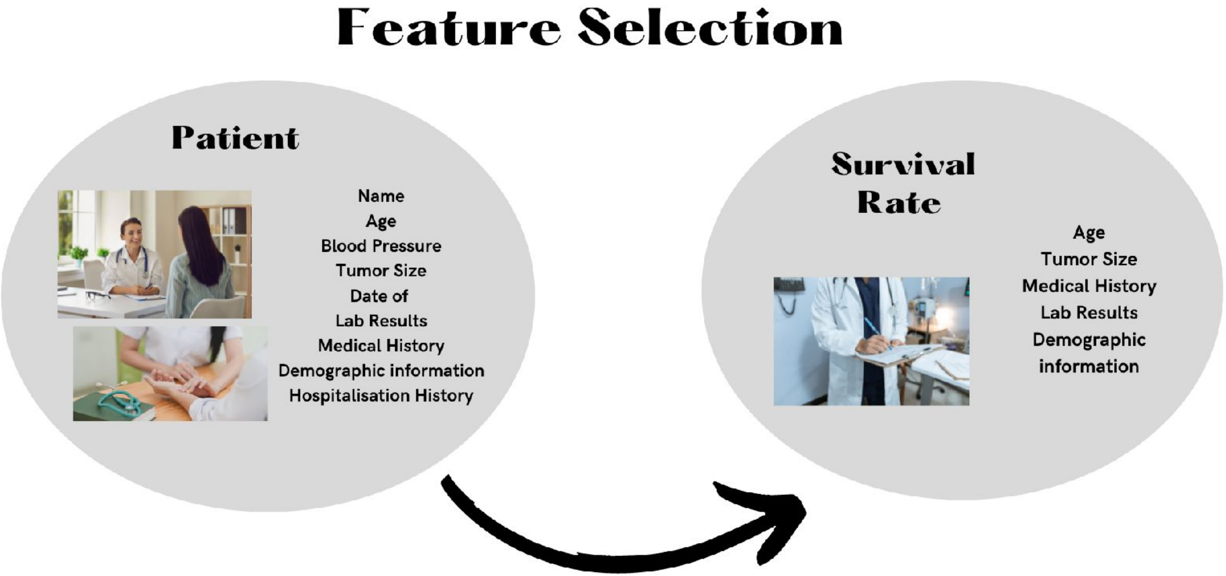
U:Pass Week 5

36106: Machine Learning Algorithms and Application
By Marisara Satrulee

Feature Selection

Feature Selection is a technique to find the **best set of features** that optimise the training of a model.



Feature Selection

Patient

Name
Age
Blood Pressure
Tumor Size
Date of
Lab Results
Medical History
Demographic information
Hospitalisation History

Survival Rate

Age
Tumor Size
Medical History
Lab Results
Demographic information

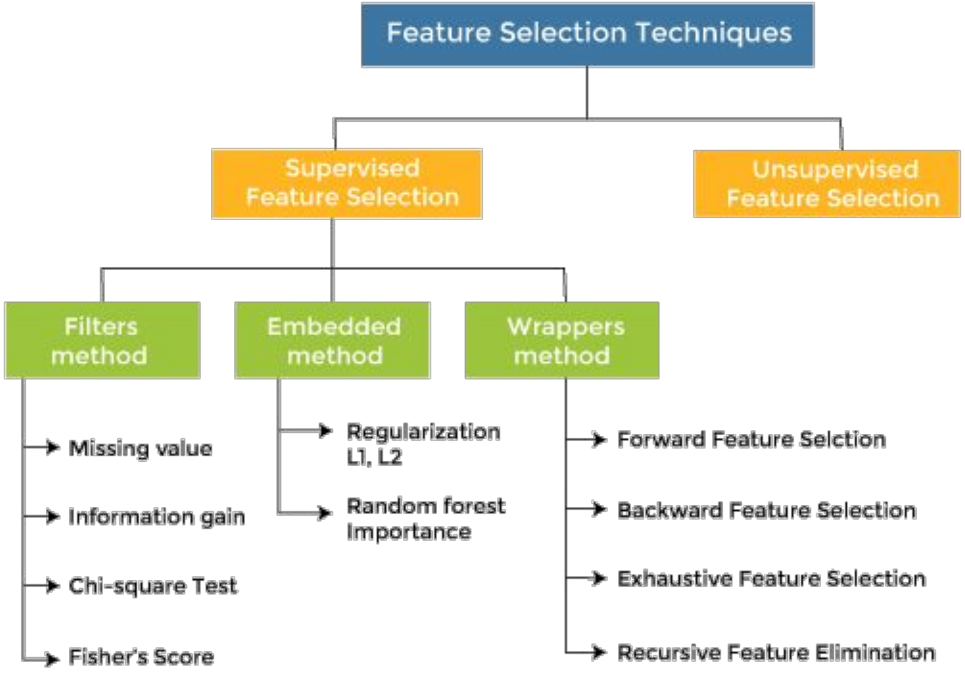
Benefits of Feature Selection

1. **Improved** Model Performance: By focusing on the most informative features, models can achieve higher accuracy, lower variance, and better generalization.
2. **Reduced** Overfitting: Feature selection helps mitigate the risk of overfitting by removing irrelevant or redundant features, allowing models to generalize well to unseen data.
3. **Faster Training**: Selecting relevant features reduces the computational complexity, resulting in faster training.
4. Enhanced **Model Interpretability**: We gain insights into the underlying patterns and relationships by selecting a subset of meaningful features, enabling better model interpretability.

Reference: https://www.linkedin.com/pulse/unleashing-power-feature-selection-machine-learning-ravi-singh/

Types of Feature Selection

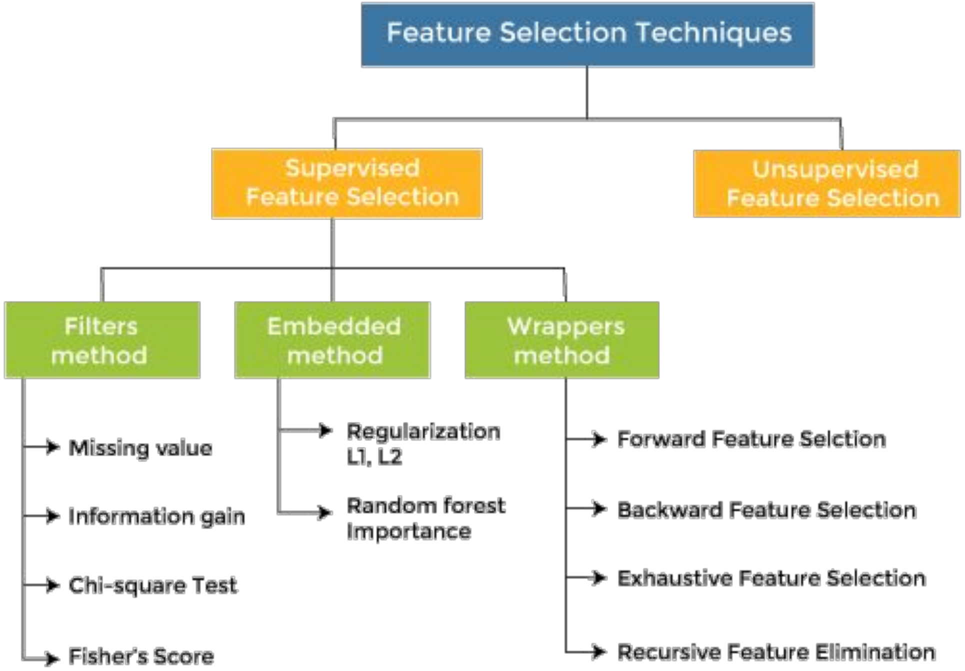
Filter Methods: Features are dropped based on their **correlation** to the output.



Reference: https://www.linkedin.com/pulse/unleashing-power-feature-selection-machine-learning-ravi-singh/

Types of Feature Selection

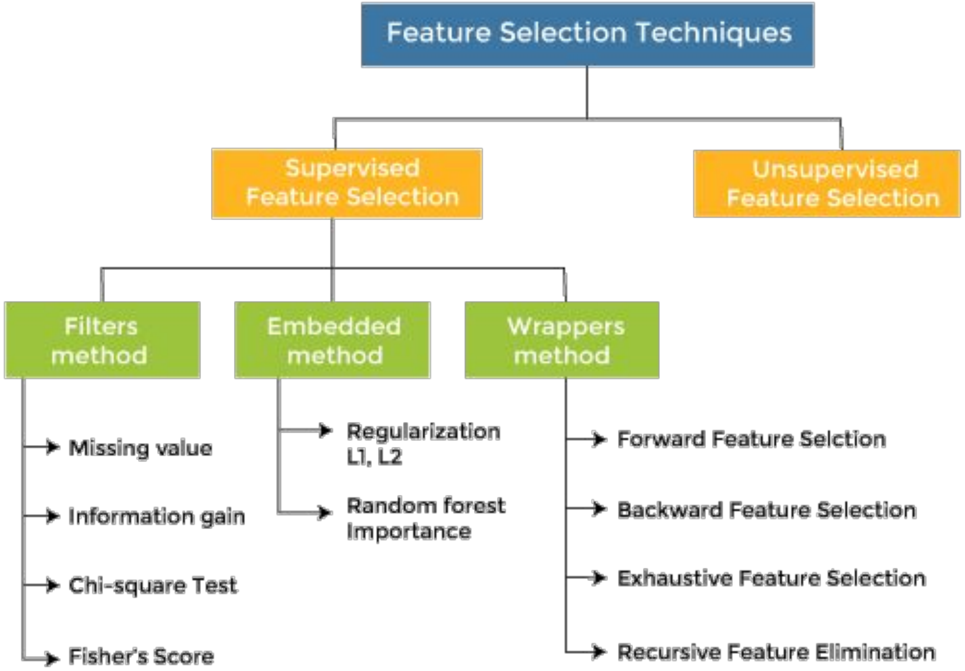
Wrapper Methods: the wrapper method is a technique that selects a subset of features by evaluating a model's performance using different subsets of features.



Reference: https://www.linkedin.com/pulse/unleashing-power-feature-selection-machine-learning-ravi-singh/

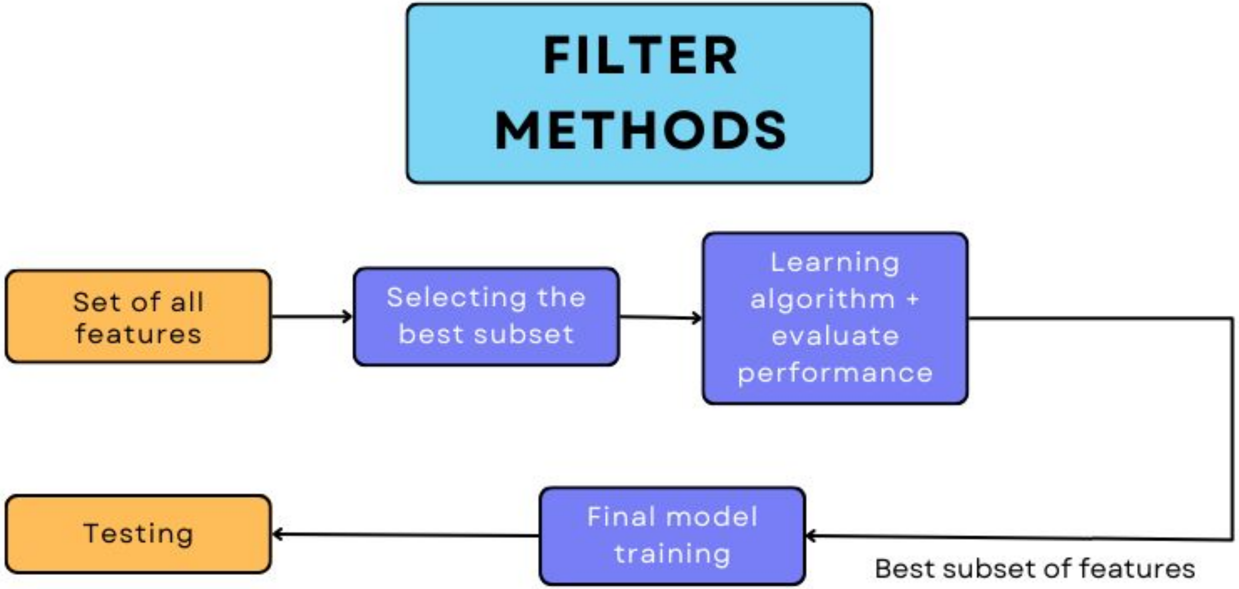
Types of Feature Selection

Embedded Methods: is the combination of Filter and Wrapper Methods



Reference: https://www.linkedin.com/pulse/unleashing-power-feature-selection-machine-learning-ravi-singh/

Feature Selection with Python



FILTER METHODS

Set of all features

Selecting the best subset

Learning algorithm + evaluate performance

Best subset of features

Final model training

Testing

Reference: https://www.analyticsvidhya.com/blog/2022/11/feature-selection-101-the-manual-for-beginners/

Feature Selection with Python

Steps in **Filter Methods** for feature selection

1. Evaluate the relevance of each feature using statistical measures
2. Rank the features based on their relevance scores
3. Select a threshold
4. Remove all features that fall below the threshold
5. Train the model, using the remaining features

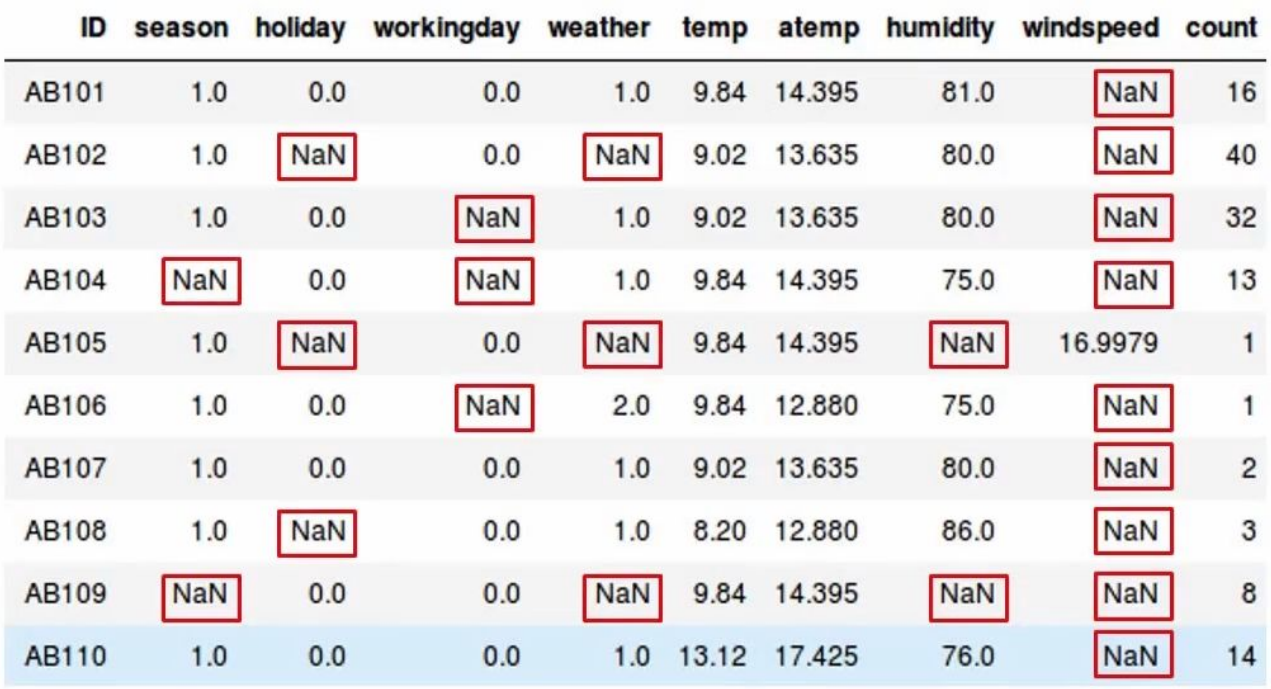
Week 5 - Code snippets

https://github.com/merrymira/UPASS_ML_WEEK5

Codes for all weeks:
https://github.com/merrymira/UPASS_ML_WEEK3/

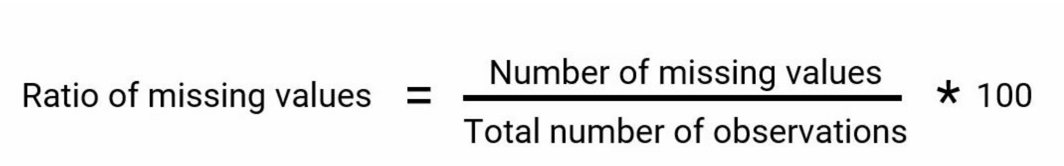
Filter Method: Missing Value

Missing Value



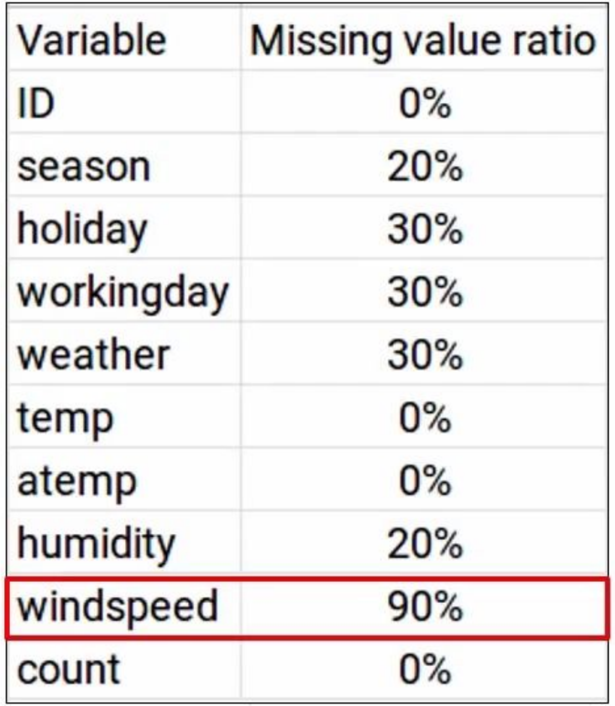
ID	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	count
AB101	1.0	0.0	0.0	1.0	9.84	14.395	81.0	NaN	16
AB102	1.0	NaN	0.0	NaN	9.02	13.635	80.0	NaN	40
AB103	1.0	0.0	NaN	1.0	9.02	13.635	80.0	NaN	32
AB104	NaN	0.0	NaN	1.0	9.84	14.395	75.0	NaN	13
AB105	1.0	NaN	0.0	NaN	9.84	14.395	NaN	16.9979	1
AB106	1.0	0.0	NaN	2.0	9.84	12.880	75.0	NaN	1
AB107	1.0	0.0	0.0	1.0	9.02	13.635	80.0	NaN	2
AB108	1.0	NaN	0.0	1.0	8.20	12.880	86.0	NaN	3
AB109	NaN	0.0	0.0	NaN	9.84	14.395	NaN	NaN	8
AB110	1.0	0.0	0.0	1.0	13.12	17.425	76.0	NaN	14

Filter Method: Missing Value



$$\text{Ratio of missing values} = \frac{\text{Number of missing values}}{\text{Total number of observations}} * 100$$

Some sources say, dropping the column when the missing values is more than 5-10%. It is also said that null value columns should be **only dropped** when the number of **records is in millions**. ***



Variable	Missing value ratio
ID	0%
season	20%
holiday	30%
workingday	30%
weather	30%
temp	0%
atemp	0%
humidity	20%
windspeed	90%
count	0%

Let’s say 70% is the threshold, so you will **drop** any feature with missing values over 70%

Filter Method: Information Gain

Information gain calculates the reduction in **entropy** from the transformation of a dataset. It can be used for feature selection by evaluating the Information gain of each variable in the context of the target variable.

Talking about “**entropy**”

In Machine Learning, entropy is a metric used to **measure the amount of information** in a dataset. It's commonly employed to evaluate the model's quality and its ability to make accurate predictions. A higher entropy value indicates a **heterogeneous** dataset with diverse classes, while a lower entropy indicates a more pure and **homogeneous** subset of data.

Reference:
https://www.analyticsvidhya.com/blog/2020/11/entropy-a-key-concept-for-all-data-science-beginners/#:~:text=In%20Machine%20Learning%2C%20entropy%20measures,ability%20to%20make%20accurate%20predictions.

Heterogeneous vs Homogeneous

We want the dataset to be homogeneous.

As a general rule, when dealing with a heterogeneous population, the population should be divided into as many groups as necessary to ensure that each subgroup is sufficiently homogeneous for the sampling purpose as defined by the audit objective and scope.

“Meaning that no one is left behind.”

Reference:
https://www.caaf-fcar.ca/en/sampling-methodology-concepts-and-context/population-characteristics/homogeneity-and-heterogeneity#:~:text=The%20more%20homogenous%20a%20population,results%20obtained%20from%20a%20sample.

Heterogeneous vs Homogeneous

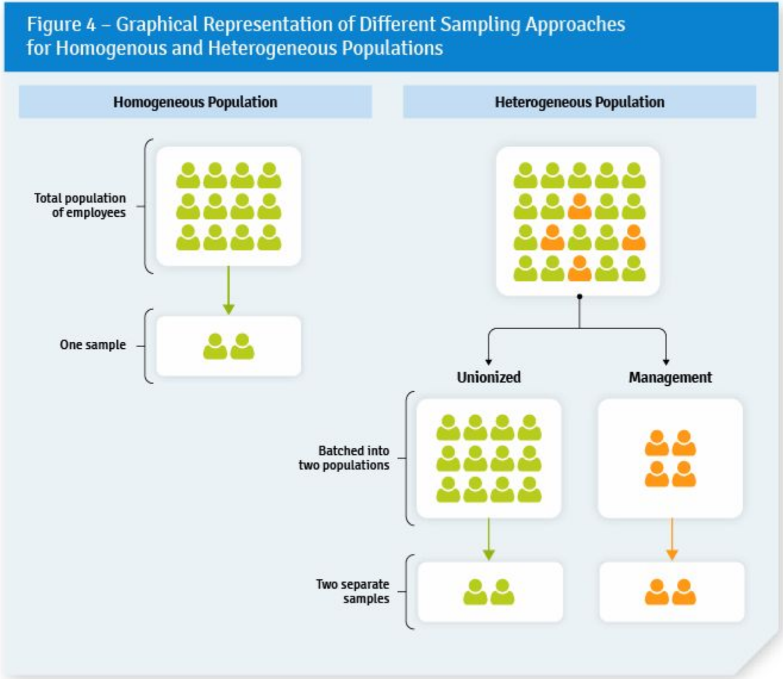


Figure 4 – Graphical Representation of Different Sampling Approaches for Homogenous and Heterogeneous Populations

In a hypothetical survey, government employees were asked about the adequacy of water and air quality in their work environment. Because they all breathe the same air and drink the same water, they form a homogenous population.

In the second example, the survey is about management style in the department. Because some employees are part of management and others are unionized, there are two distinct groups of people (two populations) that may have very different opinions about management style in the department. Therefore, it would make more sense to take two separate samples, one from each population.

Reference:
https://www.caaf-fcar.ca/en/sampling-methodology-concepts-and-context/population-characteristics/homogeneity-and-heterogeneity#:~:text=The%20more%20homogenous%20a%20population,results%20obtained%20from%20a%20sample.

Filter Method: Correlation Coefficient

Correlation is a measure of the linear relationship between 2 or more variables. Through correlation, we can predict one variable from the other. The logic behind using correlation for feature selection is that good variables correlate highly with the target. Furthermore, variables should be correlated with the target but uncorrelated among themselves.

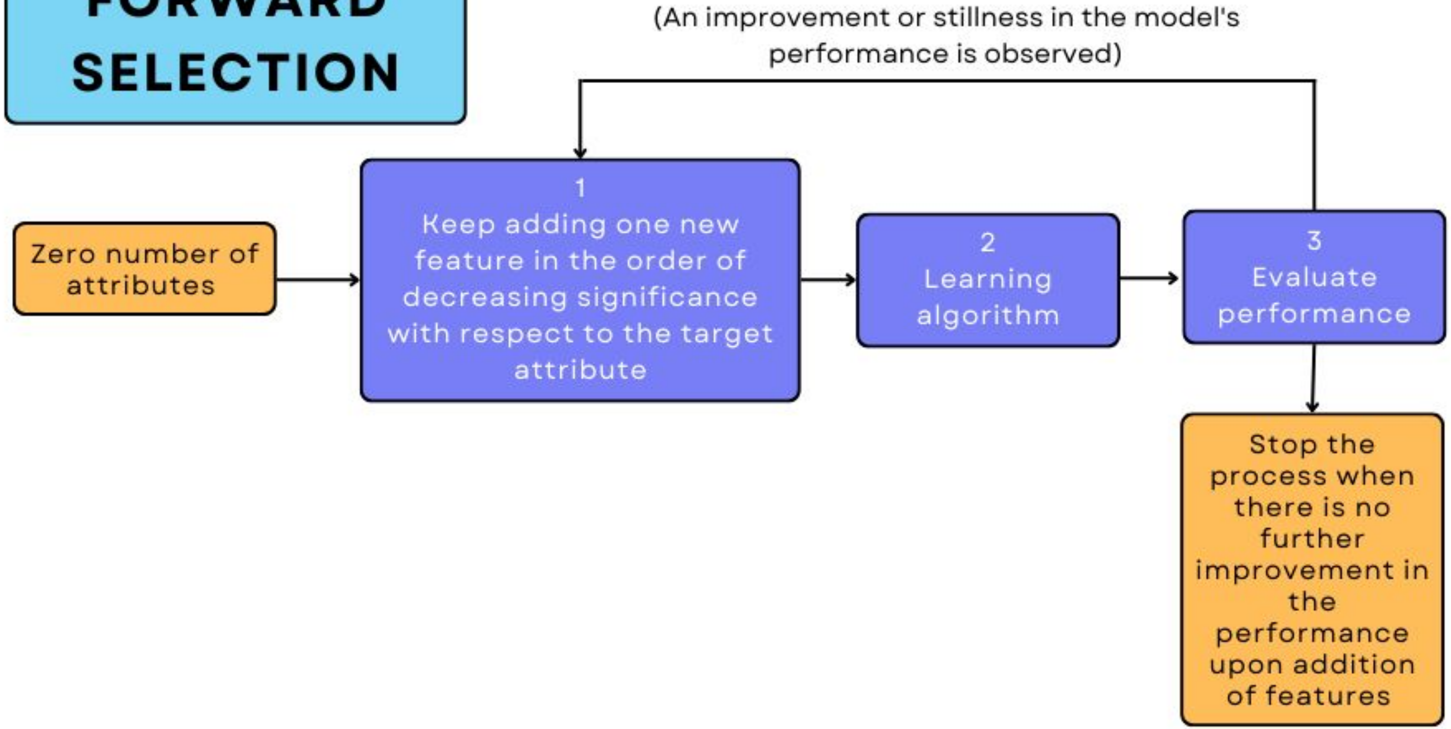
Because if two features are correlated, the model only needs one, as the second does not add additional information.

Wrapper Method

Forward Feature Selection

- **Starting from Scratch**: Begin with an empty set of features and iteratively add one feature at a time.
- **Model Evaluation**: At each step, train and evaluate the machine learning model using the selected features.
- **Stopping Criterion**: Continue until a predefined stopping criterion is met, such as a maximum number of features or a significant drop in performance.

FORWARD SELECTION



Zero number of attributes

1
Keep adding one new feature in the order of decreasing significance with respect to the target attribute

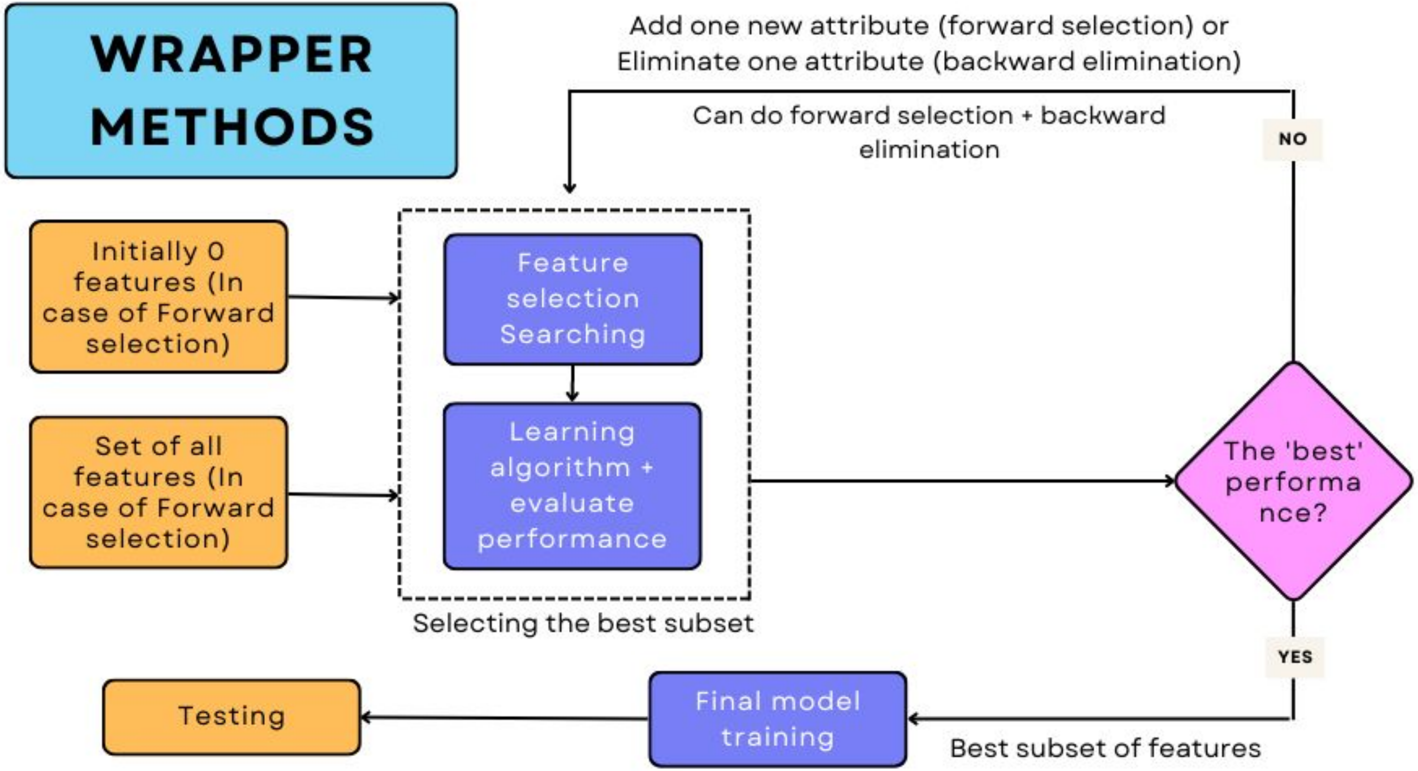
2
Learning algorithm

3
Evaluate performance

(An improvement or stillness in the model's performance is observed)

Stop the process when there is no further improvement in the performance upon addition of features

WRAPPER METHODS



Add one new attribute (forward selection) or Eliminate one attribute (backward elimination)

Can do forward selection + backward elimination

NO

Initially 0 features (In case of Forward selection)

Set of all features (In case of Forward selection)

Feature selection Searching

Learning algorithm + evaluate performance

Selecting the best subset

The 'best' performance?

YES

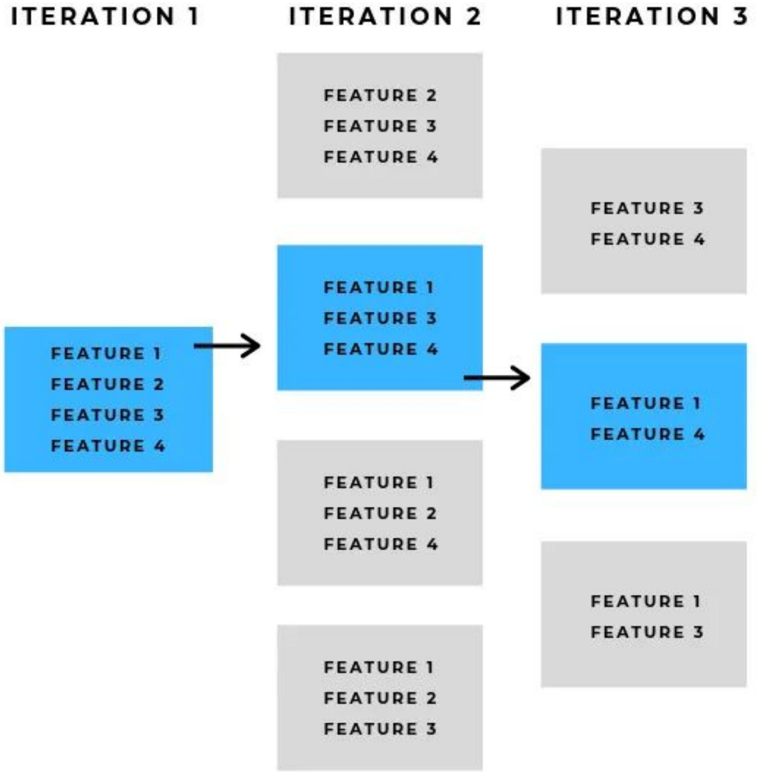
Best subset of features

Final model training

Testing

Wrapper Methods

Recursive Feature Elimination(RFE) trains the model on the original number of features, and each feature is given importance. The least important features are removed and then repeated to a specified number of features.

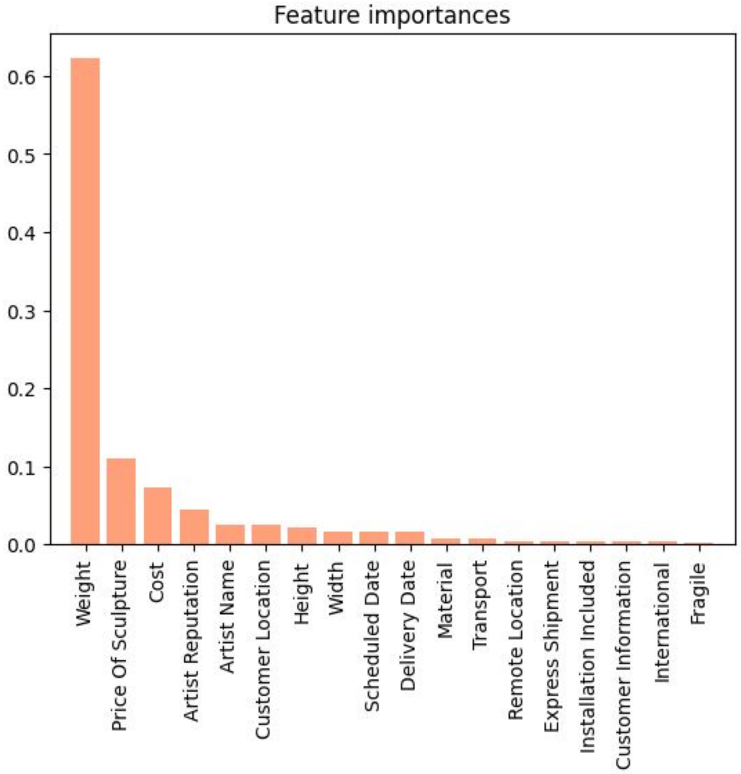


Embedded Methods

Random Forest uses mean decrease impurity (Gini index) to estimate a feature’s importance.

By looking at the above bar plot, we can conclude that not all features in the dataset contribute the same to the model.

You may drop the features with small importances based on a threshold value.



Reference: https://medium.com/data-science-365/visualizing-and-selecting-important-features-in-random-forest-5e1eb865d7c4

Try selecting your feature set

Let's conduct a small experiment to train the model on your selected function. Compare which feature sets result in better RMSE.