Predicting Students' Academic Performance

Alexandros Zervopoulos - ΠΜΕ201905

Data Mining and Knowledge Management
May 13, 2020

Introduction

- Goal: Predict academic student performance
  - Useful in personalized identification of problematic background and lackluster habits
- Methodology
  - Search for relevant publications and datasets
  - Compare different approaches
  - Reproduce their results
  - Using Machine Learning (ML) techniques

Datasets

- Two different datasets were experimented with
- Dataset 1: Student Academics Performance
  - 300 instances
  - 22 features
  - available: **UCI**[1]
  - Introduced in [1]
- Dataset 2: Students' Academic Performance
  - 480 instances
  - 17 features
  - available: **Kaggle**[2]
  - Introduced in [2]

[1] https://archive.ics.uci.edu/ml/datasets/Student+Academics+Performance

[2] https://www.kaggle.com/aljarah/xAPI-Edu-Data

Dataset 1

- Constructed in [1] from students enrolled in three colleges in India
- Includes features
  - Demographic
    - Gender, family size, family income, father/mother education and profession
  - Academic Background
    - Past grades, free/paid admission, class attendance
  - Personal
    - Studying habits, number of friends
- Performance Class
  - Best, Very Good, Good, Pass, Fail

Dataset 2

- Constructed in [2] from Kalboard, a learning management system
- Obtains behavioral features from API
  - Demographic
    - Gender, nationality, place of birth, parent responsible for student
  - Academic Background
    - Current stage, grade level, semester, topic, absence days
  - Behavioral
    - Discussion groups, visited resources, raised hand on class, viewing announcements
  - Parent Participation
    - Parent answering survey, parent school satisfaction
- Performance Class
  - Low, Medium, High

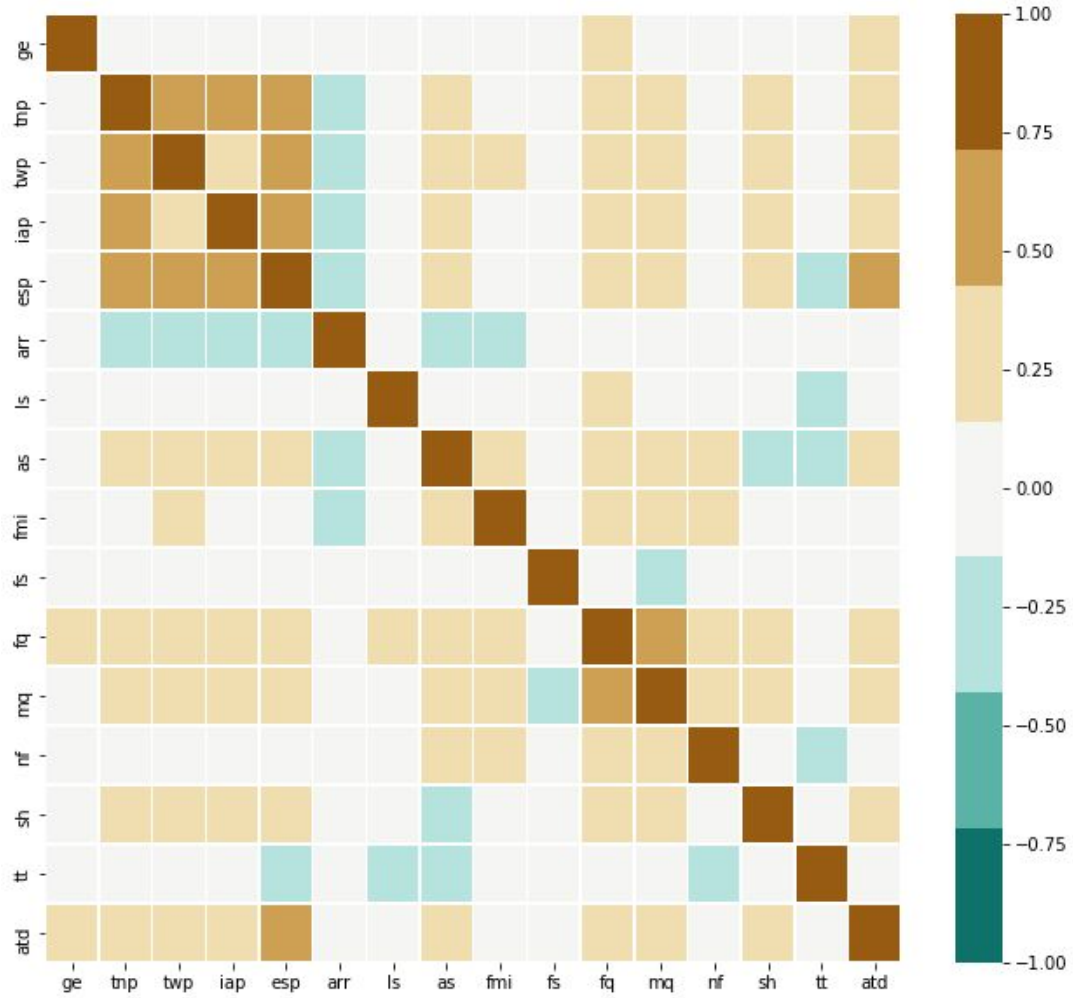
Methodology

- Weka was used in both [1, 2]
- Here, a standard Python-based toolkit is utilized
  - Scikit-learn for classification, feature selection
  - pandas for data management and preprocessing
  - seaborn and matplotlib for visualization and plotting
- All conducted experiments are available on Github[3]
  - In Jupyter notebook format

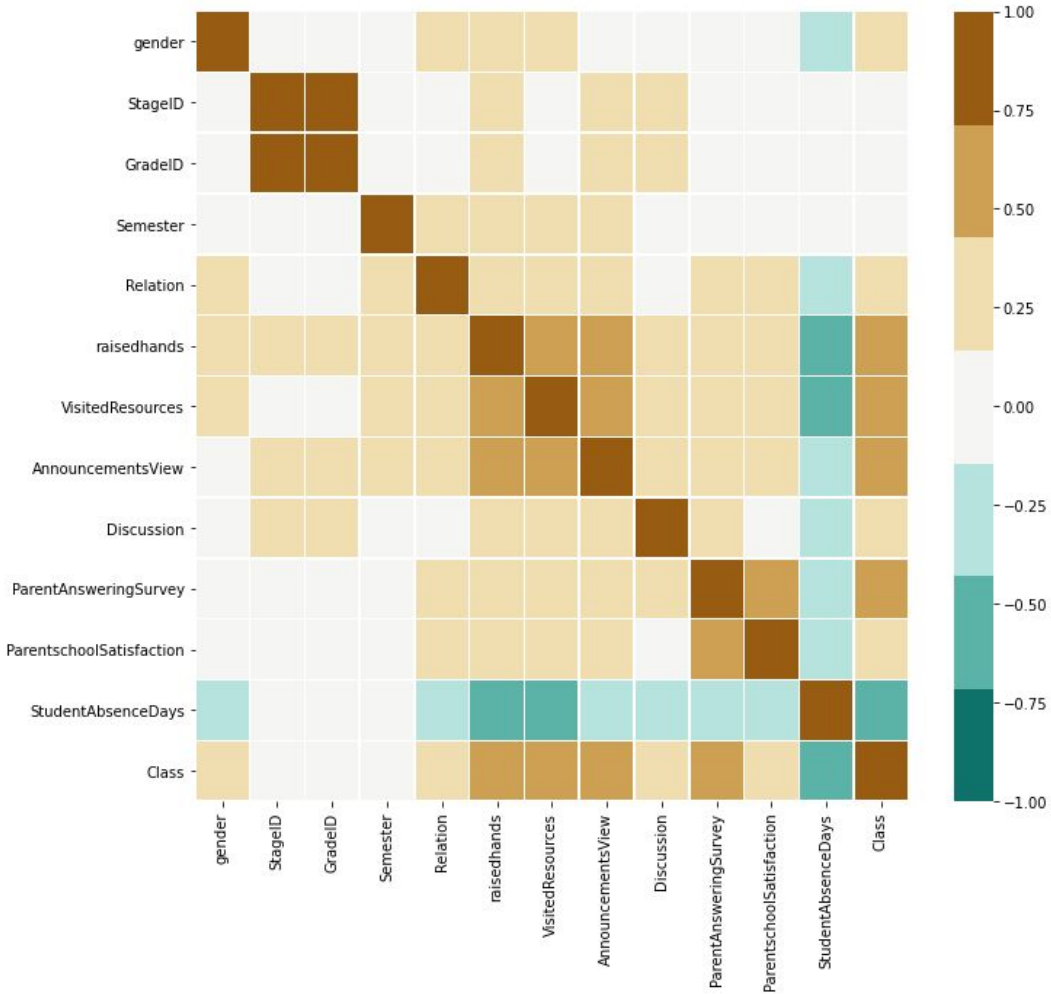
[3] https://github.com/p15zerv/academic-performance-prediction-sklearn

Preprocessing and Visualization

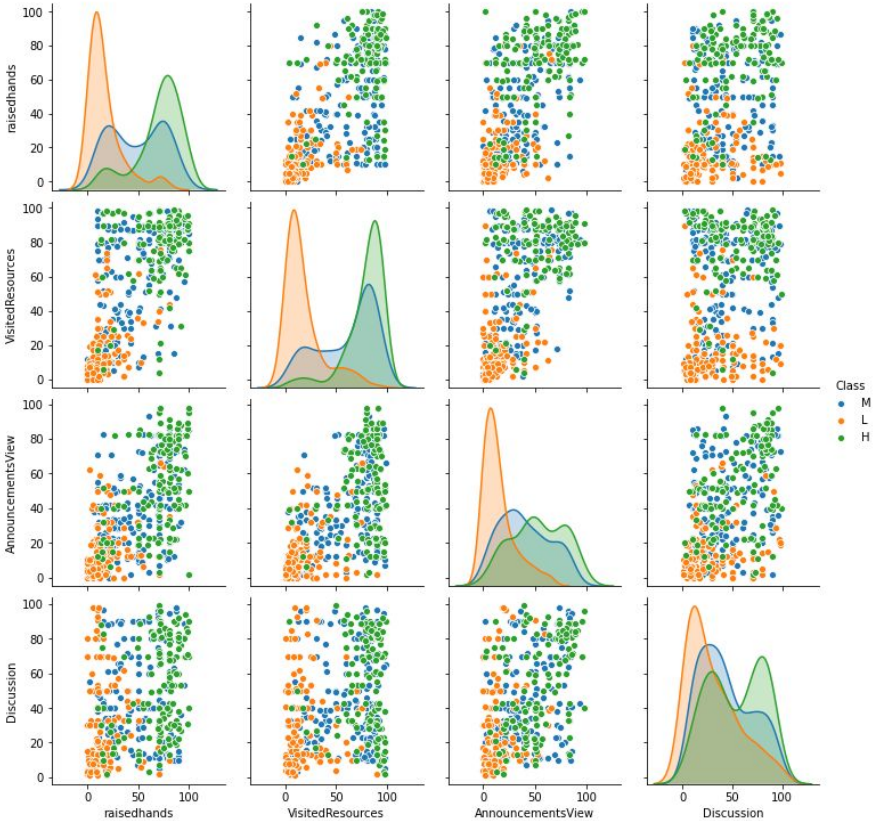
- Distribution of all features, colored by class
- For preprocessing
  - All nominal attributes are mapped to a numeric scale
  - Other categorical attributes are dropped or one-hot encoded
- Correlation heatmaps and pairwise relation plots are generated



Pairwise Pearson correlation heatmap for Dataset 1



Pairwise Pearson correlation heatmap for Dataset 2



Pairwise relation plot for numerical features of Dataset 2

Baseline Classification - Dataset 1

- No mention of train-test split or cross validation!
- Only 131 instances seem to be available in the online dataset!

Algorithm	Current Accuracy	Reported Accuracy	Corresponding algorithm [1]
Naive Bayes (NB)	48.89%	-	-
-	-	65.33%	BayesNet
Logistic Regression (LR)	61.08%	-	-
Decision Tree (DT)	57.92%	73%/74.33%	J48/PART
Random Forest (RF)	61.88%	99%	Random Forest
DT AdaBoost (AB)	58.91%	-	-

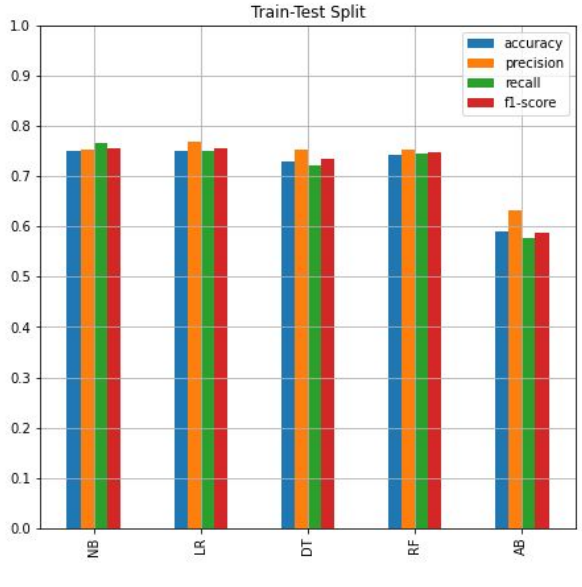
Baseline Classification - Dataset 2

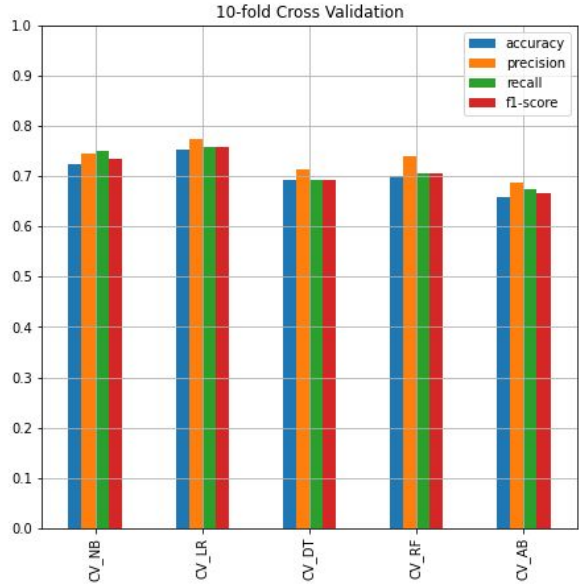
- 10-fold cross validation

Algorithm	Current Accuracy	Reported Accuracy	Corresponding algorithm [2]
Naive Bayes (NB)	72.5%	75.8%	Naive Bayes
Logistic Regression (LR)	75.21%	-	-
Decision Tree (DT)	69.17%	75.8%	J48
Random Forest (RF)	70.42%	75.6%	Random Forest
DT AdaBoost (AB)	65.83%	77.7%	Boosting - J48
Artificial Neural Network	76%	79.1%	Artificial Neural Network

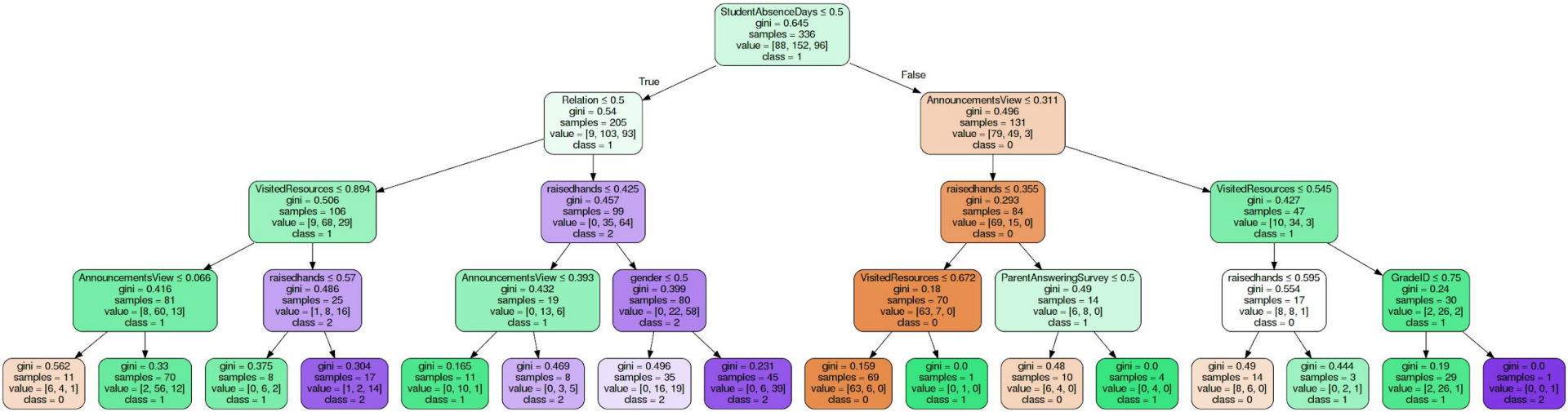
Baseline Classification - Dataset 2

- 70:30 train test split
- 10-fold cross validation

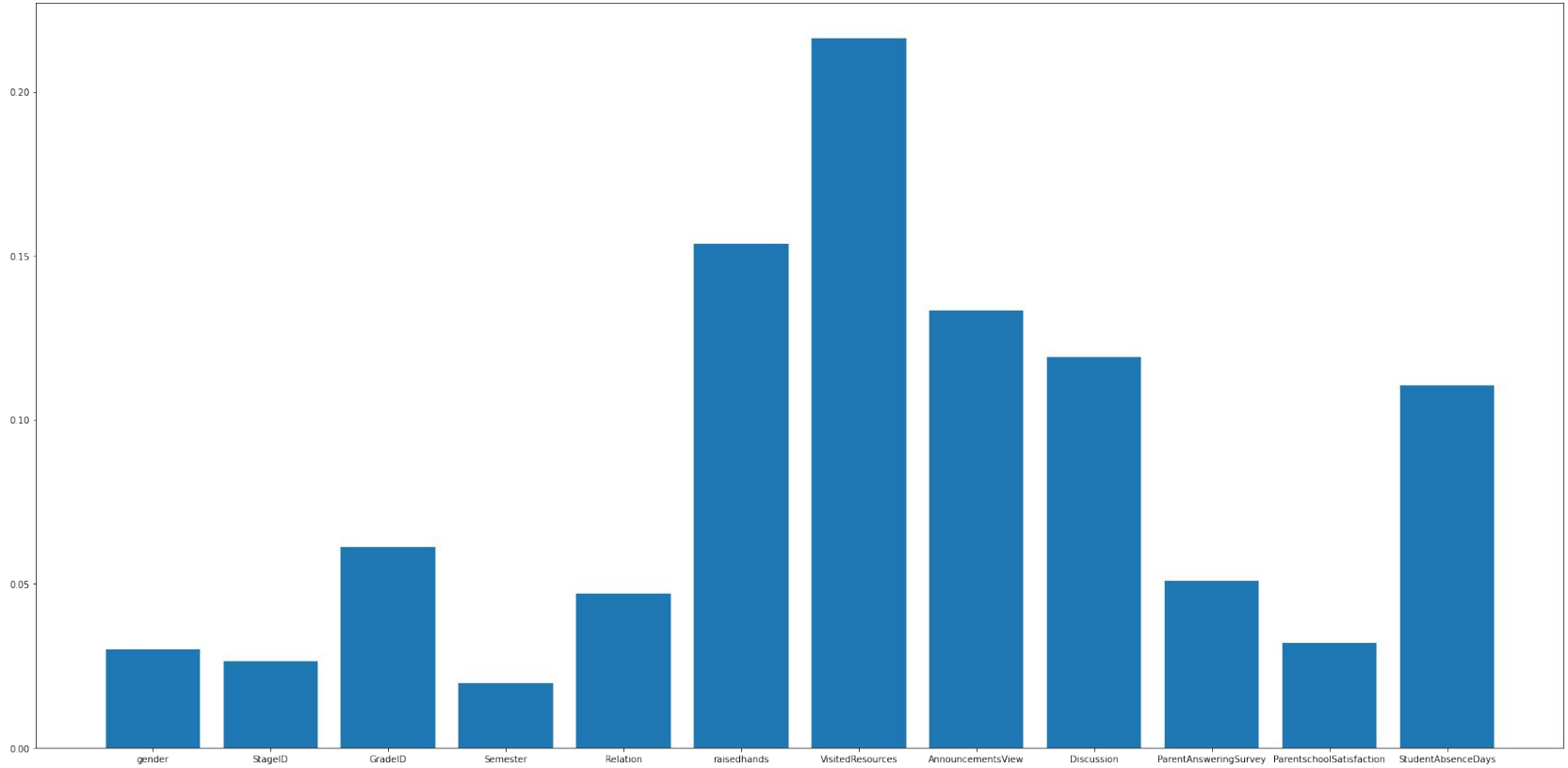




Decision Tree Model - Dataset 2

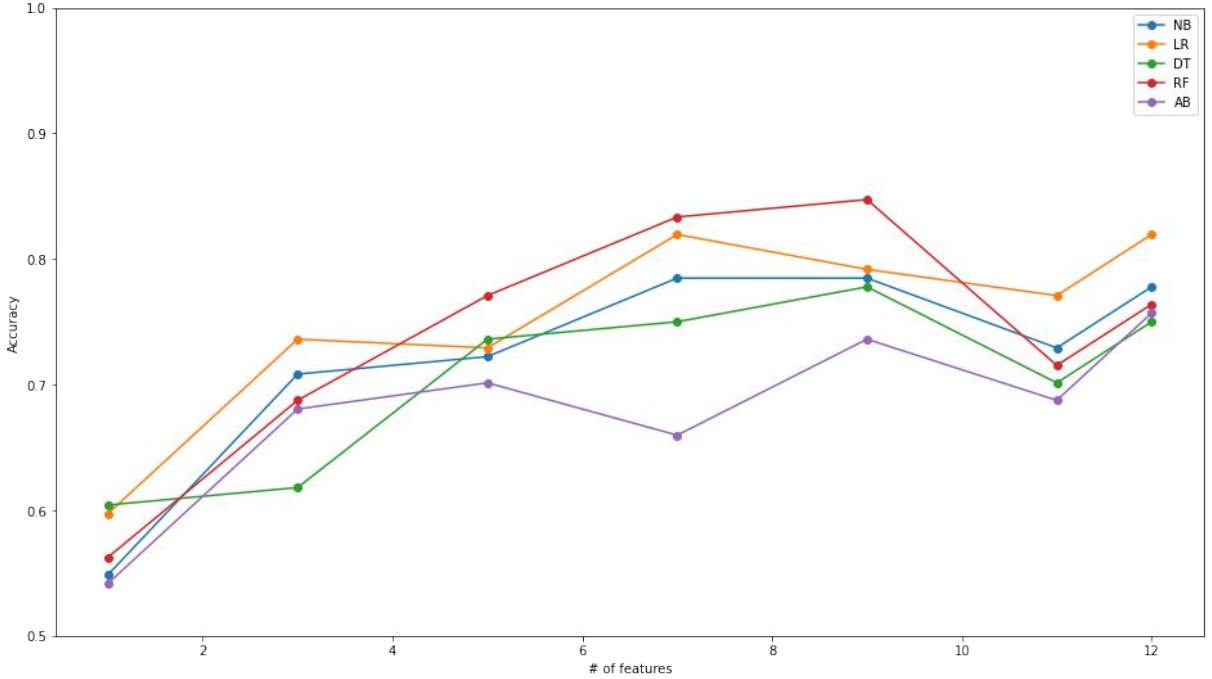


RF Features Importances - Dataset 2



Feature Selection - Dataset 2

- Select top *k* features for various k
  - Features ranked according to mutual information
- Compare performance
  - Using 70:30 train-test split



Feature Selection - Dataset 2

- 10-fold cross validation
- Using Recursive Feature Elimination
  - Except for Naive Bayes

Algorithm	Accuracy w/o Feature Selection	Accuracy w Feature Selection	Number of Features
Naive Bayes (NB)	72.5%	78.81%	7
Logistic Regression (LR)	75.21%	75.21%	10
Decision Tree (DT)	69.17%	68.54%	6
Random Forest (RF)	70.42%	70.83%	10
DT AdaBoost (AB)	65.83%	65.83%	7

Thank you for your attention!

Any questions?

References

[1] Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. International Journal of Database Theory and Application, 9(8), 119-136.

[2] Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. Indonesian Journal of Electrical Engineering and Computer Science, 9(2), 447-459.