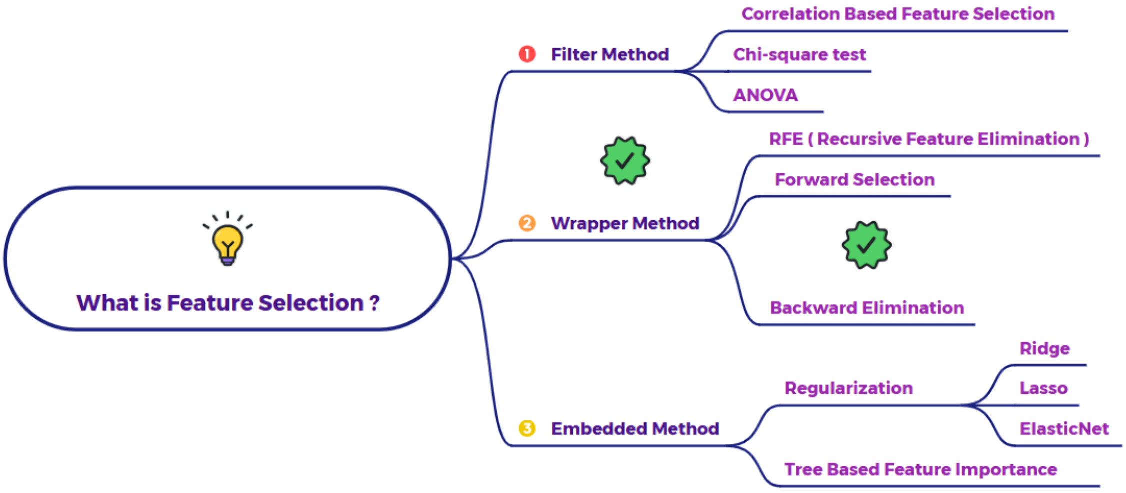
Explain Backward elimination based feature selection



Wrapper Method - Backward Elimination for feature selection

This method is the counterpart to forward selection. Instead of starting with no features and adding them, backward elimination starts with all features and iteratively removes the least beneficial ones until a stopping criterion is met.

Core Idea:

Backward elimination begins with the entire set of features in the model. It then iteratively removes one feature at a time, selecting the feature whose removal leads to the smallest decrease in the model's performance. This process continues until removing more features significantly degrades performance or a predefined number of features remains.

How it Works:

1. **Start with All Features:** Begin with the entire set of features in your dataset.
2. **Train the Model on All Features:** Train the chosen machine learning model using all the features. Evaluate its performance on a validation set or using cross-validation.
3. **Evaluate Feature Importance (or Impact of Removal):** For each feature in the current set, evaluate the impact of removing it from the model. This can be done by:
    - **Retraining the model without that specific feature and evaluating its performance.**
    - **Using model-specific feature importance scores (if available) and considering the least important feature for removal.**

4. **Remove the Least Significant Feature:** Identify the feature whose removal results in the smallest decrease in model performance (or the feature with the lowest importance score). Remove this feature from the current set of features.
5. **Repeat:** Repeat steps 2-4 on the reduced set of features. Retrain the model, evaluate the impact of removing each remaining feature, and remove the least significant one.
6. **Stopping Criterion:** Continue the process until a predefined stopping criterion is met. Common stopping criteria include:
   - Reaching a desired number of features.
   - When removing a feature leads to a significant drop in model performance (the decrease exceeds a certain threshold).
   - When the performance on the validation set starts to worsen considerably.
7. **Final Selected Features:** The set of features remaining when the stopping criterion is met is the final set of features chosen by backward elimination.

Example: Predicting Employee Salary (Regression)

Let's say we want to predict the salary (numerical target) of employees using the following numerical features:

- Years of Experience
- Education Level (encoded numerically: 1=High School, 2=Bachelor's, 3=Master's, 4=PhD)
- Number of Projects Completed
- Commute Distance (in miles)
- Employee Satisfaction Score (on a scale of 1-10)

We'll use Linear Regression as our base model and R-squared as our evaluation metric.

Steps:

1. **Start with All Features:** Selected features = {Years of Experience, Education Level, Number of Projects Completed, Commute Distance, Employee Satisfaction Score}
2. **Train on All Features:** Train a Linear Regression model using all five features and evaluate R-squared (e.g., $R^2 = 0.82$).
3. **Evaluate Feature Removal:** For each feature, we temporarily remove it and train the model on the remaining four features, noting the change in R-squared:
   - Removing "Years of Experience": $R^2 = 0.78$ (drop of 0.04)
   - Removing "Education Level": $R^2 = 0.81$ (drop of 0.01)
   - Removing "Number of Projects Completed": $R^2 = 0.79$ (drop of 0.03)

    - Removing "Commute Distance": R^2 = 0.815 (drop of 0.005)
    - Removing "Employee Satisfaction Score": R^2 = 0.75 (drop of 0.07)
4. **Remove Least Significant:** Removing "Commute Distance" results in the smallest drop in R-squared (0.005). So, we remove "Commute Distance". Selected features = {Years of Experience, Education Level, Number of Projects Completed, Employee Satisfaction Score}
5. **Repeat (Iteration 1):**
    - Train a Linear Regression model with the remaining four features and evaluate R-squared (let's say R^2 = 0.815).
    - Evaluate the impact of removing each of these four features:
        - Removing "Years of Experience": R^2 = 0.77 (drop of 0.045)
        - Removing "Education Level": R^2 = 0.80 (drop of 0.015)
        - Removing "Number of Projects Completed": R^2 = 0.785 (drop of 0.03)
        - Removing "Employee Satisfaction Score": R^2 = 0.74 (drop of 0.075)
    - Removing "Education Level" results in the smallest drop (0.015). So, we remove "Education Level". Selected features = {Years of Experience, Number of Projects Completed, Employee Satisfaction Score}
6. **Continue until Stopping Criterion:** We continue this process, evaluating the removal of each remaining feature and stopping when removing any feature leads to a significant drop in R-squared or we reach a desired number of features.

Advantages of Backward Elimination (Wrapper Method):

- **Considers Feature Interactions:** Similar to forward selection and RFE, it evaluates features in the context of the model and thus considers potential interactions between them.
- **Potentially Better Performance than Filter Methods:** By directly optimizing for the model's performance, it can often lead to better predictive accuracy.
- **Can Identify Less Useful Features Early:** It starts with all features, which might be beneficial if you suspect many features are irrelevant.

Disadvantages of Backward Elimination (Wrapper Method):

- **Computationally Expensive:** Training the model repeatedly with different subsets of features can be computationally intensive, especially with a large number of initial features.
- **Model Dependent:** The selected feature subset is specific to the chosen machine learning model.

- **Can be Sensitive to Initial Model Performance:** If the initial model with all features is poor due to noise or irrelevant features, the elimination process might be less effective in identifying the truly important ones.

In summary, Backward Elimination is a wrapper method that starts with all features and iteratively removes the least significant one based on its impact on model performance. It continues this process until a defined stopping point is reached, resulting in a subset of features deemed most important for the chosen machine learning model.