Feature Selection

(Practice the problem given in the Lab Manual section related to feature selection for better understanding)

Missing Values Ratio

1. **Diabetes Dataset**: Identify and remove features in the diabetes dataset where the percentage of missing values exceeds 30%, then analyze how the reduced feature set affects model accuracy when predicting diabetes outcomes.
2. **Melbourne Housing Dataset**: Filter out columns in the Melbourne housing dataset where more than 20% of values are missing, and determine the impact on a price prediction model's performance.

High Correlation Filter

3. **Diabetes Dataset**: Identify pairs of highly correlated features (correlation > 0.8) in the diabetes dataset, then remove one feature from each pair and assess how model performance changes in diabetes classification.
4. **Melbourne Housing Dataset**: Remove highly correlated features (correlation > 0.85) from the Melbourne housing dataset and evaluate the effect on the prediction of property prices.

Low Variance Filter

5. **Diabetes Dataset**: Apply a low variance filter to remove features in the diabetes dataset with very low variability, and observe how this affects the model's accuracy in predicting diabetes.
6. **Melbourne Housing Dataset**: Filter out features in the Melbourne housing dataset with low variance (e.g., those that are nearly constant across samples), and analyze its impact on predicting housing prices.

Forward Feature Selection

7. **Diabetes Dataset**: Use forward feature selection to iteratively select the best features from the diabetes dataset for a logistic regression model, and determine how many features are optimal for predicting diabetes outcomes.
8. **Melbourne Housing Dataset**: Implement forward feature selection on the Melbourne housing dataset to find the optimal set of features for predicting housing prices using a linear regression model.

Backward Feature Elimination

9. **Diabetes Dataset**: Perform backward feature elimination on the diabetes dataset using a decision tree classifier, removing the least important features one by one, and examine the final set of features and its effect on model performance.

10. **Melbourne Housing Dataset**: Apply backward feature elimination on the Melbourne housing dataset using a random forest model, and analyze how removing the least important features one at a time impacts the accuracy of price predictions.

Random Forest

11. **Diabetes Dataset**: Use the feature importance scores from a random forest model to rank the features in the diabetes dataset, then keep only the top 5 most important features and evaluate how well the reduced model predicts diabetes.
12. **Melbourne Housing Dataset**: Train a random forest model on the Melbourne housing dataset to determine the most important features for predicting housing prices, and assess the model's accuracy after removing the least important features.