Análise de dados de SCADA de uma turbina eólica

16/07/2025

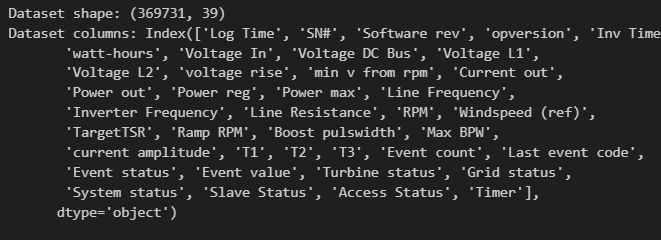
Resumo (Abstract):

O projeto tem como objetivo agrupar/classificar os dados retirados de um sistema SCADA de uma turbina eólica da universidade de São Paulo em 2022, além de identificar padrões nos períodos. Apenas os dados de 2022 foram escolhidos, já que apenas ele já possui 369.731 linhas com dados e 39 colunas, delas esse projeto irá usar Log Time, Windspeed, Power out, RPM, T1/2/3 e Voltage In para serem utilizadas para extrair dados de tempo e serem usados para Machine learning com o K-means. Com o K-means e outras analises podemos ter resultados como os melhores meses/períodos de geração de energia e agrupamento dos dados com base em seu nível de produtividade/performace.

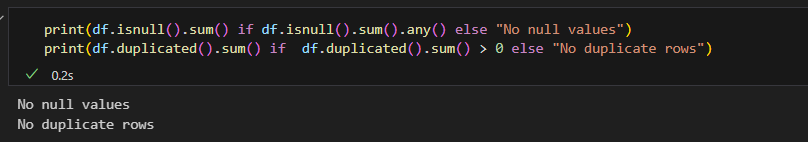
**Introdução:**

Hoje em dia grandes empresas de eólica, já possuem noção da performance e períodos de melhor eficiência para turbinas antes mesmo de serem instaladas, utilizando métricas de satélite ou de instrumentos que já existiavm no local, mas não é o mesmo para turbinas de pequeno porte como a escolhida para a análise, que foi escolhida e colocada dentro do campus da USP. Por atualmente estar em um estagio em uma empresa de turbinas eólicas, queria praticar mais meus conhecimentos e entender mais sobre a área, portanto escolhi analisar e aplicar machine learning em dados SCADA de uma turbina eólica, com objetivo de identificar melhores meses/periodos e agrupar os dados em diferentes níveis de performance.  
  
**Descrição da Base de Dados**

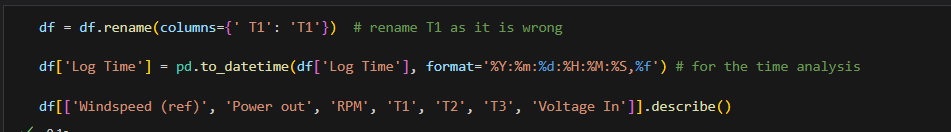
Os dados foram importados do CSV desse site <https://zenodo.org/records/7348454> , onde apenas peguei os dados de 2022, que já nos da um dataset com o shape (369731, 39), que é mais que o suficiente para analise e machine learning. Como colunas temos:

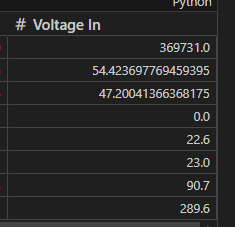
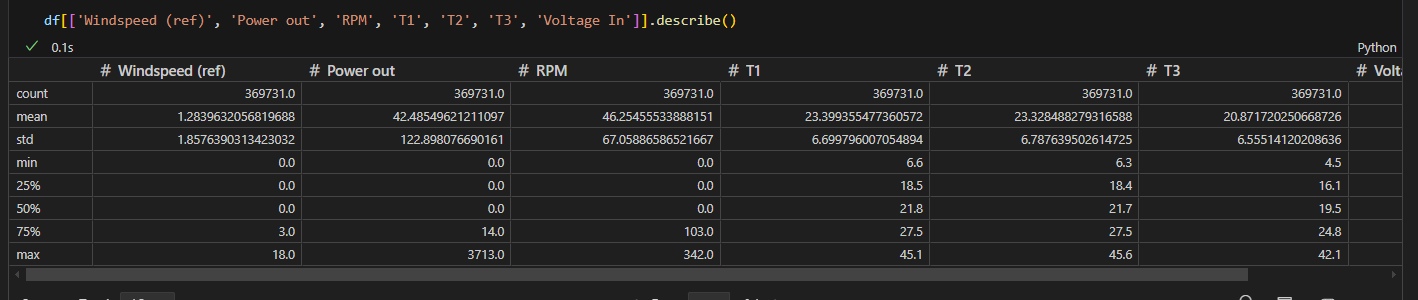


Sendo todas elas Int64 ou float64, tirando Log Time que nos da um dateTime (que está mal formato).

Não há dados nulous ou duplicados:  


Para o pre processamento apenas corrigi pequenos erros nos dados como renomear a coluna de T1 e transformar Log Time em dateTime levando em consideração que estava mal formatado.



Após isso utilizei o .describe() nas colunas mais importantes:  
  


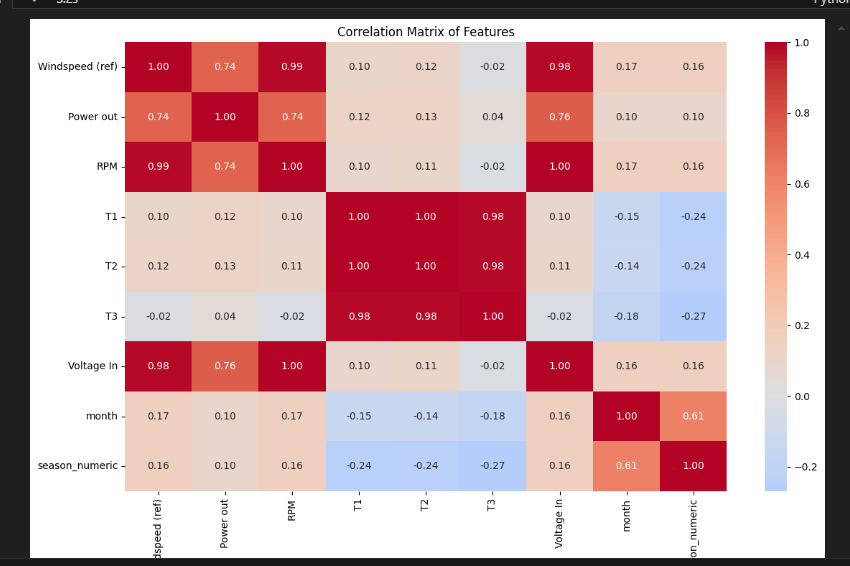
Apartir do describe acima podemos ver que maior parte do tempo essa turbina esteve parada ou em um nível muito baixa de produção (que depois será confirmado com as analises).

**Metodologia:**

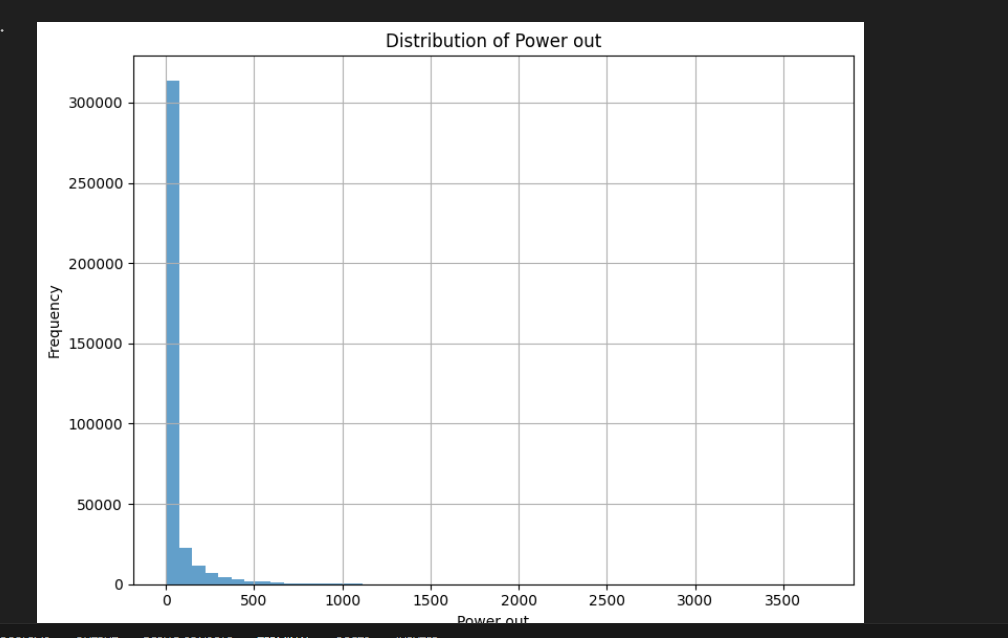
Para realizar a analise primeiro importei os dados e analisei o .info(), columns, shape e o .describe(), onde podemos ver que uma grande parte dos dados está como 0 que indica que a turbina estava parada ou em um nível muito baixo de produção.

Como mostrado antes não existe nulos ou duplicatas no dataset então nada foi usado para remove-los.

Após isso comecei a fazer as variáveis de tempos que depois serão usadas para análises e para machine learning. As variáveis foram derivadas da coluna Log Time, onde apartir dele consegui fazer uma logica onde consegui pegar o mês e ligar a um período do ano (summer, autumm, winter, spring) que seriam as colunas month, season (para melhor visualização) e season\_numeric (essa que seria usada para machine learning).

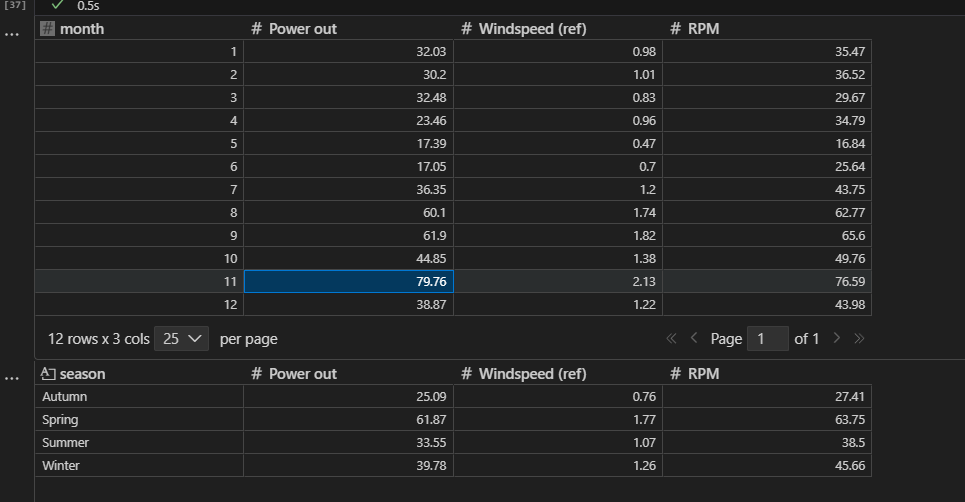
Com isso fiz uma visualização de dados e correlção dos dados:  


Uma das varias destribuições que tem esse mesmo formato:



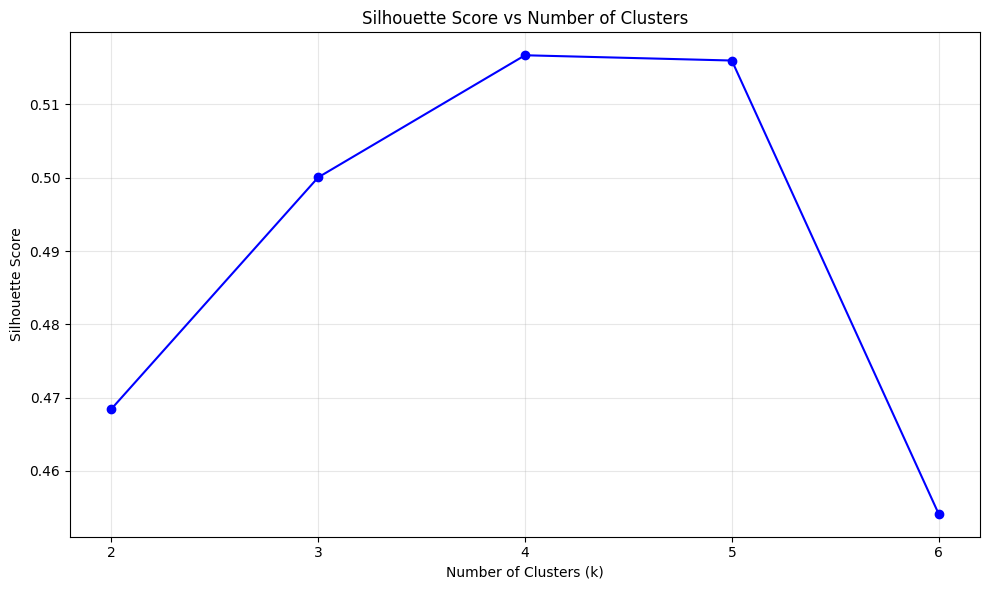
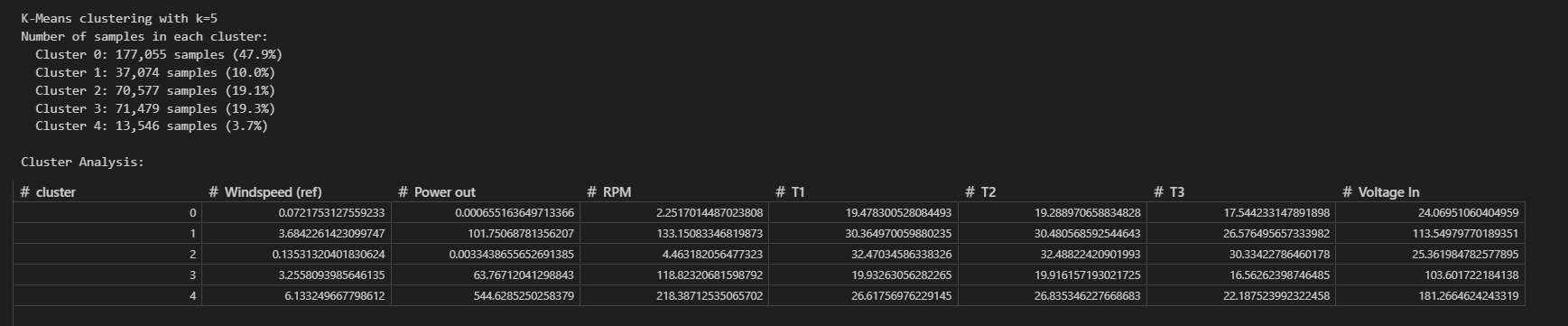
No começo eu não havia considerado Voltage-In como uma variável importante, mas após analisar melhor os dados e a correlação, resolvi voltar atras e adiciona-la a analise/ML. E novamente vemos uma grande parte dos valores como 0.

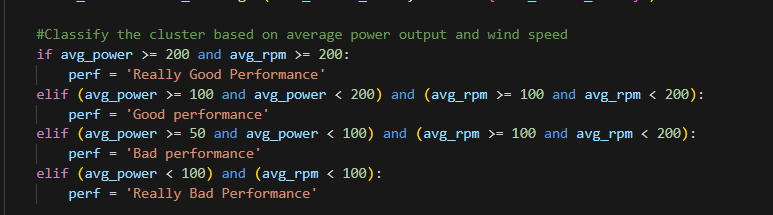
Após isso fiz a analise temporal com as medias por mês e temporada:



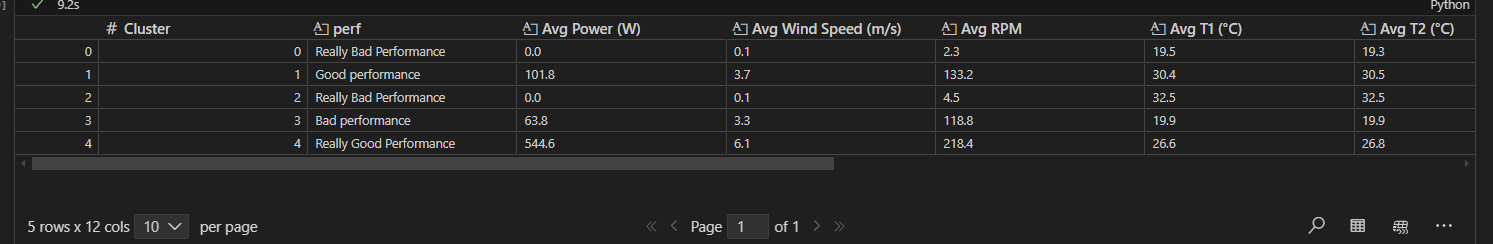
Onde podemos ver que o mês 11 é o mês com maior performace e a primavera sendo a temporada com maior performace.

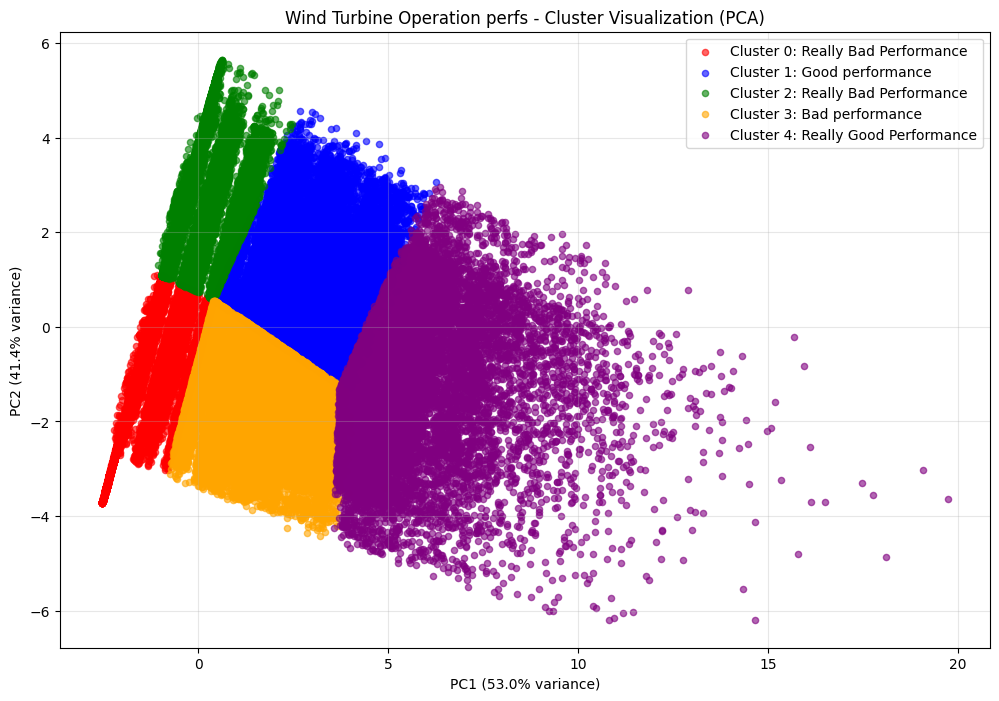
Após essa analise temporal, começei a parte de Machine Learning utilizando o K-means para poder classificar cada linha dos dados e ter diferentes grupos de performace e **normalizando os dados**. Com isso utilizei a biblioteca sklearn para fazer o K-means e calcular o silhouette score dos cluster, depois de testar diversos valores K, demorando mais ou menos 20-30m cada teste, cheguei que o melhor K seria 5 com um silhoette score de 0.5160, que é bom já que esse score vai de -1 até 1, sendo 1 o melhor possível, portanto esse é o valor usado no código final. Distribuição dos clusters:

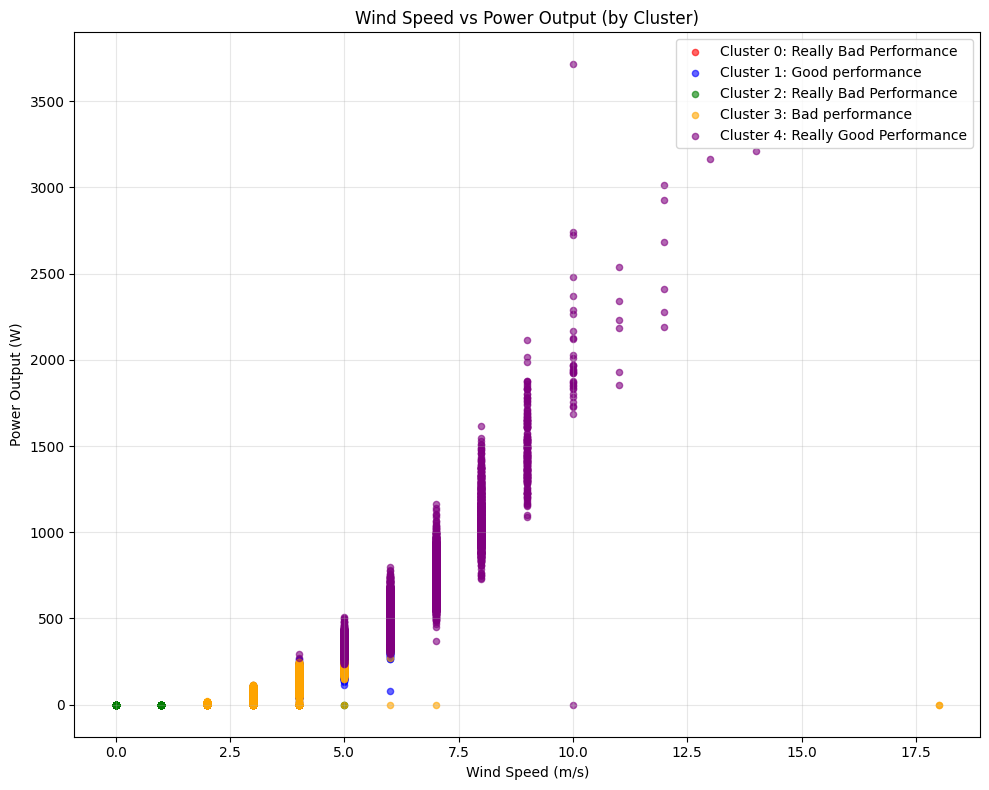
  


Após ter achado que o valor 5 seria o melhor, começei a analise e classifiquei cada cluster apartir desse codigo:  


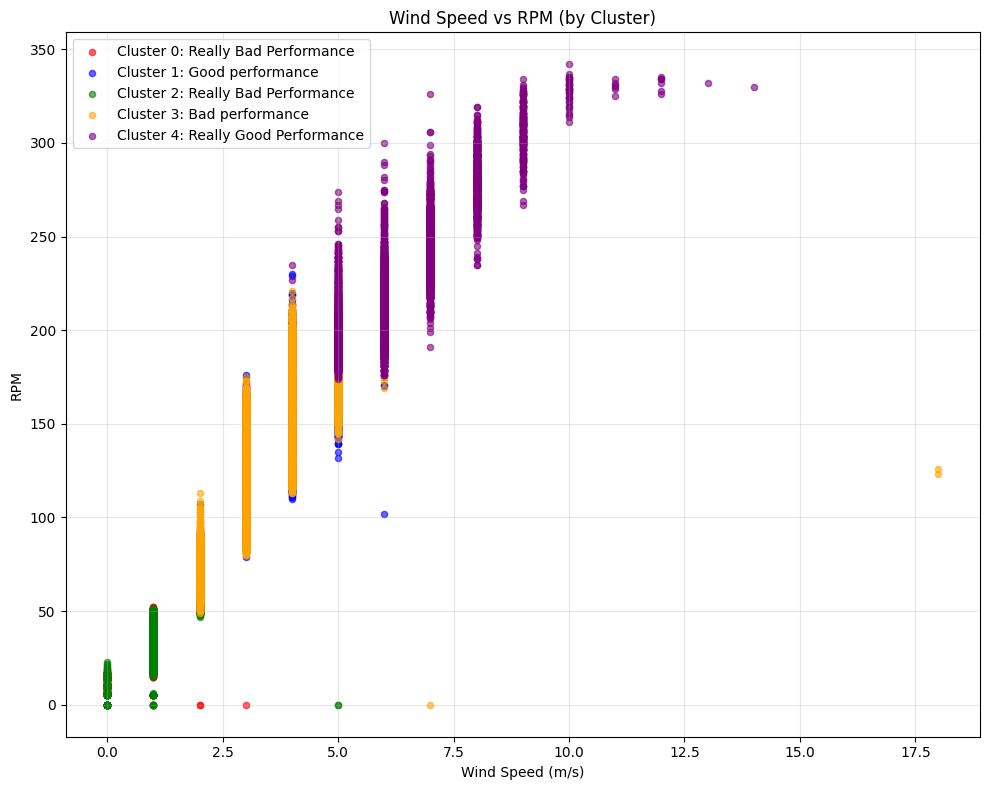
A idea seria fazer um ratio de RPM por W

Que nos da isso:  




E aqui alguns dos prints anterior mas com divisão de cor por cluster:  


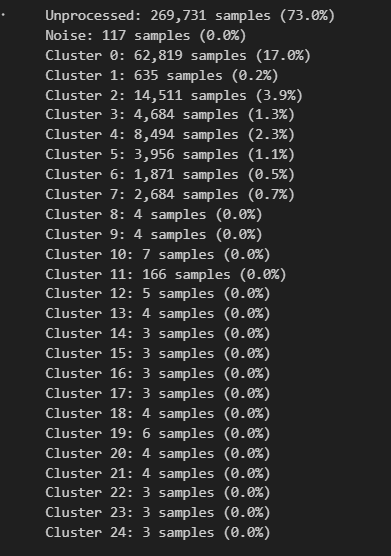


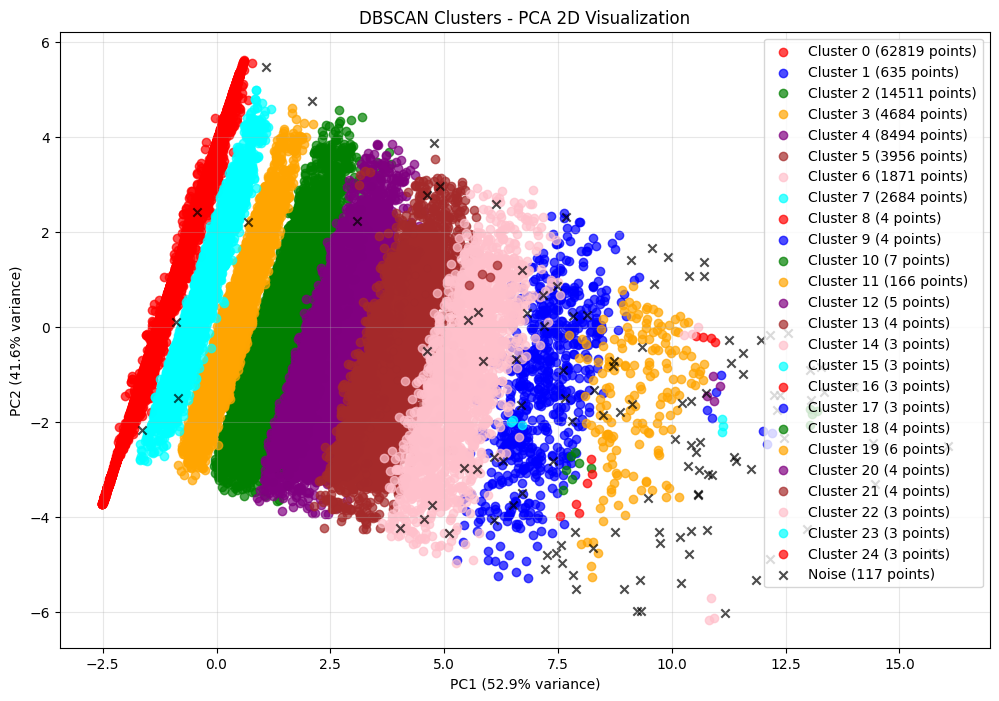


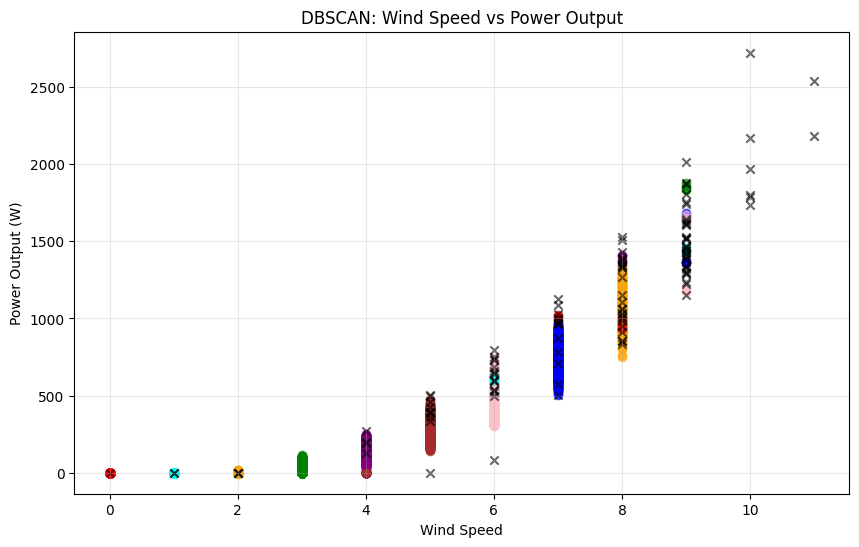
Apartir desses gráficos, acho que classifiquei os clusters corretamente.

Após isso fiz uma tentativa de aplicar DBSCAN, mas olhando o gráfico PSA anteriormente já era possível perceber que não seria uma boa analise. Mas utilizando épsilon como 0.5 e min\_samples como 3 consegui fazer com que o DBSCAN formasse mais de 2 grupos, com esses valores conseguimos 25 grupos tirando o ruido.

Tive que limitar a 100.000 linhas para essa analise já que indo mais acima disso eu tinha MemoryError por acabar a memoria para o Machine learning.







Como podemos ver no PC2 temos diversos grupos grandes e quanto mais para direita os pontos vão mais pequenos grupos são formados ou ruídos são classificados.

Como isso foi apenas uma tentativa não fiz nenhuma avaliação.

Ferramentas e bibliotecas utilizadas: pandas, matplotlib, sklearn e seaborn (apenas para a tabela de correlação).

**Resultados/Conclusão:**

Apartir da analise temporal mais simples, conseguimos ver que os meses 11,9 e 8 são os que mais produziram Watts utilizáveis enquanto os meses 4,5 e 6 foram os menores. Podemos ver que a primavera é o melhor mês para uma turbina elétrica em São Paulo ainda mais quando vemos a media de velocidade dos ventos, ela tem seu menor valor no mês 5 mas aumenta drasticamente durante os próximos meses tendo seu pico no mês 11 (que é o mês com maior produção de watts).

Para o resultado do K-means eu tinha em idea classificar os dados em:

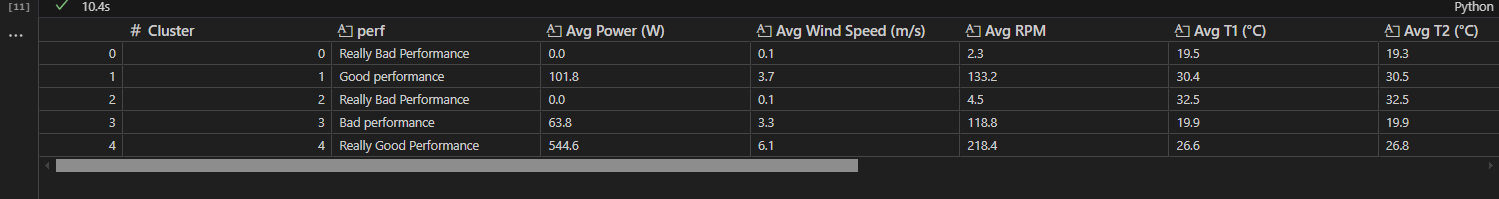
- Performace Muito boa

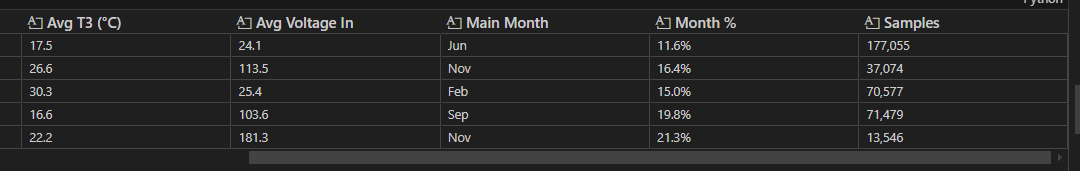
- Performace Boa

- Performace Ruim

- Performace Muito ruim

Com o k-means houve a separação em 5 grupos em que consegui uma boa variação sendo 2 grupos com performace muito ruim, 1 com perfomace ruim, 1 com performace boa e o ultimo com performace muito boa:



  
Alguns insights apartir de cada cluster:

Cluster 0 e 2 – Temos que ambos representam momentos em que a turbina estava parada/desligada/com problema ou não estavam gerando basicamente nada, como podemos ver pela media de energia gerada e media de velocidade dos ventos. Tem como meses principais em seus clusters (maior participação nos dados) janeiro e junho.

Cluster 3 – Temos nosso primeiro cluster que está realmente produzindo, mas não em um nível satisfatório para as métricas em que fiz, esse cluster gera 0.5 W por mais ou menos 2 RPM, que considero uma performace ruim. Tem como mês principal em seu cluster setembro.

Cluster 1 – Cluster que está em um nível bom em relação as métricas onde podemos ver que ele produz quase 1W para cada 1RPM, tornando esse cluster bem eficiente e tem como mês principal novembro, que vimos que foi o mês com maior produção de Watts em analises anteriores.

Cluster 4 – O melhor cluster, temos que ele produz 2.5W para cada 1RPM e vemos novamente que o mês principal sendo novembro.

Com isso nos vemos que a analise temporal anteriormente está realmente correta, por nos dizer que novembro foi o melhor mês e que a primavera é o melhor período para geração de Watts.

Assim caso queiramos otimizar uma turbina que mantê-la (manutenção em dia) em dia durante o período de primavera, principalmente nos meses de setembro e novembro que foram nossos principais contribuidores nos cluster de boa performace. E quem sabe (caso isso fosse uma turbina comercial e não da faculdade de são paulo) uma mudança de local para uma região com ventos mais constantes durante o ano como aqui na região do RN, onde existe uma boa quantidade de turbinas eólicas.

Agora para o resultado do DBSCAN, como dito anteriormente, ajustando os parâmetros em maior parte das vezes resultava em 2/3 cluster com vários ruídos, mas eu consegui encontrar uma combinação de parâmetros que nos da mais diversos clusters (25) mas analisando as imagens e as separações não há insights, por ser uma grande quantidade de cluster e muitos formarem pequenos grupos que não fazem muito sentido, por isso não continuei essa análise.

Por fim, esse trabalho ajudou bastante a entender métricas e oque o sistema SCADA analisa no geral, além de entender muito mais sobre machine learning, oque me ajudara muito em futuras analises em meu próprio estagio e como futuro estudo com certeza irei mais afundo em deep learning e como utiliza-la.

**Referências:**

Alguns dos códigos relacionados a K-means e principalmente ao gráfico PCA 2D e DBSCAN vinheram da matéria de aprendizado de máquina (do próprio IMD).