CHECKPOINT 01 – Data Science e Machine Learning no Python e Orange Data Mining (Essa atividade é composta de 4 partes)

PARTE 1 – Exercícios iniciais com Individual Household Electric Power Consumption

1. Carregue o dataset e exiba as 10 primeiras linhas.

2. Explique a diferença entre as variáveis Global_active_power e Global_reactive_power.

3. Verifique se existem valores ausentes no dataset. Quantifique-os.

4. Converta a coluna Date para o tipo datetime e crie uma nova coluna com o dia da semana correspondente.

5. Filtre os registros apenas do ano de 2007 e calcule a média de consumo diário de Global_active_power.

6. Gere um gráfico de linha mostrando a variação de Global_active_power em um único dia à sua escolha.

7. Crie um histograma da variável Voltage. O que pode ser observado sobre sua distribuição?

8. Calcule o consumo médio por mês em todo o período disponível no dataset.

9. Identifique o dia com maior consumo de energia ativa global (Global_active_power).

10. Compare o consumo médio de energia ativa global em dias de semana versus finais de semana.

11. Calcule a correlação entre as variáveis Global_active_power, Global_reactive_power, Voltage e Global_intensity.

12. Crie uma nova variável chamada Total_Sub_metering que some Sub_metering_1, Sub_metering_2 e Sub_metering_3.

13. Verifique se há algum mês em que Total_Sub_metering ultrapassa a média de Global_active_power.

14. Faça um gráfico de série temporal do Voltage para o ano de 2008.

15. Compare o consumo entre os meses de verão e inverno (no hemisfério norte).

16. Aplique uma amostragem aleatória de 1% dos dados e verifique se a distribuição de Global_active_power é semelhante à da base completa.

17. Utilize uma técnica de normalização (Min-Max Scaling) para padronizar as variáveis numéricas principais.

18. Aplique K-means para segmentar os dias em 3 grupos distintos de consumo elétrico. Interprete os resultados.

19. Realize uma decomposição de série temporal (tendência, sazonalidade e resíduo) para Global_active_power em um período de 6 meses.

20. Treine um modelo de regressão linear simples para prever Global_active_power a partir de Global_intensity. Avalie o erro do modelo.

PARTE 2 – Exercícios adicionais no dataset inicial

21. Séries temporais por hora

- Converta Date e Time em índice datetime.

- Reamostre os dados em intervalos de 1 hora, calculando a média de Global_active_power.

- Identifique os horários de maior consumo médio ao longo do dia.

Observação: uma série temporal é um conjunto de dados registrados em ordem cronológica, útil para identificar padrões de comportamento ao longo do tempo.

22. Autocorrelação do consumo

- Use a série temporal de Global_active_power.

- Calcule a autocorrelação em lags de 1h, 24h e 48h.

- Pergunta: existem padrões repetidos diariamente?

23. Redução de dimensionalidade com PCA

- Selecione Global_active_power, Global_reactive_power, Voltage e Global_intensity.

- Aplique PCA para reduzir para 2 componentes principais.

- Analise a variância explicada por cada componente.

24. Visualização de clusters no espaço PCA

- Combine os resultados do PCA com K-Means (3 clusters).

- Plote os pontos resultantes e pinte cada grupo por cluster.

- Pergunta: os grupos se separam de forma clara?

25. Regressão polinomial vs linear

- Modele Global_active_power em função de Voltage.

- Compare Regressão Linear Simples com Regressão Polinomial (grau 2).

- Analise RMSE e a curva ajustada.

PARTE 3 – Novo dataset Appliances Energy Prediction

Dataset escolhido:

- Appliances energy prediction dataset
https://archive.ics.uci.edu/dataset/374/appliances+energy+prediction

- Contém consumo de energia (Appliances) e variáveis ambientais (temperatura, umidade, condições internas/externas).

26. Carregamento e inspeção inicial

- Carregue o dataset no Pandas.

- Liste tipos de dados e estatísticas descritivas (.info() e .describe()).

27. Distribuição do consumo

- Crie histogramas e séries temporais para a variável Appliances.

- Pergunta: o consumo tende a se concentrar em valores baixos ou altos?

28. Correlações com variáveis ambientais

- Calcule correlações entre Appliances e variáveis como temperatura e umidade.

- Pergunta: quais fatores têm mais relação com o consumo?

29. Normalização dos dados

- Aplique Min-Max Scaling às variáveis numéricas.

- Reutilize esses dados em modelos posteriores.

30. PCA

- Aplique PCA e reduza para 2 componentes principais.

- Plote os dados resultantes.

- Pergunta: aparecem padrões ou agrupamentos naturais?

31. Regressão Linear Múltipla

- Modele Appliances em função das variáveis ambientais.

- Avalie R^2 e erro médio.

32. Random Forest Regressor

- Treine um modelo de Random Forest para prever Appliances.

- Compare o RMSE com a regressão linear.

33. K-Means clustering

- Aplique K-Means com 3 a 5 clusters.

- Interprete os perfis de consumo.

34. Classificação binária

- Crie uma variável: alto vs baixo consumo (Appliances maior/menor que a mediana).

- Treine Logistic Regression e Random Forest Classifier.

35. Avaliação de classificação

- Gere matriz de confusão e métricas (accuracy, precision, recall, F1-score).

- Pergunta: o modelo erra mais para alto ou para baixo consumo?

PARTE 4 – Exercícios no Orange Data Mining

36. Importação e visualização inicial

- Use o widget CSV File Import para carregar o dataset Individual Household Electric Power Consumption.

- Conecte ao widget Data Table para visualizar as primeiras linhas.

- Pergunta: quantas variáveis e registros aparecem?

37. Amostragem de dados (1%)

- Use o widget Sample Data para selecionar uma amostra de 1% dos registros.

- Pergunta: a distribuição de Global_active_power na amostra é semelhante à base completa?

- Consulte: https://orangedatamining.com/widget-catalog/

38. Distribuição do consumo

- Conecte ao widget Distribution e visualize Global_active_power.

- Pergunta: o consumo é concentrado em valores baixos ou há muitos registros de alto consumo?

39. Relação entre variáveis elétricas

- Use o widget Scatter Plot para analisar Voltage (X) vs Global_intensity (Y).

- Pergunta: existe correlação visível?

40. Clustering com K-Means

- Aplique o widget k-Means com 3 clusters.

- Use como atributos Sub_metering_1, Sub_metering_2, Sub_metering_3.

- Visualize os grupos no Scatter Plot.

- Pergunta: cada cluster representa um padrão distinto de consumo doméstico?