Employee Attrition Classification

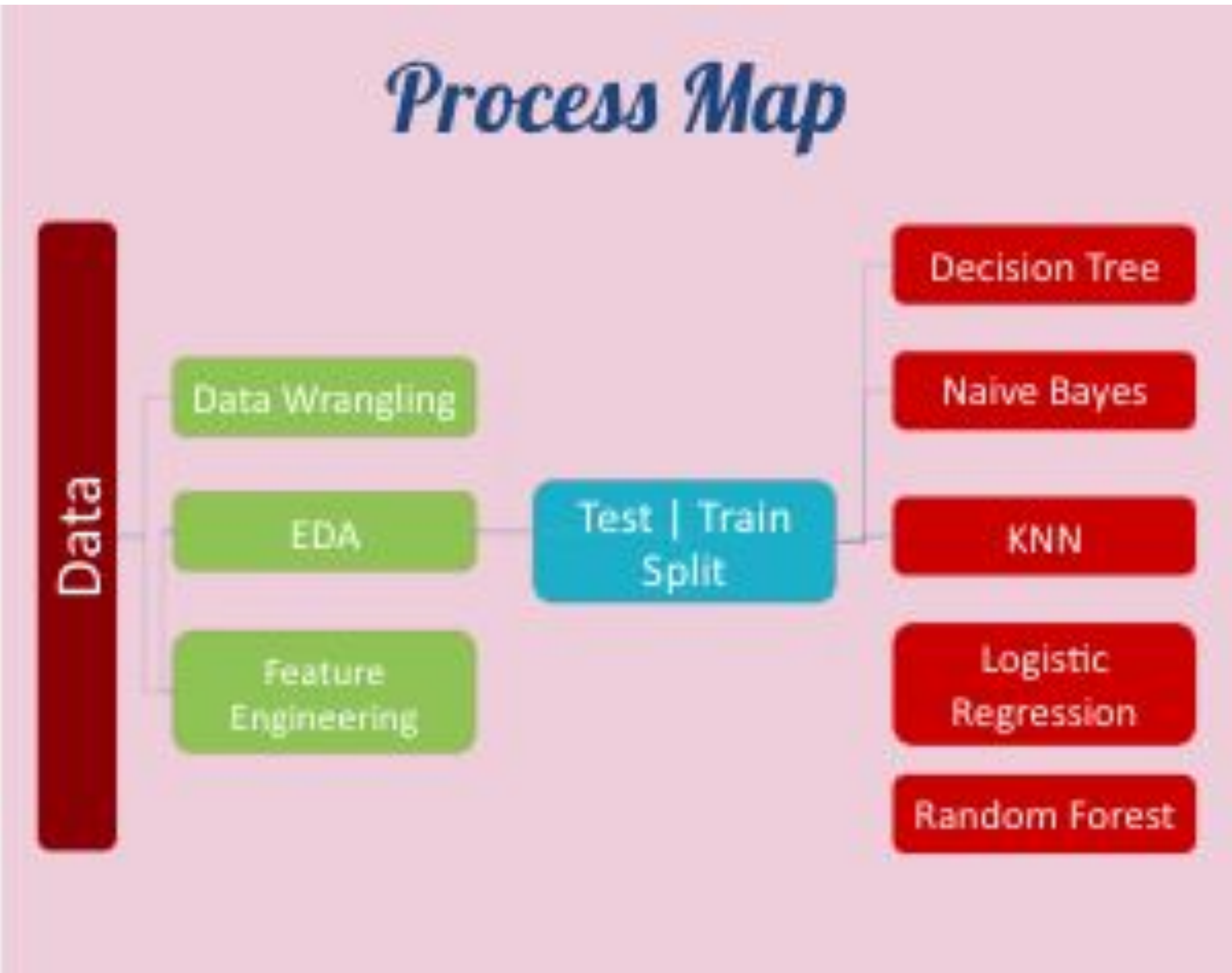
STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction:

- **Real world problem:** Airbnb has to set the right price for the property which satisfies not only the host but also is competitive enough for the market and demands of the customers.
- **Purpose:** This projects aims to perform analysis to figure out the likelihood of an active employee leaving the company and the key indicators surrounding it.
- We make predictions on the time period when employees will end the job, understanding the main drivers of employee churn.

Experiment:

- We try to delve deep in the dataset to identify patterns and trends in the dataset.
- We have used feature engineering to create new features in order to enhance modeling.
- We created a variable "EXPERIENCE_AT_COMPANY" based on the number of years of experience that a person has in the past five years at the company.
- The PCA (Principal Component Analysis) is performed on dataset after removing two columns: EMP_ID and TERMINATION_YEAR. It is done using 116 columns.



Results:

- From the PCA, we infer that 106 out of 116 are ideal for explaining 100% variance. But the accuracies are not high. So, PCA not useful.
- In Random Forest, there was high accuracy. Hyperparameters are "max_depth" of 20 & "n_estimators" as 100.

- In Logistic Regression, "ANNUAL_RATE" and "HRLY_RATE" decreased accuracy by 2 percent.
- In Decision Trees, "ANNUAL_RATE" was removed as it correlated with "HRLY_RATE".
- The KNN and the Naive Bayes models have performed equally. Removing a lot of features didn't reduce the accuracy below 60%.

Conclusion:

- As we have applied a lot of models for analysis, we are certain that the boosting algorithms predict the maximum accuracy.
- Our analysis state that factors which are the most influential ones causing attrition are:

1. Job_Groups
2. Performance Ratings
3. Job Code
4. Experience
5. Annual Rate

- These inline with general intuitions.