# CAMDA Drug Safety Challenge: PC3 Dataset

Presented by:
Bharathi Manoharan
Tejaswi Gorrepati
Spoorthy Balasubrahmanya
Ratna Kameswari Prabhala

# Presentation Overview

Problem Overview/Motivation

Libraries Used

Approach

Results

Conclusions

Division of Labor

# Introduction

Problem Statement: To predict drug induced liver injury (DILI) in humans by building machine learning models using human cell line gene expression data

Description of Data:

CMap gene expression responses of PC3 cancer cell line to 276 drug compounds is given.
Dimensions: 276 rows x 22278 columns
There are 276 drug compounds and 22278 genes

Training Vs Validation split : 190 : 86

In addition, the class labels of DILI for 190 training samples are available.

# Libraries Used

Class – for  knn
Impute – used to impute missing data with knn
e1071  – for svm and Naive Bayes
Iterative BMA - for BMA
ROCR - To draw ROC curves that visualise classifier performance

**New method:**
Naive Bayes, 10-fold cv

# Approach:

- Data processing steps:

Used RMA normalisation to remove noise in the data

- Feature selection and classification methods attempted

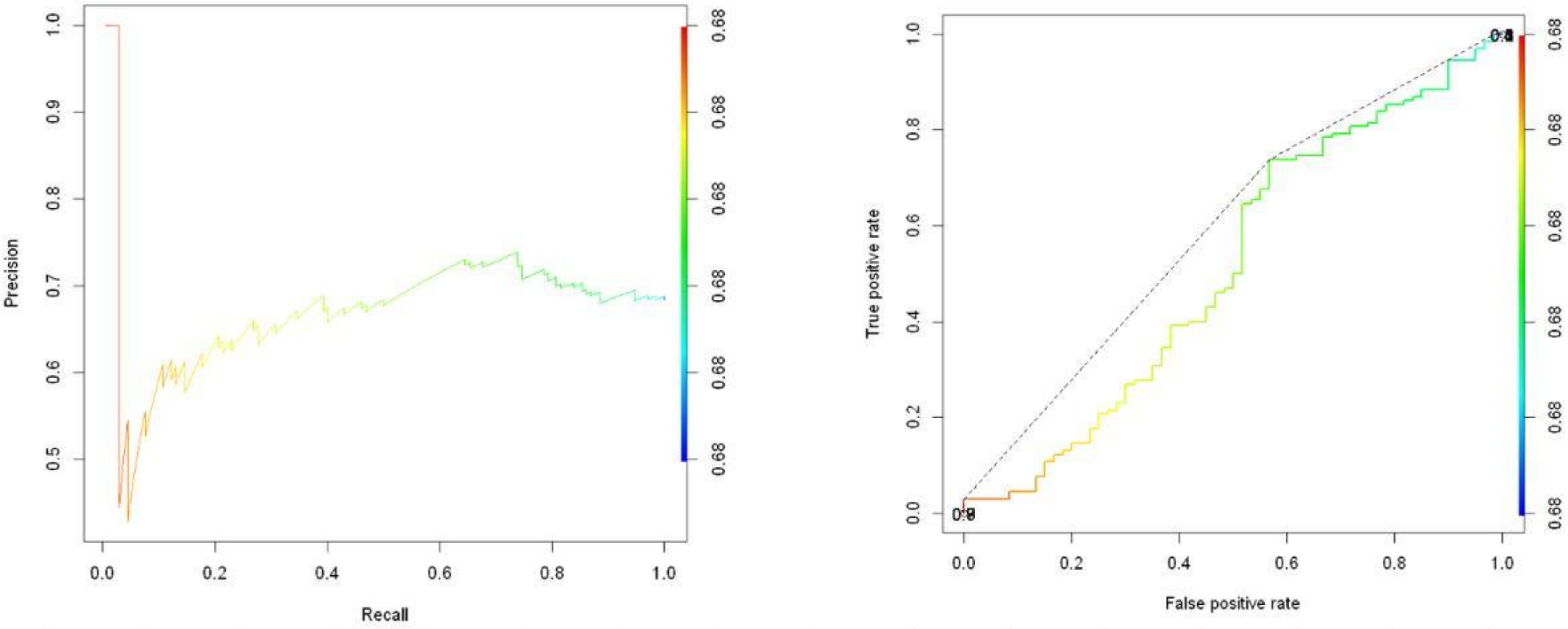
| Classification Method | Feature Selection |
|---|---|
| KNN with k=10 | Low Correlation(0.75),High Correlation(0.95) |
| KNN with k=13 | Low Correlation(0.65),High Correlation(0.98) |
| SVM | Low Correlation(0.75),High Correlation(0.9) |
| Random Forest | High Correlation(0.9) |
| Naive Bayes | Low Correlation(0.01),High Correlation(0.9) |
| BMA | P value < 0.01,0.05,0.005,0.001 |

# Approach

- Used 10 fold cross-validation with 3 repeats
- Mean function was used to get the accuracies from the predictions
- Used AUC as an assessment measure for BMA combinations

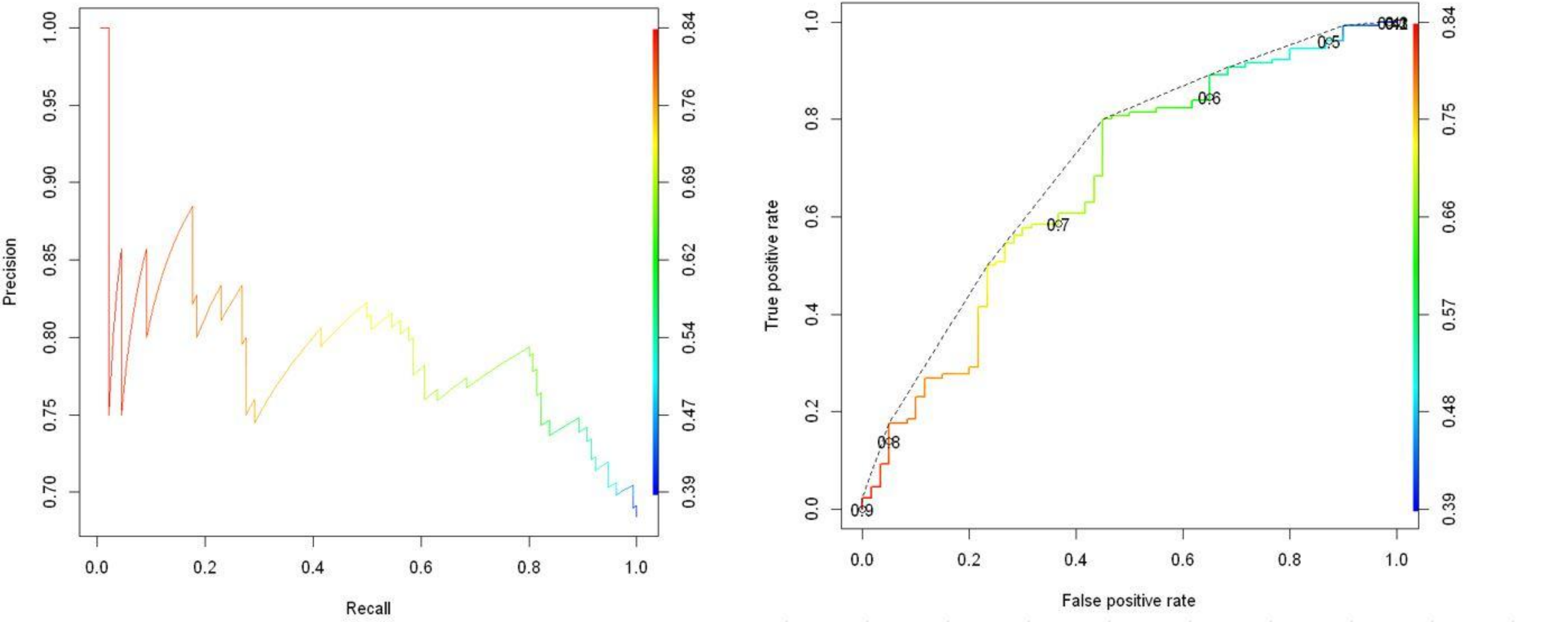
# Visualization - Performance trade-offs in 2D plot - BMA Combinations

## BMA with correlation - p value < 0.005



# Visualization - Performance trade-offs in 2D plot - BMA Combinations

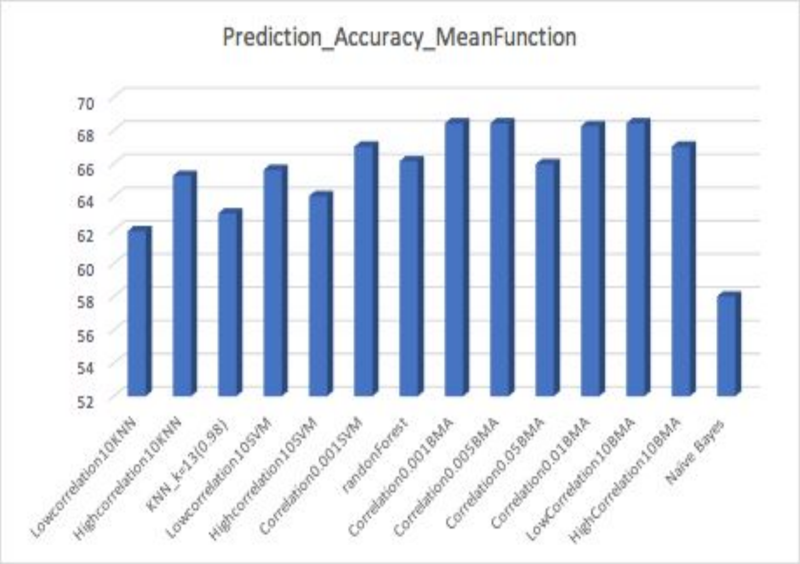
BMA with correlation - p value < 0.25



# Key Takeaway

1. What are the main concerns for a Data Scientist while building a good predictive model?
2. Performed feature selection to improve model prediction accuracy and that makes the process faster.
3. Gained knowledge about the importance of data preprocessing.
4. Investigated and tried various classifiers and understood machine learning concepts
5. Understood which method works well for what kind of classification. E.g, SVM worked well for binary classification.

# Results Visualization



BMA with Correlation 0.001 = 68.4%

SVM with Correlation 0.001= 67.0%

Random Forest = 66.1%

High Correlation KNN (0.95) = 65.2%

KNN(k=13) with correlation(0.98) =63%

Naive Bayes with correlation() = 58%

# Challenges

What worked?

1. Figuring out the best p value was challenging since the data is huge, So we had to check with various p values to see how models fit.
2. BMA (68.4%)and SVM (67%)stood out as the top two classifiers among several tried methods such as  KNN, Naive Bayes and  Random Forest.

What did not work?

1. Random forest model did not fit as expected, the accuracy was not accurate.
2. Faced an error when training and testing data size did not match.

# Future Work and Conclusion

Naive Bayes and random forest did not performed as per expectation.

Naive Bayes did not performed well when tried with both high and low correlation values.

BMA outperformed with a prediction accuracy of 68.4%.

Accuracy could be improved by applying ensemble technique.

Other feature selection techniques can be applied to see how models fit.

# Division of labour:

Bharathi Manoharan- BMA and ROCR Assessment + coding and documentation

Spoorthy Balasubrahmanya- SVM, Random Forest + coding and documentation

Tejaswi Gorrepati- KNN + coding and documentation

Ratna Kameswari Prabhala- Naive Bayes + coding and documentation

# Thank You!!