

Name – Jayashri Trembak Musale

Email- [jayashri.musale@mitaoe.ac.in](mailto:jayashri.musale@mitaoe.ac.in)

## Problem statement - Predicting Employee Attrition

- **Performance Aanalysis**

### A.Comparing Accuracy of Base Models

After training the model with 60% of the dataset on base learners, the highest accuracy is achieved by the **Linear regression, reaching 86.73%** on the testing data. Following exactly, SVM demonstrates the best accuracy on the testing dataset. Conversely, DT yields the lowest accuracy on the testing dataset, with a value of 81.46 %.

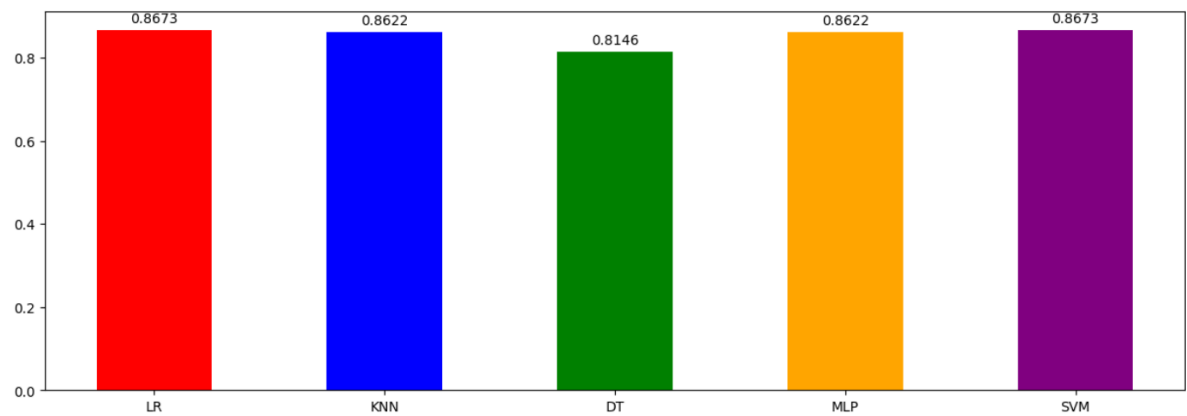


Fig.1. Comparing accuracy of base models

<b>Classifier</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>F1Score</b>
<b>LR</b>	<b>0.867</b>	<b>0.836</b>	<b>0.867</b>	<b>0.830</b>
<b>KNN</b>	<b>0.862</b>	<b>0.833</b>	<b>0.862</b>	<b>0.839</b>
<b>DT</b>	<b>0.810</b>	<b>0.808</b>	<b>0.810</b>	<b>0.809</b>
<b>MLP</b>	<b>0.857</b>	<b>0.822</b>	<b>0.857</b>	<b>0.830</b>
<b>SVM</b>	<b>0.867</b>	<b>0.842</b>	<b>0.867</b>	<b>0.819</b>

## B.Comparing Accuracy of Ensemble Models

After testing the data with the ensemble model, it was observed that **Bagging using a decision tree as the base learner achieved the highest accuracy of 87.41%**, followed by CatBoost with an accuracy of 87.07%. Gradient Boosting had the lowest accuracy among the ensemble models, with an accuracy of 82.65%.

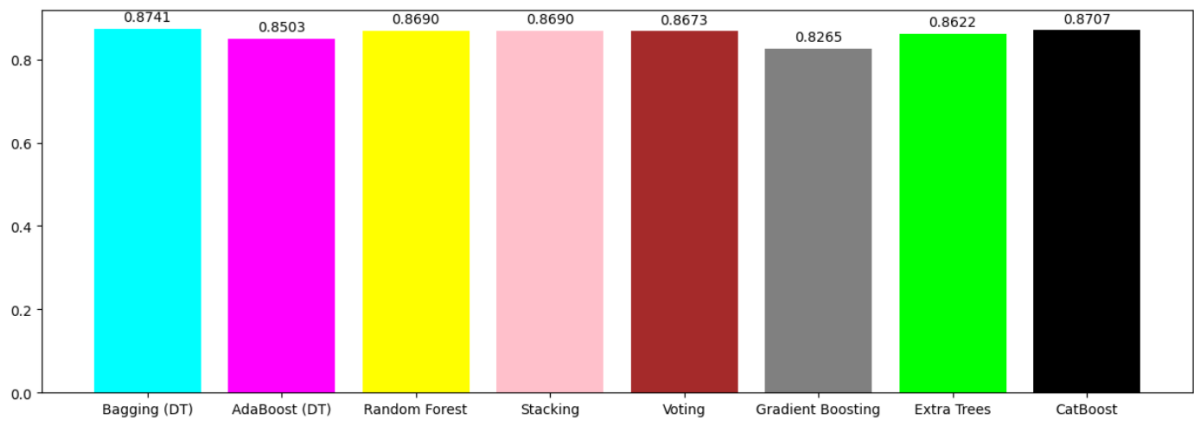


Fig.2. Comparing accuracy of ensemble models

Model	Accuracy	Precision	Recall	F1 Score
Bagging(DT)	0.874	0.851	0.874	0.850
AdaBoost(DT)	0.850	0.813	0.850	0.824
Random Forest	0.869	0.840	0.869	0.835
Stacking	0.866	0.833	0.866	0.833
Voting	0.866	0.832	0.866	0.829
Gradient Boosting	0.827	0.803	0.827	0.813
Extra Trees	0.862	0.823	0.862	0.825
CatBoost	0.871	0.844	0.871	0.838

### C. Comparison Of All Model Accuracy

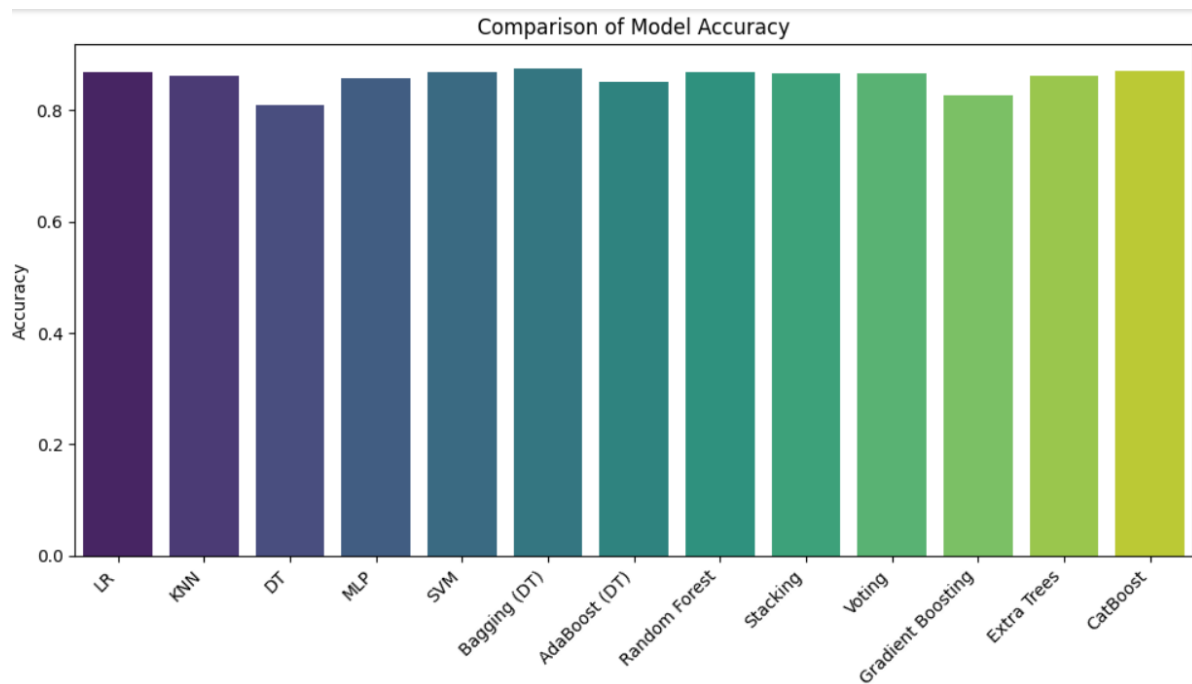


Fig.3. Comparing accuracy of all models

The best model based on accuracy is: **Bagging (DT)**

### D. ROC Curve

In ROC curve analysis, the ideal scenario is when the curve is far from the diagonal line towards the upper-left corner, indicating better model performance with higher true positive rates and lower false positive rates. A curve closer to the diagonal line signifies poor performance, resembling random guessing. Comparing the ROC curves of base and ensemble learners, the ensemble curve is notably away from the diagonal line, indicating superior accuracy compared to the base learner.

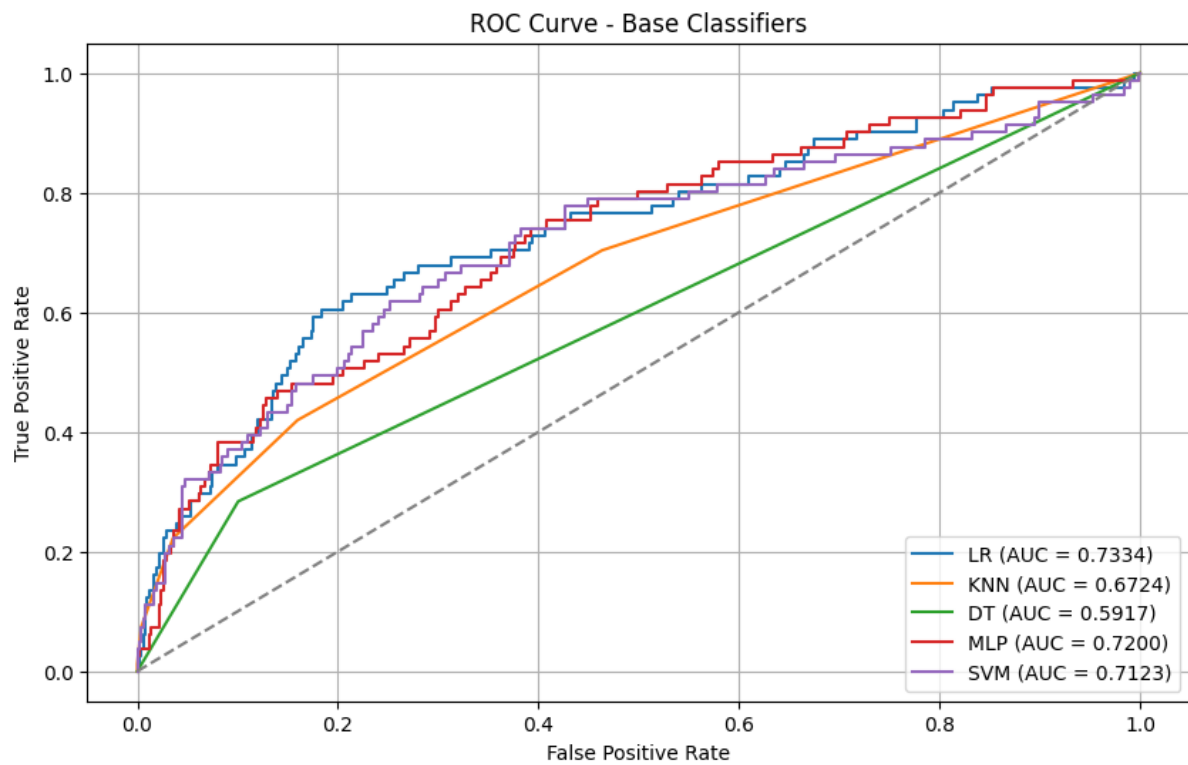


Fig.4.ROC Curve - Base classifiers

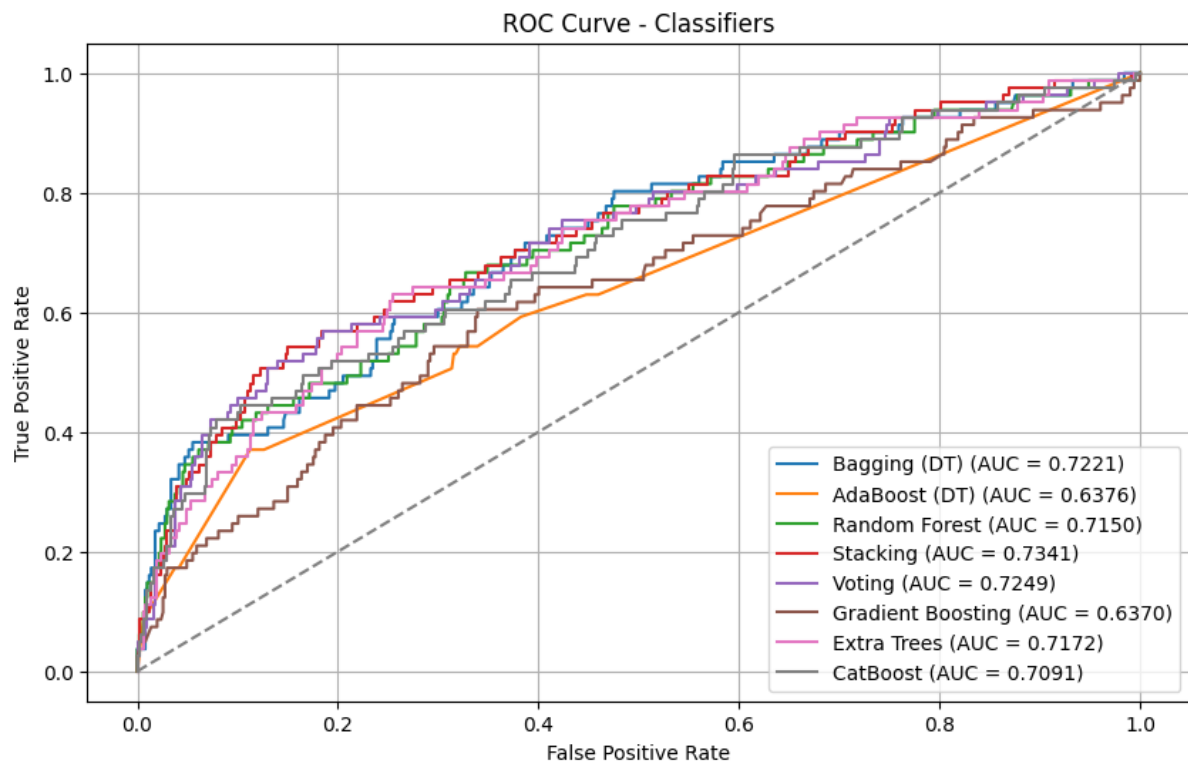


Fig.5. ROC Curve - classifiers