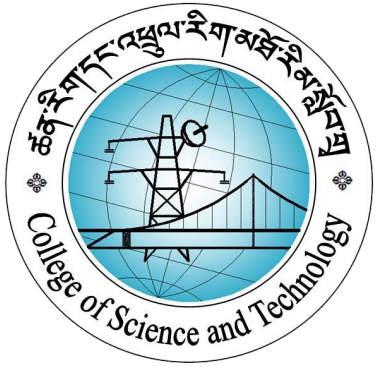


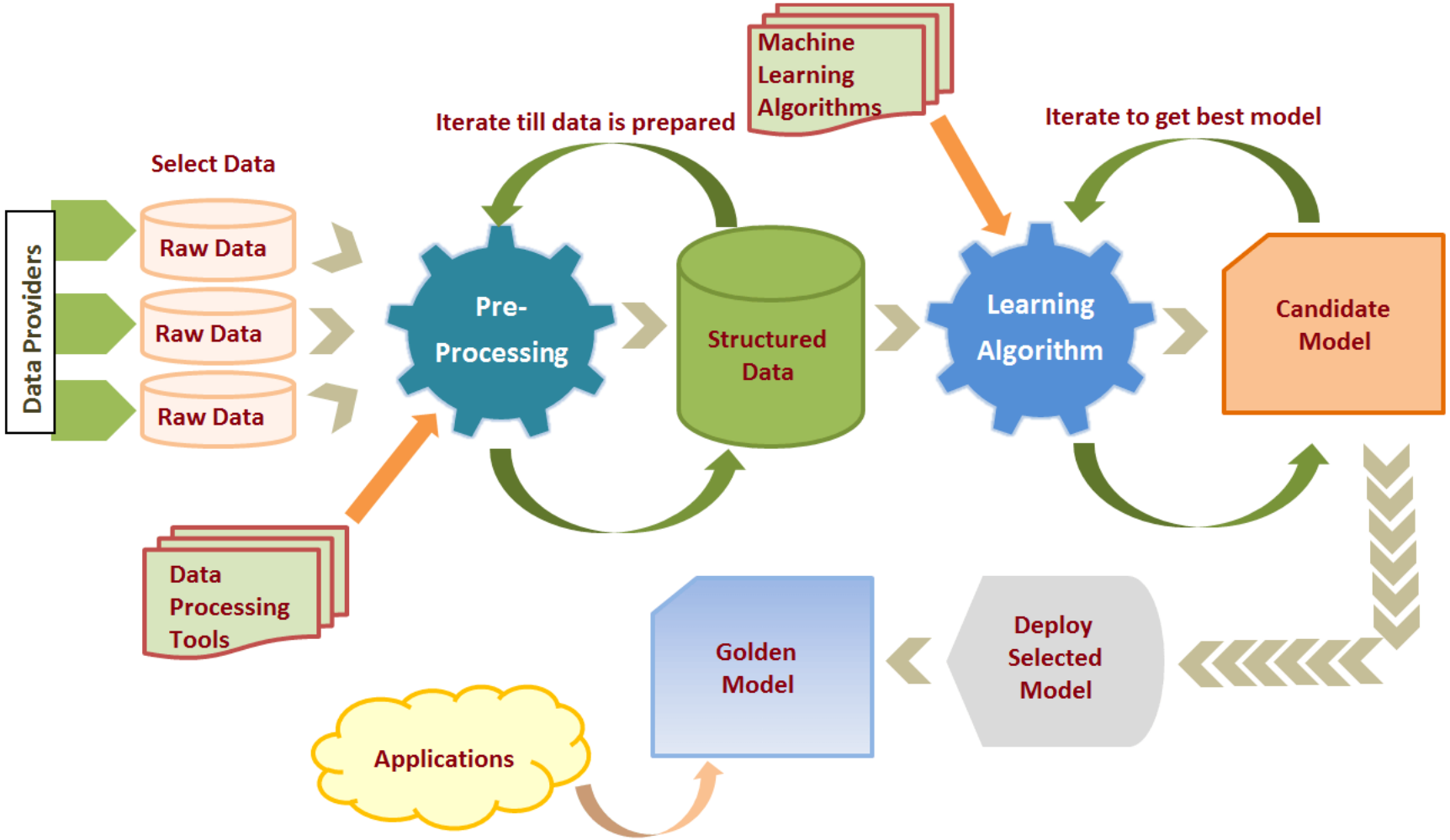
Royal University of Bhutan

Unit II: Feature Management

CTE309 Machine Learning

AS2024: BE Information Technology

Machine Learning Pipeline



Overview

- Introduction
- Why feature selection?
- What is feature selection?
- Feature selection models

Introduction

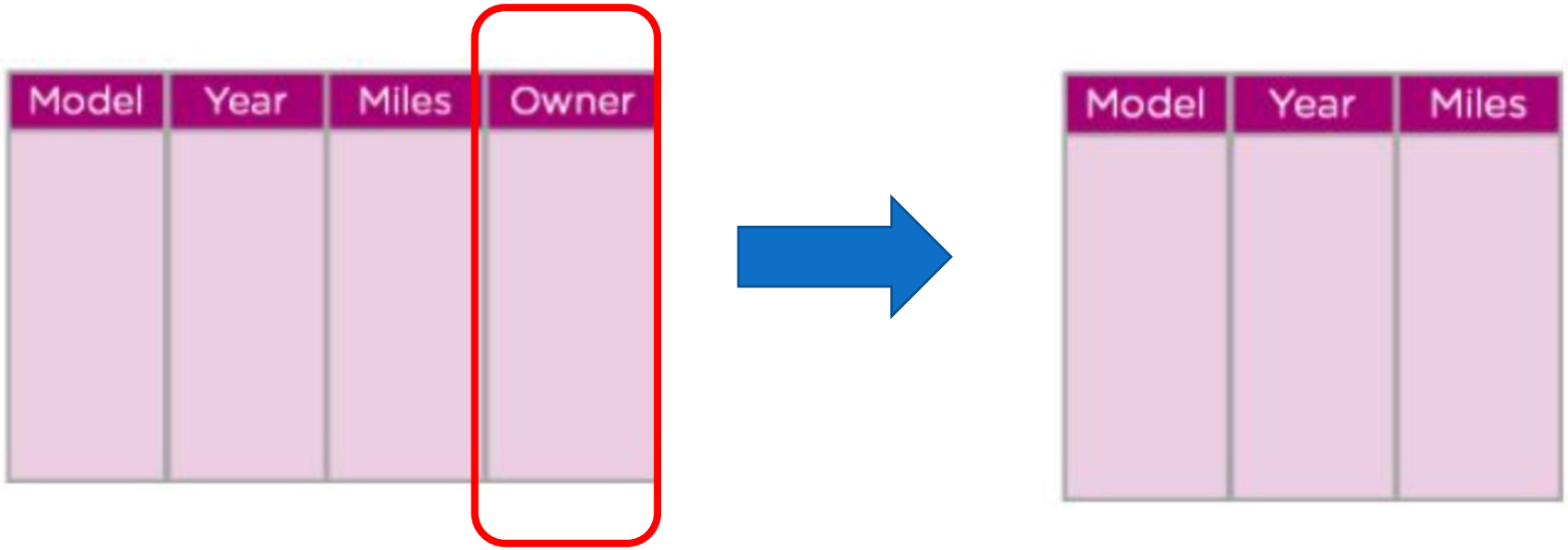
- The input variables that we give to our machine learning models are called features.
- Each column in our dataset constitutes a feature.
- To train an optimal model, we need to make sure to use only the essential features.
- If we have too many features, the model can capture the unimportant patterns and learn from noise.
- The method of choosing the important parameters of our data is called Feature Selection.

Why feature selection?

- ML follows simple rule: *Whatever goes in, comes out.*
- To train a model, we collect enormous quantities of data to help the machine learn better.
  - Usually, a good portion of the data collected is noise, while some might not contribute significantly to the performance of our model.
  - Further, having a lot of data can slow down the training process and cause the model to be slower.
  - The model may also learn from this irrelevant data and be inaccurate.
- Feature selection is what separates good data scientists from the rest.

Why feature selection?

- Feature selection is what separates good data scientists from the rest.

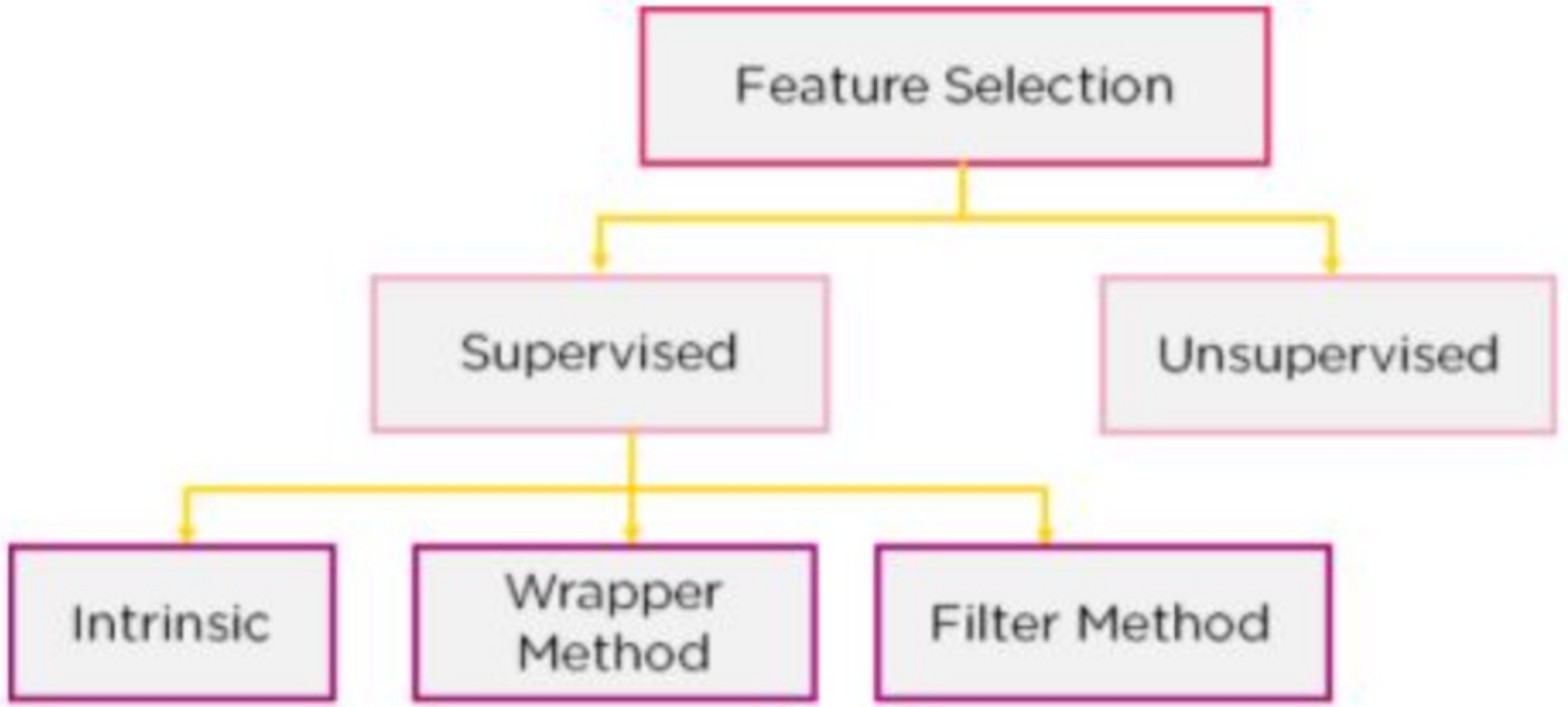


Old cars dataset

What is feature selection?

- Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data.
- helps in cutting down the noise in our data and reducing the size of our input data.

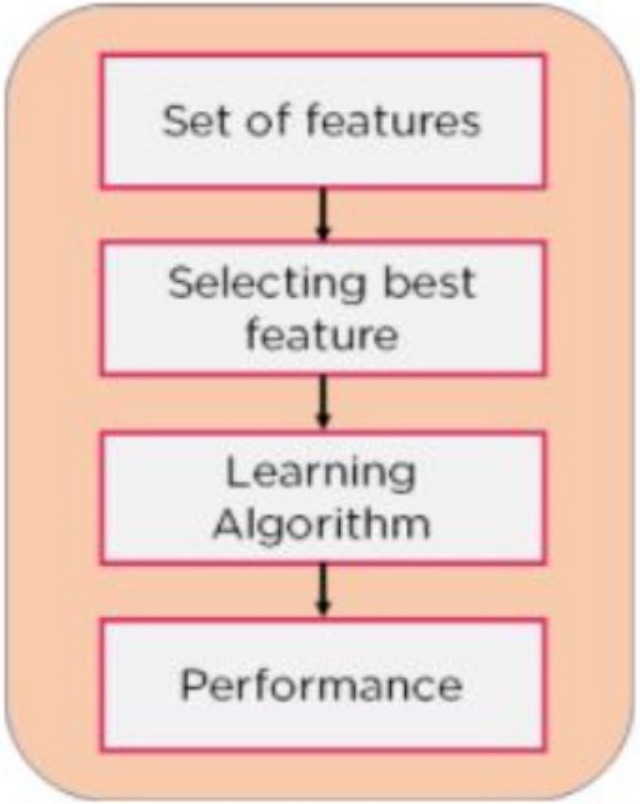
Models



Feature Selection Models

Filter method - Supervise selection model

- features are dropped based on their relation to the output, or how they are correlating to the output.
- We use correlation to check if the features are positively or negatively correlated to the output labels and drop features accordingly.
- Filter Methods
  - Univariate Feature Selection
  - Correlation-based Feature Selection
  - Variance Thresholding



Filter method – Univariate Feature Selection

- is a feature selection technique where the relationship between dependent variables and independent variables is evaluated.
- Independent variables with the strongest relationship with dependent variables are chosen.
- Chi-squared
  - evaluates the independence of a categorical characteristic from the target variable, establishing its relevance and usefulness for classification tasks
- Analysis of Variance (ANOVA)
  - employed in regression tasks
  - examines statistically significant variations in the means of the target variable across categorical features.
- Mutual Information
  - measures the amount of information shared by two random variables, quantifying the relationship between each feature and the goal variable.

Filter method – Correlation based Feature Selection

- a technique that focuses on finding and selecting the most relevant features from a dataset.
- It accomplishes this by analyzing the relationship between each feature and the target variable.
- identifies the most significant aspects that play a critical role in predicting the target variable by selecting features with strong correlations to the target.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \times (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \times \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Filter method – Variance Thresholding

- is a feature selection technique that eliminates low-variance features from a dataset since they may not be relevant for prediction.
- These features are nearly constant, which means their values do not vary much and are less likely to contribute considerably to the model's performance.
- choose an acceptable variance threshold to determine which features are kept and which are removed based on their variability between samples.

References

- https://medium.com/@jdkiptoon/feature-selection-in-machine-learning-20417d052b80

Thank you