Essential Question #3 - NK

JD's AI/ML Laboratory

Due: 07/11/2025, 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. How is Random Forest Classification model different from Random Forest Regression? Explain using a decision tree based on your project features/variables.

Turbidity > 4.7

Yes

- TDS > 250
    - Yes -> predict
    - No -> predict

No

- pH < 6.8
    - Yes -> predict
    - No -> predict

Flow discharge < 35

Yes

a. Depth > 2
    i. Yes -> predict
    ii. No -> predict

No

b. Turbidity < 3.5
    i. Yes -> predict
    ii. No -> predict

Random classification is for categorized classes; for example, short, medium, and long.

2. Add to your table and learn about these models Elastic Net Regression, Gradient Boosting Regressor, Support vector regressor, Multilayer Perceptron Regressor, Gaussian Process Regression (GPR) and Bayesian Regression.

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 Regression	Represents a non-linear relationship between dependent and independent variables	Captures nonlinear patterns, interpretable	Very sensitive to outliers, can't handle complex multivariable interactions well	Adds powers of the original features to model curves	Relationship is 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	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	Complex or nonlinear relationships, many features, when you want accuracy over interpretability

		decision tree, provides feature importance		features (bootstrapped).	
Elastic Net Regression	Combines L1 (Lasso) and L2 (Ridge) penalties for regularization.	Handles multicollinearity; performs variable selection;	Can be sensitive to hyperparameter tuning; may underperform with non-linear relationships.	Performs linear regression with both L1 and L2 regularization to shrink coefficients and select variables.	When you suspect multicollinearity and want both feature selection and regularization.
Gradient Boosting Regressor	Ensemble technique that builds models sequentially to correct previous errors.	High predictive accuracy; handles mixed data types; good at capturing non-linearities	Computationally intensive; prone to overfitting without proper tuning.	Trains decision trees sequentially where each new tree reduces the residuals of the previous model.	When accuracy is critical and computation time is less of a concern.
Support vector regressor	Extension of SVM for regression tasks.	Resistant to outliers; effective in high-dimensional spaces.	Sensitive to choice of kernel and parameters; not scalable to large datasets.	Finds a function within a margin (epsilon) that best approximates the data while minimizing model complexity.	When the dataset is small to medium-sized and potentially non-linear.
Multilayer perceptron regressor	Feedforward neural network for regression.	Captures complex non-linear patterns; flexible architecture.	Requires large data and tuning; risk of overfitting; slow training.	Uses multiple layers of neurons with non-linear activation functions to model complex relationships.	When you have a large dataset and expect complex non-linear relationships.
Gaussian process regression	Bayesian non-parametric model that provides	Provides uncertainty estimates; very flexible.	Not scalable to large datasets; computationa	Assumes a distribution over functions and computes	When uncertainty quantification is important and

	probabilistic predictions.		lly expensive.	the posterior given data using kernels.	the dataset is small.
Bayesian regression	Linear regression with Bayesian inference for coefficients.	Handles uncertainty well; avoids overfitting.	Computationally intensive; interpretation can be complex.	Estimates a posterior distribution for the coefficients instead of point estimates.	When you need probabilistic inference and interpretability for small to medium datasets.

3. Review article, focus on ML models, SHAP analysis, data curation for ML analysis and ML analysis workflow https://www.sciencedirect.com/science/article/pii/S259012302400183X#ab0010
4. Select 4 most relevant models for your project.
    - Random forest regression
    - Support vector regression
    - Gradient boosting regression
    - Multilayer perceptron (MLP) regression
5. Learn about multi objective optimization using ML, in your case you are trying to optimize both flow rate and filter life span.
6. Train a machine learning model that would help you answer your scientific question.
7. Draft PowerPoint with Background, Research Question, Hypothesis, research plan, EDA, Analysis Flowchart, ML model analysis, results, conclusions.
8. Prepare for oral presentation on Friday.
9. Draft research paper outline.