

**Engenharia de Computação**

**Iuri Ramon Cervantes Santos**

Relatório : Mineração de dados - Algerian Forest Fires Daset.

Birigui-SP

2021

**Introdução**

Este trabalho tem como objetivo aplicar os conhecimentos adquiridos na disciplina de Mineração de Dados em uma base de dados afim de extrair as informações nela contida e resolver o problema proposto pela base de dados.

1. **Apresentação da base de dados**

A base utilizada neste trabalho foi a Algerian Forest Fires Daset.

Esta base de dados apresenta informações de uma região florestal da Algéria, do período de junho de 2012 a setembro de 2012, a base de dados possui 11 atributos mais 1 atributo de saída (classe).

A base possui 122 objetos, classificados em “ocorreu incêndio” e “não ocorreu incêndio”.

**Atributos:**

* Data: Dia, mês e ano das informações coletadas;
* Temperatura: Temperatura máxima no dia em graus Celsius (22°C - 42°C);
* Humidade relativa: Humidade relativa no dia em porcentagem (21% - 90%);
* Velocidade do vento: velocidade do vento em km/h (6km/h – 29km/h);
* Chuva: total de chuva no dia em milímetros (0mm – 16,8mm);
* FFMC: Índice de humidade de combustíveis finos, representa o teor de humidade dos combustíveis finos mortos, de secagem rápida, constituindo um bom indicador do seu grau de inflamabilidade. (28,6 – 92,5; índice FWI);
* DMC: Índice de húmus, representa o teor de humidade de uma fina e pouco compactada camada de solo (até cerca de 8 cm de profundidade), dando a indicação do estado da matéria orgânica decomposta (húmus) e materiais lenhosos de tamanho médio que aí se encontram. (1,1 – 65,9; índice FWI);
* DC: Índice de Seca, representa o teor de humidade dos combustíveis florestais (húmus e materiais lenhosos de maiores dimensões), que se encontram abaixo da superfície do solo, entre 8 e 20 cm de profundidade e é um bom indicador dos efeitos do período seco sazonal. (7 – 220,4; índice FWI);
* ISI: Índice de Propagação Inicial, resulta da combinação do FFMC e da intensidade do vento, representando a taxa de propagação inicial do fogo, sem incluir a influência de quantidades variáveis de combustível. (0 – 18,5; índice FWI);
* BUI: Índice de Combustível Disponível, resulta da combinação do DMC e do DC, representando a quantidade total de combustíveis disponível para propagação do fogo. (1,1 – 68; índice FWI);
* Risco Incêndio - Índice de Perigo de Incêndio (FWI). (0 – 31,1);
* Classe: 1 para incêndio, 0 para não incêndio.

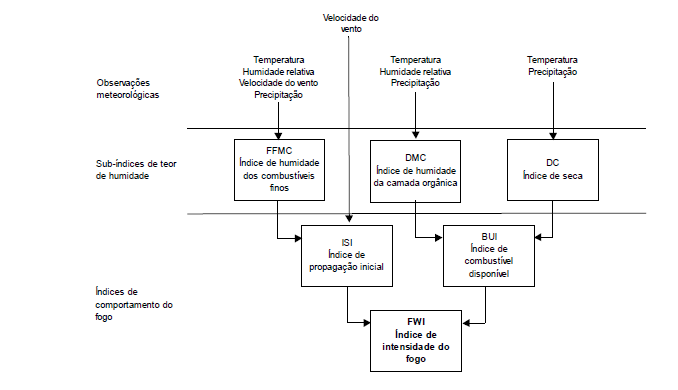


Figura 1 – Estrutura do índice meteorológico de risco de incêndio florestal FWI.

A base de dados apresentava dados faltantes em alguns atributos, tal problema é resolvido no pré-processamento, na fase de limpeza de dados.

1. **Pré-Processamento**

* Limpeza:

A limpeza da base de dados foi executada por meio do preenchimento dos valores faltantes, para tal foi considerada a média de cada atributo, média foi escolhida pois a sensibilidade aos valores extremos (desde que não sejam outliers) é importante para a detecção de um incêndio.

Seguem as informações sobre o procedimento:

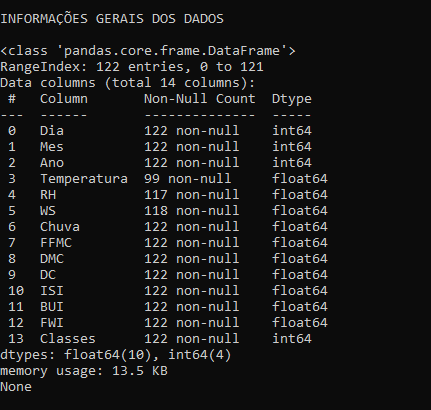


Figura 2 – Classificação geral dos atributos.

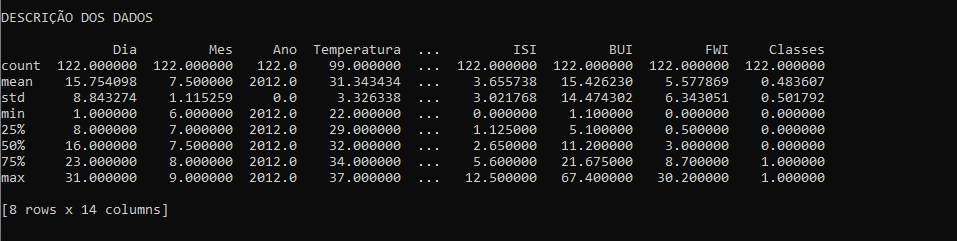


Figura 3 – Descrição dos dados antes da limpeza.

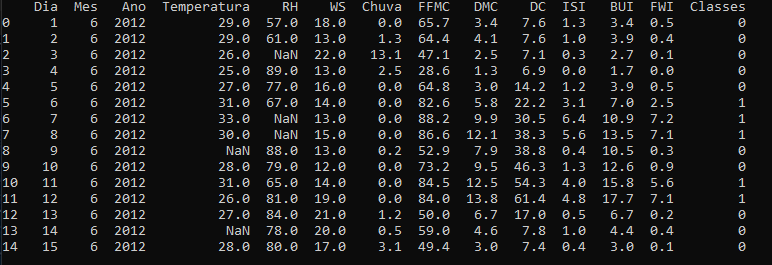


Figura 4 – 15 Linhas iniciais antes de executar a limpeza.

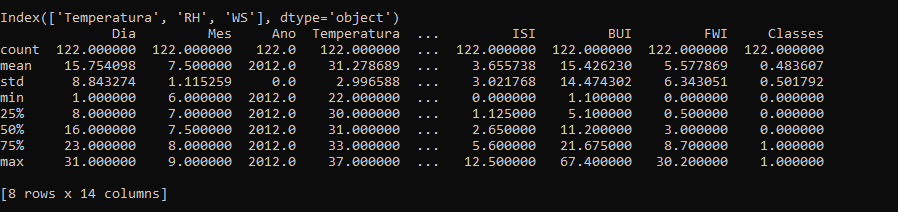


Figura 5 – Descrição dos dados pós limpeza.

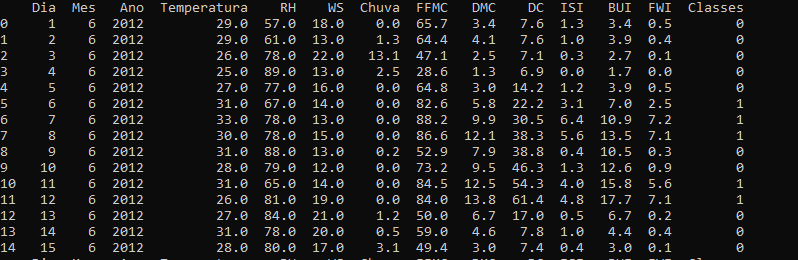


Figura 6 – 15 Linhas iniciais após executar a limpeza.

* Normalização:

Após a efetuação da limpeza dos dados, partiu-se para a etapa de normalização, para critério de estudo, foram implementados dois métodos diferentes de normalização, Z-score e Min-Max, que depois acompanharão a base de dados nos procedimentos seguintes, e estes foram os resultados:

Normalização Z-Score

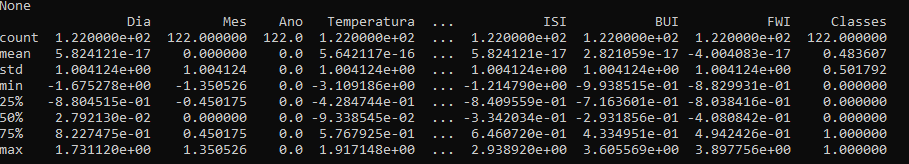


Figura 6 – Dados da normalização Z-score.

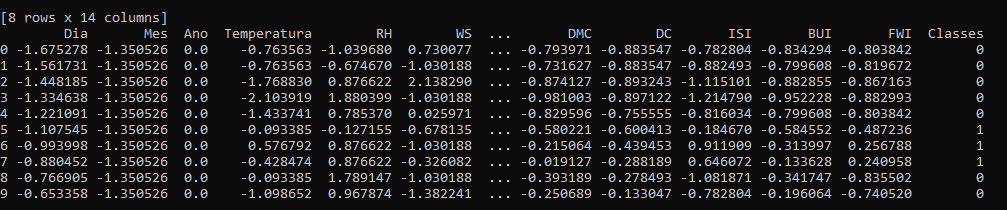


Figura 7 – Dados normalizados (Z-Score).

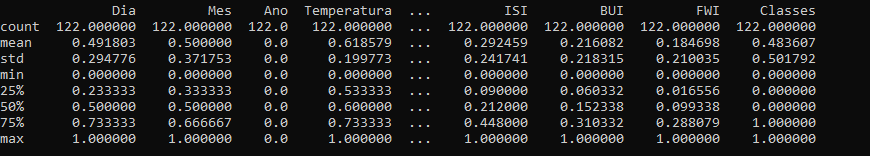


Figura 8 – Dados da normalização Min-Max.

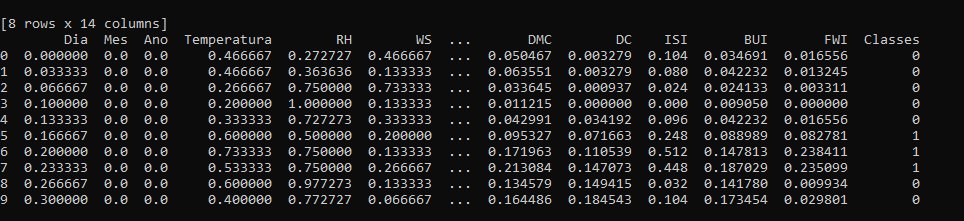


Figura 9 – Dados normalizados (Min-Max).

* Redução:

Para o processo de redução de dados, foram consideradas ambas as estratégias de normalização mencionadas anteriormente, e ambas apresentaram resultados diferentes quando expostas ao procedimento PCA, a variável alvo foi a Classe, responsável pela classificação do objeto em incêndio ou não.

* + Min-Max

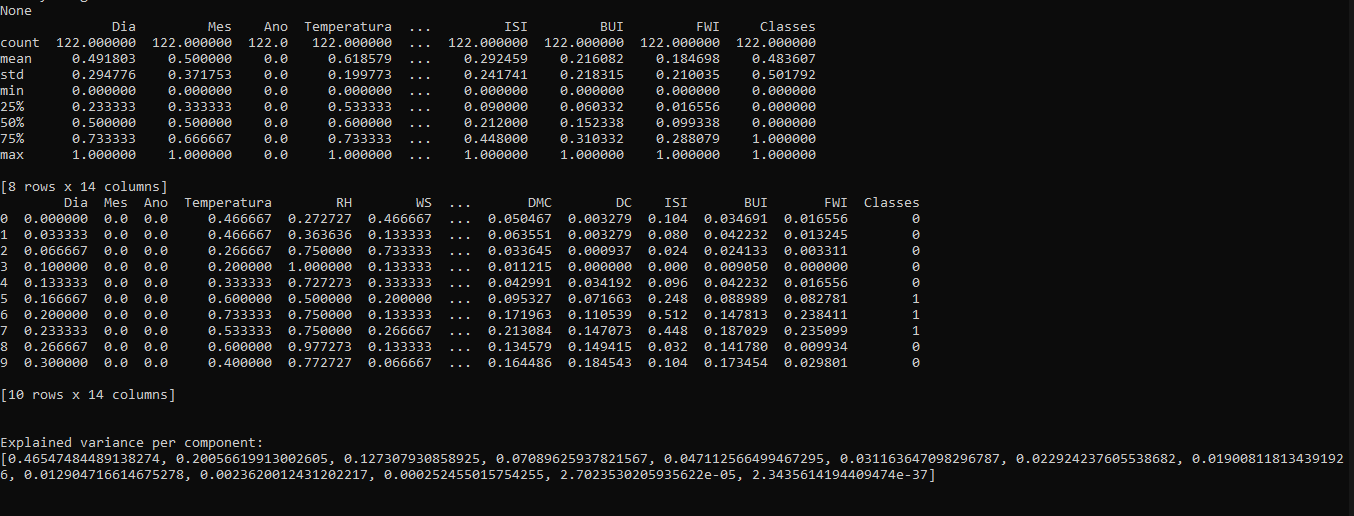


Figura 10 – Normalização Min-Max

Pela variância por componente podemos perceber que os atributos que apresentaram uma maior variância após a normalização foram BUI (Índice de Combustível Disponível) e FWI (Índice de risco de incêndio).

Após a normalização foi aplicada a técnica PCA, proporcionando os seguintes resultados:

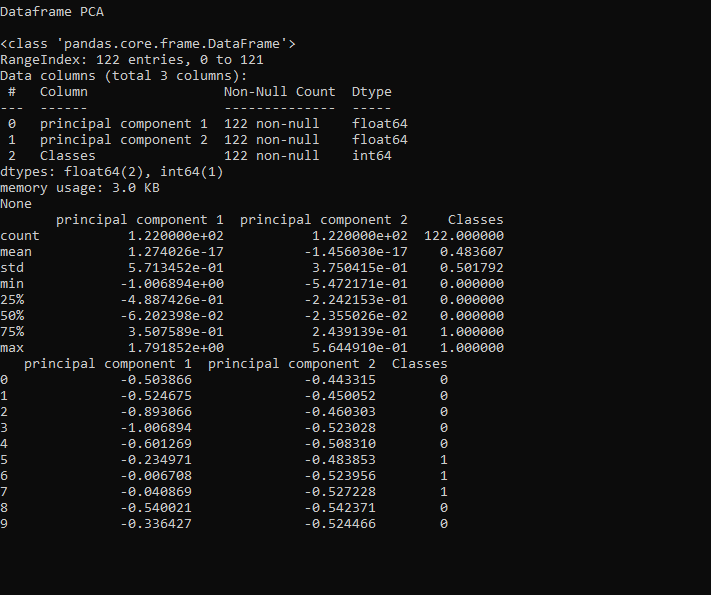


Figura 11 – PCA para Min-Max.

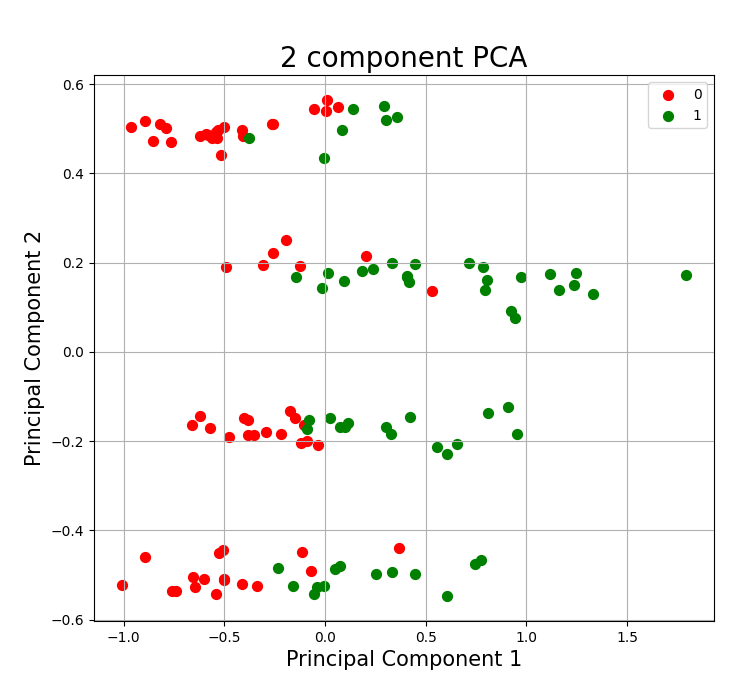


Figura 12- Plot de PCA para Min-Max.

Analisando estes resultados, é perceptível a divisão dos resultados, enquanto os valores classificados como 0 (não incêndio), se encontram em sua maioria com valores negativos para o componente 1, enquanto em relação ao componente 2, variam ao longo de todo o eixo.

* + Z-Score

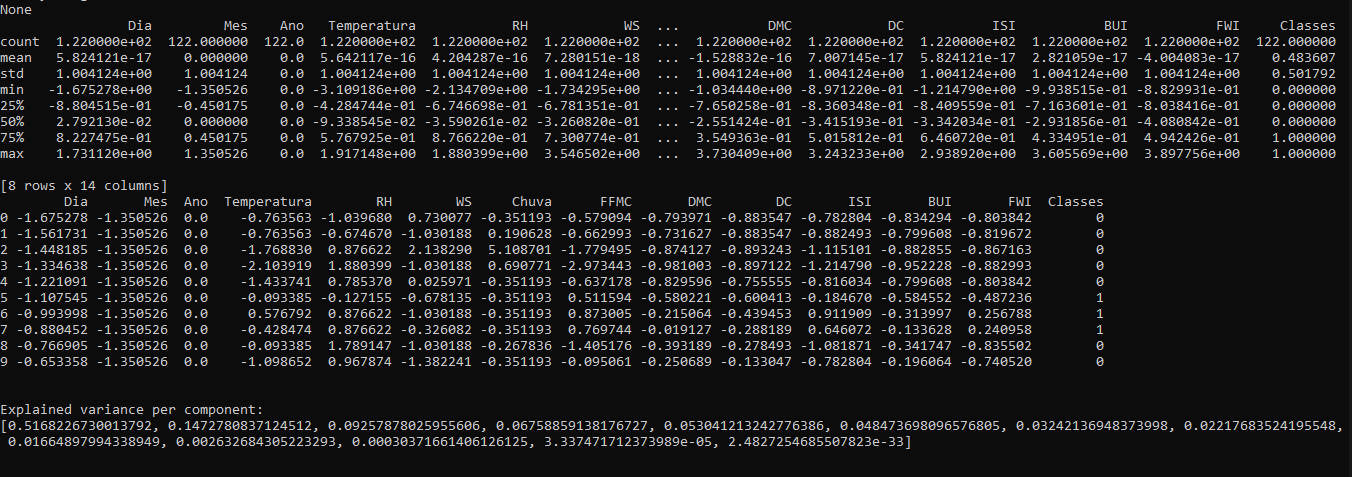


Figura 13 – Normalização Z-Score.

Pela variância por componente podemos perceber que os atributos que BUI (Índice de Combustível Disponível) e FWI (Índice de risco de incêndio), continuam sendo os que apresentaram uma variância maior do que os outros atributos.

Após a normalização foi aplicada a técnica PCA, proporcionando os seguintes resultados:

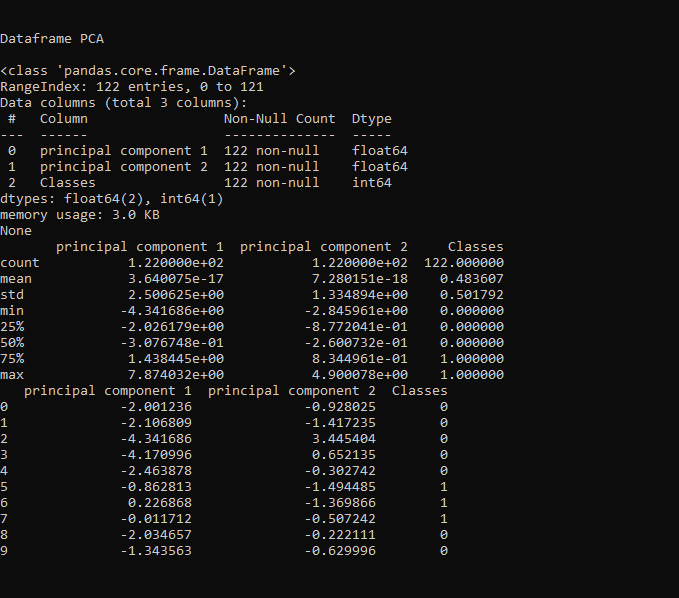


Figura 14 – PCA para Z-Score.

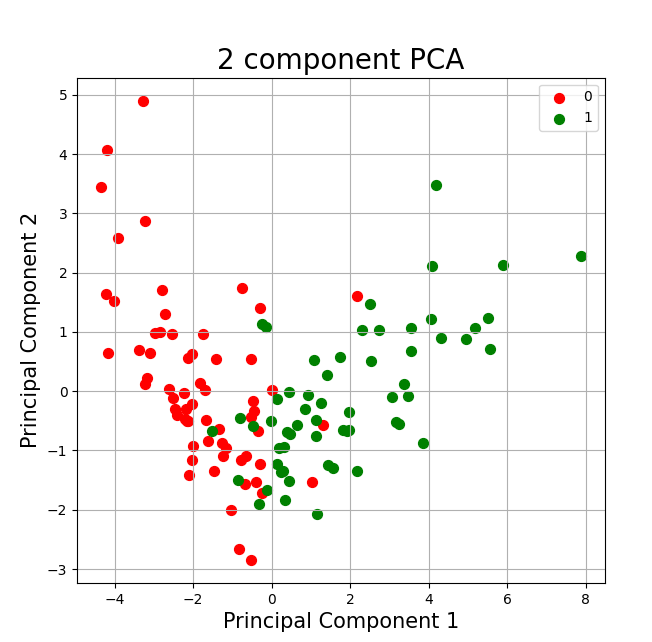


Figura 15 – Plot de PCA para Z-Score.

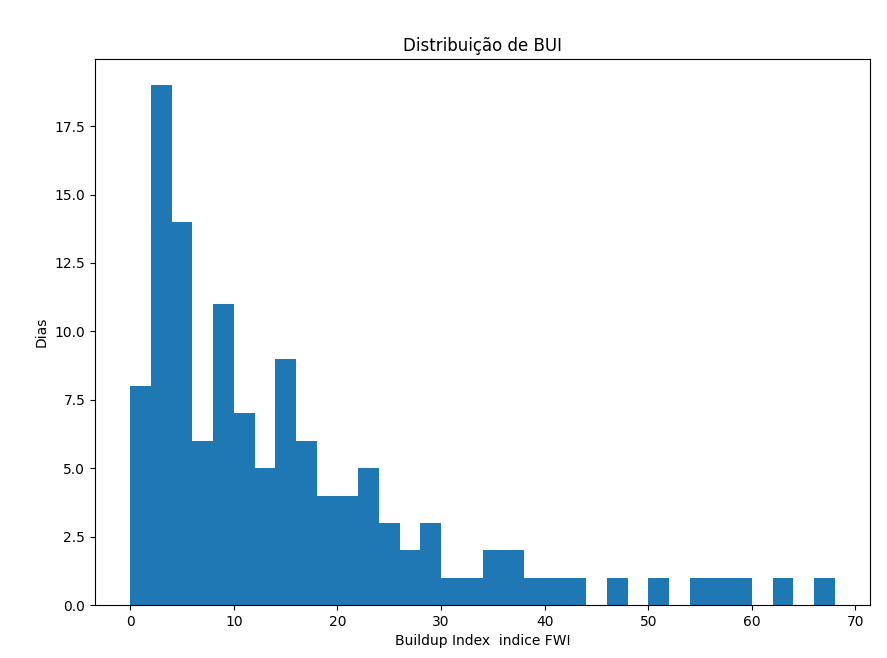
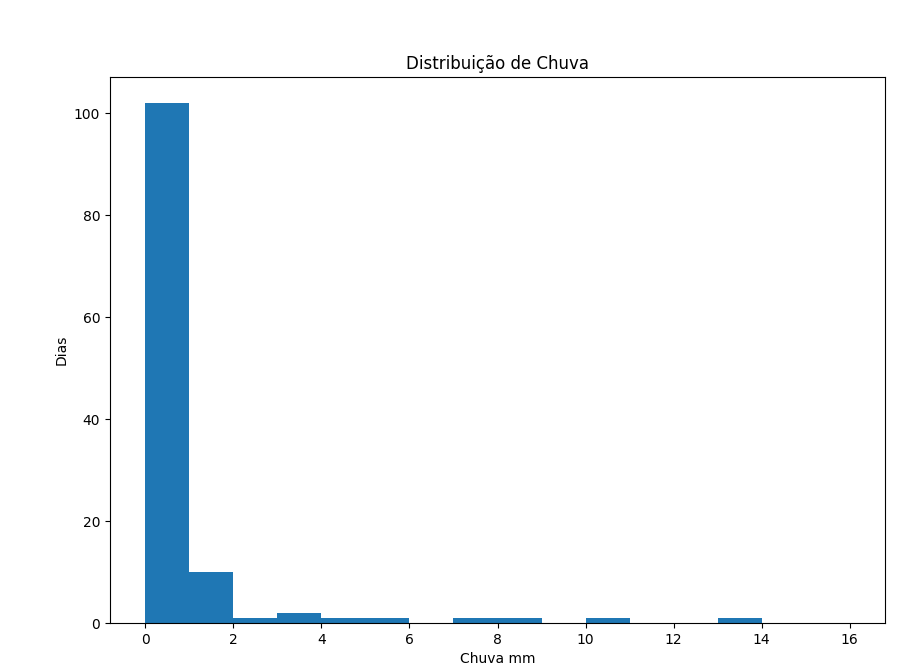
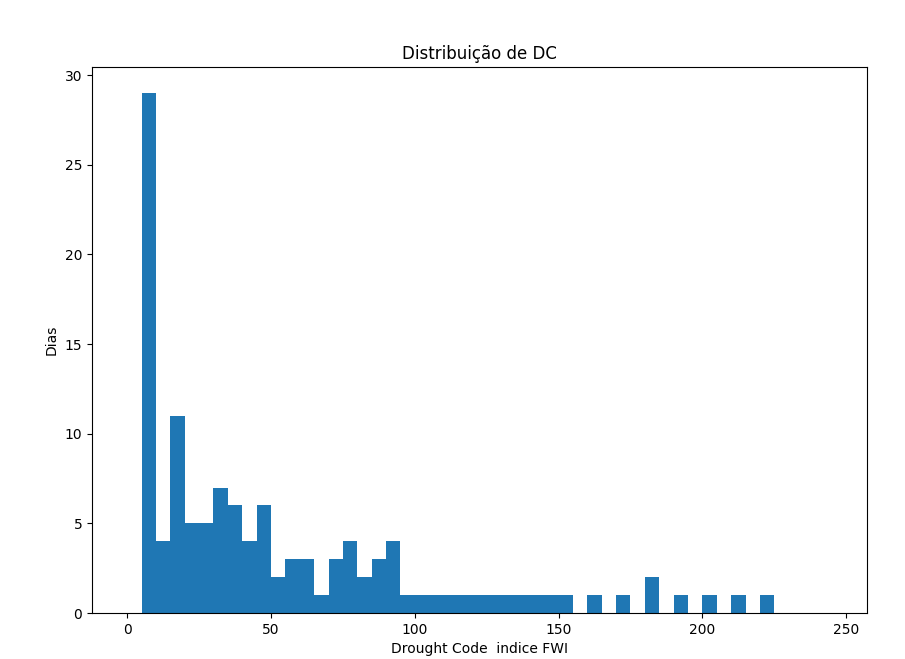
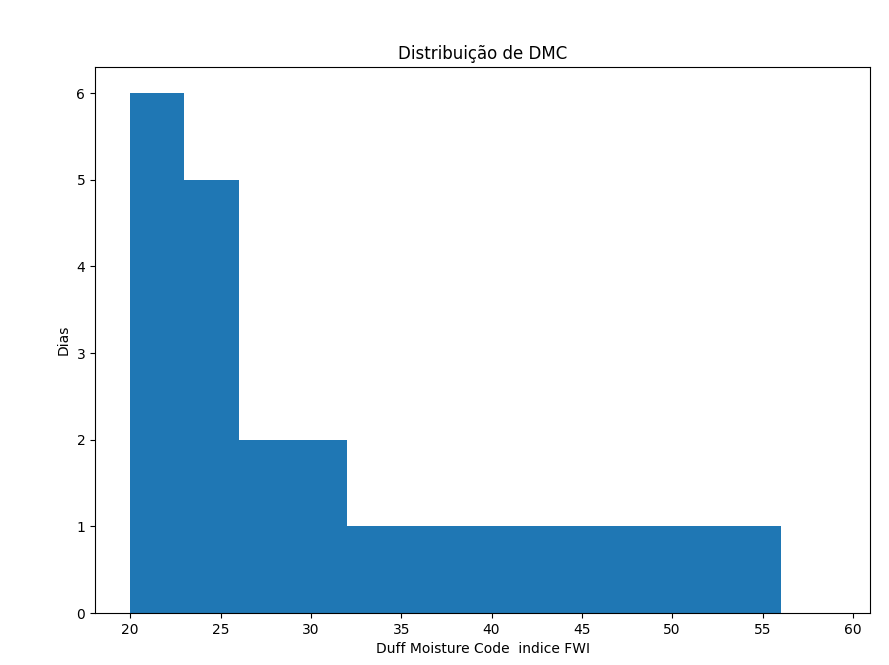
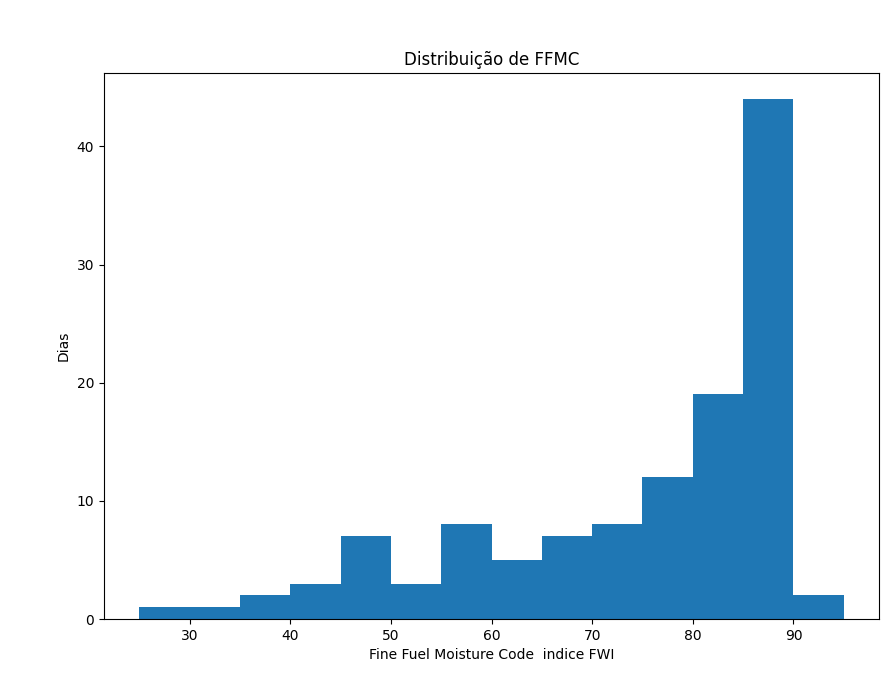
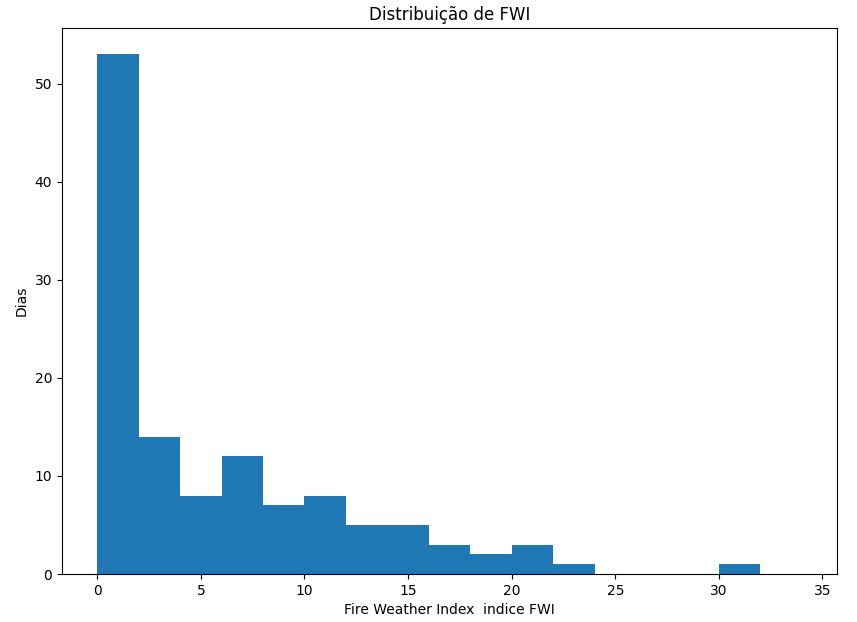
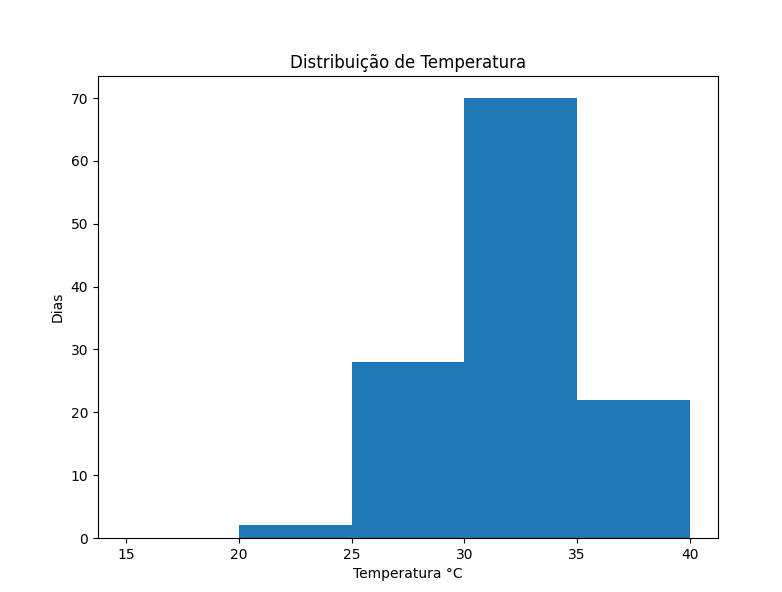
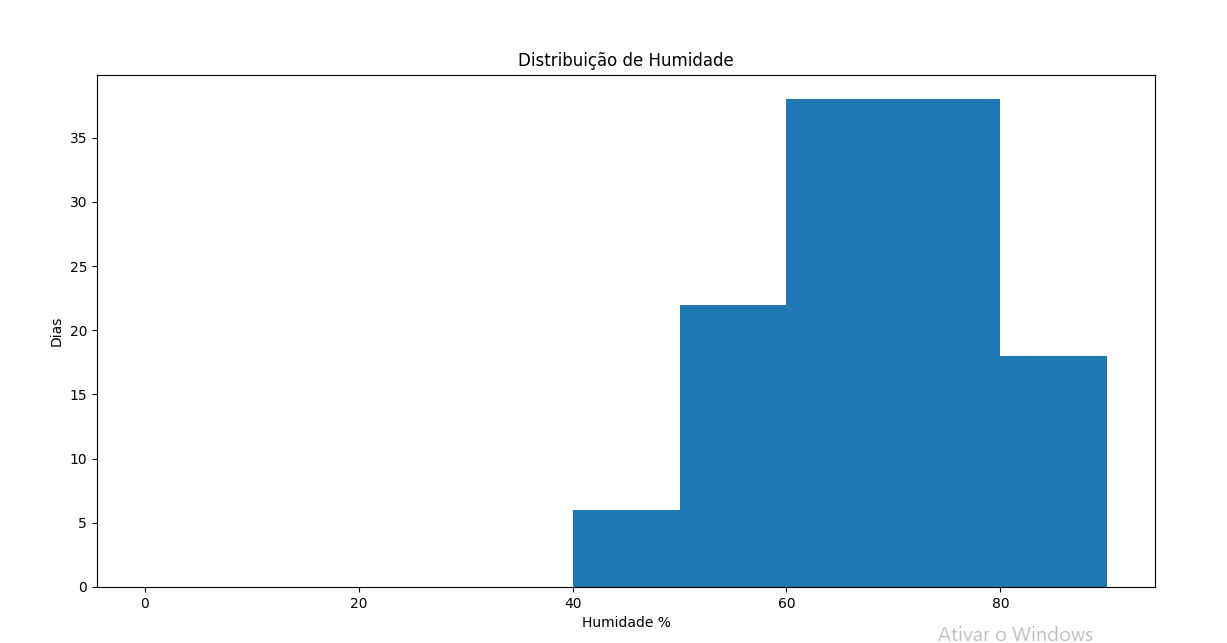
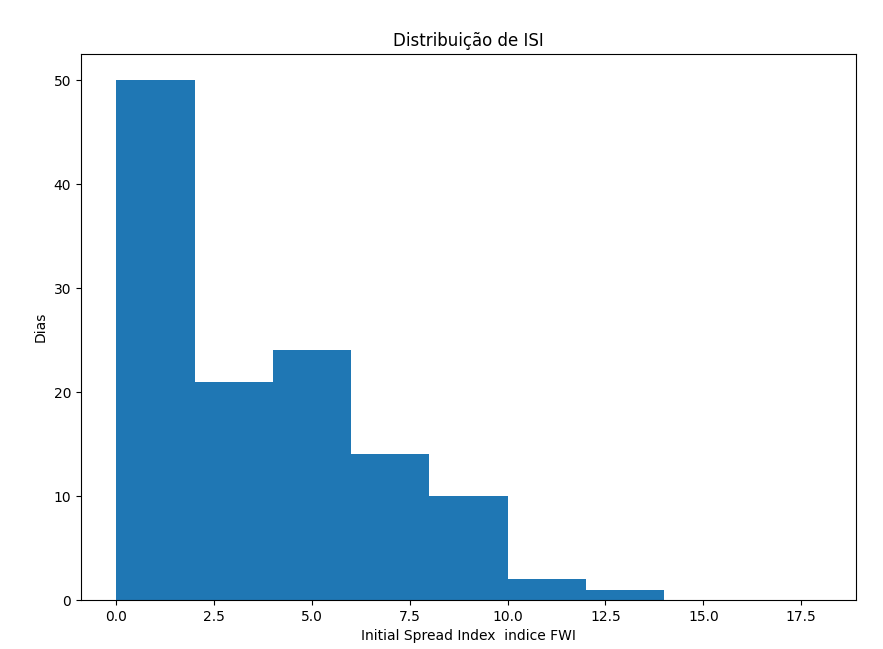
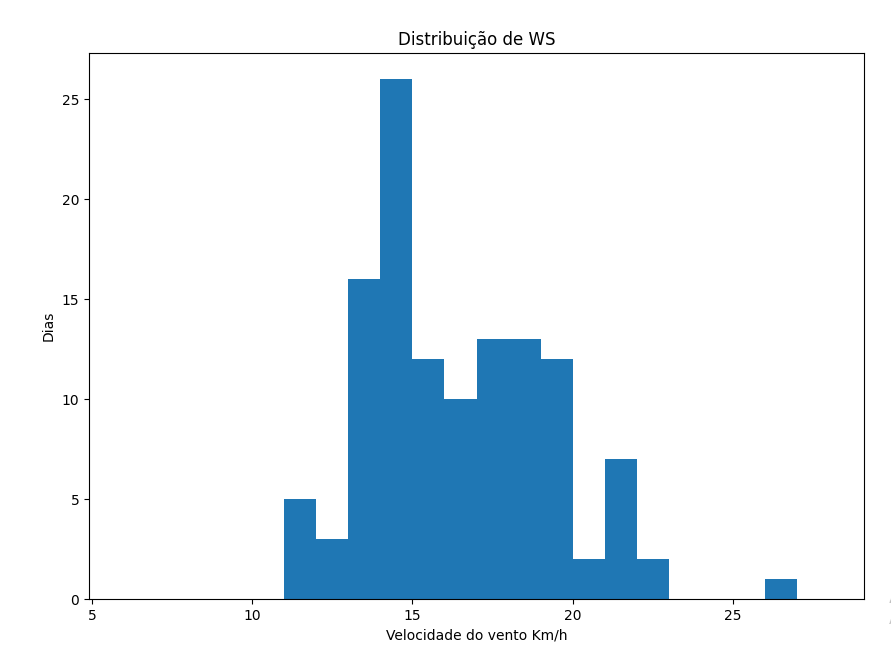
Analisando os resultados obtidos podemos perceber que a normalização Z-Score proporciona uma distribuição diferente dos pontos, enquanto no gráfico de PCA para Min-Max, podemos notar um agrupamento em patamares em relação ao componente 2, na normalização Z-Score temos os pontos mais distribuídos por esse eixo.

Porém o padrão de classificação permaneceu o mesmo, 0 (não incêndio), assumindo valores negativos para o componente 1, enquanto 1(incêndio) apresentou valores positivos para este componente.

1. **Análise descritiva dos dados**

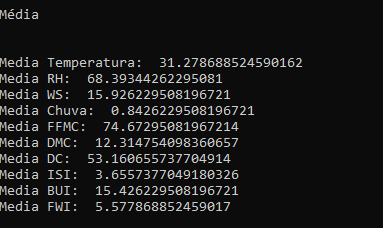
Para uma melhor análise, a base de dados foi exposta à diversas medidas de resumo:

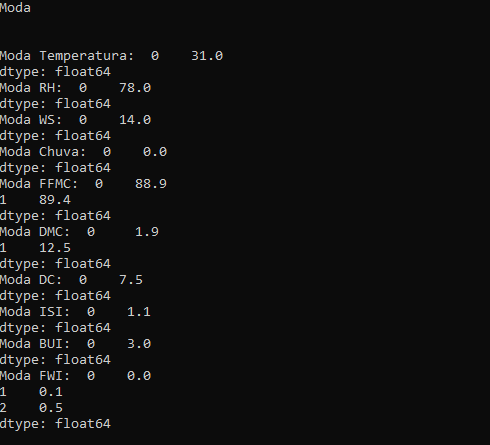
**Distribuição de frequência dos dados**

****

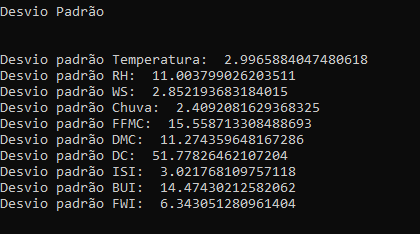
**Medidas de tendência central**

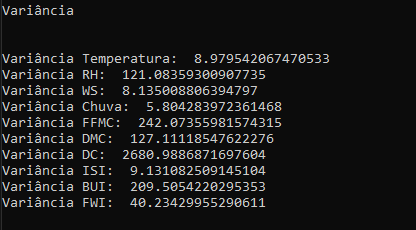
Para medidas de tendência central foram escolhidas média e moda, pois com a média podemos analisar como os extremos estão afetando a medição em relação à moda.



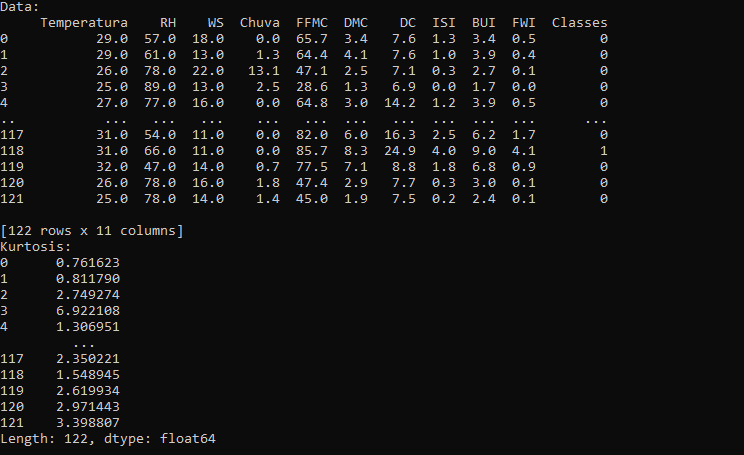


**Medidas de dispersão**

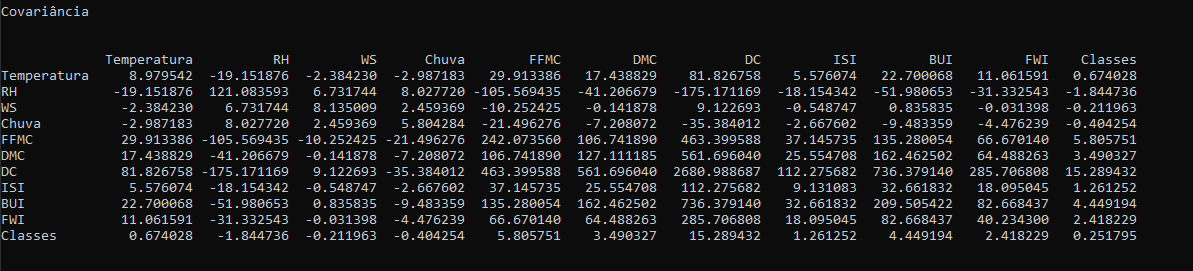
****

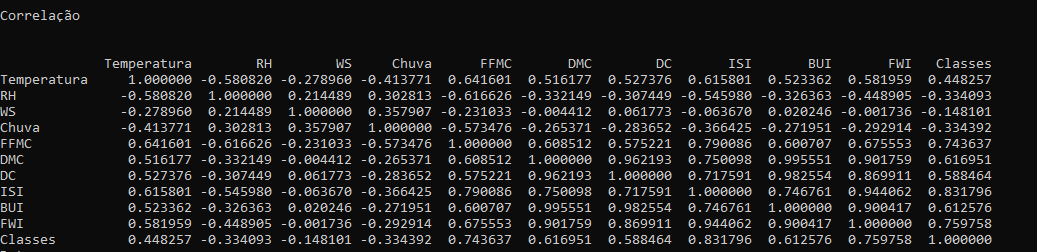
****

**Medidas de posição relativa**

****

**Medidas de associação**

****

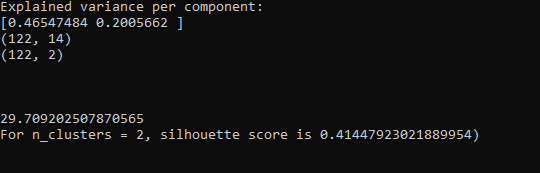
****

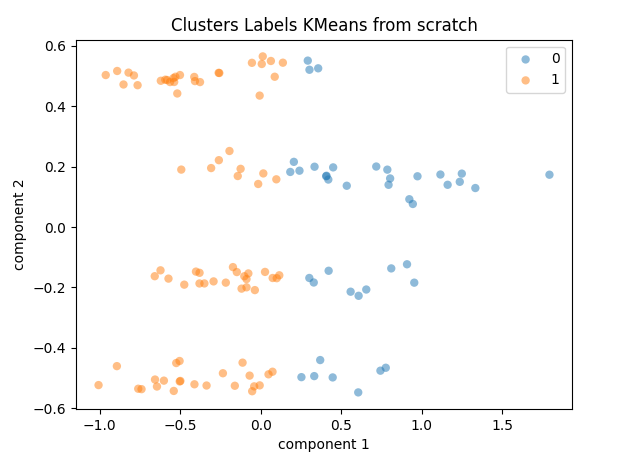
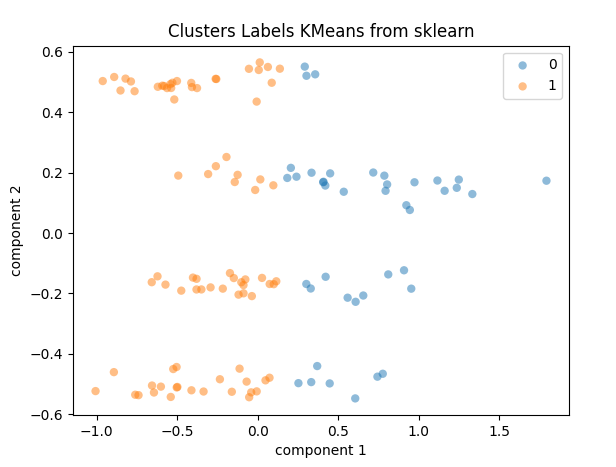
1. **Análise de grupos**

Para o agrupamento, foi utilizado o algoritmo K-Means, este algoritmo foi aplicado na base de dados, tanto normalizada pelo método Min-Max, quanto pelo método Z-Score.

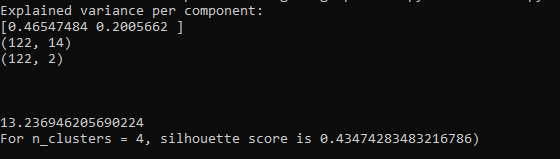
Para cada uma das normalizações o algoritmo foi rodado duas vezes, uma fazendo o agrupamento em dois grupos, e outro em quatro grupos (k = 2 e k = 4).

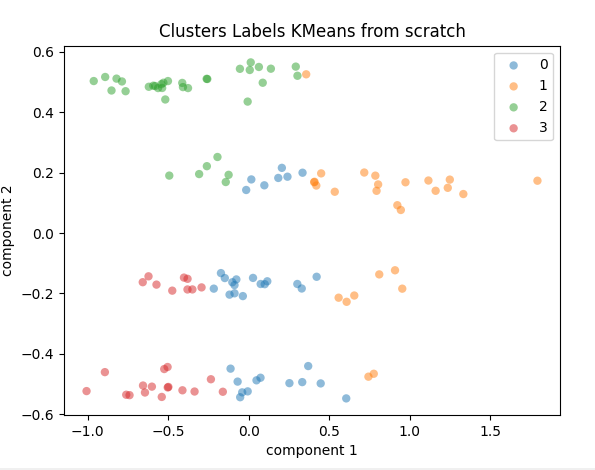
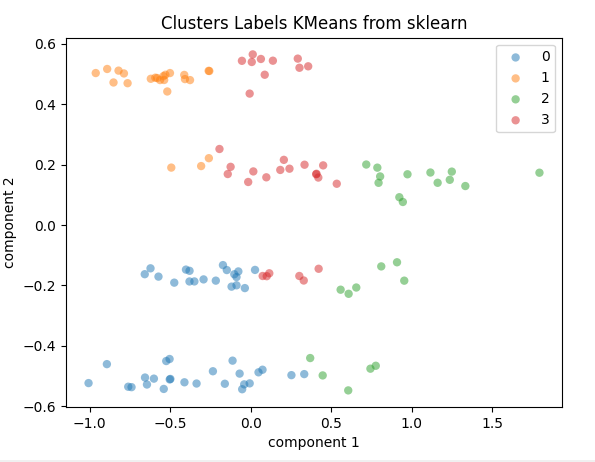
* K-Means com normalização MinMax, com k = 2



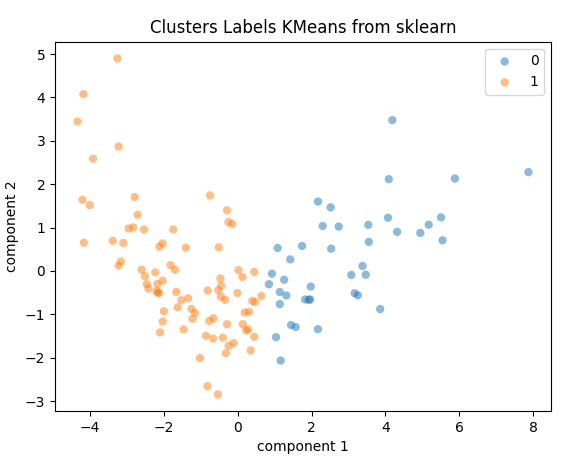
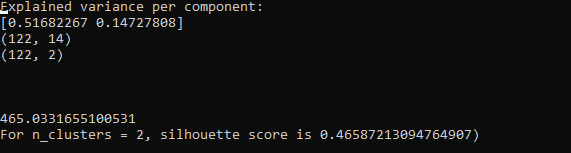


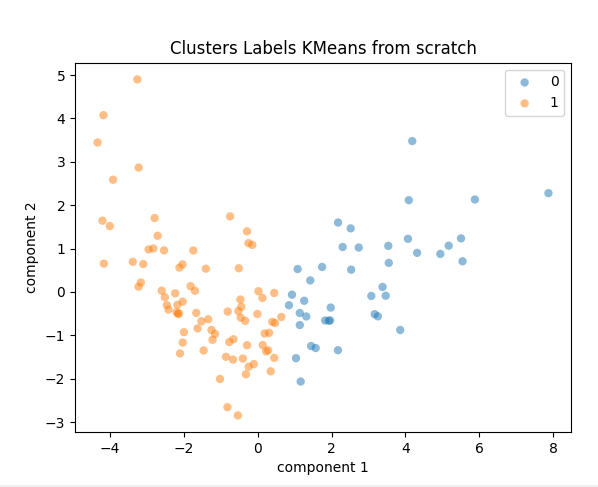
* K-Means com normalização MinMax, com k=4.



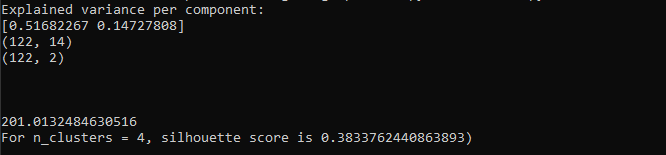


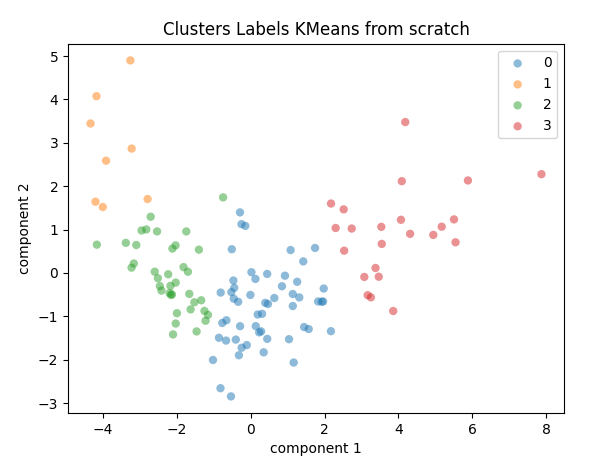
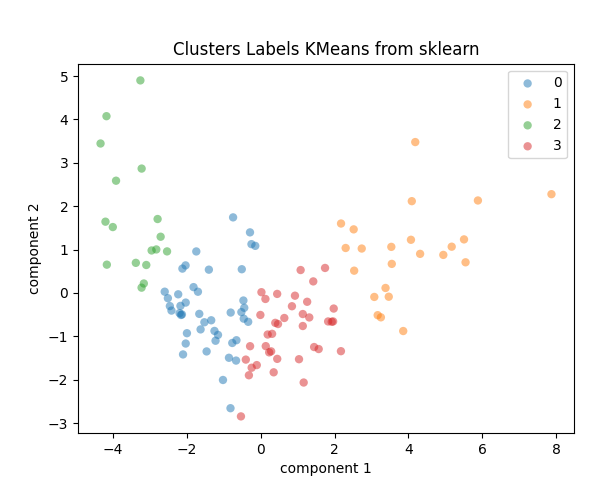
* K-Means com normalização Z-Score, com k=2.





* K-Means com normalização Z-Score, com k=4.



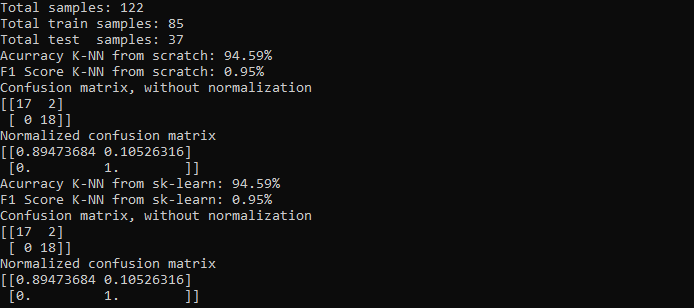


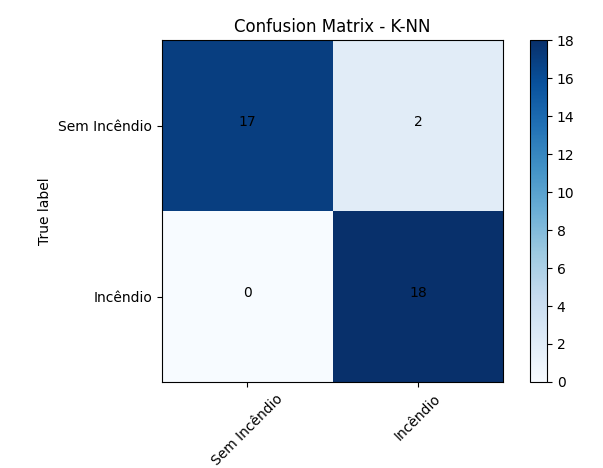
Após a análise dos resultados, é possível observar que a execução que mostrou uma maior efetividade, dada pelo valor de silhueta apresentado (0,46), foi a normalização Z-Score com dois grupos (k=2).

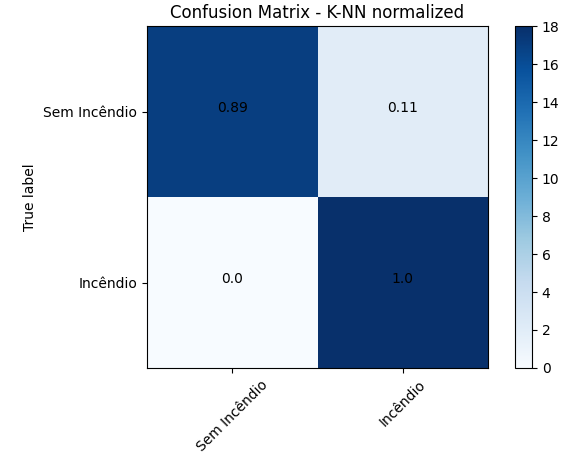
1. **Classificação**

Para a classificação foi utilizado o método de normalização Z-Score, pois o mesmo apresentou melhores resultados nas etapas realizadas previamente.

**KNN**



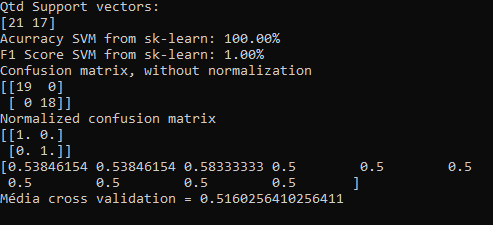


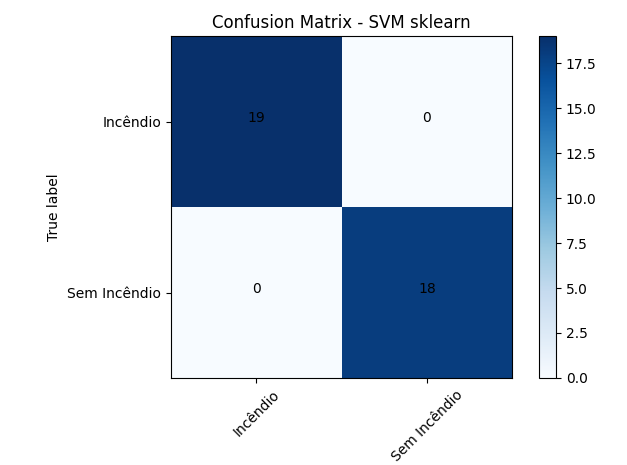


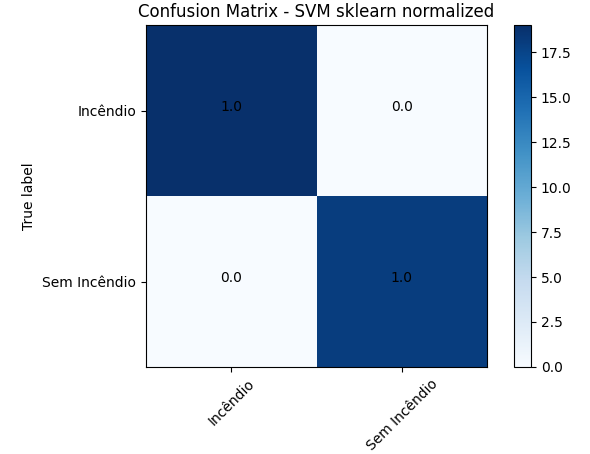
Como foi possível observar nos dados obtidos, a matriz de confusão gerada utilizando sklearn, para k=5, apresentou a mesma acurácia na predição.

**SVM**









Como foi possível observar nos dados obtidos, o algoritmo obteve acurácia de 100% dos testes de predição de incêndios feitos, utilizando o kernel = rfb como classificador SVM.

1. **Discussão**

Como podemos comparar durante todo os processos realizados durante o trabalho, alguns métodos se sobressaíram a outros, para o tratamento desta base de dados específica, para este problema tivemos uma vantagem do método de normalização Z-Score, sobre o Min-Max, quando aplicado o agrupamento K-Means.

Outra comparação que podemos fazer é a da acurácia fornecida pelo método SVM, que alcançou 100% de acurácia nas predições, quando configurado com kernel=rfb, enquanto o KNN, com k=5 (n° de objetos próximos para comparação), alcançou 94,59 %.