Project Presentation

Group 16: Aarthi Vasudevan, Akruthi Srikanth, Sravya Gopireddy

Supervised Learning

Summary of Models

Six Classifiers:

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Naive Bayes
- Random Forest
- Support Vector Machines (SVM)
- XGBoost

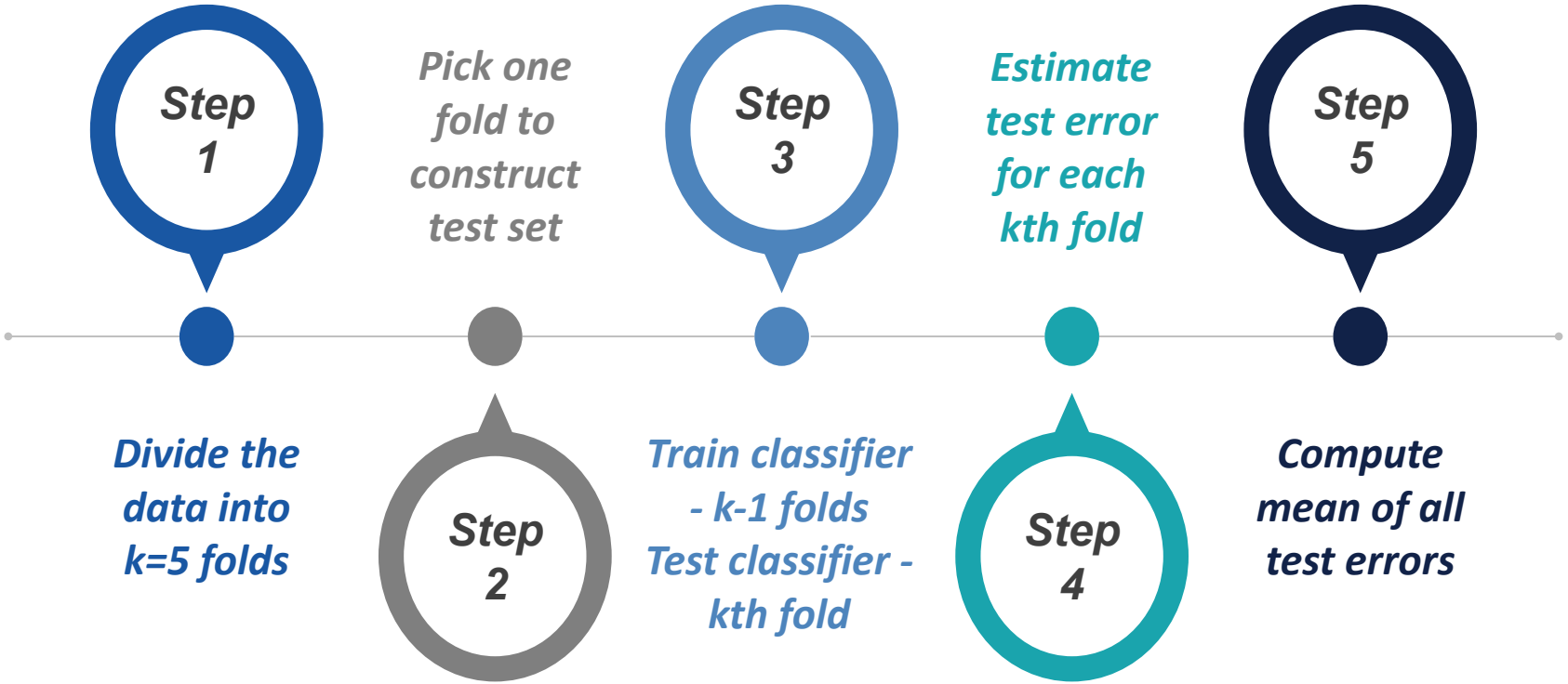
XGBoost

- XGBoost (Extreme Gradient Boosting) - machine learning algorithm that uses an ensemble of decision trees to model and predict data.

- It iteratively trains weak decision trees on the residuals of the previous tree and gradually improving the accuracy of the model.

- It includes advanced features like regularization, cross-validation, and handling missing data.

Implementation of K-fold Cross Validation (CV)



Data pre-processing

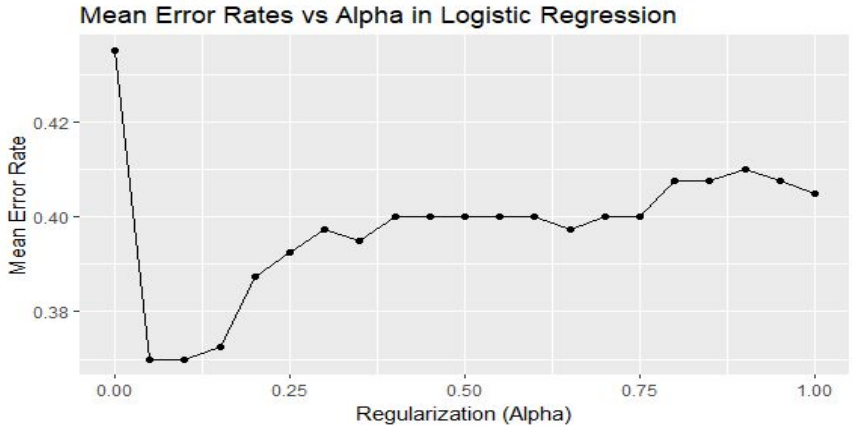
- Response variable ***y*** is factorized
- Threshold of 0.5 is used

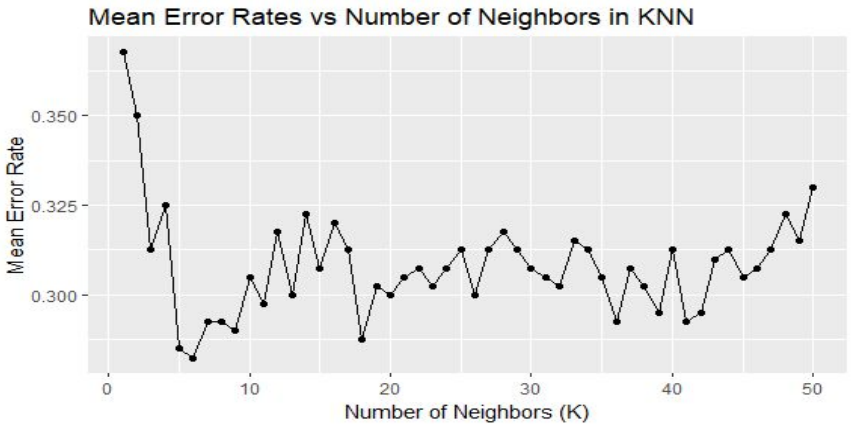
$$y = \begin{cases} \text{Yes} & \text{When threshold} > 0.5 \\ \text{No} & \text{Otherwise} \end{cases}$$

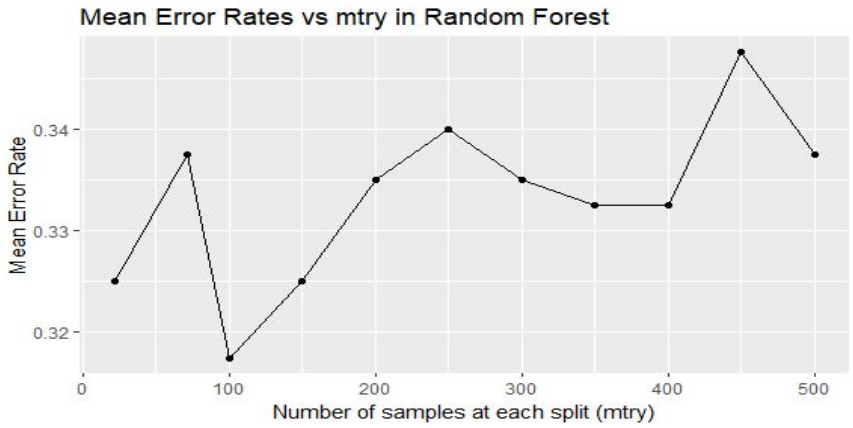
Tuning parameters

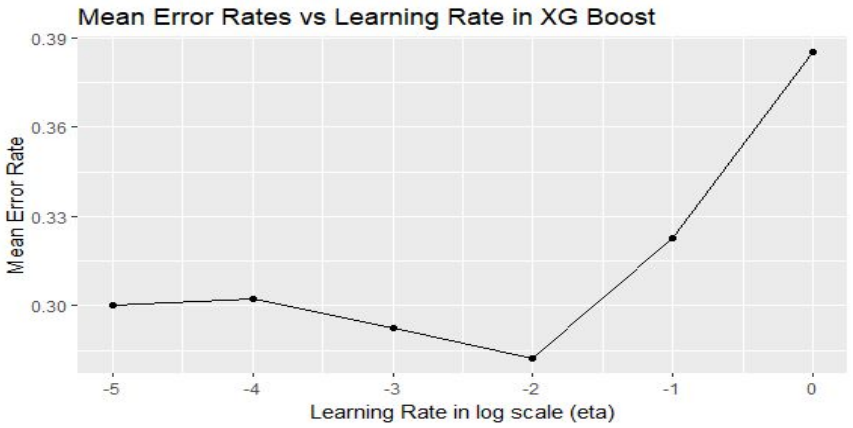
MODEL	*TUNING PARAMETER*
Logistic Regression	alpha
KNN	k
Naive Bayes	-
Random Forest	mtry
SVM	cost, gamma, kernel, degree
XGBoost	colsamp, gamma, nrounds, max_depth, eta

Results: Logistic Regression, KNN, Random Forest, XGBoost









Estimated Test Error & Optimal Tuning Parameters

MODEL	*ESTIMATED TEST ERROR*	*TUNING PARAMETER*
Logistic Regression	37%	alpha=0.05
KNN	28.25%	k=6
Naive Bayes	37.25%	-
Random Forest	31.75%	mtry=100
SVM	29.75%	cost=0.1, gamma=0.1, kernel=poly, degree=4
XGBoost	28.25%	colsamp=0.25, gamma=0, nrounds=500, max_depth=9, eta=0.01

Unsupervised Learning

Summary of Models

Four Clustering Techniques:

- Principal Component Analysis (PCA)
- K-Means Clustering
- t-Distributed Stochastic Neighbor Embedding (t-SNE)
- Agglomerative Clustering

t-SNE

- t-SNE (t-Distributed Stochastic Neighbor Embedding) - machine learning algorithm used for data visualization and dimensionality reduction.

- It takes high-dimensional data and reduces it to a low-dimensional representation that can be easily visualized.

- It models the high-dimensional data as a set of pairwise similarities, and finds a lower-dimensional representation that preserves these pairwise similarities as closely as possible.

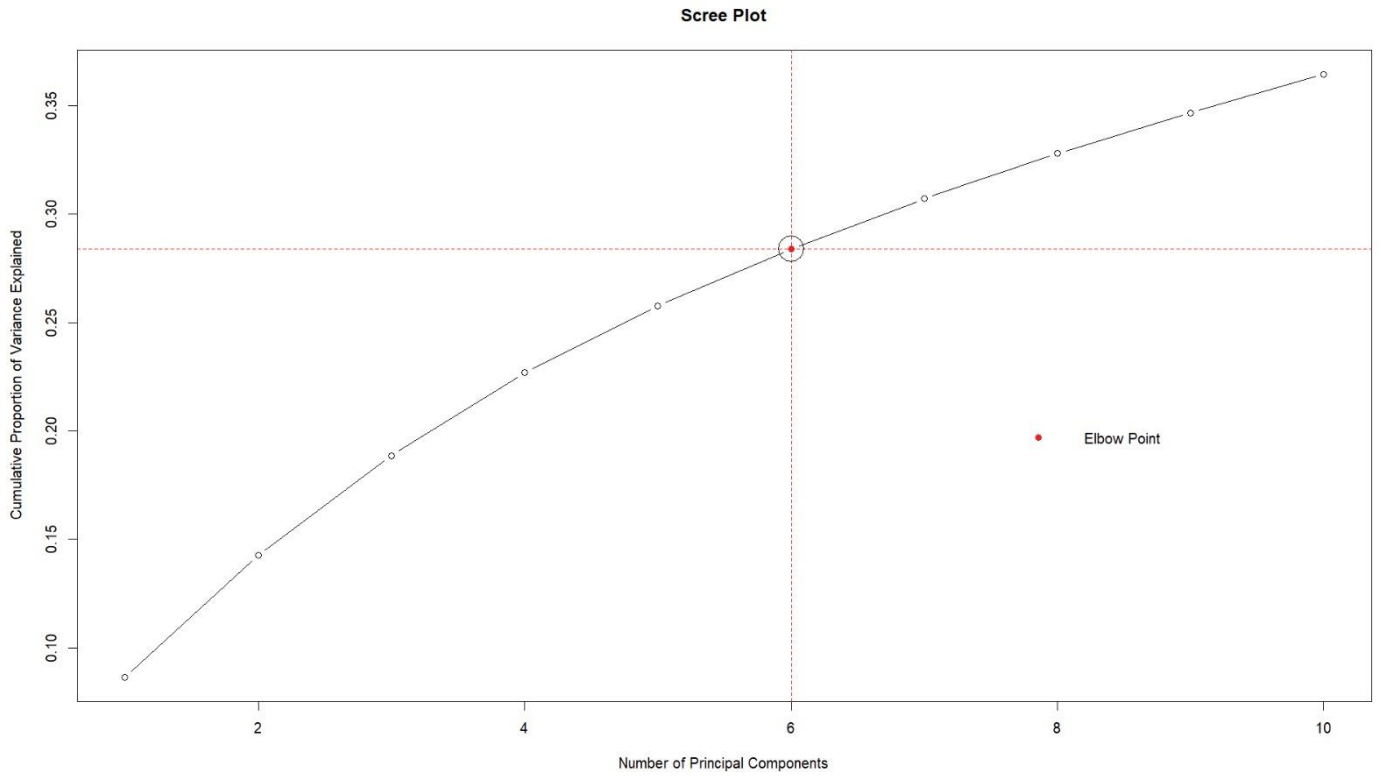
Measure of Evaluation

- Scree Plot (for PCA)
- Silhouette Score
- Within Cluster Sum of Squares (WCSS)

Optimal number of clusters is determined with maximum Silhouette Score and minimum WCSS

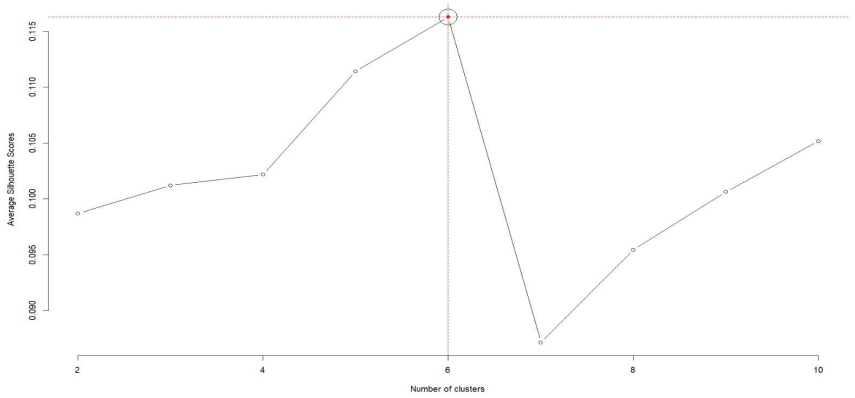
Results: PCA

Scree Plot

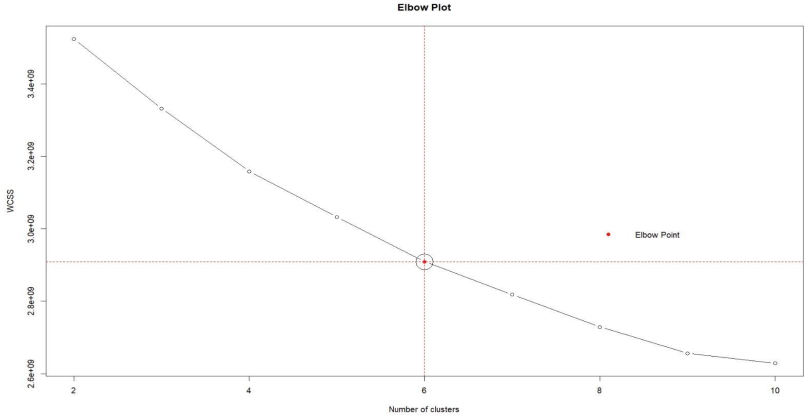


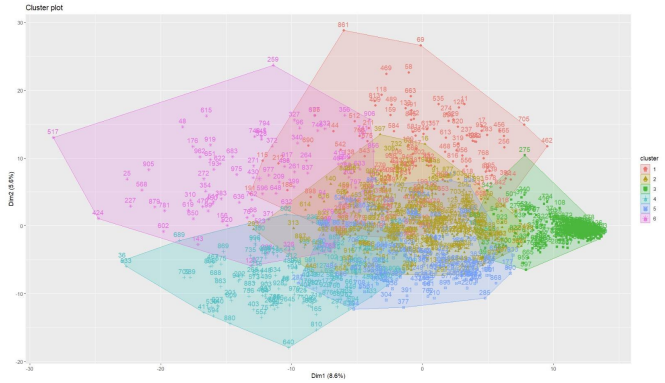
Results: K-Means Clustering

Number of clusters Vs Average Silhouette Scores



Number of clusters Vs WCSS

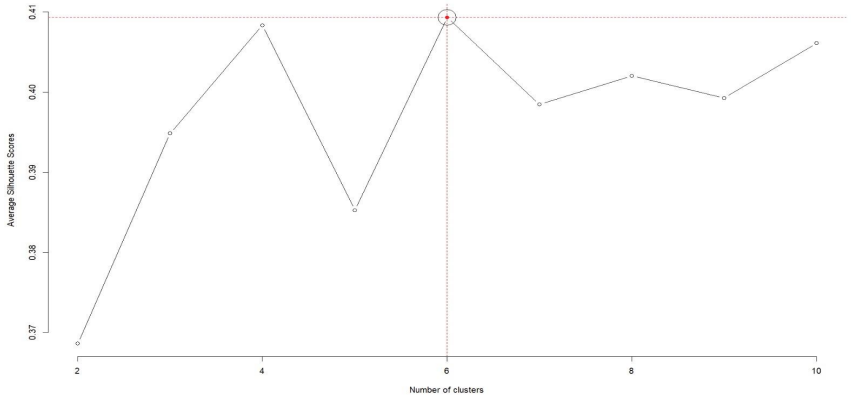




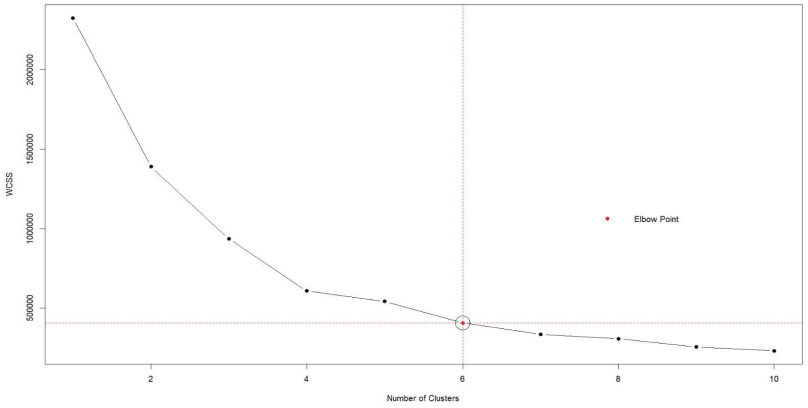
Cluster plot

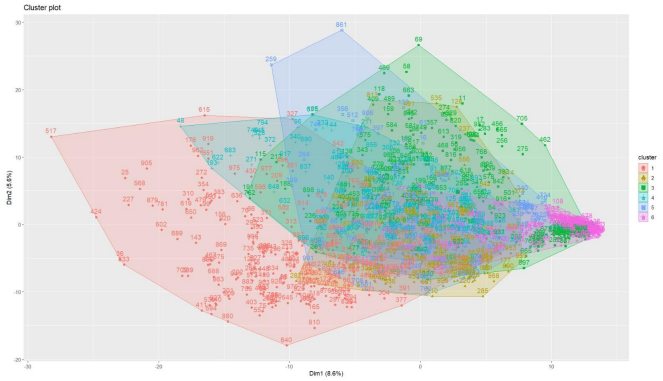
Results: t-SNE

Number of clusters Vs Average Silhouette Scores



Number of clusters Vs WCSS

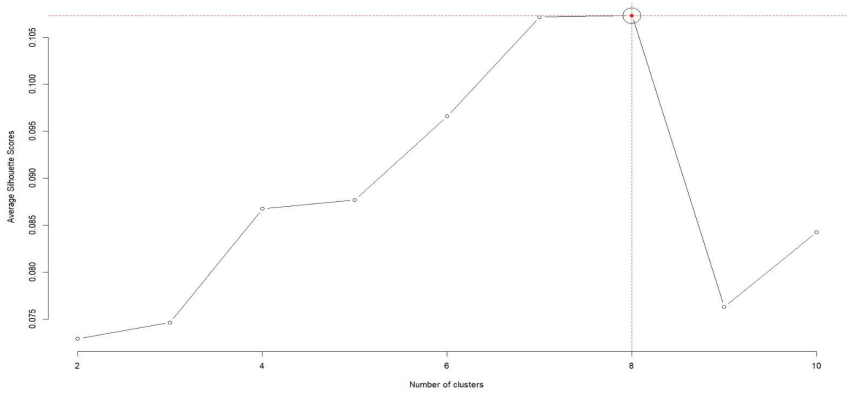




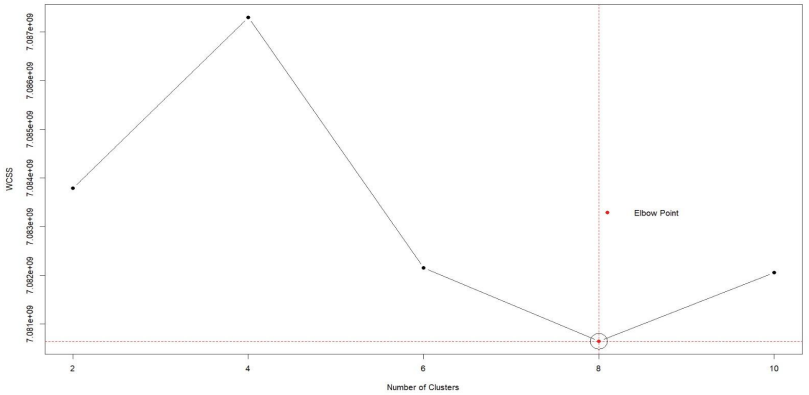
Cluster plot

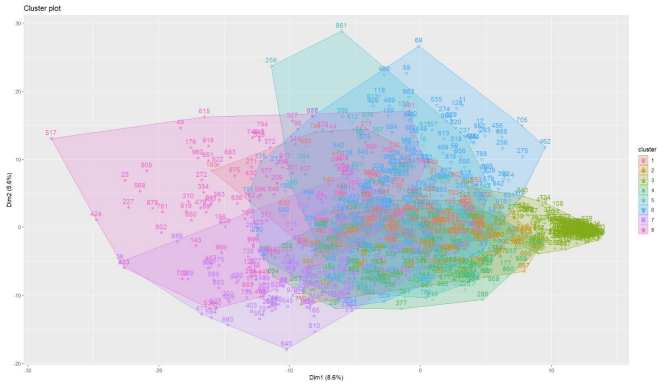
Results: Agglomerative Clustering

Number of clusters Vs Average Silhouette Scores



Number of clusters Vs WCSS





Cluster plot

Results

MODEL	*NUMBER OF CLUSTERS*
PCA	6
K-Means Clustering	6
t-SNE	6
Agglomerative Clustering	8

Thank You !!