Final Project Presentation

CS 7641 Machine Learning

Predicting Probabilities of getting into the playoffs in NBA

TEAM - 42

Akshay Jadiya

Carlos Aguilar

Saurabh Aggarwal

Yashwant Singh

Agenda

- Motivation
- Literature Review
- Data Collection and EDA
- Process Flow
- Methods Used
- Results and Discussion
- Conclusions

Motivation and Problem Statement

- Being a sports analyst requires one to efficiently predict outcomes.
- Fantasy Gaming enthusiasts may bet on teams more effectively.

We employed various ML approaches to figure out which teams would make it to the playoffs, given regular season statistics for NBA

Agenda

- Motivation
- Literature Review
- Data Collection and EDA
- Process Flow
- Methods Used
- Results and Discussion
- Conclusions

What has already been done?

1

Hybrid Fuzzy-SVM (HFSVM) to predict outcomes of basketball games and assess variable importance

2

Using Naïve Bayes to predict wins and regression to predict spread in 2009-2010 NBA

3

Applying Bayesian Belief Networks and Neural Networks to predict game outcomes from dataset of 650 NBA matches

Agenda

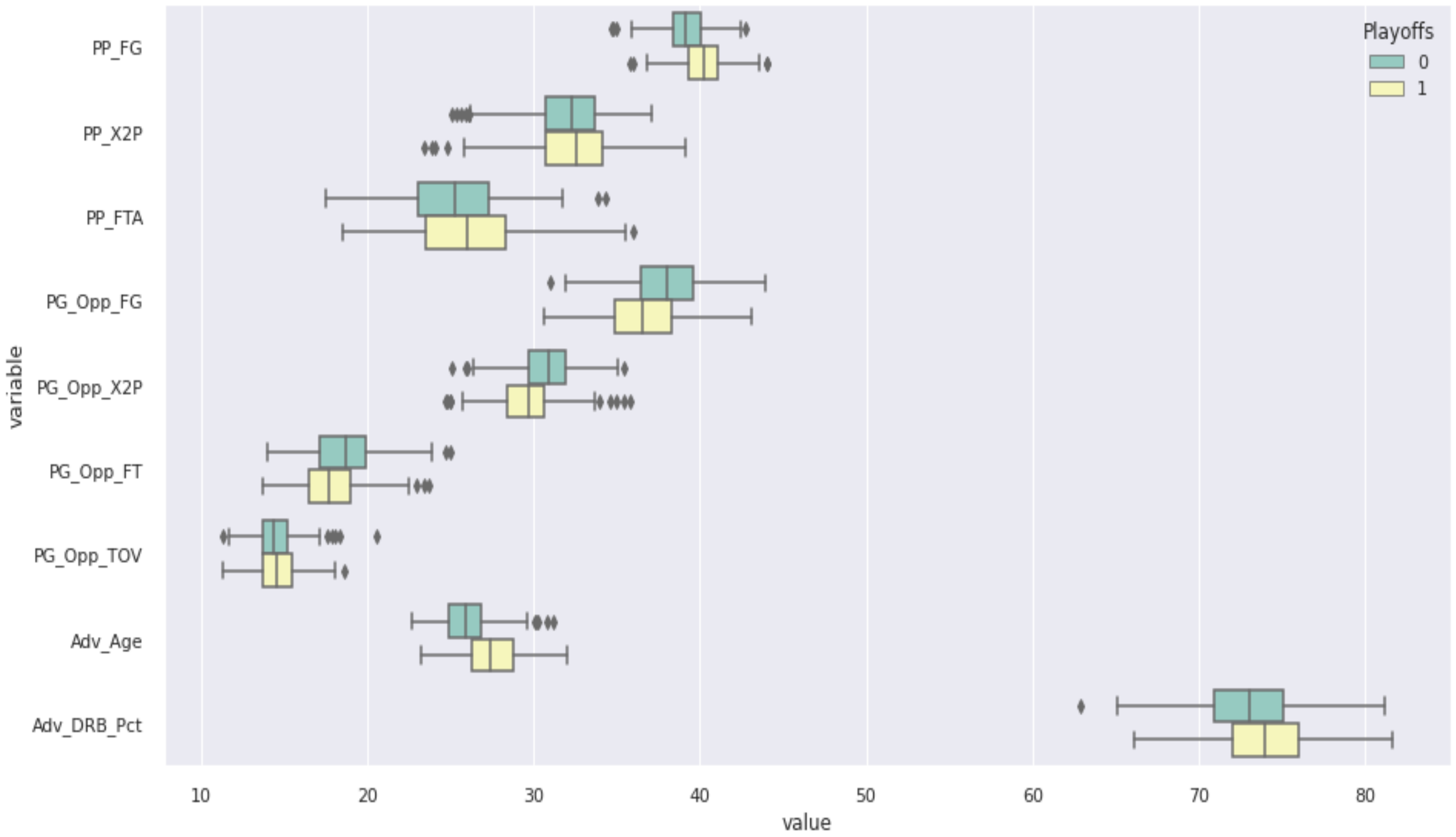
- Motivation
- Literature Review
- **Data Collection and EDA**
- Process Flow
- Methods Used
- Results and Discussion
- Conclusions

Data Collection and EDA

- Collected from https://www.basketball-reference.com/

- Focus on data after the 97-98 season.

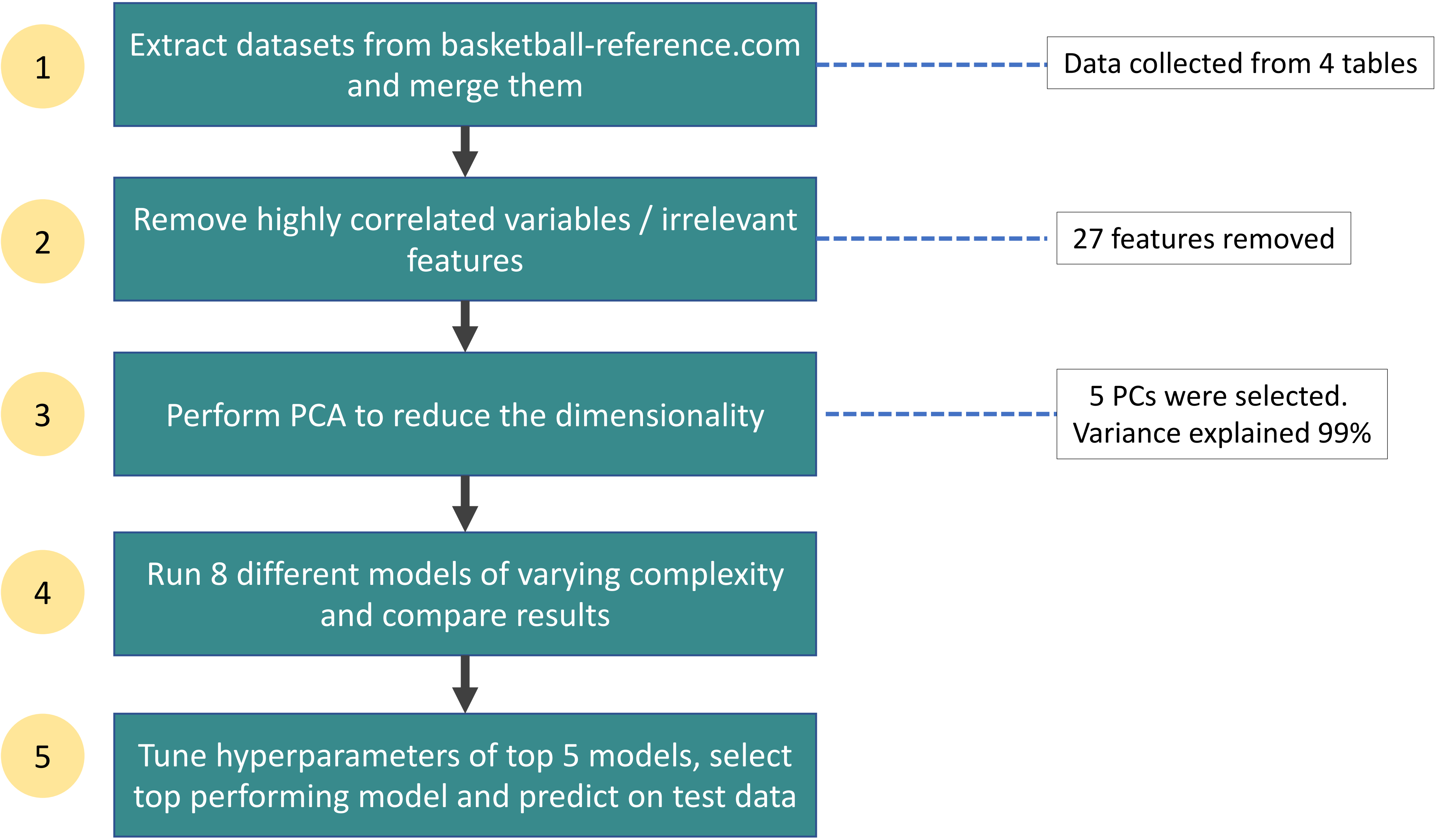
- Regular and Advanced Statistics.



Agenda

- Motivation
- Literature Review
- Data Collection and EDA
- Process Flow
- Methods Used
- Results and Discussion
- Conclusions

Process Flow



1 Extract datasets from basketball-reference.com and merge them

Data collected from 4 tables

2 Remove highly correlated variables / irrelevant features

27 features removed

3 Perform PCA to reduce the dimensionality

5 PCs were selected.
Variance explained 99%

4 Run 8 different models of varying complexity and compare results

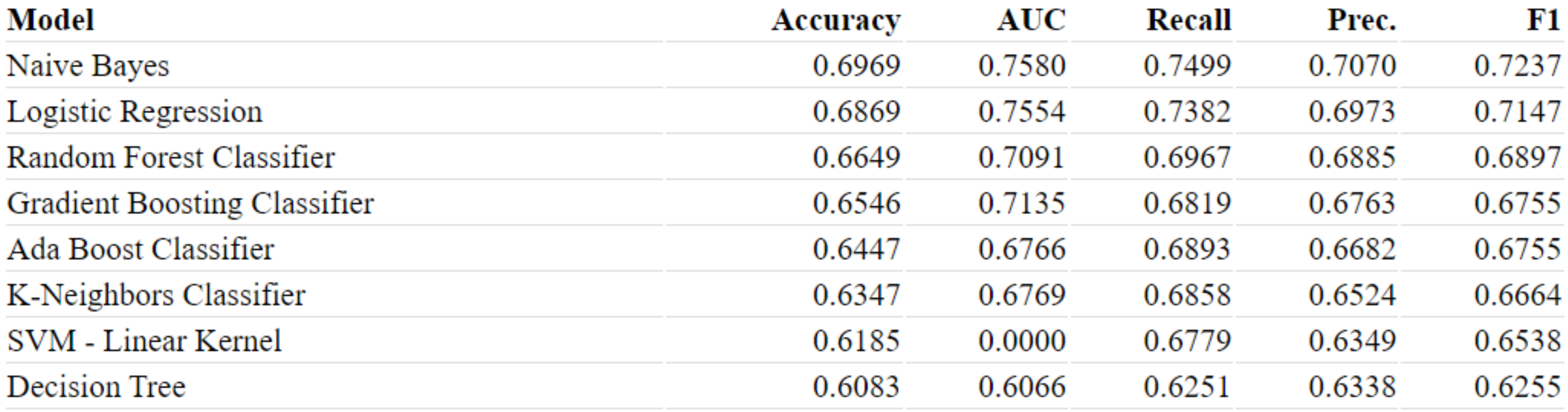
5 Tune hyperparameters of top 5 models, select top performing model and predict on test data

Agenda

- Motivation
- Literature Review
- Data Collection and EDA
- Process Flow
- **Methods Used**
- Results and Discussion
- Conclusions

Methods

- We ran 8 different models to classify whether a team would make it to the playoffs or not.



Model	Accuracy	AUC	Recall	Prec.	F1
Naive Bayes	0.6969	0.7580	0.7499	0.7070	0.7237
Logistic Regression	0.6869	0.7554	0.7382	0.6973	0.7147
Random Forest Classifier	0.6649	0.7091	0.6967	0.6885	0.6897
Gradient Boosting Classifier	0.6546	0.7135	0.6819	0.6763	0.6755
Ada Boost Classifier	0.6447	0.6766	0.6893	0.6682	0.6755
K-Neighbors Classifier	0.6347	0.6769	0.6858	0.6524	0.6664
SVM - Linear Kernel	0.6185	0.0000	0.6779	0.6349	0.6538
Decision Tree	0.6083	0.6066	0.6251	0.6338	0.6255

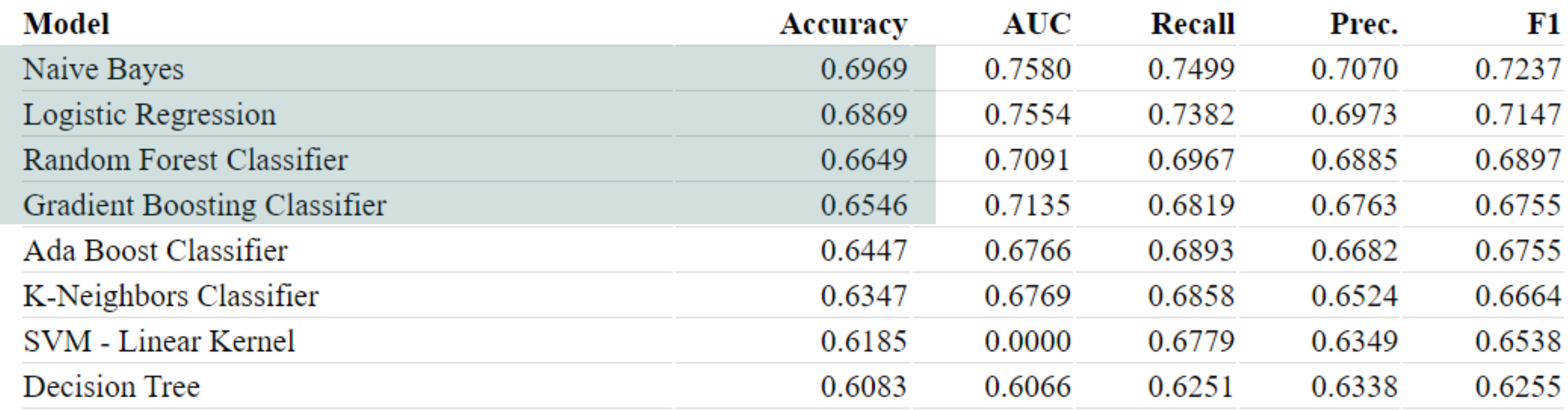
- As we can see from the above table, **Naive Bayes, Logistic Regression, Random Forest Classifier and Gradient Boosting Classifier** are top performing models.

- We select these 4 models for hyperparameter tuning along with **SVM**.

- We are comparing and selecting the model based on accuracy because there is **no class imbalance** in the dataset and, hence, **accuracy is a fair metric** to make comparisons on.

Methods

- We ran 8 different models to classify whether a team would make it to the playoffs or not.



Model	Accuracy	AUC	Recall	Prec.	F1
Naive Bayes	0.6969	0.7580	0.7499	0.7070	0.7237
Logistic Regression	0.6869	0.7554	0.7382	0.6973	0.7147
Random Forest Classifier	0.6649	0.7091	0.6967	0.6885	0.6897
Gradient Boosting Classifier	0.6546	0.7135	0.6819	0.6763	0.6755
Ada Boost Classifier	0.6447	0.6766	0.6893	0.6682	0.6755
K-Neighbors Classifier	0.6347	0.6769	0.6858	0.6524	0.6664
SVM - Linear Kernel	0.6185	0.0000	0.6779	0.6349	0.6538
Decision Tree	0.6083	0.6066	0.6251	0.6338	0.6255

- As we can see from the above table, **Naive Bayes, Logistic Regression, Random Forest Classifier and Gradient Boosting Classifier** are top performing models.

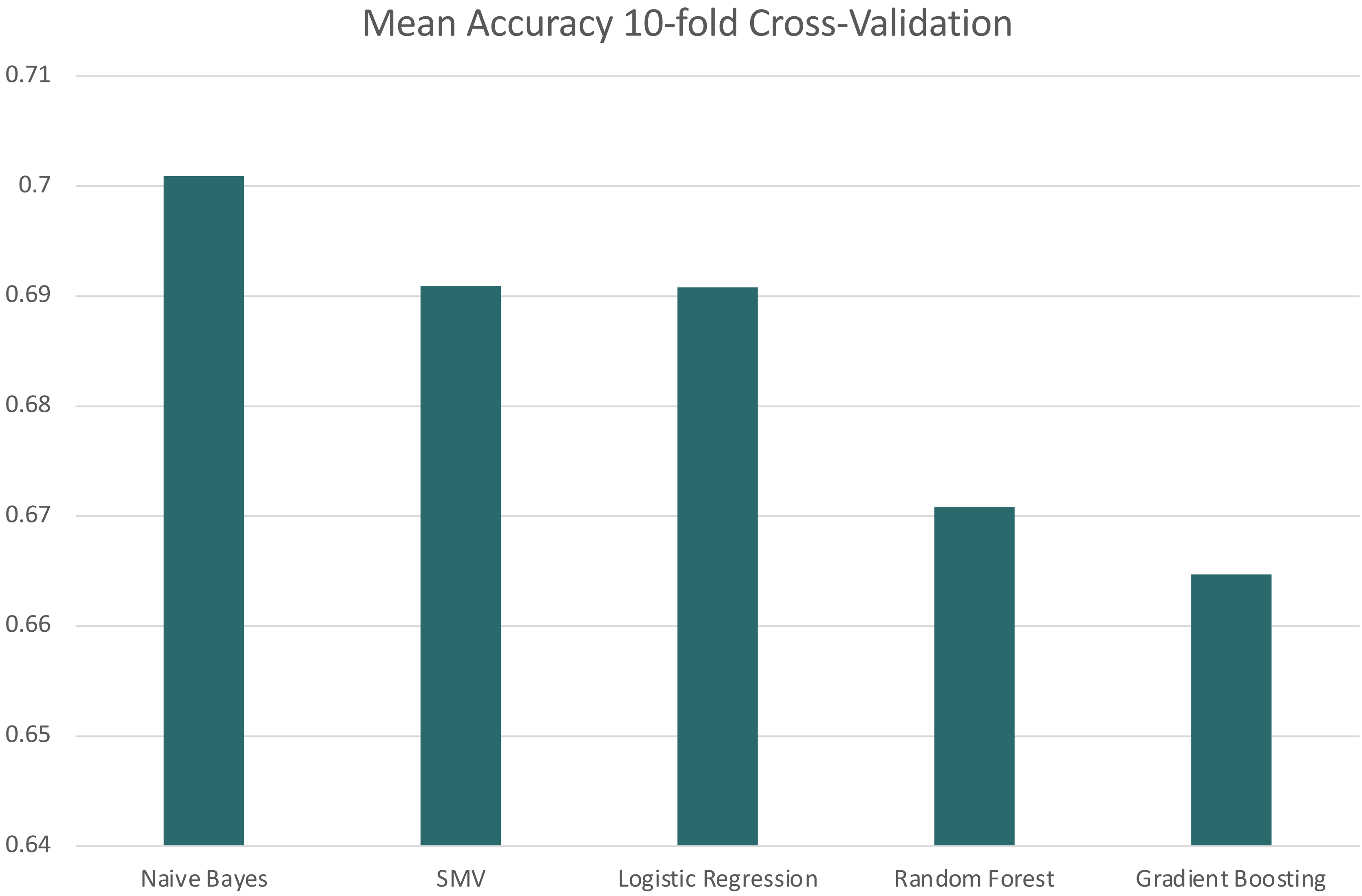
- We select these 4 models for hyperparameter tuning along with **SVM**.

- We are comparing and selecting the model based on accuracy because there is **no class imbalance** in the dataset and, hence, **accuracy is a fair metric** to make comparisons on.

Agenda

- Motivation
- Literature Review
- Data Collection and EDA
- Process Flow
- Methods Used
- **Results and Discussion**
- Conclusions

Hyperparameter Tuning and Model Comparison



Agenda

- Motivation
- Literature Review
- Data Collection and EDA
- Process Flow
- Methods Used
- Results and Discussion
- Conclusions

Conclusion

- Accuracy on the test set is **63.55%**.
- **Naive Bayes** performed the best.
- Why?

  1. Simple dataset.

  2.Naive Bayes fewer assumptions.

- After performing PCA, we get 5 **independent** predictors which satisfy Naive Bayes' assumption of feature independence.

Thank you