Para a previsão de chuva, precipitação, estamos utilizando as seguintes variáveis independentes, features, pressão atmosférica, pressão atmosférica máxima, pressão atmosférica mínima, temperatura em Celsius, temperatura do ponto de orvalho, temperatura máxima, temperatura mínima, temperatura do ponto de orvalho máxima, temperatura do ponto de orvalho mínima, umidade relativa do ar máxima, umidade relativa do ar mínima, umidade relativa do ar, velocidade do vento.

Ao utilizar o coeficiente de correlação de Pearson, uma medida estatística que avalia a força e a direção da relação linear entre duas variáveis contínuas, obtive a seguinte relação das variáveis, vemos que os dados não possuem uma correlação linear alta:

Calendário

Descrição gerada automaticamente

Já usando uma tentativa do coeficiente de correlação de Spearman, medida estatística que avalia a relação monotônica, não necessariamente linear entre duas variáveis, obtive uma relação um pouco maior:

Calendário

Descrição gerada automaticamente

Como a atmosfera possui uma natureza muito dinâmica e caótica, ela não possui um padrão linear, as mudanças em uma métrica podem resultar em respostas não proporcionais ou até mesmo não previsíveis em outras métricas.

Para ter certeza de que não possuí um padrão linear utilizei os métodos de ACF, Função de Autocorrelação, e PACF, Função de Autocorrelação Parcial. O ACF ajuda a identificar padrões sazonais ou cíclicos na série temporal, ao examinar os picos e vales na ACF, é possível inferir a presença de padrões recorrentes. Já o PACF ajuda a determinar quais defasagens diretamente influenciam a observação atual. Como podemos ver os valores oscilam bem próximos a 0 indicando que não possuí um padrão linearGráfico, Gráfico de linhas

Descrição gerada automaticamente

Para dar ainda mais certeza da falta de um padrão linear utilizei o teste de Dickey-Fuller Aumentado, ADF, tendo os valores de um ADF de -32.43 e um valor de P de 0.0 isso é uma forte evidência contra a presença de tendência nos dados.

Como apresentado nos gráficos e cálculos acima não possuímos uma relação e padrões lineares, o que me fez optar a utilizar um modelo de árvores de decisão que lidam muito bem com dados não lineares, apesar dele ter uma instabilidade maior e ter uma propenção maior a overfitting, quando o modelo se ajusta excessivamente aos dados de treinamento, capturando padrões específicos que não são generalizáveis para novos dados, seria um modelo com a implementação mais fácil no momento