**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS  
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**

**Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data**

**Romulos Machado**

**Phillip Furtado**

**Análise de dados de temperatura ambiente versus logs de alerta em UM Data Center**

Belo Horizonte

2020

**Romulos Machado**

**Phillip Furtado**

**Análise de dados de temperatura ambiente versus logs de alerta em UM Data Center**

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte

2020

**SUMÁRIO**

[1. Introdução 4](#__RefHeading___Toc633_3942287182)

[1.1. Contextualização 4](#__RefHeading___Toc635_3942287182)

[1.2. O problema proposto 4](#__RefHeading___Toc637_3942287182)

[2. Coleta de Dados 6](#__RefHeading___Toc639_3942287182)

[3. Processamento/Tratamento de Dados 7](#__RefHeading___Toc641_3942287182)

[4. Análise e Exploração dos Dados 11](#__RefHeading___Toc643_3942287182)

[5. Criação de Modelos de Machine Learning 11](#__RefHeading___Toc645_3942287182)

[6. Apresentação dos Resultados 11](#__RefHeading___Toc647_3942287182)

[7. Links 12](#__RefHeading___Toc649_3942287182)

[REFERÊNCIAS 13](#__RefHeading___Toc651_3942287182)

# 1. Introdução

## 1.1. Contextualização

Para que data centers e seus respectivos equipamentos (servidores, grandes unidades de armazenamento e ativos de rede) consigam atender à necessidade dos seus usuários, é imperativo que permaneçam em funcionamento constante 24 horas por dia, 7 dias por semana. Consequentemente, assim como a maioria dos equipamentos que trabalham com processamento de informações, a temperatura nos data centers está diretamente relacionada com o bom funcionamento dos servidores, sendo fundamental que eles não superaqueçam durante a execução de suas atividades. Pois, caso isso aconteça, estes equipamentos correm o risco de se auto desligarem ou queimarem, o que pode causar uma parada não programada do data center, consequentemente a interrupção dos serviços por ele provido. Com isso, se faz necessário uma análise estatística e preditiva dos eventos de elevação da temperatura de data centers e suas consequências no funcionamento dos equipamentos.

## 1.2. O problema proposto

No que concerne a alta disponibilidade de serviços, identificar se registros de eventos de alta temperatura dentro de um data center estão ou não relacionados a danos registrados através de logs, a partir de dados coletados sobre o próprio ambiente e usando conceitos de ciências de dados para analisá-los, ajuda não somente a entender a extensão de problemas em equipamentos, como também a definir sistematicamente que determinadas condições estabelecidas, possivelmente irão gerar impactos negativados decorrentes e, por isso, deve ser evitados ao máximo.

Para desenvolvedor este trabalho, foram utilizados dados de temperatura de ambiente registrados por sensores internos de 30 servidores Dell, de modelo R720, de um determinado data center de Manaus, entre os anos de 2014 a 2020, com o objetivo de compará-los com dados de problemas ou erros de hardware apresentados dentro desse mesmo período. Onde esta comparação determinará se existe correlação entre os eventos de superaquecimento do data center e os problemas registrados nos equipamentos, dentro de um mesmo intervalo de tempo.

# 2. Coleta de Dados

A base de dados é constituída por informações de 30 servidores de mesma especificação e característica de hardware e, partir deles, estudar dois tipos de logs de cada um desses servidores, que foram armazenados ao longo de aproximadamente 04 anos, através de um dispositivo chamado de Integrated Dell Remote Access Controller (IDRAC). Esse componente também instalado dentro de cada um desses servidores com total independência de acesso e isolado do sistema operacional, permitiu que ficassem registrados vários alertas relacionados às mais diversas falhas relativas ao hardware dessas máquinas, assim como também registrou dados de temperatura externa a estes servidores, ao longo desses 04 anos.

Sendo assim, como foi indicado no parágrafo anterior, tivemos a oportunidade de trabalhar com dois tipos de logs:

***-*** ***Log de Alertas (log\_servidorXX.csv)***: contendo dados sobre alertas de falhas do equipamento, tendo a seguinte estrutura:

|  |  |  |
| --- | --- | --- |
| **Nome da coluna/atributo** | **Descrição** | **Tipo** |
| Severity | Indica qual a severidade do alerta (“Normal”, “Warning”,” Critical”) | Texto |
| Date | Indica a data e horário do alerta | Data/horário |
| Description | Indica a mensagem de alerta | Texto |

***- Log de Temperaturas (servidorXX\_inlettemp.csv)***: contendo dados de temperaturas registrados de hora em hora dos equipamentos, tendo a seguinte estrutura:

|  |  |  |
| --- | --- | --- |
| **Nome da coluna/atributo** | **Descrição** | **Tipo** |
| Average | Indica a temperatura média em Celsius, dentro do intervalo de tempo. | Numérico |
| Peak | Indica o maior pico de temperatura em Celsius, dentro do intervalo de tempo | Numérico |
| Time | Indica a data e hora a qual foi feito a captura de temperaturas, usada para os cálculos de pico e média | Data/horário |

# 3. Processamento/Tratamento de Dados

Para obtermos uma base mais consistente de dados e facilitar as associações, resolvemos adicionar o atributo “Server”, de tipo “Texto”, em cada um dos logs de alertas e também nos logs de temperaturas, para preenchê-los com as respectivas informações de nome dos servidores.

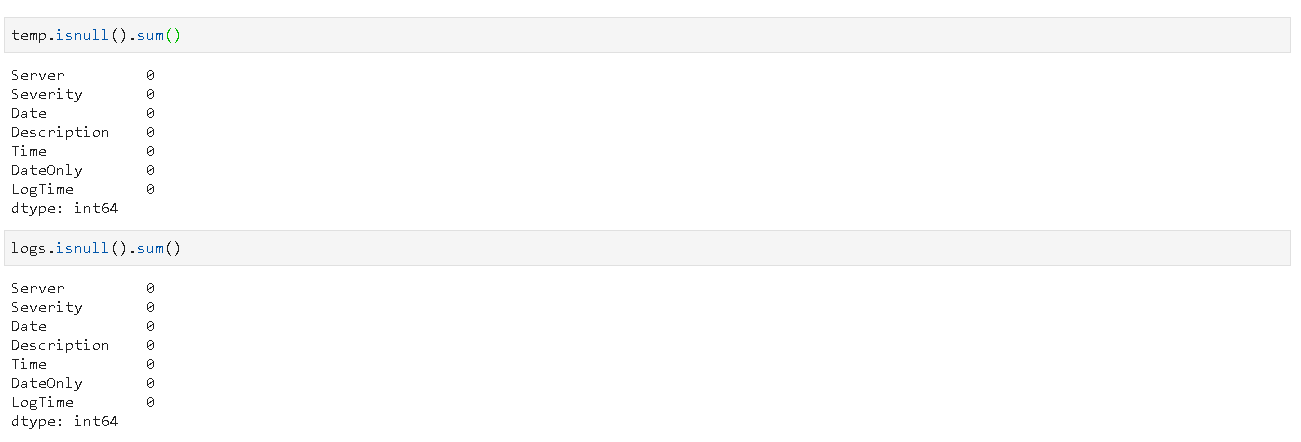
Isto facilitou o processo de criação de um *dataset* unificado, contendo todos os registros de temperaturas referentes aos 30 servidores. Da mesma forma, foi criado um *dataset* unificado de registros de alertas de todos os respectivos servidores. Respectivamente, o *dataset* de temperatura foi nomeado de “servidores-inlettemp.csv” (contendo 1541376 linhas, 4 colunas e sem valores nulos ou vazios) e o *dataset* de logs de alertas foi chamado de “logs\_servidores.csv” (contendo 6368 linhas, 4 colunas e sem valores nulos ou vazios).

Usando a ferramenta *Jupyter Notebook*, e com isso facilitar as nossas interações com os *datasets* através do Python, assim tratamos e processamos os mesmos conforme a sequência descrita a seguir:

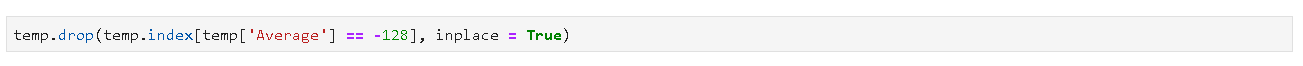
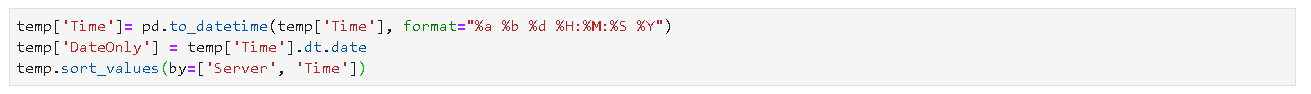
Carregar arquivos originais de temperatura e log de servidores.



Se avalia a ocorrência de linhas com valores nulos nos *datasets* de temperaturas e logs de alertas.

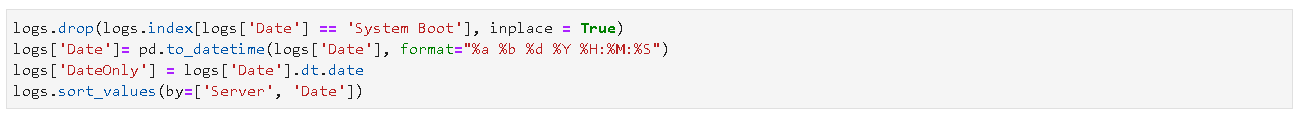


É necessário realizar alguns procedimentos de ajustes e limpeza nos dados, o primeiro passo é remover as não-ocorrências de temperatura no *dataset* de temperaturas, pois evidencia que o servidor estava desligado e isso é caracterizado quando o atributo *Average* indica o valor de -128. Esse tipo de registro de máquina desligada não é relevante para o nosso estudo, pois não fica viável comparar temperatura com qualquer outra variável. O segundo passo é transformar o atributo Time em tipo *date* e cria um atributo *DateOnly* para posteriores comparações entre bases. Esse novo atributo irá nos ajudar a indexar e juntar as bases. O terceiro e último passo de ajuste é ordenar a base de temperaturas em função do nome do servidor e do dia/hora.

Agrupar as linhas em função do nome de servidor e data, além de sumarizar através do agrupamento os atributos *Average* e *Peak* com informações de Mínimo, Máximo, Média, Mediana, Variância e Desvio Padrão. Para isso, é criado um novo *dataset* resultante chamado *temp\_summary* que agora manter uma ocorrência de evento de temperatura por dia/servidor sem que os valores originais de *Average* e *Peak* sejam perdidos com a utilização das variáveis estatísticas. Essa operação de *groupby* adiciona uma nova linha de índices (nomes de colunas), pra isso é necessário realizar um ajuste para se manter uma só linha de índice sem perda de identidade das colunas. 

No *dataset* de logs é necessário também realizar um processo limpeza, para isso, é necessário remover as ocorrências da base de logs aonde a data tem formato inválido pois indica um processo de reinicialização da máquina (tecnicamente chamado de System Boot) , após desligamento, e transforma o atributo Date em tipo Date. Além de, criar o atributo DateOnly e ordernar a base de logs em função do nome do servidor e o dia do registro.



Agora com ambos datasets processados é necessário combinar as duas bases em uma só, usando como índice o nome do servidor e o dia do registro. Neste caso levou-se em consideração o fato de que nem todas as vezes que houve um registro de temperatura, necessariamente, houve um registro de alerta. Para isso, é criado um novo *dataset* resultante chamado *merged*.



O que se percebeu no novo *dataset* gerado no passo anterior, é a frequência de valores vazios, para é realizada a substituição dos valores vazios (ausência de registro de log) da coluna *Severity* com o valor *NoneAlert*, indicando que nessa data não houve registros de alerta. Posteriormente, isso irá nos ajudar a efetuar comparações entre dados de temperatura e severidade de alertas.



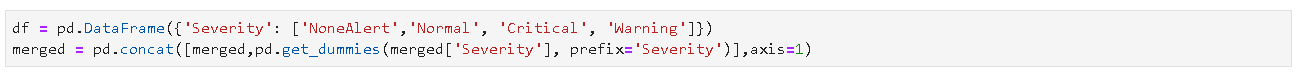
Adiciona o atributo categórico *TempSala* que indica se a sala de servidores estava quente, esquentando ou fria, baseado nos valores de *Peak\_max*.



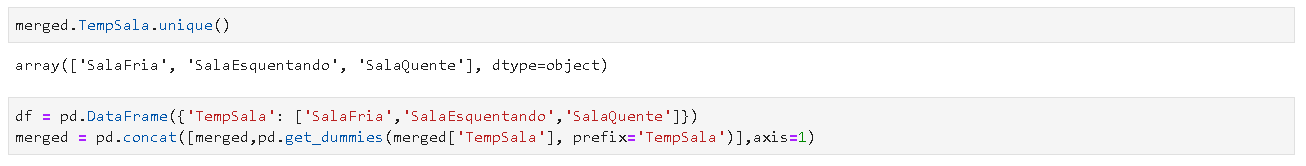
O atributo *Severity* será útil ao executar alguns modelos como o PCA na base *merged*, mas a maioria dos modelos requer que atributos categóricos sejam transformados em inteiros. Para isso, utilizaremos a técnica de *One-Hot Encoding*, na qual as categorias se transformaram em colunas (variáveis) onde o número 1 representa o valor afirmativo e o 0 negativo.



Cria-se então um *dataset* com os valores possíveis e utilizamos a função *pd.get\_dummies()* para gerar as colunas numéricas.



Semelhante ao tratamento dado ao atributo *Severity*, também fizemos uso do mesmo processo no atributo *TempSala*.



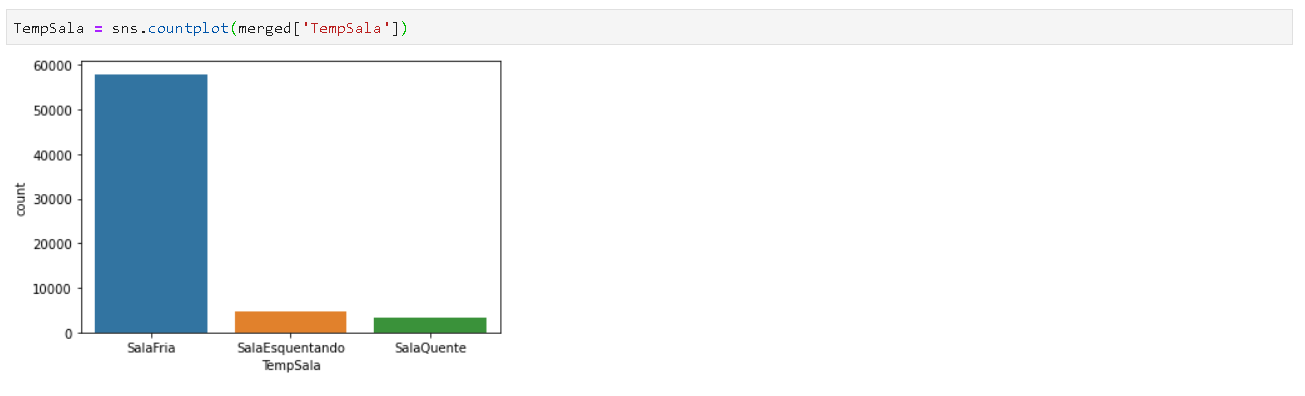
Com estes passos todos executados com sucesso, exportamos as bases para arquivos do tipo CSV, os quais serão utilizados no próximo tópico (merged.csv) deste trabalho e também usados como evidências de resultados de nosso tratamento de dados (temperatura.csv e logs.csv).

# 4. Análise e Exploração dos Dados

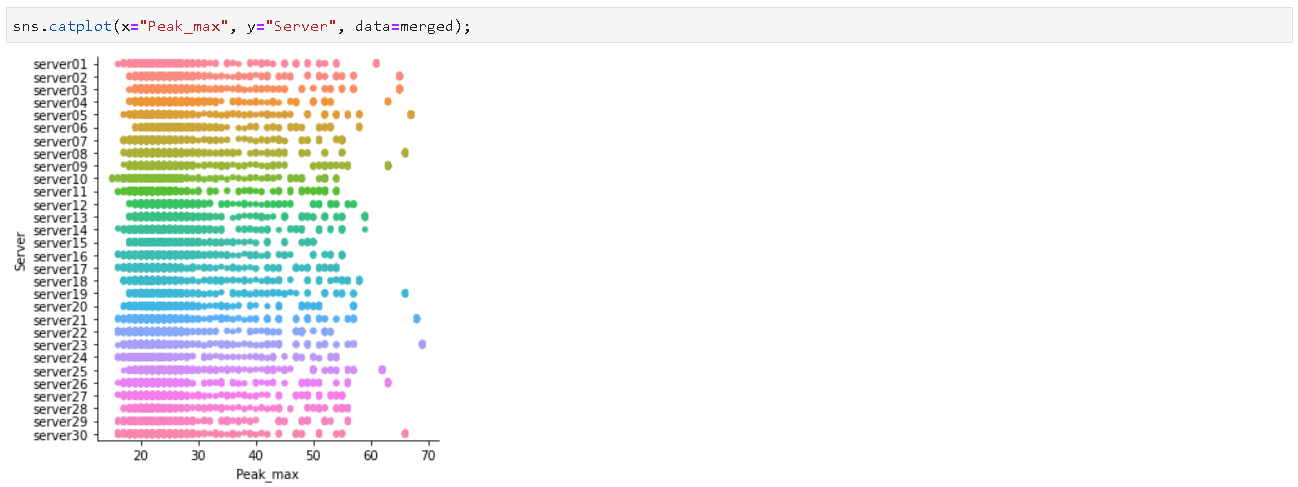
Como primeiro passo no processo de exploração de dados verificamos a frequência de alguns atributos importantes, como o atributo *Severity* com valor **NoneAlert** apresenta a maior frequência de todas, isso era esperado, pois como foi mencionado na seção de coleta a frequência de linhas de medidas de temperatura sem ocorrências de logs de alerta é muito alta.

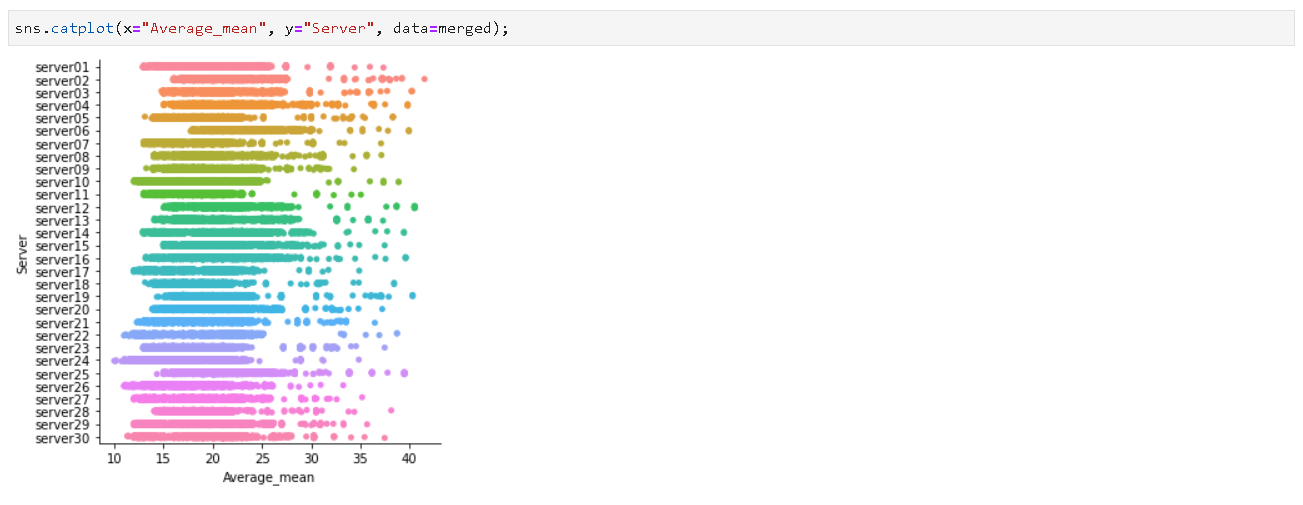


Podemos verificar também que a frequência de eventos de **SalaFria** é bem maior que **SalaEsquentando** e **SalaQuente**, respectivamente, o que é esperado de uma ambiente de data center com controle de temperatura adequado.



Nos dois gráficos seguintes é possível verificar a distribuição de temperaturas, em pico máximo (Peak\_max) e média (Average\_mean), para cada um dos servidores. O que se pode perceber é o que o pico máximo não chega na margem dos 10 a 15 graus normalmente alçando pela temperatura média, é possível visualizar a existência de uma sobreposição densa na faixa de 15 a 25 graus entre pico e média, aonde podemos assumir que boa parte do tempo o pico máximo se mantem neste faixa de temperatura.



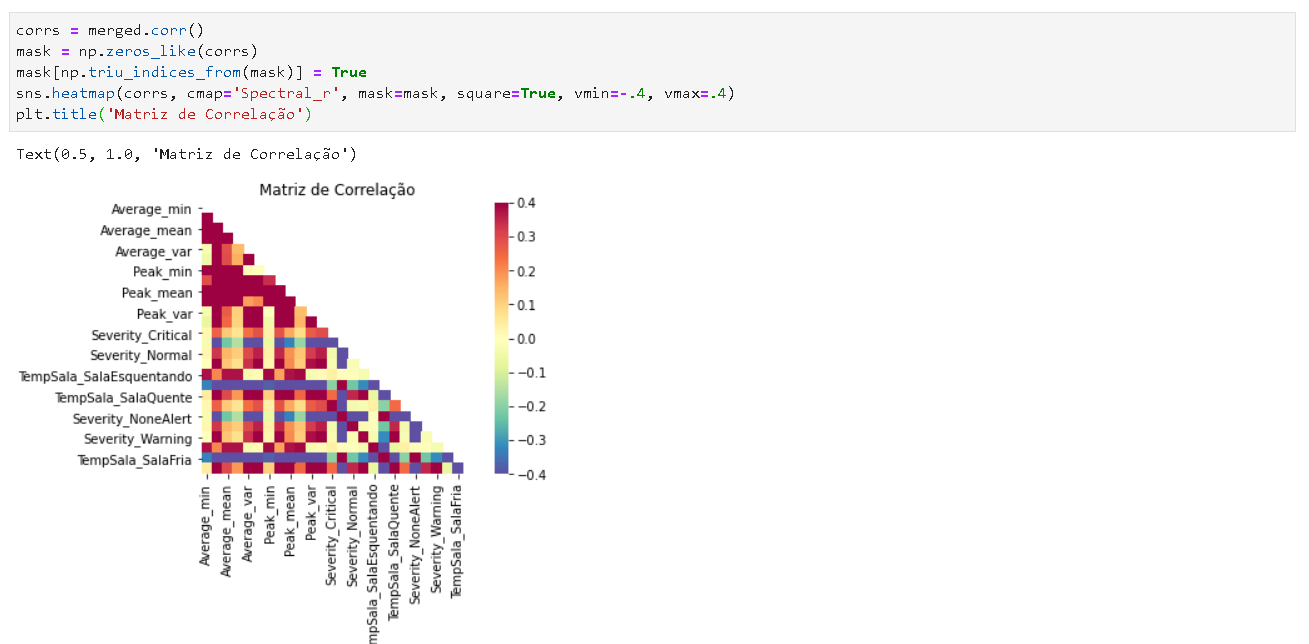


No gráfico abaixo se verifica a relação de picos máximos de temperatura e a severidade de um log de alerta, esse gráfico levanta a principal hipótese deste trabalho que é: **existe relação entre alta temperatura e logs de alerta com erro (Critical)?** Essa hipótese será verificada no próximo capítulo.

Mas podemos verificar que o tipo de log de alerta NoneAlert permanece nas temperaturas mais baixas. Outro aspecto importante é a distribuição dos tipos Normal, Critical e Warning possuem semelhanças gerais, como exemplo a ocorrência dos três tipo mesmo em altas temperaturas.



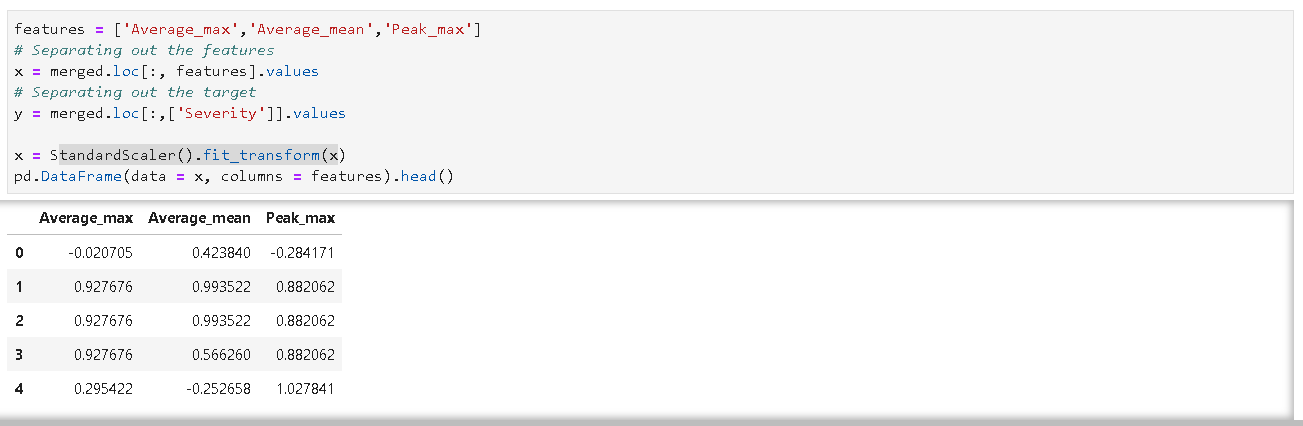
A matriz de correção é uma estatística muito utilizada no mercado, pois é ela ajuda a entender melhor como os atributos se relacionam e ajuda também no processo de manutenção de umas das máximas da estatística, correlação não implica em causalidade. O principais atributos de Average\_mean e Peak\_max tem uma correlação baixa com Severity\_Critical o que não significa que não seja causa, mas os gráficos anteriores indicam que o atributo Severity\_Critical é influenciado pela temperatura em diversos níveis altos e baixos.



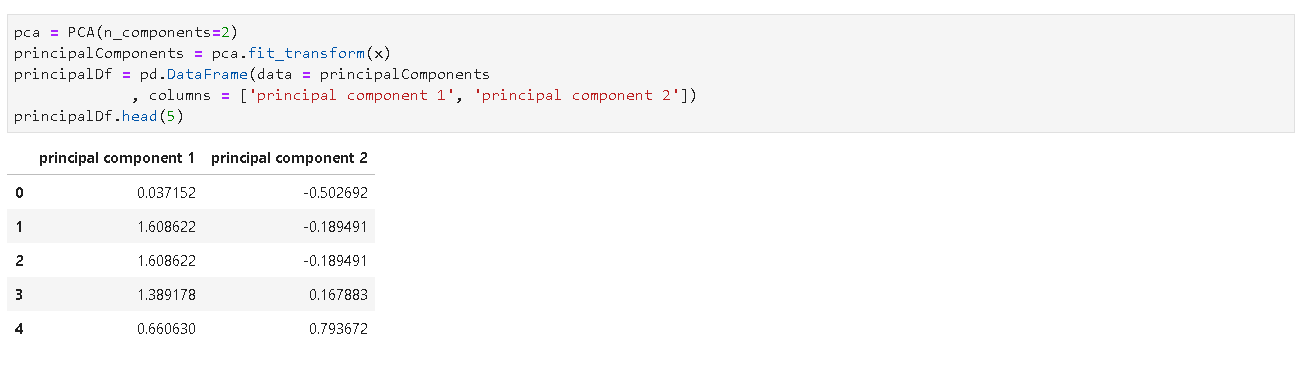
# 5. Criação de Modelos de Machine Learning

Para compreender os modelos de aprendizado de máquina, utilizamos o PCA e árvore de decisão como principais modelos neste estudo. O PCA (principal component analisys) ajuda a visualizar os dados em um plano de duas dimensões através da redução de dimensão de um conjunto de dados sem perder as características destes dados necessariamente.

O modelo original tem uma oferta limitada de atributos, para executar o PCA utilizamos os atributos *Average\_max*, *Average\_mean* e *Peak\_max*. Foi criado um *dataset* X somente os atributos de teste e um *dataset* Y com as classes, neste caso utilizamos o atributo *Severity*. O PCA requer que os atributos estejam normalizados para isso utilizarmos a função *fit\_transform(x)*.

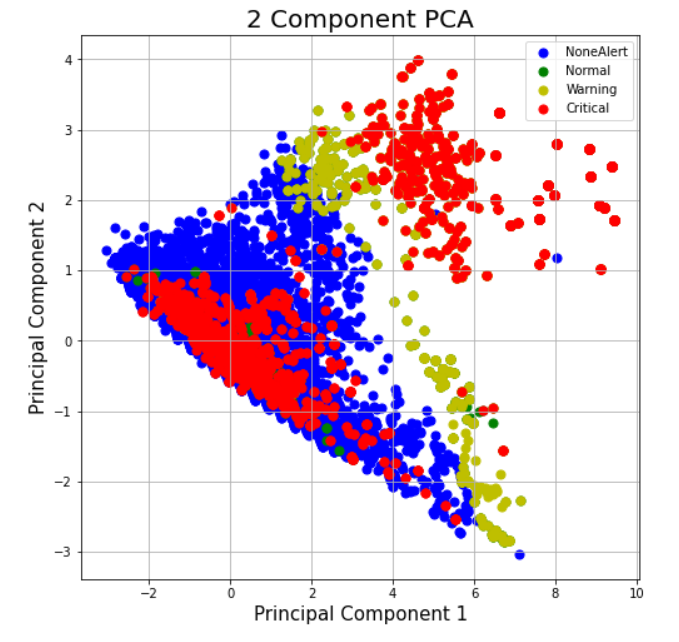


TODO...comentar sobre a execução e dimensões...

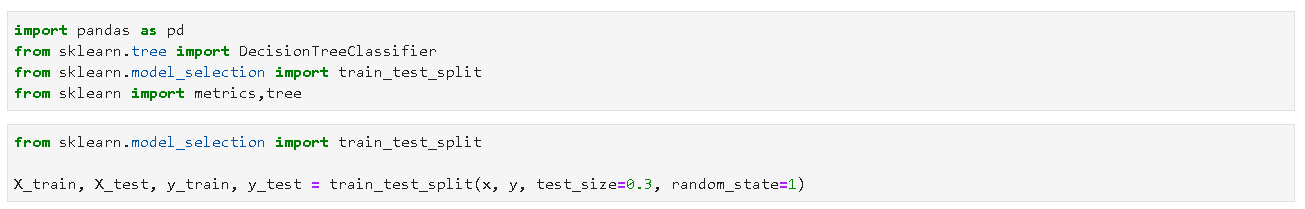


TODO...falar sobre o processo de plot e comentar os resultados dos clusters...

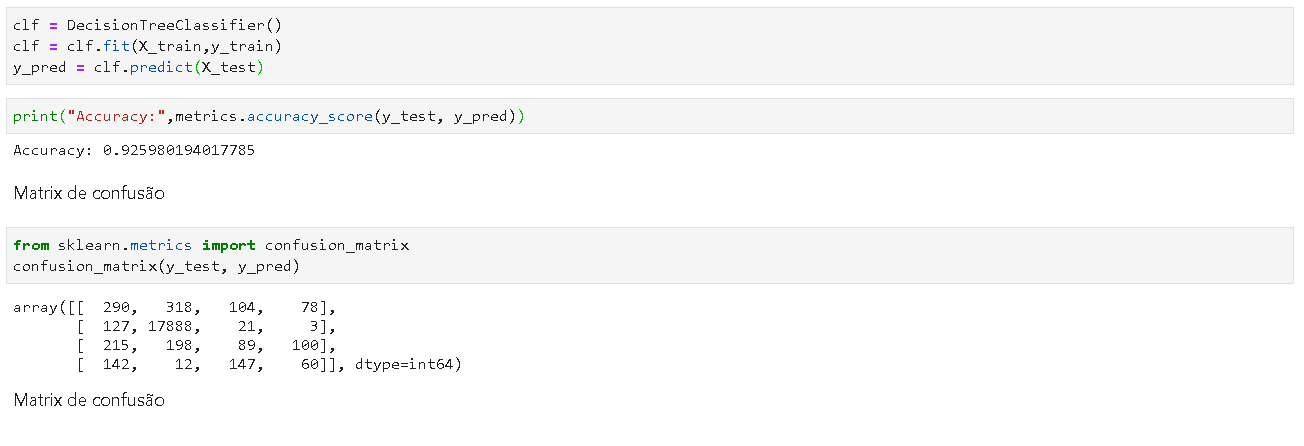




TODO – Falar sobre arvore... treino, teste....



Falar sobre acurácia e matriz.



# 6. Apresentação dos Resultados

Nessa seção você deve apresentar os resultados obtidos. Apresente gráficos, dahsboards, conte a sua história de forma bastante criativa. Aqui você pode utilizar os modelos de Canvas propostos por Dourard (clique [aqui](https://www.louisdorard.com/machine-learning-canvas)) ou por Vasandani (clique [aqui](https://towardsdatascience.com/a-data-science-workflow-canvas-to-kickstart-your-projects-db62556be4d0)).



# 7. Links

Aqui você deve disponibilizar os links para o vídeo com sua apresentação de 5 minutos e para o repositório contendo os dados utilizados no projeto, scripts criados, etc.

# REFERÊNCIAS

Um projeto de Ciência de Dados não requer revisão bibliográfica. Portanto, a inclusão das referências não é obrigatória. No entanto, caso você deseje incluir referências relacionadas às tecnologias ou às metodologias usadas em seu trabalho, relacione-as de acordo com o modelo a seguir.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.