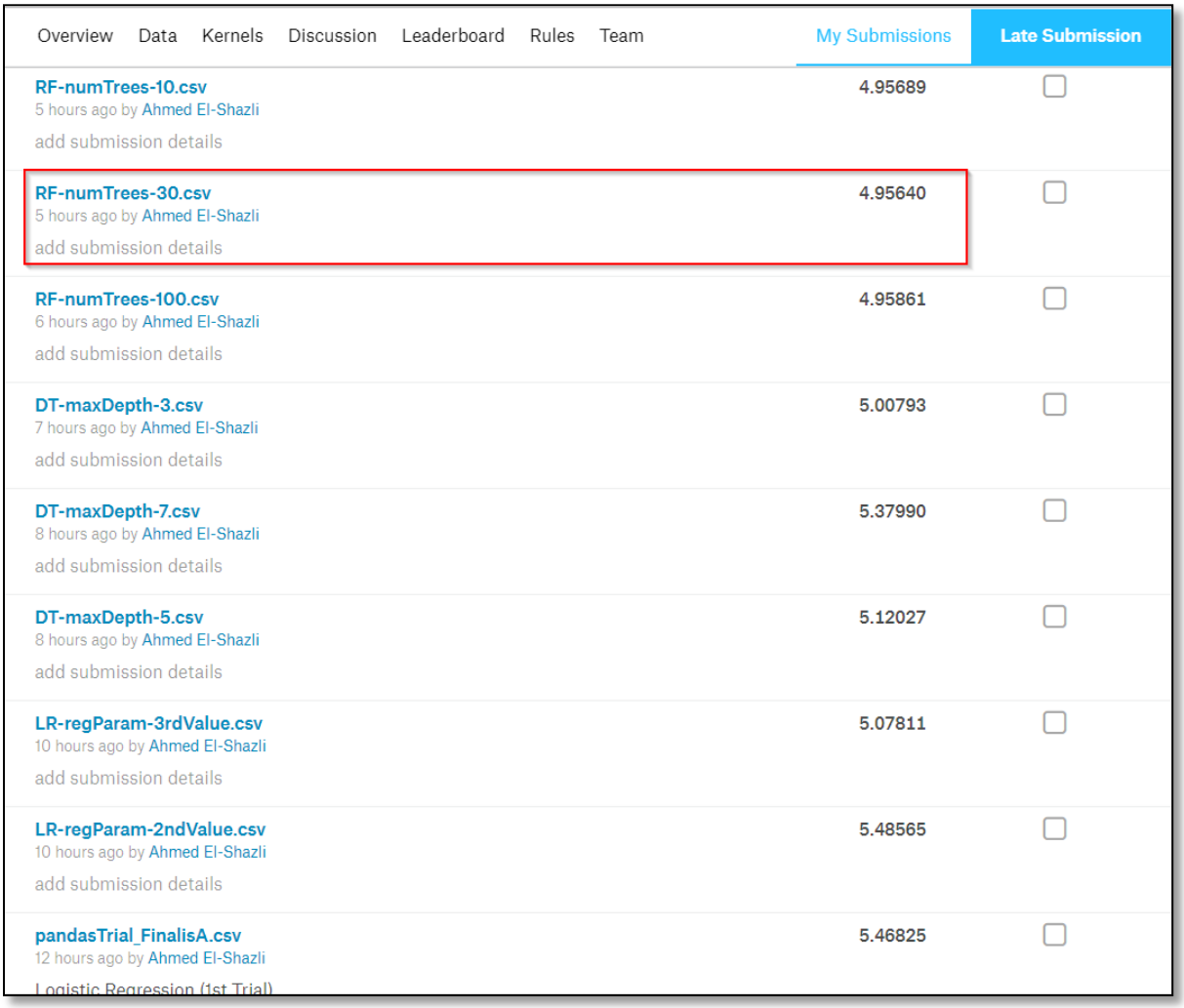
Big Data 2 Project
San Francisco Crime Classification

- **Name: Ahmed Mohamed Elshazli**
- **Student ID: 181104**

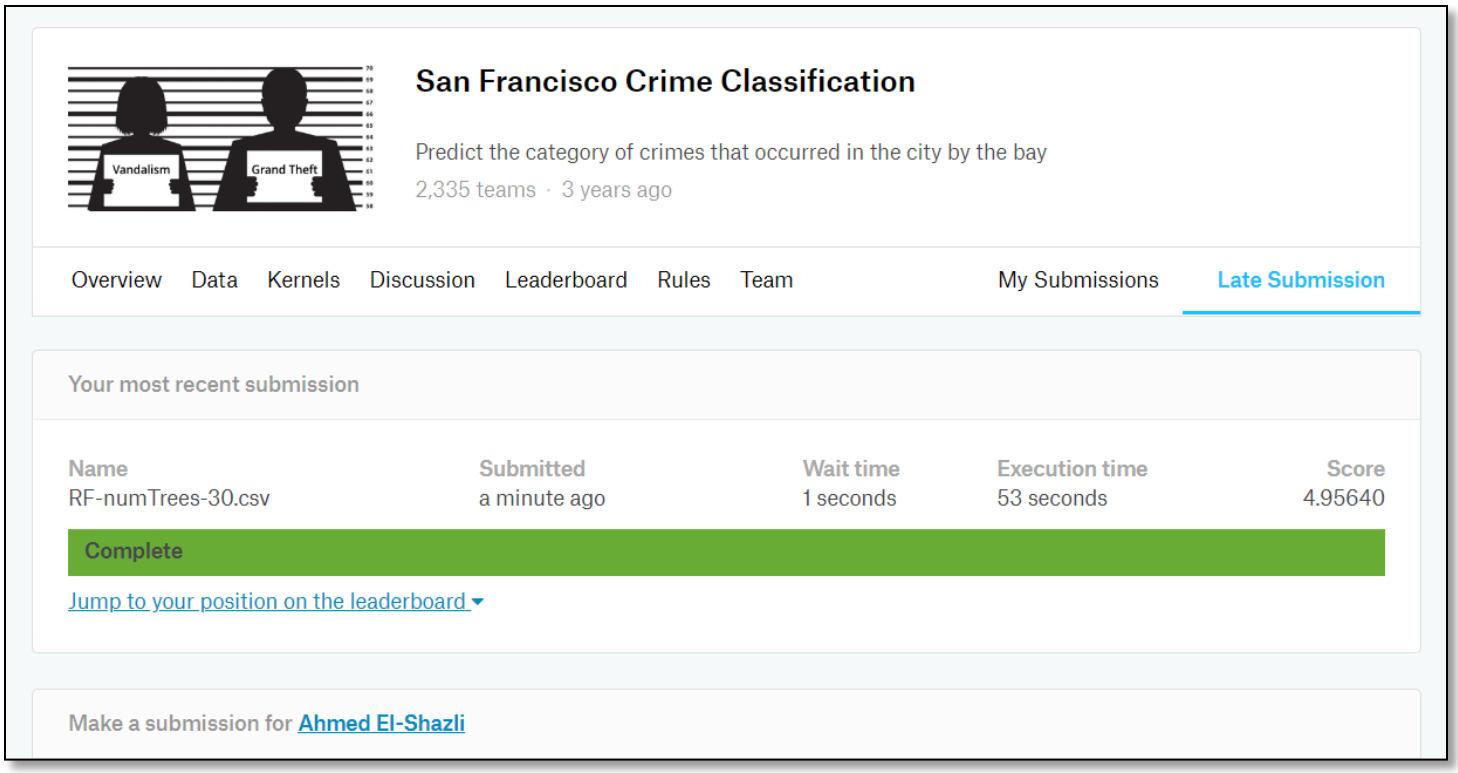
➢ Three classifiers were used as mentioned below (with the corresponding tuned parameter & final score):

Classifier	Hyper Parameter	Score
Logistic Regression	regParam = 0.01	5.46825
	regParam = 0.1	5.48565
	regParam = 1	5.07811
Decision Tree	maxDepth = 3	5.00793
	maxDepth = 5	5.12027
	maxDepth = 7	5.37990
Random Forest	numTrees = 10	4.95689
	numTrees = 30	4.95640
	numTrees = 100	4.95861

➢ **Kaggle Submissions:**



- **Kaggle Best Submission:** obtained through Random Forest Classifier with numTrees set to 30.



- Why Random Forest is the best model?

  Answer:

  1- Normally, when the available independent variables (features) are categorical, **random forest** tends to perform better than **logistic regression**.

  2- Additionally, compared to **Decision Tree**; A **random forest** is simply a collection of **decision trees** whose results are aggregated into one final result. Their ability to limit overfitting without substantially increasing error due to bias is why they are such powerful models.

  3- **For numTrees = 30**: In general, the more trees you use the better get the results. However, the improvement decreases as the number of trees increases, i.e. at a certain point the benefit in prediction performance from learning more trees will be lower than the cost in computation time for learning these additional trees.