Análise de Dados Faltantes

Nossa base de dados tem vários problemas com dados faltantes, onde em alguns casos é perdido tanta informação que impede a classificação correta dos indicadores de Fadiga.

Para remoção desses casos foi visualizado 3 informações, Quantidade de Dados, Tamanho total da Série e %de Valores não nulos. Abaixo vemos suas distribuições na base de dados

Gráfico, Histograma

Descrição gerada automaticamenteGráfico, Histograma

Descrição gerada automaticamente

Gráfico, Histograma

Descrição gerada automaticamente

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamenteGráfico, Gráfico de caixa estreita

Descrição gerada automaticamente

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamente

Primeiro Visualizamos como eram os exemplos que tinham < 85% de dados não nulos, como vera abaixo todos os exemplos apresentam faltas muito grande de dados que ocasiona em duas situações, perda quase total da série ou identificar várias series como uma só:

Gráfico, Gráfico de linhas, Gráfico de dispersão

Descrição gerada automaticamenteGráfico, Gráfico de linhas

Descrição gerada automaticamenteGráfico, Gráfico de dispersão

Descrição gerada automaticamente

Gráfico, Gráfico de linhas

Descrição gerada automaticamenteGráfico, Histograma

Descrição gerada automaticamenteGráfico, Gráfico de linhas

Descrição gerada automaticamente

Acima de 85% já era possível visualizar casos em que uma interpolação simples já resolveria o problema de dados faltantes:

Gráfico, Gráfico de linhas

Descrição gerada automaticamenteGráfico, Gráfico de linhas

Descrição gerada automaticamenteGráfico, Gráfico de linhas

Descrição gerada automaticamente

Com isso fizemos nossa primeira filtragem da base, exemplos com < 85% de dados vazios serão excluídos da nossa base.

Vamos visualizar como ficou a distribuição depois do corte:

Gráfico

Descrição gerada automaticamenteGráfico

Descrição gerada automaticamente

Como podemos ver ainda temos alguns casos com muitos pontos, vamos visualizar o comportamento das series com > 500 pontos

Gráfico

Descrição gerada automaticamenteGráfico, Gráfico de linhas

Descrição gerada automaticamenteGráfico

Descrição gerada automaticamenteGráfico

Descrição gerada automaticamenteGráfico, Gráfico de linhas

Descrição gerada automaticamenteGráfico, Gráfico de linhas

Descrição gerada automaticamente

Como podemos visualizar temos alguns casos de platô que devem ter mais pontos que uma série normal, mas temos casos de series com mais de uma subida/descida em uma só. Então mostra um filtro eficiente para remover exemplos de series com mais de uma subida/descida e de casos com classificações incorretas.

Olhamos também para casos muito pequenos, < 80 pontos

Gráfico, Gráfico de linhas

Descrição gerada automaticamenteGráfico, Gráfico de linhas, Gráfico de dispersão

Descrição gerada automaticamenteGráfico, Gráfico de linhas

Descrição gerada automaticamenteGráfico, Gráfico de linhas

Descrição gerada automaticamenteGráfico

Descrição gerada automaticamenteGráfico, Gráfico de linhas

Descrição gerada automaticamente

Esses casos pequenos mostram alguns casos que foram classificados como movimentos normais, mas que possuem variações no movimento.

Por fim verificamos que uma das pessoas da base possuía muitos dados faltantes:

Interface gráfica do usuário, Histograma

Descrição gerada automaticamente

Como Podemos visualizar a pessoa além de ter poucos movimentos em relação as outra pessoas, está também possui vários casos com muitos dados faltantes o que impossibilita o seu uso para classificação