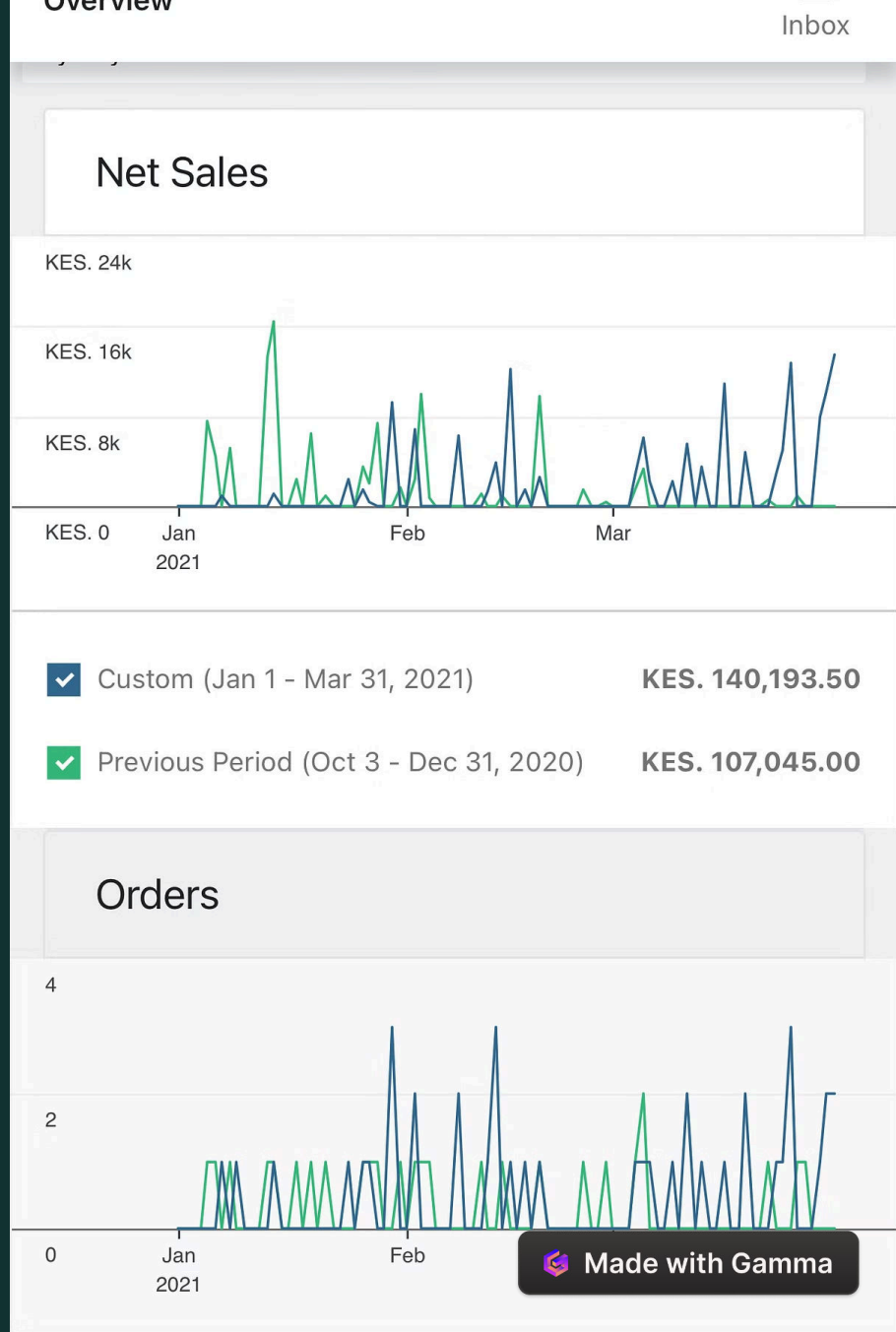
employee data prediction by using ensemble booster

TA by Tejaswini



PROBLEM STATEMENT

Prepare a Model by utilizing machine learning techniques, this dataset aims to provide valuable insights into the factors influencing the organisation and enhance the efficiency of the employees.

This presentation explores the application of exploratory data analysis (EDA) and classification algorithms to employee data. Understanding these techniques is crucial for HR decision-making and workforce management.

DATA ANALYSIS:

· Collected data in CSV format from kaggle

- Dataset consists of 6599 observation of 11 variables in which one is dependent or target variable and other 10 variables are independent.
- dataset contains variables like job_title , experience_level, employment_type , work_models , work_year, employee_residence, salary,salary_currency ,salary_in_usd ,company_location, company_size.

Target Variable : company_size of employees

Data Cleaning and Preprocessing

Handling Missing Values

The process of imputing or removing missing data to ensure dataset completeness.

Duplicates Removal

Duplicate entries can skew results and therefore must be carefully removed to uphold the integrity of the dataset.

Outliers handling and removing

Detecting and removing outliers is an essential step in data preprocessing to ensure that the data used for analysis or modeling is accurate and representative.

Winsorization: Adjusting extreme values to a more moderate level.

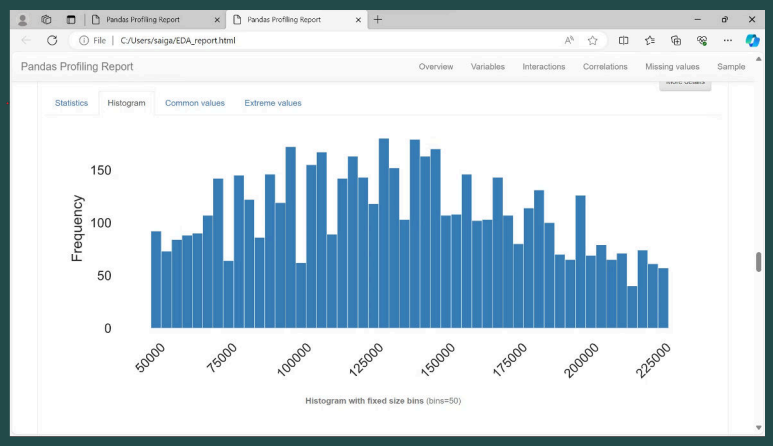
Importance of Exploratory Data Analysis (EDA)

Understanding Dataset Characteristics

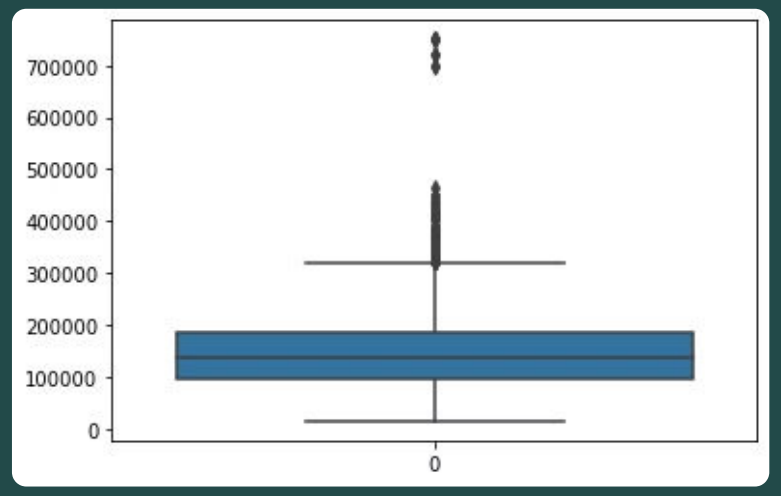
EDA helps in comprehending the distribution, relationships, and key statistical measurements of employee from the dataset

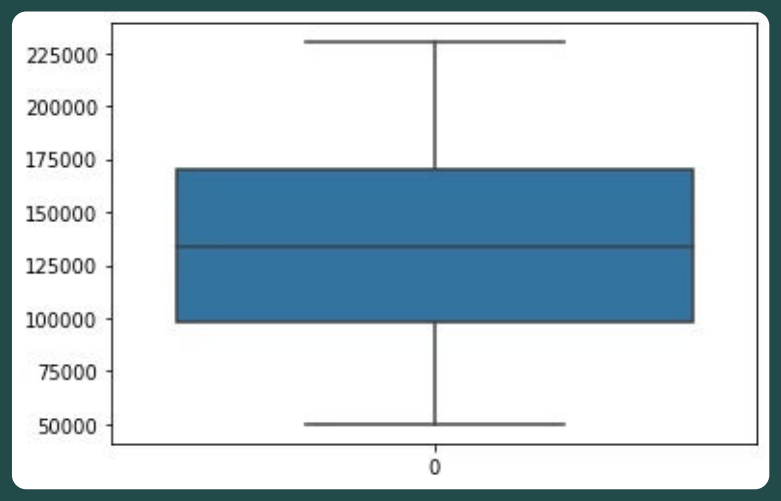
Visual Exploration of Data

Histograms

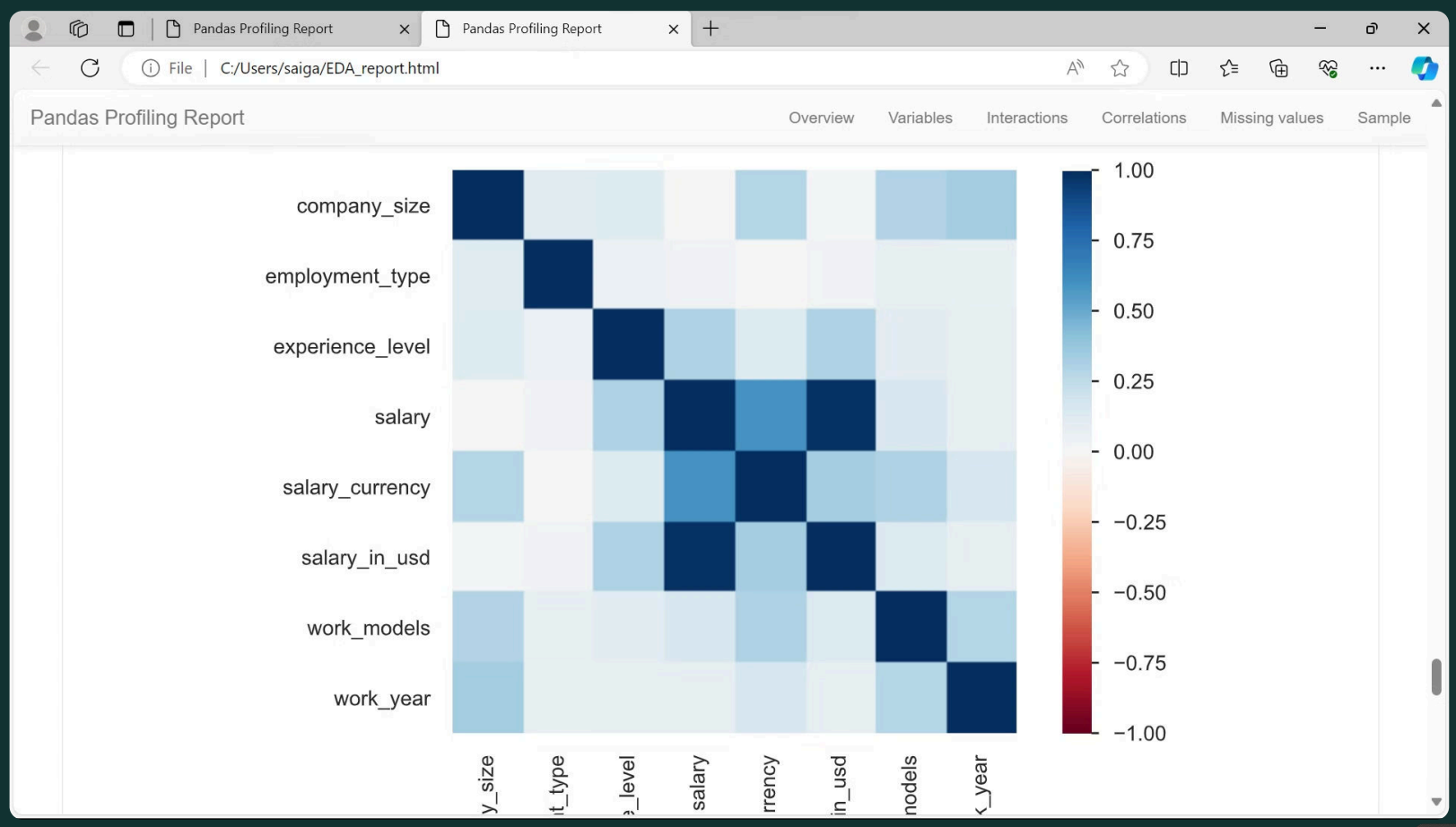


Boxplotts





Heatmap

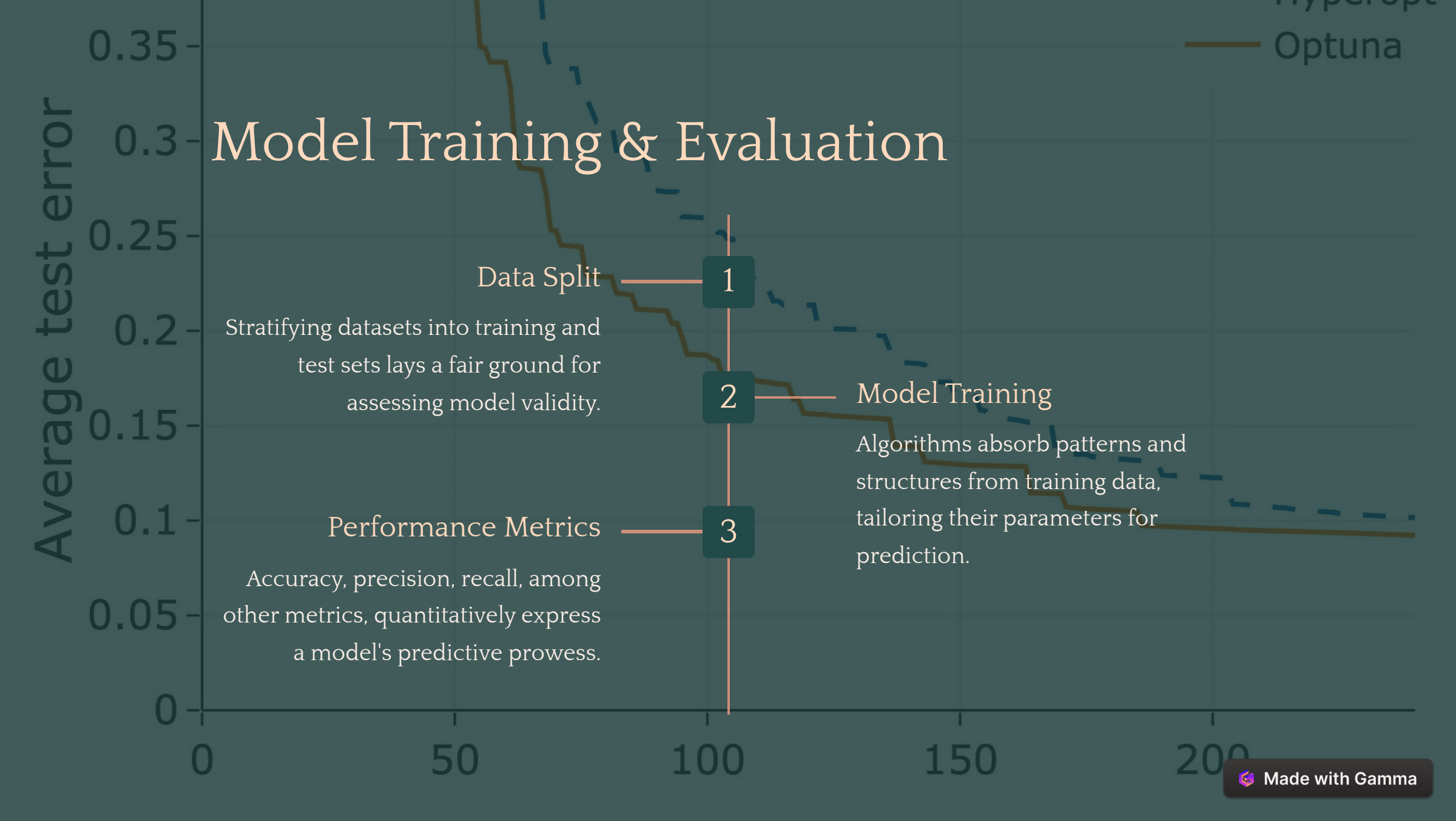


Feature Engineering Strategies

1

Encoding Categoricals

Turning categorical data into numerical formats like one-hot encoding further accommodates algorithmic processing.



Model Training & Evaluation

Data Split

1

Stratifying datasets into training and test sets lays a fair ground for assessing model validity.

Model Training

2

Algorithms absorb patterns and structures from training data, tailoring their parameters for prediction.

Performance Metrics

3

Accuracy, precision, recall, among other metrics, quantitatively express a model's predictive prowess.

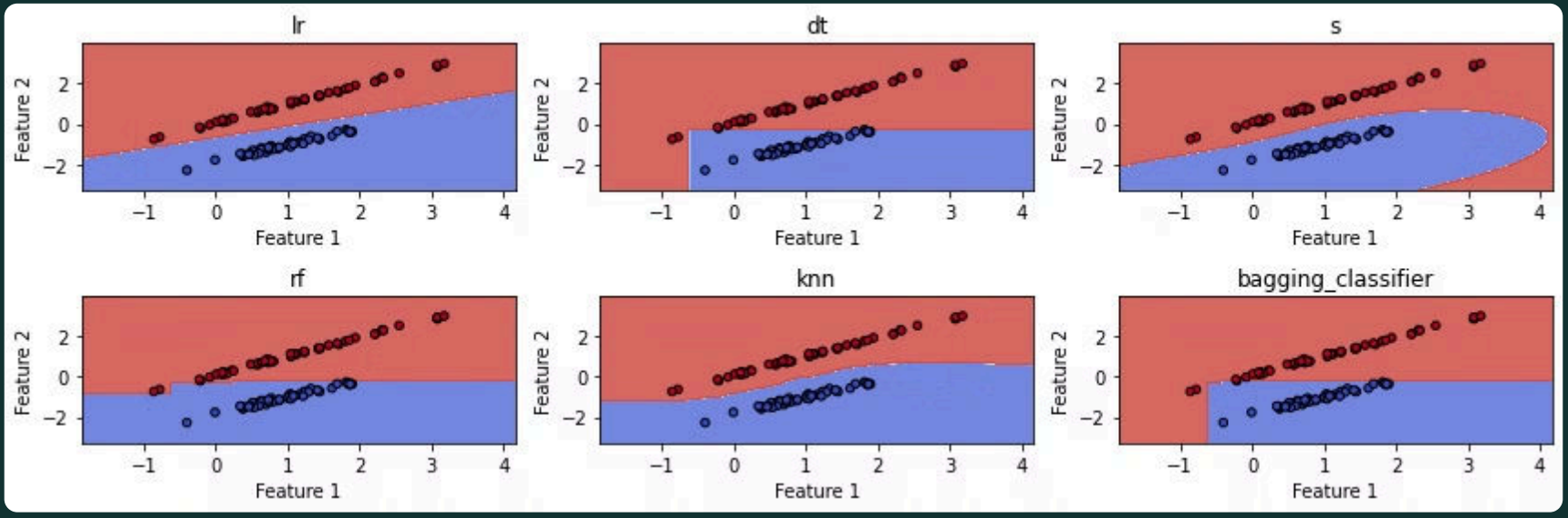
Classifying the Classifiers

Technique	Type	Description
Logistic Regression	Classification	Used for binary classification tasks, modeling the probability of an input belonging to a particular class.
Decision Trees	Both	Recursive partitioning of data into subsets based on input features, creating a tree-like structure for decision making.
Random Forests	Ensemble	Ensemble learning method combining multiple decision trees to improve predictive performance.
Support Vector Machines	Both	Finds the hyperplane that best separates classes in the feature space, used for both classification and regression tasks.
K-Nearest Neighbors	Classification	Classifies new data points based on the majority class of their nearest neighbors in the feature space.
Naive Bayes	Classification	Probabilistic classifier based on Bayes' theorem with the assumption of independence between features.
Gradient Boosting Machines	Ensemble	Builds a sequence of weak learners (often decision trees) and combines their predictions to make a final prediction.

Classifiers	accuracy
Random Forest	0.90
Logistic Regression	0.89
Support Vector Machines	0.89
Decision Tree	0.88
Support Vector Machines	0.89
knn	0.89
navies bayes	0.84
Ensemble Booster	0.91

conclusion: comparing all the accuracy randam forest classifier gives good accuracy so for ensemble booster the base_estimator is random forest classifier

visualising the classifiers



Thank you