Essential Question #2 NK

JD's AI/ML Laboratory

Due: 07/06/2025, 11:59 PM EST

General Instruction: Please **answer all questions** and provide relevant support from literature using **APA references**. You can answer the questions using bullet points with citations (preferably from peer-review journals). Keep in mind that most science competitions focus on participants having a very strong grasp of the literature. I would highly recommend printing out the relevant journal articles and do a deep dive. Using a highlighter and writing notes on the journal article can help you narrow down important content for future reference. I will be thorough in my review of your response and expect all the answers to be well thought-out and clearly explained with supporting literature.

1. Change the data source for the coding exercise to the local iris.csv data in the Coding Exercise folder.
    - Done
2. Create summary stats table for the iris dataset, mean, median, standard deviation. Add density plots to the feature histogram plots.
    - Done
3. What is your research question, Hypothesis and how are you going to test the hypothesis using ML?
    - Can an AI model be used to accurately predict water filter lifespan?
    - AI can predict water filter lifespan based on flow discharge, turbidity, total dissolved solids (TDS), pH, and depth.
    - I will process and analyze a dataset with various water quality parameters, then split it into a training set (80%) and a testing set (20%). I will use a random forest regression model, since it combines predictions from multiple decision trees to produce a more accurate prediction (Medium, 2023). I will then train the model using input features (flow discharge, turbidity, total dissolved solids, pH, and depth) to predict the water filter lifespan. Then I will test the model ___
4. List the datasets you have reviewed?
    - Water Quality Metrics & Filter Performance Dataset
    - UWO - Field observations (2019 to 2021)
5. Describe the dataset you are going to use for answering the research question.
    - 19,656 rows, 7 columns (features)
    - Features: total dissolved solids (TDS) mg/l, turbidity (NTU), pH, depth (m), flow discharge (L/min), filter lifespan (hrs), filter efficiency (%)

-

```
Mean:
 TDS (mg/l)                  249.621741
Turbidity (NTU)                4.996001
pH                             7.250062
Depth (m)                      2.753614
Flow Discharge (L/min)        50.300216
Filter Life Span (hours)    4274.433573
Filter Efficiency (%)         88.740132
dtype: float64

Median:
 TDS (mg/l)                  249.4460
Turbidity (NTU)                5.0130
pH                             7.2500
Depth (m)                      2.7560
Flow Discharge (L/min)        50.5825
Filter Life Span (hours)    4274.4035
Filter Efficiency (%)         88.7285
dtype: float64

Standard Deviation:
 TDS (mg/l)                  144.273284
Turbidity (NTU)                2.876936
pH                             0.722374
Depth (m)                      1.292128
Flow Discharge (L/min)        28.688768
Filter Life Span (hours)    226.300799
Filter Efficiency (%)          1.946349
dtype: float64
```

```
Minimum:
 TDS (mg/l)                    0.006
Turbidity (NTU)                0.000
pH                             6.000
Depth (m)                      0.501
Flow Discharge (L/min)         1.003
Filter Life Span (hours)    3547.166
Filter Efficiency (%)         82.340
dtype: float64

Maximum:
 TDS (mg/l)                  499.962
Turbidity (NTU)                9.999
pH                             8.500
Depth (m)                      4.999
Flow Discharge (L/min)        99.999
Filter Life Span (hours)    4989.885
Filter Efficiency (%)         94.933
dtype: float64
```

6. Perform EDA for your dataset, including data visualization and generating summary statistics and correlations. Explore if there are any missing values.
    - Done
7. Create an analysis dataset for your project. What is your outcome variable? Filter life span?
    - Filter lifespan is my outcome variable
8. Learn about regression ML analysis and perform Regression analysis for predicting Filter Life Span.
    - Done
9. Exhaustive search of regression models, create a table with the description, pros and cons, how they work, when to use. Short list of 3-4 regression models

Model	**Description**	**Pros**	**Cons**	**How**	**When**
Linear Regression	Assumes a linear connection between a dependent variable and an independent variable.	Very simple and fast, easy to interpret, works well with linearly separable data	Cannot work with nonlinear relationships, sensitive to outliers, poor generalization with more complex and larger datasets	Fits a straight line to the data, minimizes the sum of squared errors between the actual and predicted values	Relationship between variables in linear, want a model that is easy to interpret and fast
Polynomial	Represents a	Captures	Very sensitive	Adds powers of	Relationship is

Regression	non-linear relationship between dependent and independent variables	nonlinear patterns, interpretable	to outliers, can't handle complex multivariable interactions well	the original features to model curves	nonlinear and follows a smooth curve, small to medium sized dataset
Ridge Regression	Independent variables are highly correlated, reduces standard errors by introducing a small bias to the regression estimates	Reduces variance and overfitting, handles collinear features better than linear reg., efficient and interpretable	Assumes linear relationship, introduces bias to reduce variance, can't capture complex linear patterns	Modifies linear regression by adding L2 regularization (penalty on the size of coefficients)	High multicollinearity, many correlated or weak predictors, a more regularized and stable linear model
Random Forest Regression	Combines predictions from multiple decision trees to produce a more accurate prediction	Handles nonlinearities and feature interactions, robust to outliers and noise, less overfitting than a single decision tree, provides feature importance	Less interpretable, slower to train and predict, may require more memory	Builds many decision trees and averages their predictions. Each tree is trained on a random subset of the data and features (bootstrapped).	Complex or nonlinear relationships, many features, when you want accuracy over interpretability

10. Create a workflow for the machine learning analysis.
    —

11. How would you optimize filter flow rate and life span?
    Find the ranges of the other variables that lead to maximum flow rate and life span.

12. Search for existing tools and models related to estimating filter lifespan.
    None really out there for estimating filter lifespan specifically.

13. **Optional for this week**: Train a machine learning model that would help you answer your scientific question.