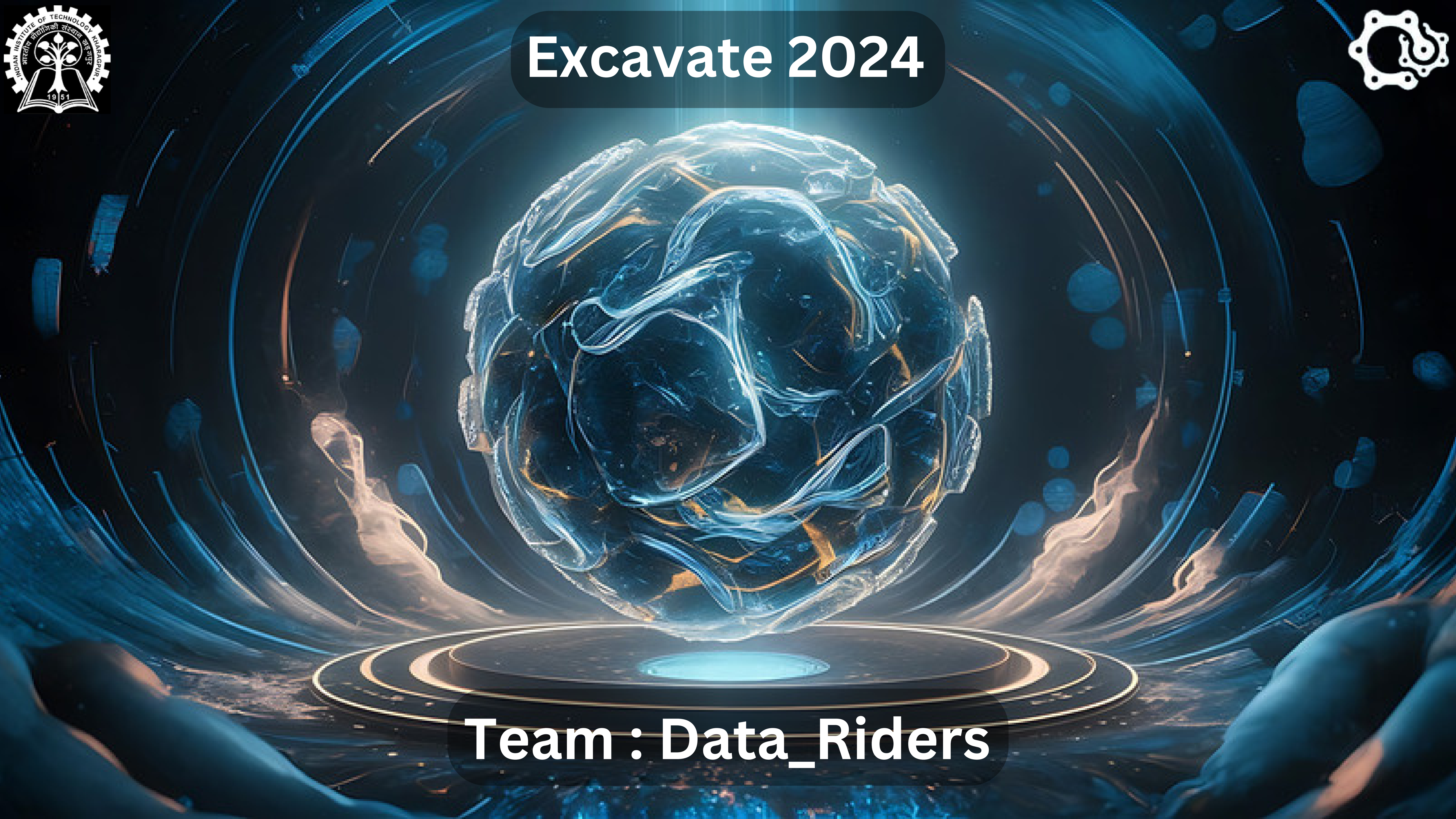
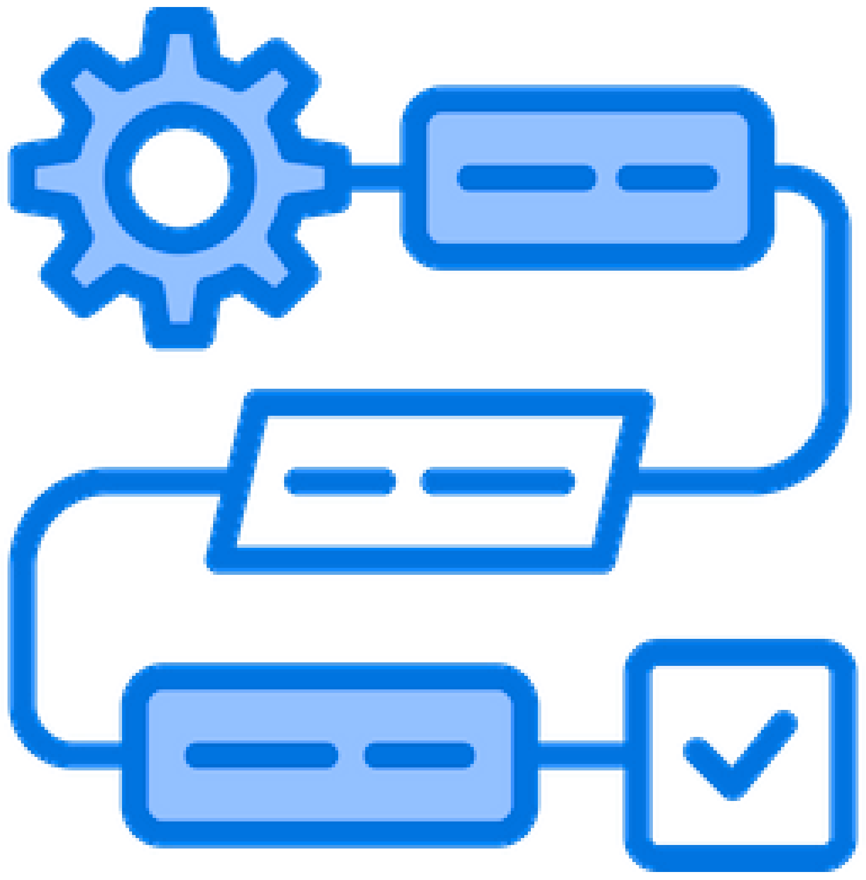
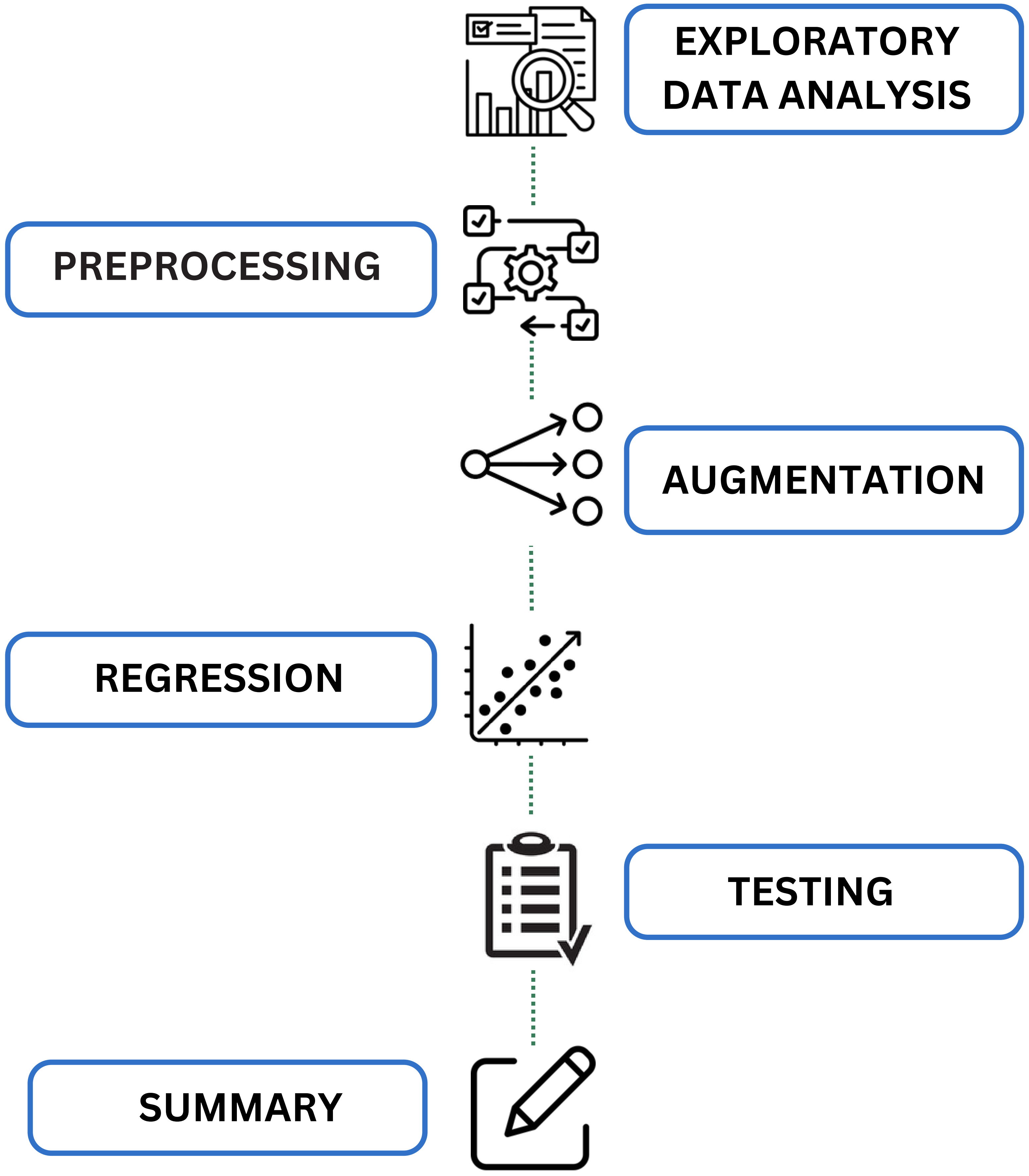


Excavate 2024

Team : Data_Riders



WORK FLOW



EXPLORATORY DATA ANALYSIS

PREPROCESSING

AUGMENTATION

REGRESSION

TESTING

SUMMARY

Exploratory Data Analysis : Finding Insights from Data

EXPLORATORY DATA ANALYSIS

PREPROCESSING

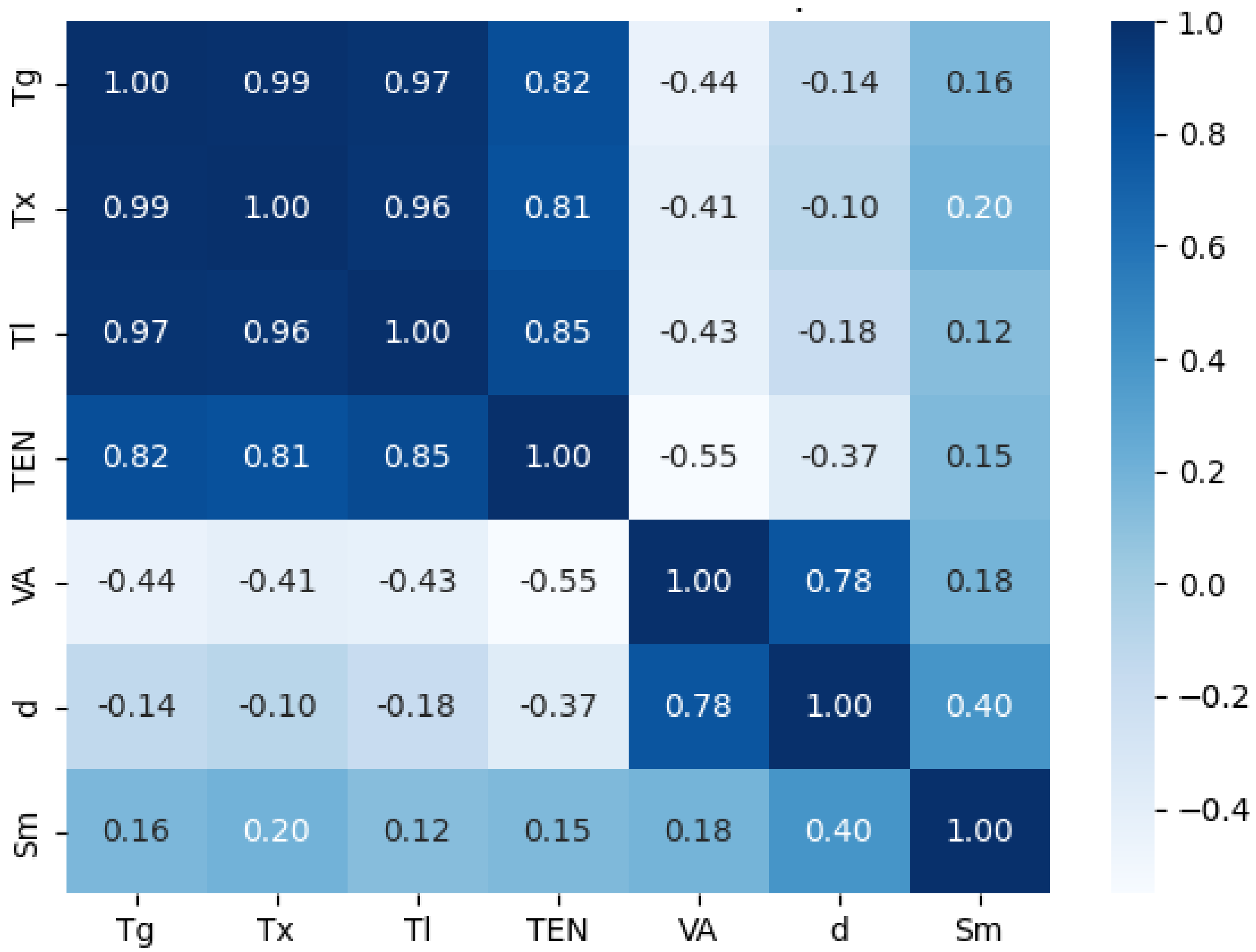
AUGMENTATION

REGRESSION

TESTING

SUMMARY

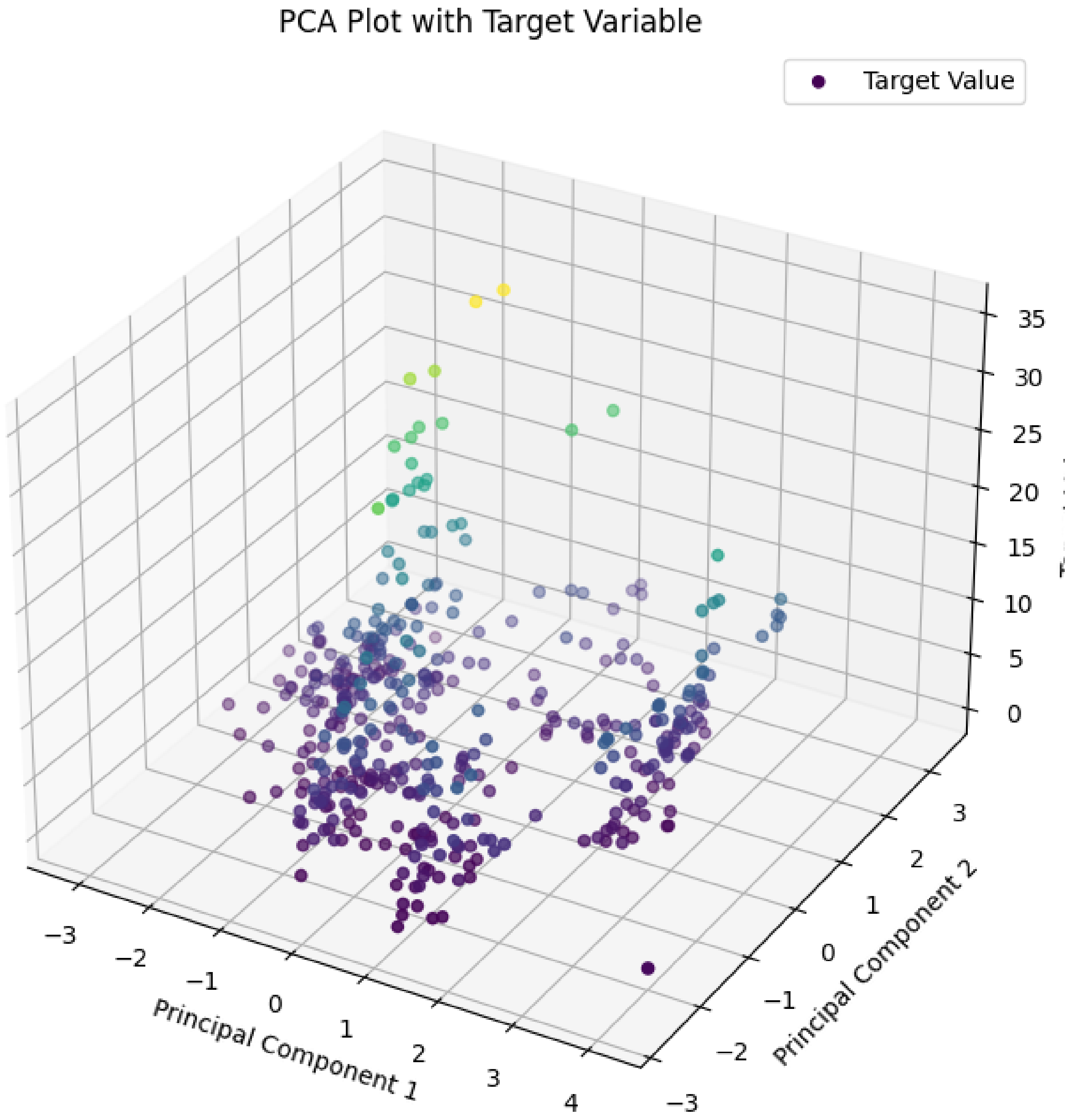
CORRELATION AMONG FEATURES



	Tg	Tx	Tl	TEN	VA	d	Sm
Tg	1.00	0.99	0.97	0.82	-0.44	-0.14	0.16
Tx	0.99	1.00	0.96	0.81	-0.41	-0.10	0.20
Tl	0.97	0.96	1.00	0.85	-0.43	-0.18	0.12
TEN	0.82	0.81	0.85	1.00	-0.55	-0.37	0.15
VA	-0.44	-0.41	-0.43	-0.55	1.00	0.78	0.18
d	-0.14	-0.10	-0.18	-0.37	0.78	1.00	0.40
Sm	0.16	0.20	0.12	0.15	0.18	0.40	1.00

- Plot demonstrates very high correlation (>0.9) between Tg, Tx and Tl.
- So Tx and Tl are dropped and only Tg is kept in the final dataset

Principle Component Analysis(PCA)



- Visualize the variation of Target variable with 2 Principle Component variables formed from 7 features.
- Useful for dealing with high dimensional data patterns

Exploratory Data Analysis : Finding Insights from Data

EXPLORATORY DATA ANALYSIS

PREPROCESSING

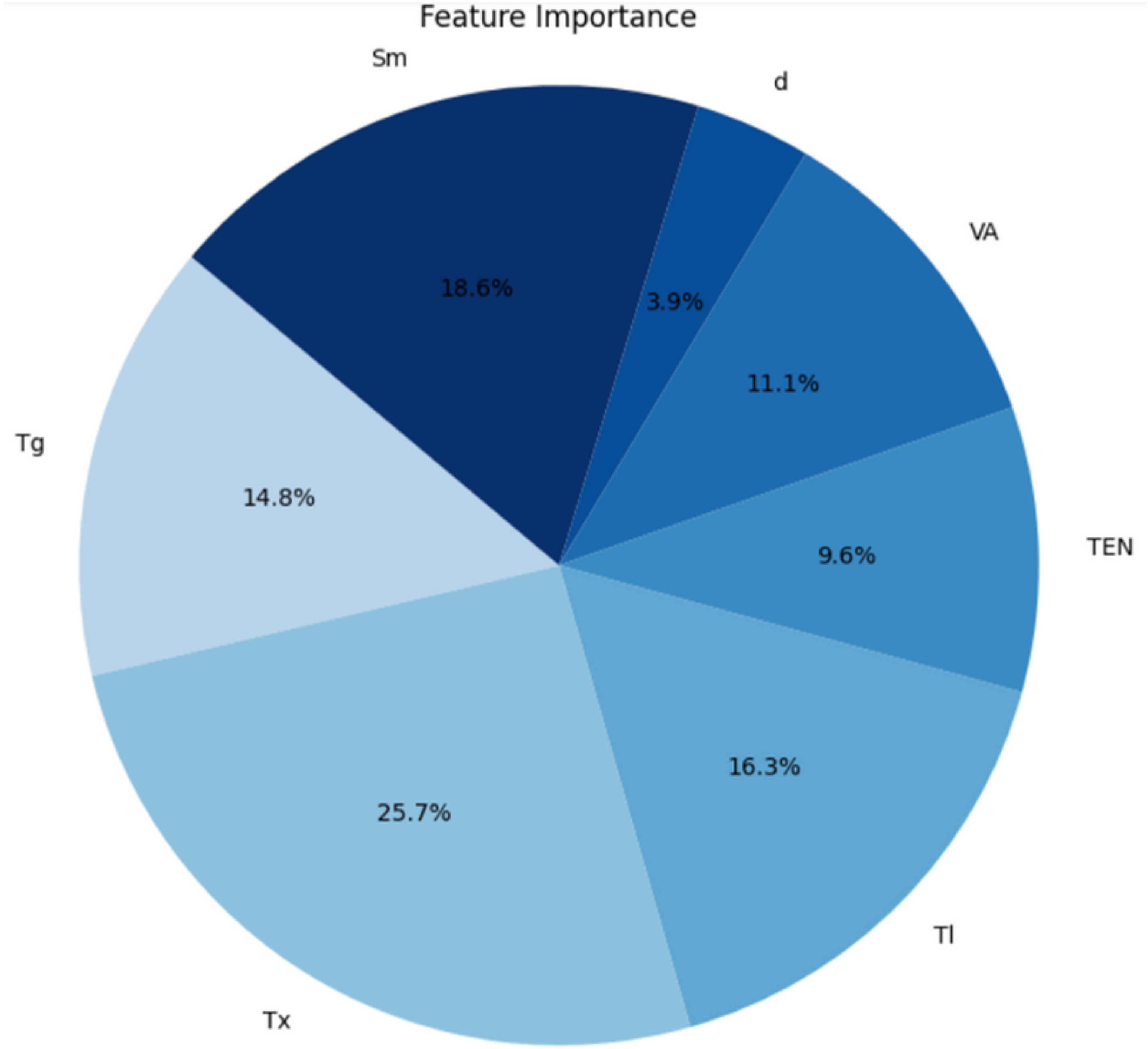
AUGMENTATION

REGRESSION

TESTING

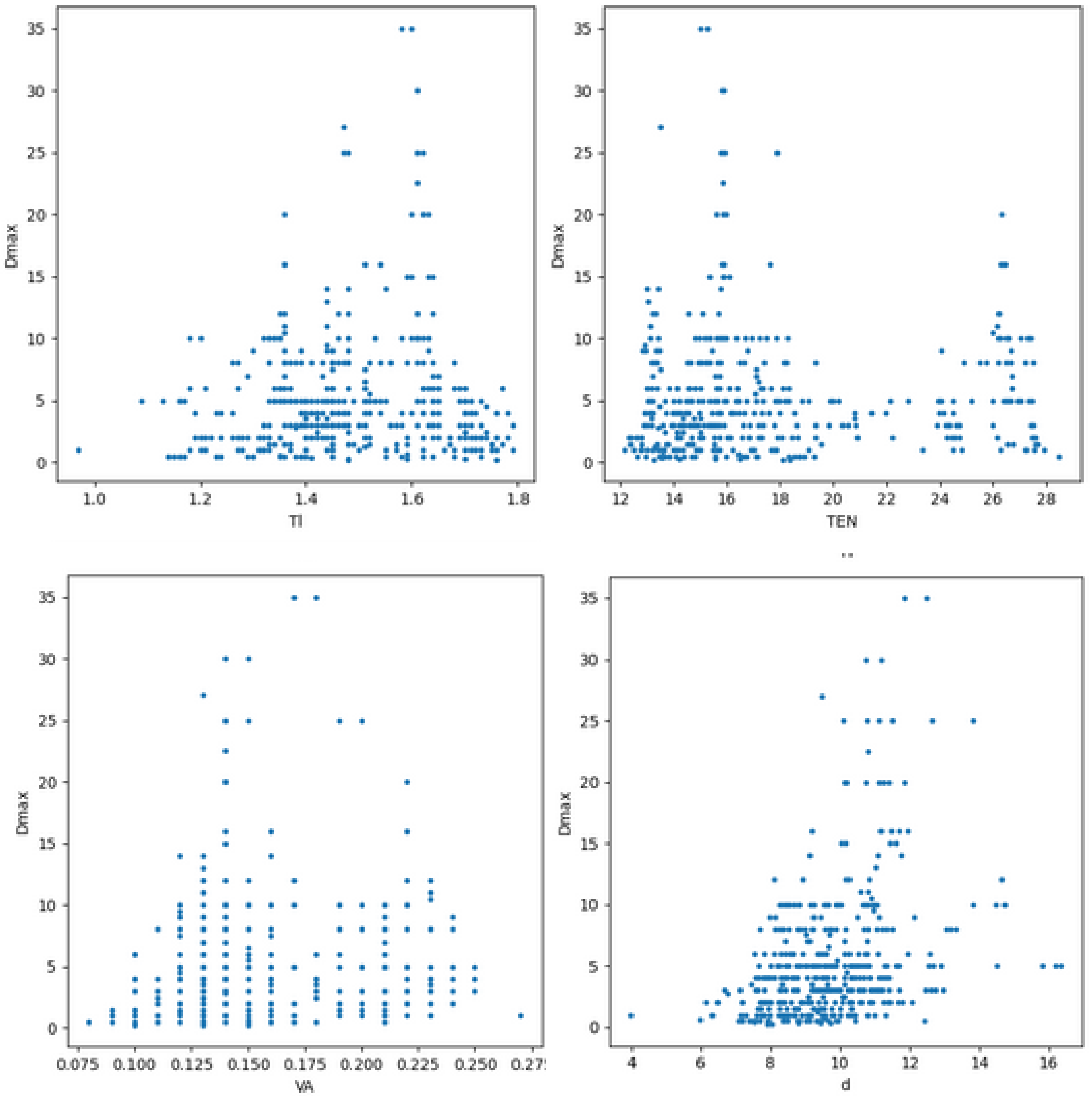
SUMMARY

IMPORTANCE OF FEATURES



- This Pie Chart shows the percentage importance of all the feature in predicting the output(Dmax).

VARIATION OF FEATURES WITH DMAX



Preprocessing : Enhance the Dataset

EXPLORATORY DATA ANALYSIS

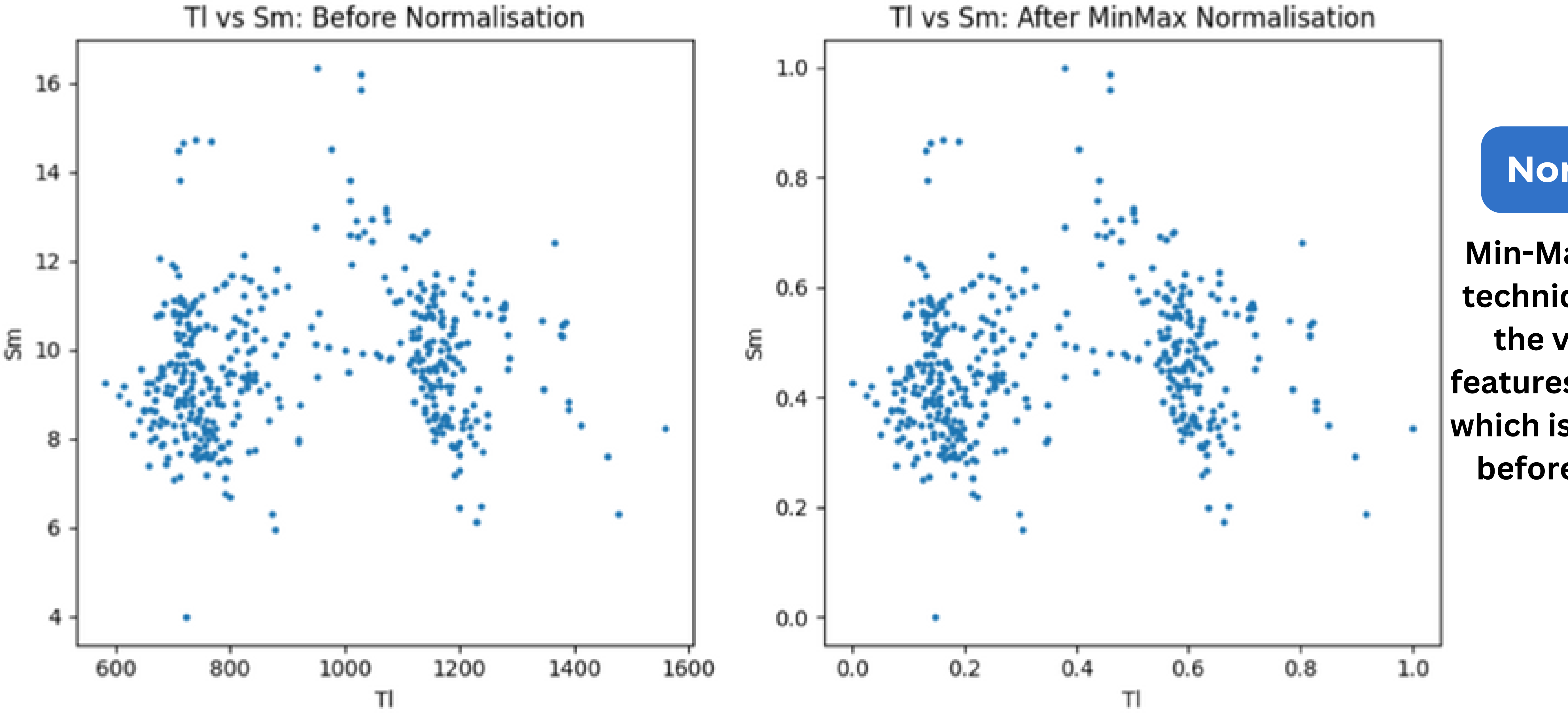
PREPROCESSING

AUGMENTATION

REGRESSION

TESTING

SUMMARY



Normalisation

Min-Max Normalisation technique used to scale the values of all the features between 0 and 1 which is evident from the before and after plots

Missing Values

No missing values in the datset

Outliers

Outliers present which should not be removed for this given dataset. Outliers handled using data augmentation

EXPLORATORY DATA ANALYSIS

PREPROCESSING

AUGMENTATION

REGRESSION

TESTING

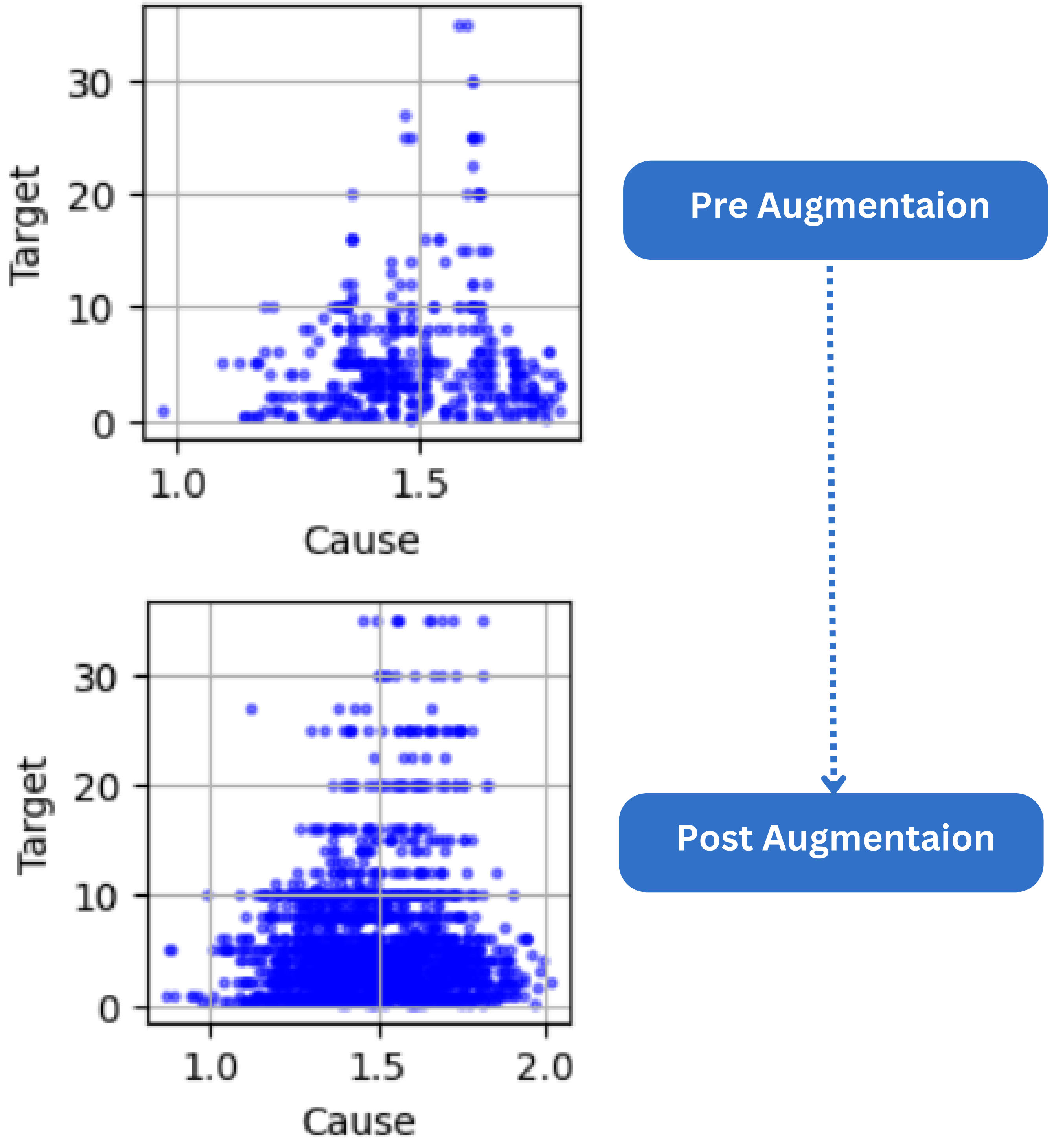
SUMMARY

Data Augmentation : Artificially expand dataset

Data augmentation is a technique used to artificially expand a dataset by creating new, modified versions of existing data samples. The goal is to improve the model's generalization, robustness, and performance on unseen data by introducing variations and reducing overfitting.

How is Data Augmented?

- **Adding Noise**: Introduce random noise to the input features to create variations in the data.
- **SMOTE(**Synthetic Minority Over-sampling Technique): Oversampling technique used to balance the class distribution of a dataset by creating synthetic minority class samples.



Pre Augmentaion

Post Augmentaion

The plots shows that the augmented dataset closely resembles the original dataset(Only 1 feature shown here)

EXPLORATORY DATA ANALYSIS

PREPROCESSING

AUGMENTATION

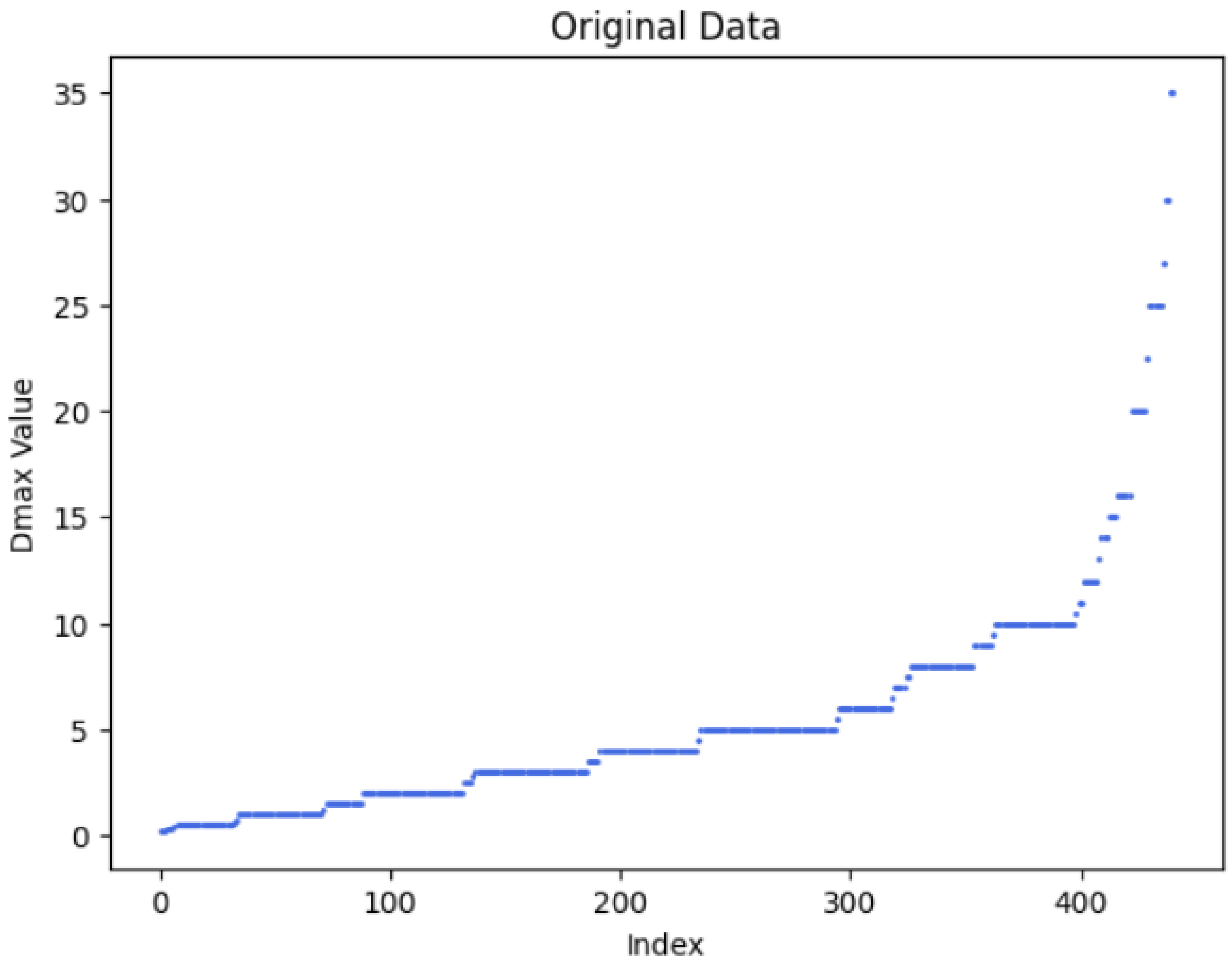
REGRESSION

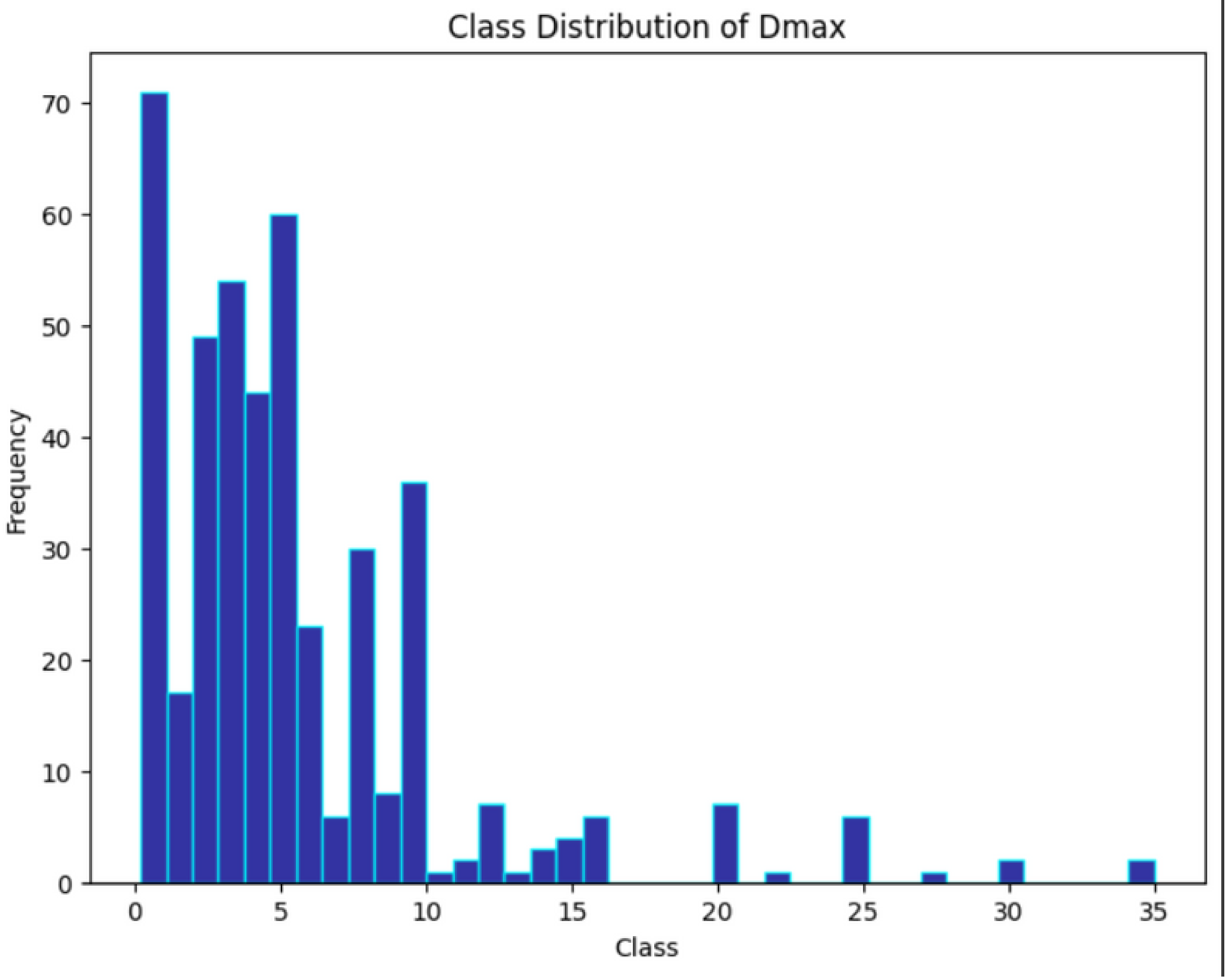
TESTING

SUMMARY

Data Augmentation : Artificially expand dataset

Why Data Augmentation is so necessary in this Dataset?

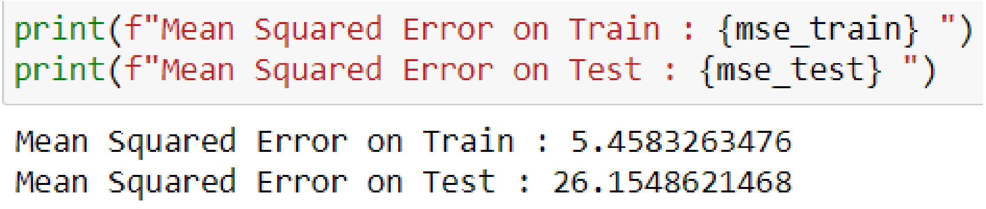




- Original Dataset is very small with discontinuous data points(which may seem like outliers), for higher values of Dmax which is evident fromthe plot

- Skewed Dataset with high number of readings for lower values of Dmax and very low number of readings for higher values of Dmax

- Problem of Overfitting due to small dataset



```
print(f"Mean Squared Error on Train : {mse_train} ")
print(f"Mean Squared Error on Test : {mse_test} ")
```

```
Mean Squared Error on Train : 5.4583263476
Mean Squared Error on Test : 26.1548621468
```

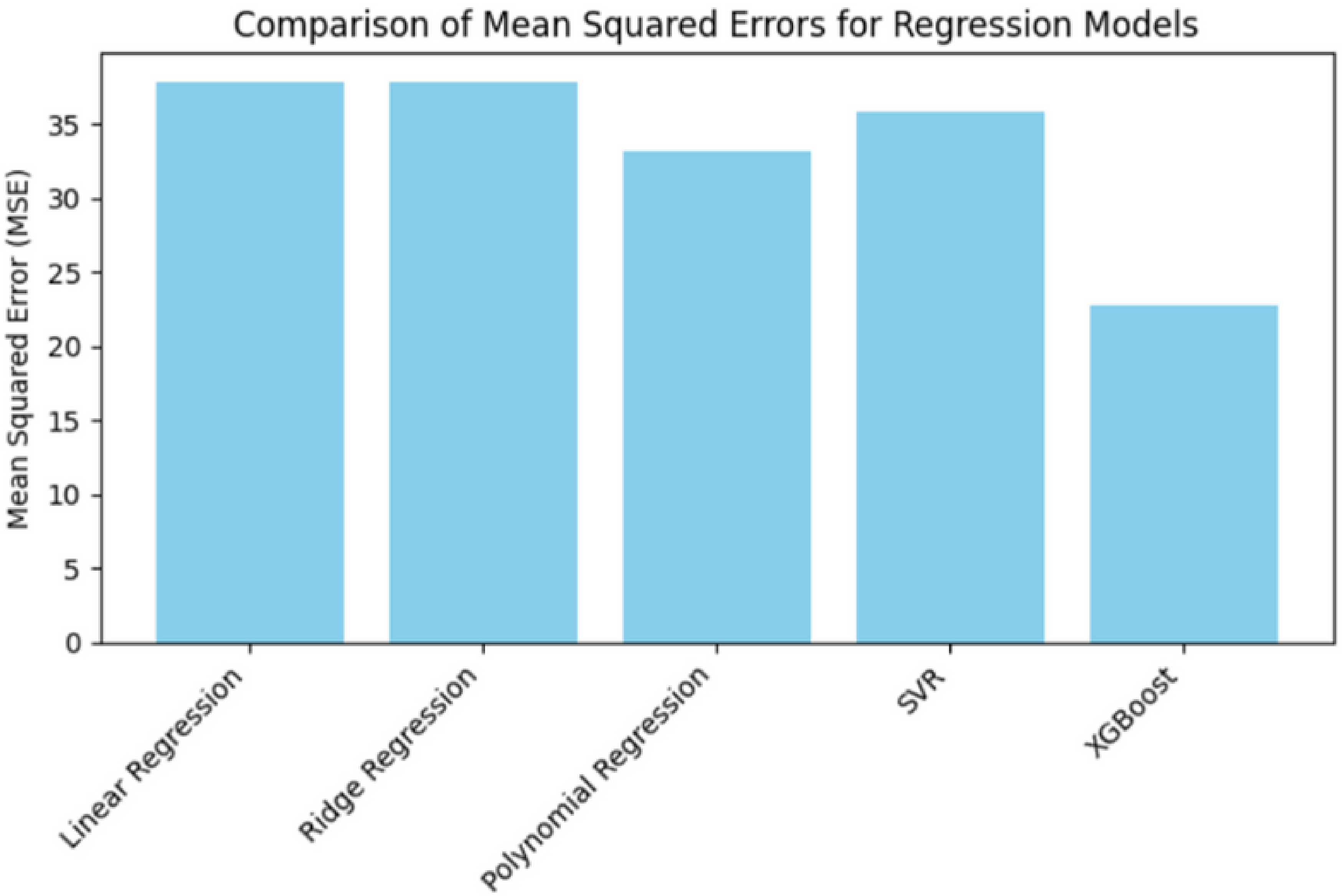
Regression: Machine Learning Algorithm

Models Used

- Linear Regression
- Polynomial Regression
- Ridge Regression
- SVR
- XGBoost

Evaluation Metrices

- MSE: Mean Squared Error
- RMSE: Root Mean Squared Error
- MAPE: Mean Absolute Percentage Error
- MAE: Mean Absolute Error
- R Squared Score(R2)



We see from the plot that XGBoost model gives the lowest MSE on the original data. We finally apply XGBoost on the Augmented + Resampled Dataset.

Regression: Machine Learning Algorithm

EXPLORATORY DATA ANALYSIS

PREPROCESSING

AUGMENTATION

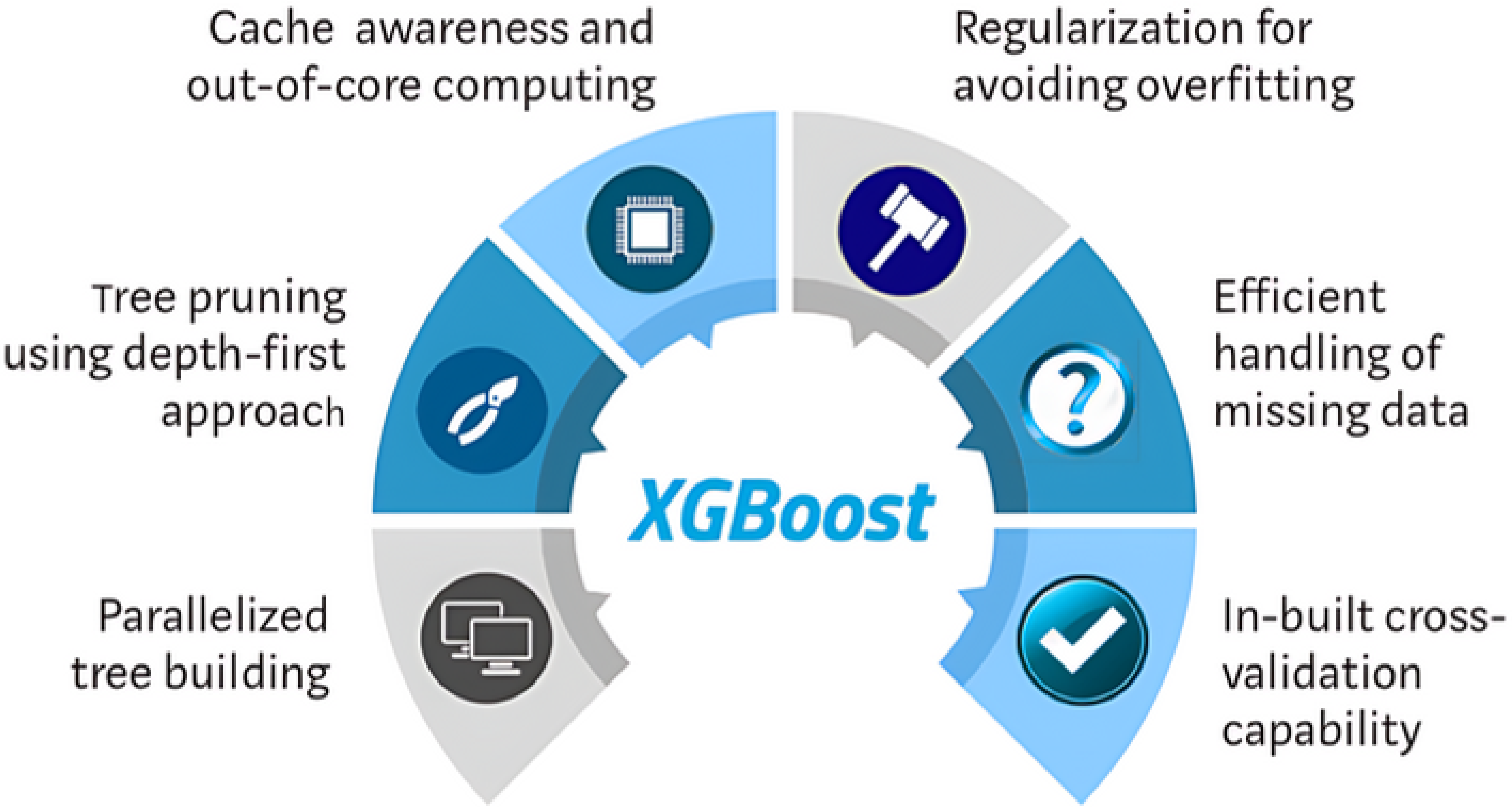
REGRESSION

TESTING

SUMMARY

XGBoost(Extreme Gradient Boosting)

XGBoost is an ensemble learning algorithm that uses decision trees in a boosting framework whose optimized architecture speeds up training and improves accuracy. Using gradient boosting, it adds trees sequentially to correct errors. XGBoost includes regularization like L1/L2 to prevent overfitting and parallelized processing for efficiency, making it versatile and powerful for diverse machine learning tasks.



Cache awareness and out-of-core computing

Regularization for avoiding overfitting

Tree pruning using depth-first approach

Efficient handling of missing data

Parallelized tree building

In-built cross-validation capability

EXPLORATORY
DATA ANALYSIS

PREPROCESSING

AUGMENTATION

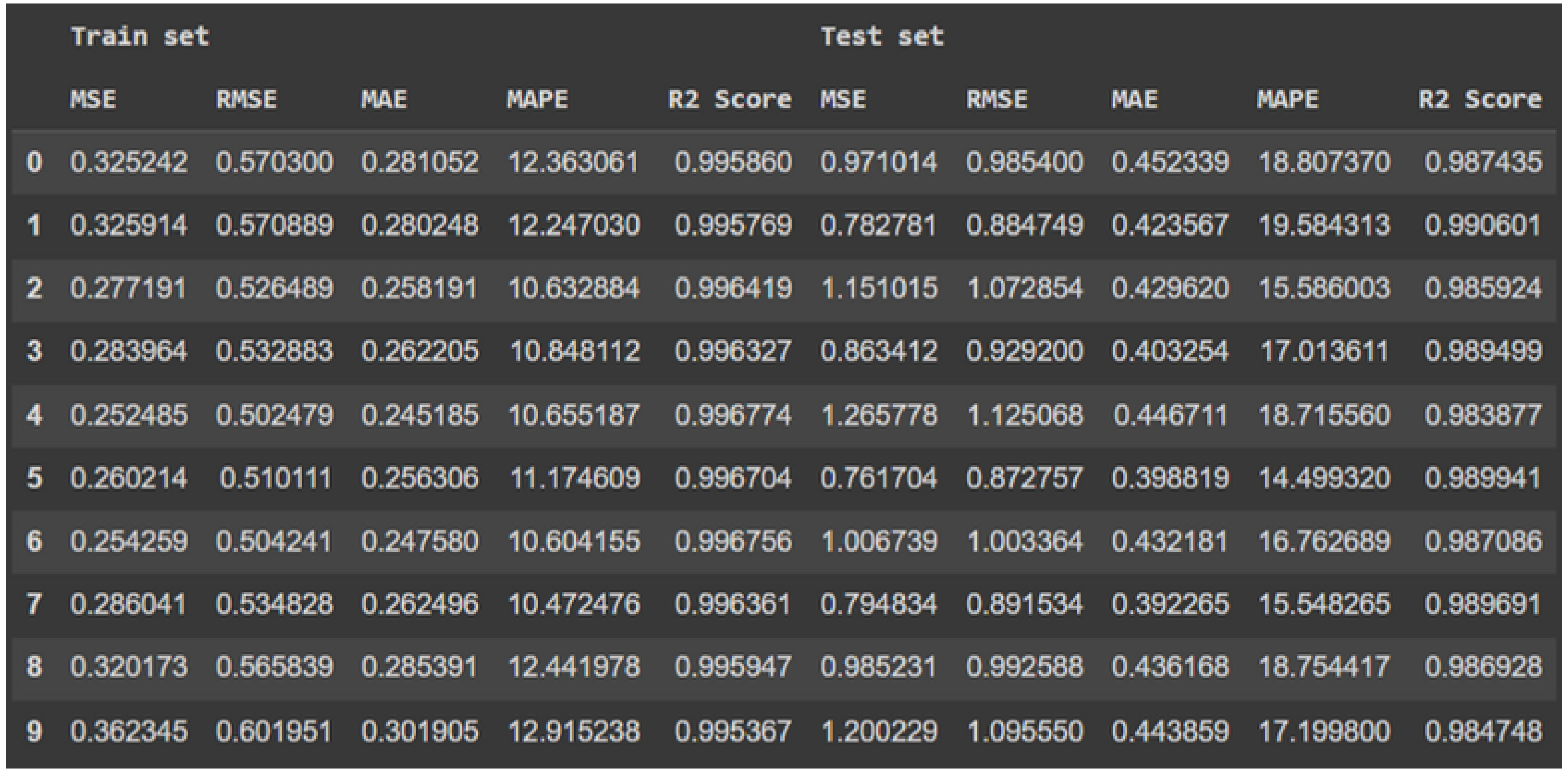
REGRESSION

TESTING

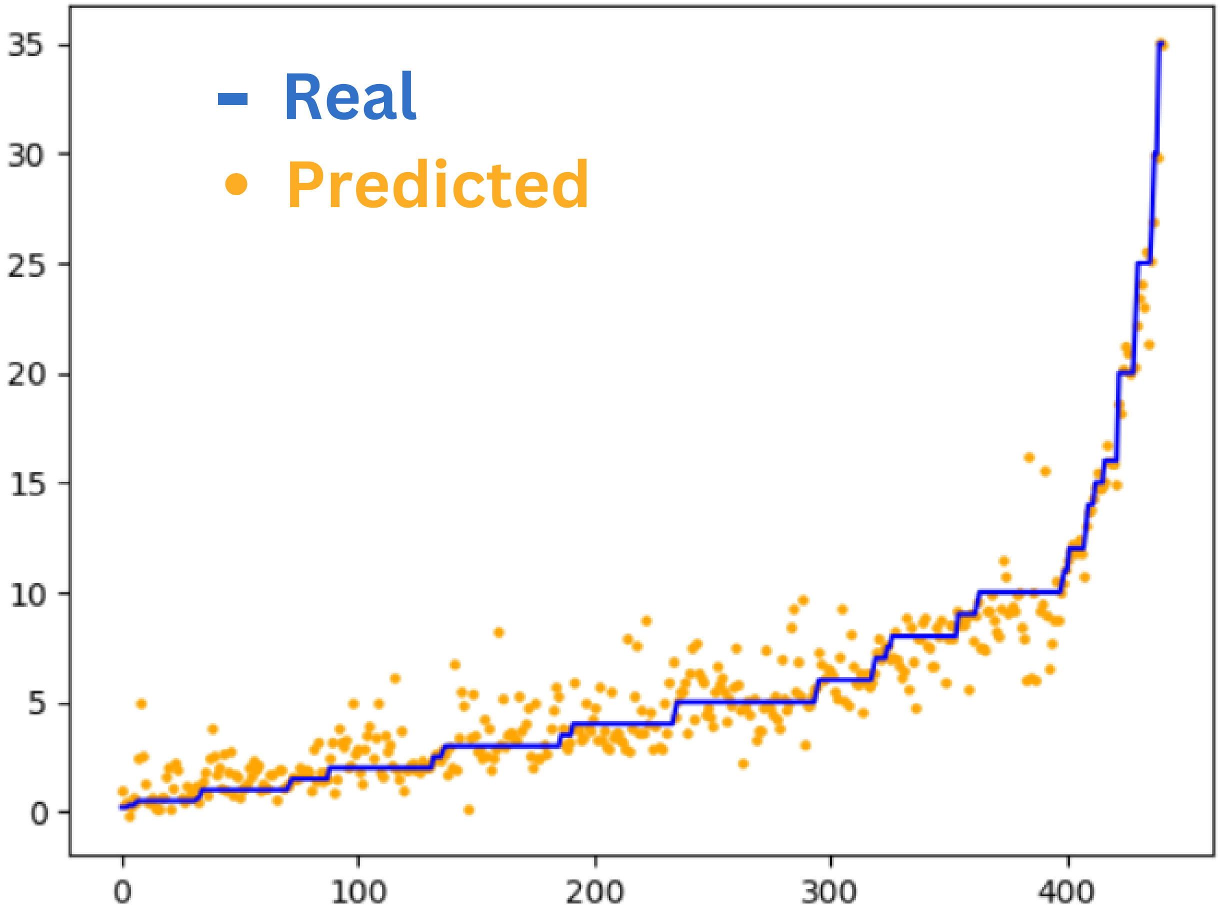
SUMMARY

Testing

The table shows the relevant metrices after training 10 times with a random Training-Testing Partition



	Train set					Test set				
	MSE	RMSE	MAE	MAPE	R2 Score	MSE	RMSE	MAE	MAPE	R2 Score
0	0.325242	0.570300	0.281052	12.363061	0.995860	0.971014	0.985400	0.452339	18.807370	0.987435
1	0.325914	0.570889	0.280248	12.247030	0.995769	0.782781	0.884749	0.423567	19.584313	0.990601
2	0.277191	0.526489	0.258191	10.632884	0.996419	1.151015	1.072854	0.429620	15.586003	0.985924
3	0.283964	0.532883	0.262205	10.848112	0.996327	0.863412	0.929200	0.403254	17.013611	0.989499
4	0.252485	0.502479	0.245185	10.655187	0.996774	1.265778	1.125068	0.446711	18.715560	0.983877
5	0.260214	0.510111	0.256306	11.174609	0.996704	0.761704	0.872757	0.398819	14.499320	0.989941
6	0.254259	0.504241	0.247580	10.604155	0.996756	1.006739	1.003364	0.432181	16.762689	0.987086
7	0.286041	0.534828	0.262496	10.472476	0.996361	0.794834	0.891534	0.392265	15.548265	0.989691
8	0.320173	0.565839	0.285391	12.441978	0.995947	0.985231	0.992588	0.436168	18.754417	0.986928
9	0.362345	0.601951	0.301905	12.915238	0.995367	1.200229	1.095550	0.443859	17.199800	0.984748



This plot shows the correspondence of the real and predicted values of the target variable Dmax

Avg MSE of 10 performances **= 0.978274**
Std deviation of 10 performances = **0.180935**

EXPLORATORY DATA ANALYSIS

PREPROCESSING

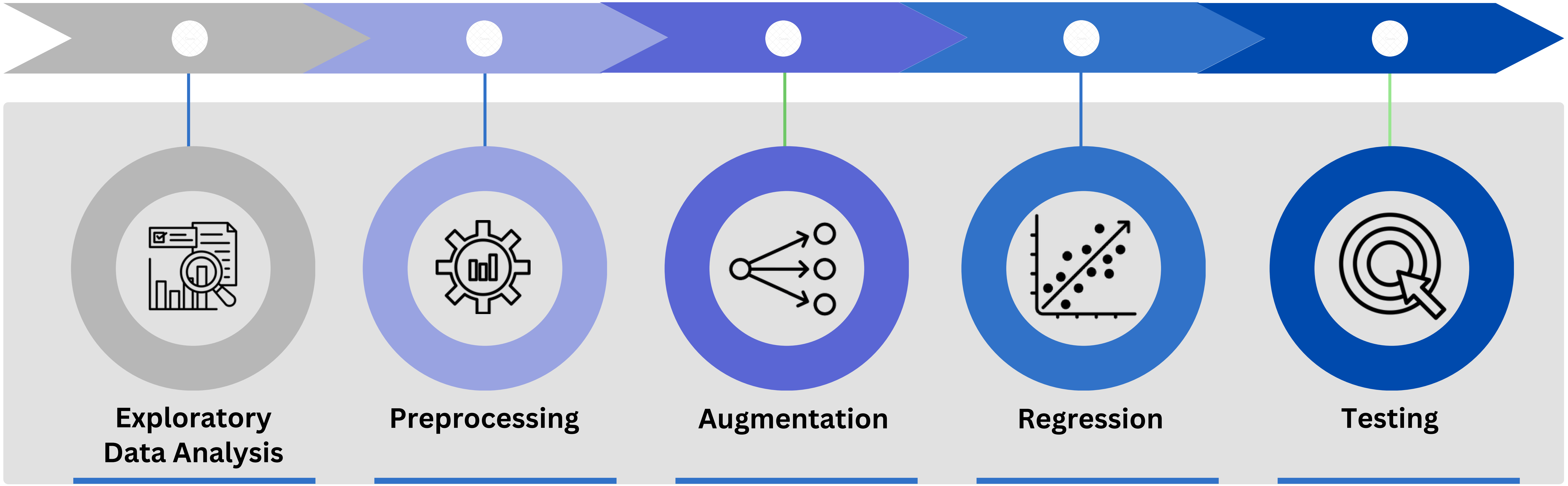
AUGMENTATION

REGRESSION

TESTING

SUMMARY

Summary



Exploratory Data Analysis

Grasp fundamental characteristics of data.

Preprocessing

Normalisation, Checking for Outliers and missing values

Augmentation

Add noise in features and use SMOTE to increase dataset

Regression

Use various models to choose best model

Testing

Train Test Split 10 times and apply XGBoost

THANK YOU