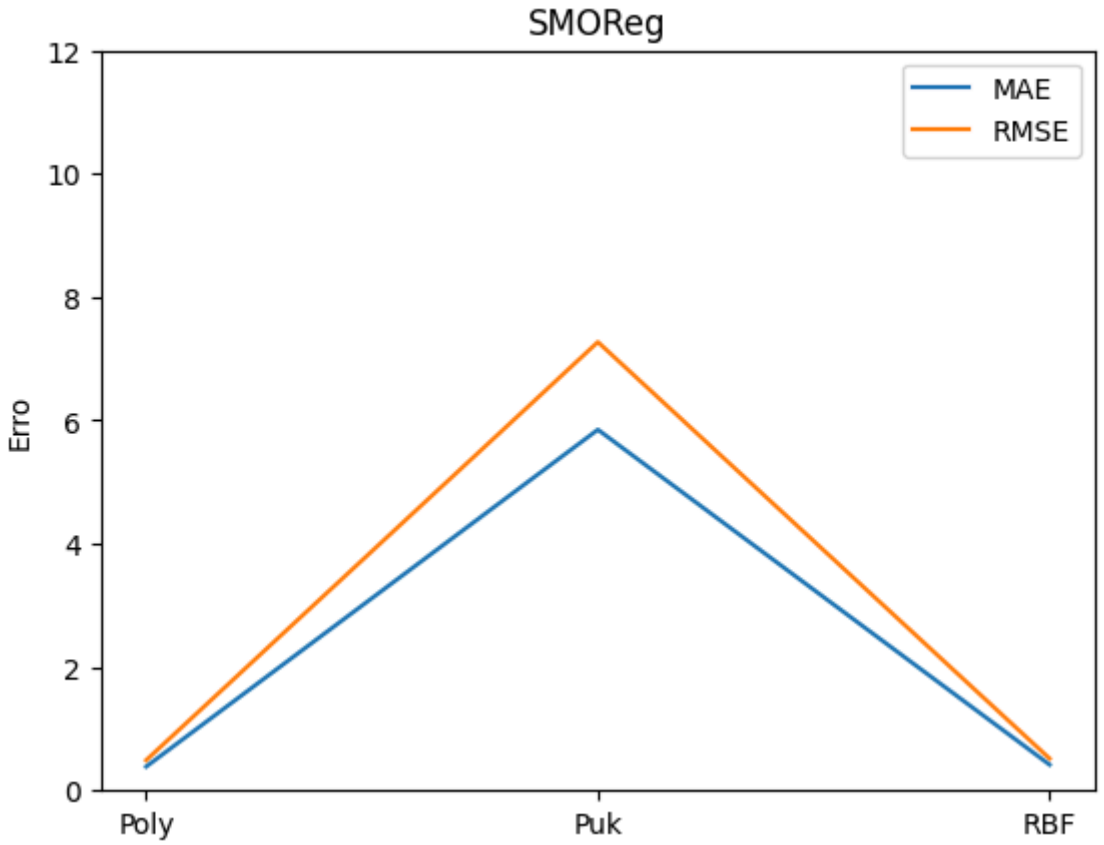


Fonte: Resultados obtidos no WEKA.

Tabela 8 – Resultados Linear Regression

K	MAE	RMSE
1	5,7427	7,493
2	5,2187	6,6942
3	4,6125	5,8283
7	4,0094	4,9825

Figura 9 – Resultados obtidos com o algoritmo SMOReg

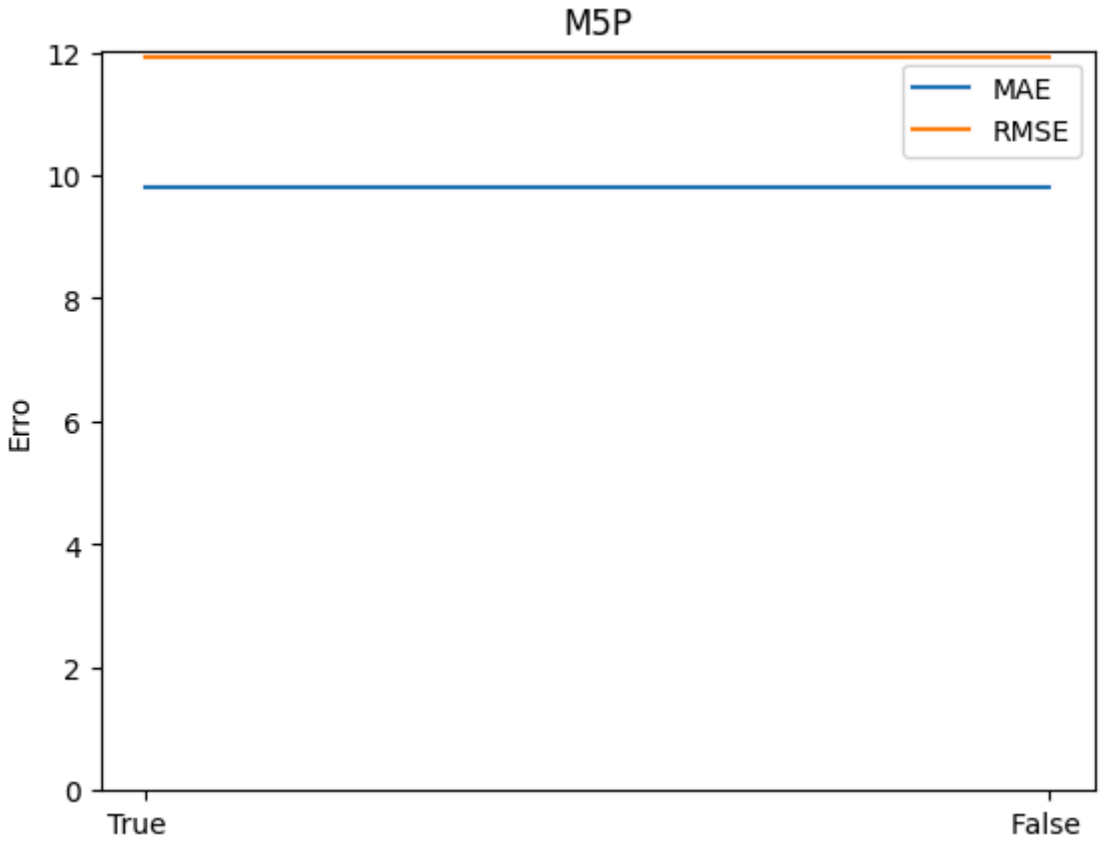


Fonte: Resultados obtidos no WEKA.

Tabela 9 – Resultados SMOReg

Kernel	MAE	RMSE
Poly	0,3869	0,4907
Puk	5,8507	7,2719
RBF	0,418	0,5175

Figura 10 – Resultados obtidos com o algoritmo M5P

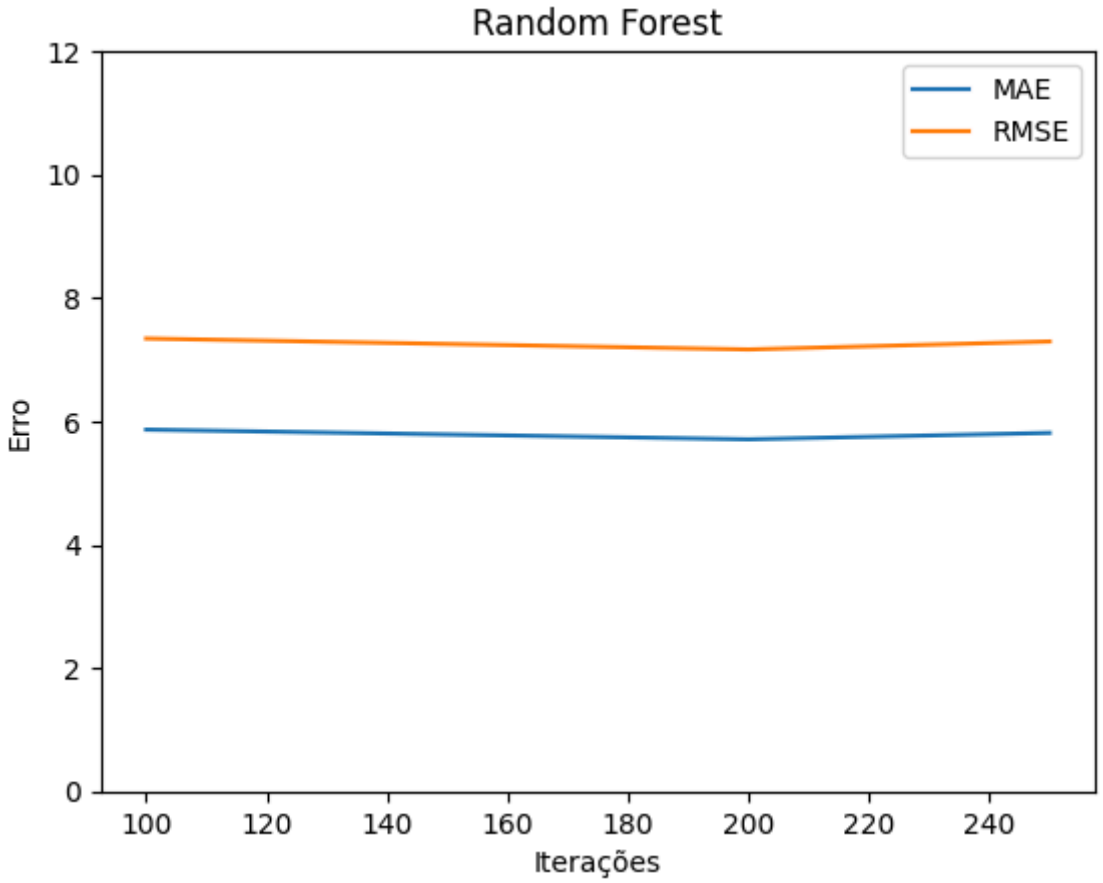


Fonte: Resultados obtidos no WEKA.

Tabela 10 – Resultados M5P

Unpruned	MAE	RMSE
True	9,8055	11,9219
False	9,8055	11,9219

Figura 11 – Resultados obtidos com o algoritmo Random Forest

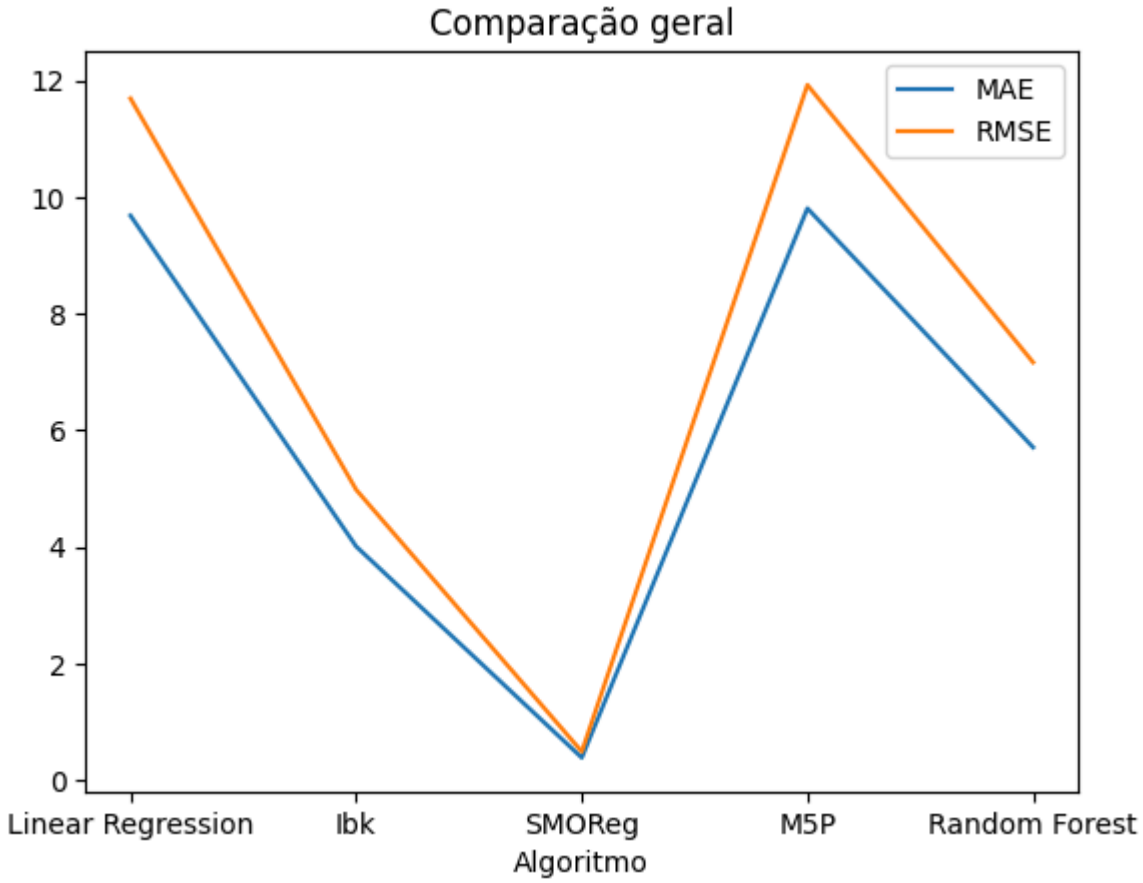


Fonte: Resultados obtidos no WEKA.

Tabela 11 – Resultados Random Forest

Iterações	MAE	RMSE
100	5,8652	7,3429
200	5,708	7,1655
250	5,8122	7,2943

Figura 12 – Resultados obtidos com os algoritmos



Através dos resultados obtidos, foi visto que o melhor algoritmo para prever o lucro obtido em uma transação é o algoritmo SMOReg (Poly Kernel) que obteve o resultado de 0,3869 no Mean absolute error, e 0,4907 no Root mean squared error, resultados estes que são próximos ao cenário desejado, onde os erros seriam iguais a zero. Logo, esse seria o algoritmo ideal para prever lucro em transações.

3.3 SELEÇÃO DE ATRIBUTOS

Para diminuir os valores de erro, a base foi submetida a um processo de seleção de atributos usando os algoritmos Cfs e Relief.

O algoritmo Cfs selecionou os seguintes atributos: Unit price, Quantity, Tax 5%, Total, Date, cogs, Rating e o atributo gross income (alvo da regressão). Os testes foram realizados e foi obtido uma boa melhora nas médias de erro. Para tentar melhorar mais os valores, foi analisado que a data (Date) e a avaliação (Rating) não têm relação com o lucro total, portanto esses dois atributos foram retirados e foi adicionado o Product Line, que pode ser um atributo relevante que o algoritmo julgou errado. Realizando os testes, novamente obtivemos uma melhora, dessa vez não tão acentuada quanto a

primeira. Para identificar cada seleção, iremos chamar a seleção feita pelo Cfs de “Seleção 1” e a outra de “Seleção 2”, os resultados foram os seguintes:

Tabela 12 – Resultados seleção de atributos

Algoritmo	MAE	RMSE
SMOReg (Poly)	0,3869	0,4907
Selecão 1	0,0268	0,0345
Seleção 2	0,0197	0,0259

Também foram realizados testes com o Relief, que ranqueia os atributos, para realizar os testes, foram utilizados todos os atributos que tem o peso >=0.

Tabela 13 – Resultados seleção de atributos

Algoritmo	MAE	RMSE
SMOReg (Poly)	0,3869	0,4907
Relief	0,0206	0,0293

Portanto a regressão feita deu muito certo, com erros baixíssimos, os melhores resultados foram os com a Seleção 2 e com o Relief. Para entender o quão baixo são essas médias de erro, a unidade de medida é o centavo de dólar já que estamos trabalhando com o lucro, logo, estamos falando de algo em torno de 2 centavos de média de erro.

4. REGRAS DE ASSOCIAÇÃO

4.1 O QUE É UMA REGRA DE ASSOCIAÇÃO

Uma regra de associação é uma técnica importante em mineração de dados e análise exploratória de dados. Ela revela padrões de relacionamento entre diferentes itens em conjuntos de dados. A regra de associação mais comum é expressa na forma de "se X, então Y", onde certos itens (ou conjunto de itens) estão associados a outros itens. As regras de associação são frequentemente quantificadas e avaliadas com base em métricas como confiança, suporte e lift, que serão explicadas posteriormente. Essas métricas ajudam a determinar a força e a relevância das associações descobertas.

4.2 A BASE DE DADOS

A base de dados escolhida para realizar a tarefa de regras de associação foi uma relacionada a distúrbios do sono. A base contém 374 instâncias e 13 atributos.

Quadro 3 – Dicionário de Dados dos campos presentes na base utilizada

ATRIBUTO	DESCRIÇÃO
Person ID	Número identificador para cada pessoa.
Gender	Gênero da pessoa.
Age	Idade da pessoa em anos.
Occupation	Ocupação ou profissão da pessoa.
Sleep Duration	Duração do sono em horas.
Quality of Sleep	Uma classificação subjetiva da qualidade do sono, variando de 1 a 10.
Physical Activity Level	O número de minutos que a pessoa pratica atividade física diariamente.
Stress Level	Uma classificação subjetiva do nível de estresse vivenciado pela pessoa, variando de 1 a 10.
BMI Category	A categoria de IMC da pessoa (por exemplo, Abaixo do Peso, Normal, Sobrepeso).
Blood Pressure	A medição da pressão arterial da pessoa, indicada como pressão sistólica sobre pressão diastólica.
Heart Rate	A frequência cardíaca em repouso da pessoa em batimentos por minuto.
Daily Steps	O número de passos que a pessoa dá por dia.
Sleep Disorder	A presença ou ausência de um distúrbio do sono na pessoa (Nenhum, Insônia, Apneia do Sono).

4.3 CONFIANÇA

Refere-se à probabilidade condicional de que a presença de um item em uma transação esteja associada à presença de outro item. Em outras palavras, a confiança mede a força de uma relação entre os itens em uma regra de associação. A confiança é a proporção de vezes em que a regra de associação $X \rightarrow Y$ é verdadeira em relação ao número de vezes em que o conjunto X aparece na base de dados.

4.4 SUPORTE

O suporte é uma métrica importante em regras de associação que mede a frequência com que um determinado item ou conjunto de itens ocorre na base de dados. Em outras palavras, o suporte indica a proporção de ocorrências de um determinado item ou conjunto de itens aparecendo na base de dados.

4.5 LIFT

O lift é uma medida de quanto mais frequentemente os itens ocorrem juntos do que seria esperado se eles fossem independentes um do outro. Um valor de lift maior que 1 indica que os itens ocorrem juntos com mais frequência do que o esperado ao acaso.

4.6 CONVERSÃO

A conversão é uma métrica que mede a probabilidade de que uma regra seja válida. Ela é calculada como (1 - confiança) / (1 - suporte). Um valor alto de conversão indica que a regra é mais provável de ser útil e verdadeira.

4.7 EXECUÇÃO DO ALGORITMO APRIORI

Para obter as regras de associação, foi utilizado o algoritmo Apriori, antes da execução foi preciso preparar a base de dados com o filtro Discretize que transforma os atributos numéricos em um atributo nominal que trabalha com intervalos.

Na execução do algoritmo o único parâmetro alterado foi o “numRules” que diz o número de regras a serem encontradas, o número do parâmetro foi modificado para 10000. Depois foi realizado um processo de filtragem a fim de encontrar as melhores regras de associação.

4.8 REGRAS ENCONTRADAS

Tabela 14 – Regras de associação obtidas com o algoritmo Apriori

Regra	Confiança	Lift
{Sleep Duration='(7.6-inf)', Stress Level='(-inf-3.5]'} ⇒ {Quality of Sleep='(8.5-inf)'}	1,00	5,27
{Stress Level='(7.5-inf)'} ⇒ {Sleep Duration='(-inf-6.7]'}	1,00	2,54
{Blood Pressure=140/95} ⇒ {Sleep Disorder=Sleep Apnea}	0,91	4,35
{Stress Level='(-inf-3.5]'} ⇒ {Quality of Sleep='(8.5-inf)'}	0,97	5,12
{BMI Category=Normal, Heart Rate='(67.1-69.2]'} ⇒ {Quality of Sleep='(7.5-8]', Sleep Disorder=None}	0,93	3,45

Com as regras obtidas podemos identificar fatores de risco, como ocorreu na terceira regra, onde as pessoas com 140/95 de pressão arterial têm apneia do sono, uma pressão arterial considerada alta. Além dos fatores de risco, podemos identificar fatores que nos ajudam a melhorar o sono como demonstrado na primeira regra da tabela, onde as pessoas que dormem mais de 7,6 horas por dia e têm um nível de estresse inferior a 3,5 tendem a ter uma melhor qualidade do sono.

5. CLUSTERIZAÇÃO

5.1 O QUE É CLUSTERIZAÇÃO

A clusterização, também conhecida como análise de agrupamento, é uma técnica de aprendizado não supervisionado em machine learning e mineração de dados. O objetivo da clusterização é agrupar um conjunto de objetos em subconjuntos, ou clusters, onde os objetos dentro de um mesmo cluster são mais semelhantes entre si do que com os objetos de outros clusters, de acordo com algum critério de similaridade ou dissimilaridade.

5.2 A BASE DE DADOS

A base utilizada para o algoritmo de clusterização foi a base sobre vendas em supermercados, a mesma utilizada na tarefa de regressão. Com essa base podemos, por exemplo, agrupar clientes que podem ter interesses parecidos. Dividir os clientes de supermercado em clusters por meio de técnicas de clusterização pode ser extremamente interessante por motivos como: segmentação de mercado, personalização de produtos, estratégias de precificação, gerenciamento de inventário, identificação de tendências e oportunidades de mercado, fidelizar e reter clientes.

5.3 MÉTODO DE ELBOW

O método de Elbow, em clusterização, é uma técnica utilizada para determinar o número ideal de clusters em um conjunto de dados. O termo "elbow" (cotovelo, em inglês) refere-se à aparência de um gráfico de "número de clusters" versus “Total within sum of squares”, onde o ponto de inflexão se assemelha a um cotovelo.

A TWSS representa a soma das distâncias quadradas de cada ponto para o centróide do cluster ao qual ele foi atribuído. Em outras palavras, mede o quão compactos são os clusters. Quanto menor for a TWSS, mais compactos são os clusters e, idealmente, melhor é a clusterização e maior o conhecimento que poderemos adquirir com estes grupos.

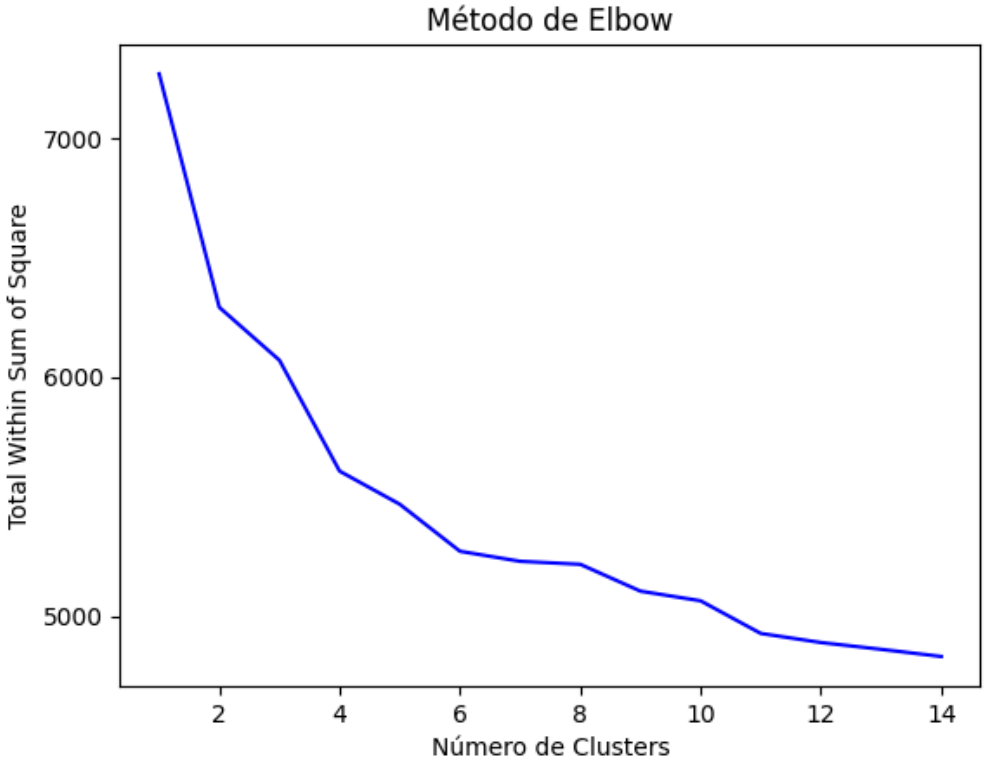
5.4 EXECUÇÃO DO ALGORITMO SIMPLEKMEANS

O algoritmo SimpleKMeans é uma implementação do algoritmo K-Means no Weka e é utilizado para fazer clusterização.

Tabela 15 – Resultados SimpleKMeans

Clusters	TWSS
1	7268,03
2	6291,67
3	6070,20
4	5606,51
5	5468,08
6	5271,44
7	5229,42
8	5216,97
9	5104,65
10	5064,33
11	4927,85
12	4889,80
13	4861,38
14	4831,50

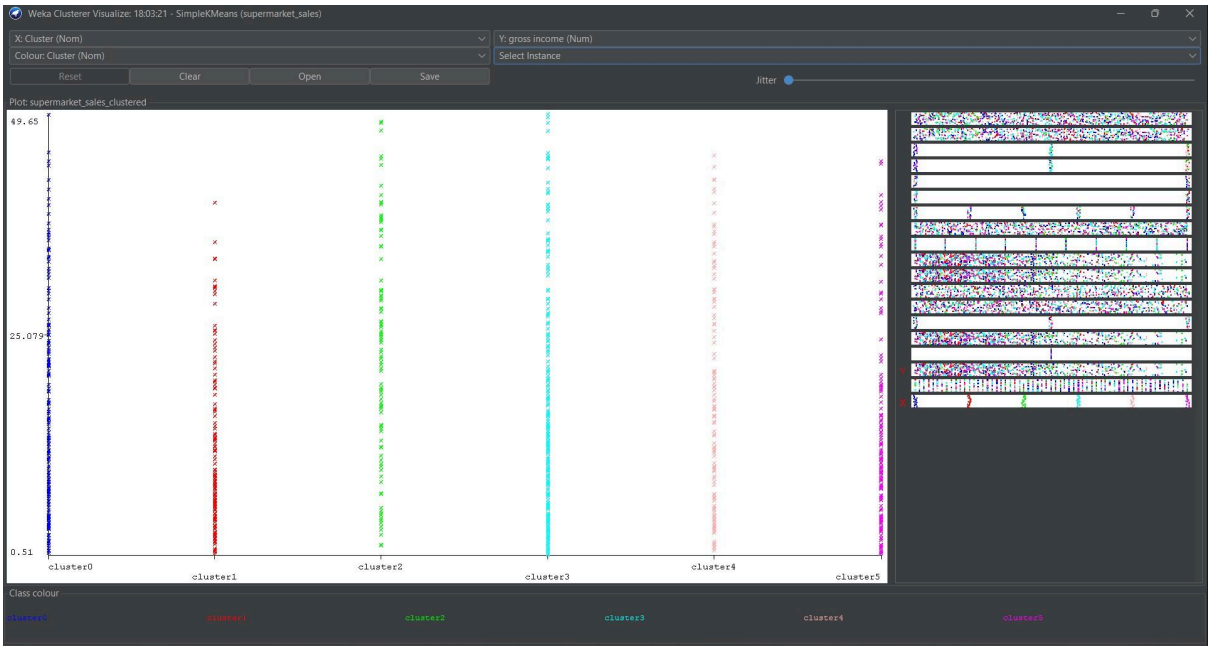
Figura 13 – Resultados obtidos com o algoritmo SimpleKMeans



Fonte: Resultados obtidos no WEKA.

Analisando o gráfico acima, vemos que o número ideal de clusters está no intervalo [4-6], a escolha desse número depende da avaliação de quem vai utilizar os dados, pois deve ser avaliado se o número de clusters faz sentido no domínio do problema, além de que as vezes um número menor de grupos pode ser melhor, por ser menos custoso, mais fácil de interpretar, mais útil na prática. Porém esse não foi o caso, pois existem 6 categorias de produto na base, então o melhor número de clusters para a base seria 6, o gráfico mais interessante foi o que representa o lucro total vs clusters, onde os grupos tinham números máximos diferentes.

Figura 14 – Gráfico Lucro Total x Clusters



6. CONCLUSÃO

Diante das comparações feitas e dos algoritmos executados, vemos o quão abrangente é a mineração de dados que nos dá conhecimento para tomadas de decisões baseadas em bancos de dados gigantes, nos livrando das análises intuitivas que por muitas vezes podem estar incorretas. A mineração de dados complementa as análises intuitivas ao fornecer insights mais profundos, identificar padrões complexos e permitir a análise de grandes volumes de dados de forma eficiente e baseada em evidências.

A mineração de dados é uma disciplina multidisciplinar amplamente aplicada em diversas áreas. Nos negócios e marketing, a mineração de dados é essencial para análise de mercado, segmentação de clientes, previsão de vendas, análise de tendências, recomendação de produtos e otimização de campanhas de marketing. Na área da saúde, essa técnica é utilizada na análise de registros médicos eletrônicos, diagnóstico assistido por computador, detecção de fraudes em seguros de saúde, pesquisa médica e descoberta de padrões em grandes conjuntos de dados de saúde. No setor financeiro, a mineração de dados é crucial para a detecção de fraudes em transações financeiras, previsão de riscos de crédito, modelagem de séries temporais para previsão de mercado, detecção de padrões de investimento e análise de sentimentos em dados financeiros. Em resumo, a mineração de dados é uma ferramenta poderosa para descobrir insights valiosos e tomar decisões informadas em uma variedade de domínios.