



# **Analysis Report on Kagglers' Job Satisfaction Level and Building a Machine Learning Model**

**SABANCI UNIVERSITY**

**Course-DA514**

**Group 7**

**Gizem Güneş, Esra Akarçay, Ayda Gizem Kumkumoğlu**

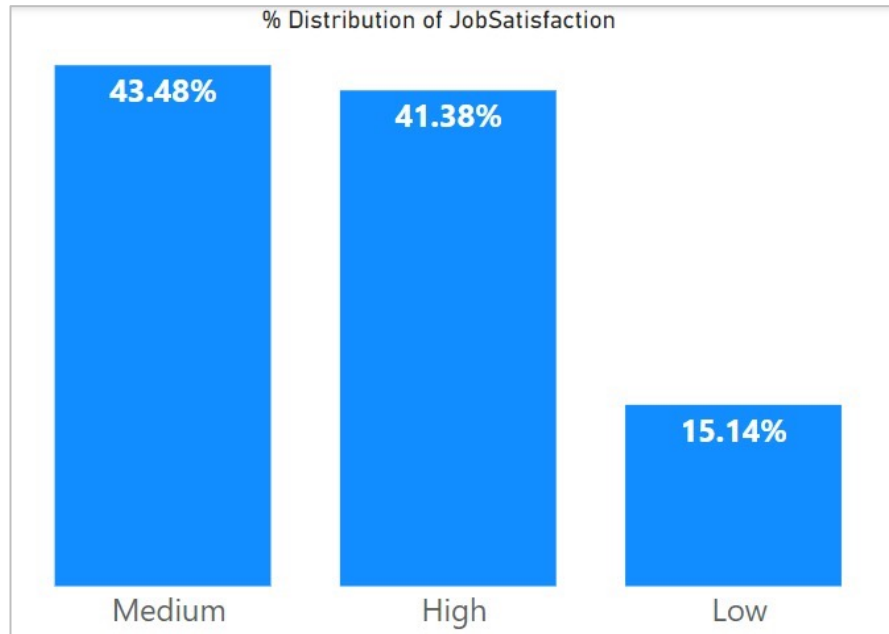


# CONTENTS

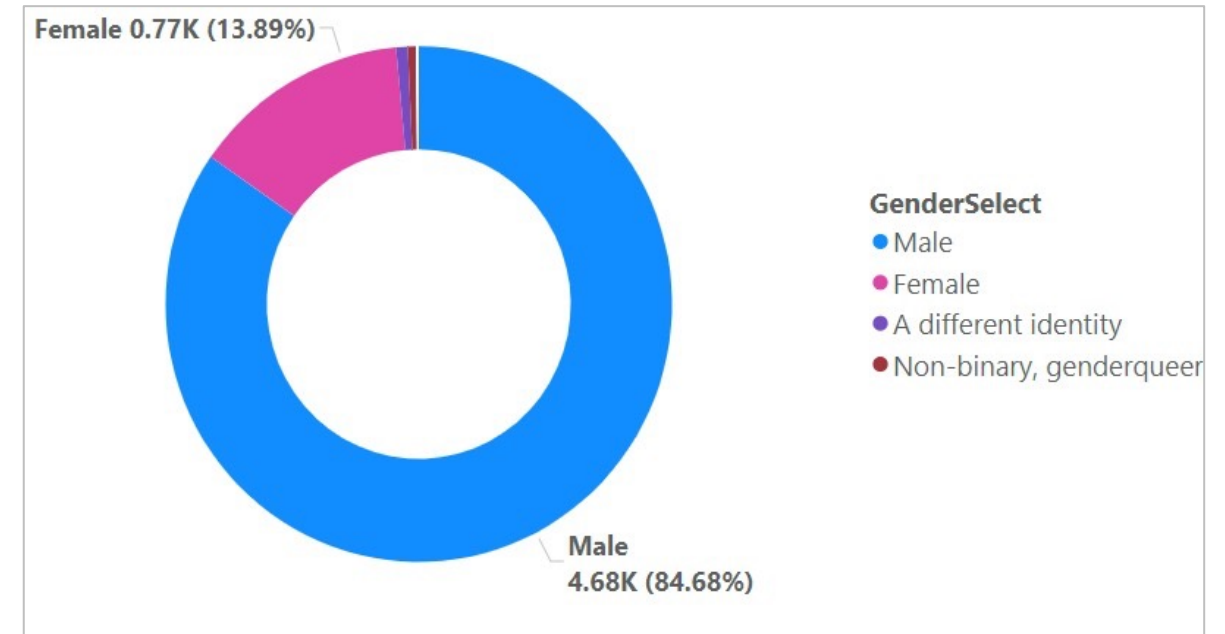
**01** Statistics About Kagglers

**02** Model Building

# Statistics About Kagglers

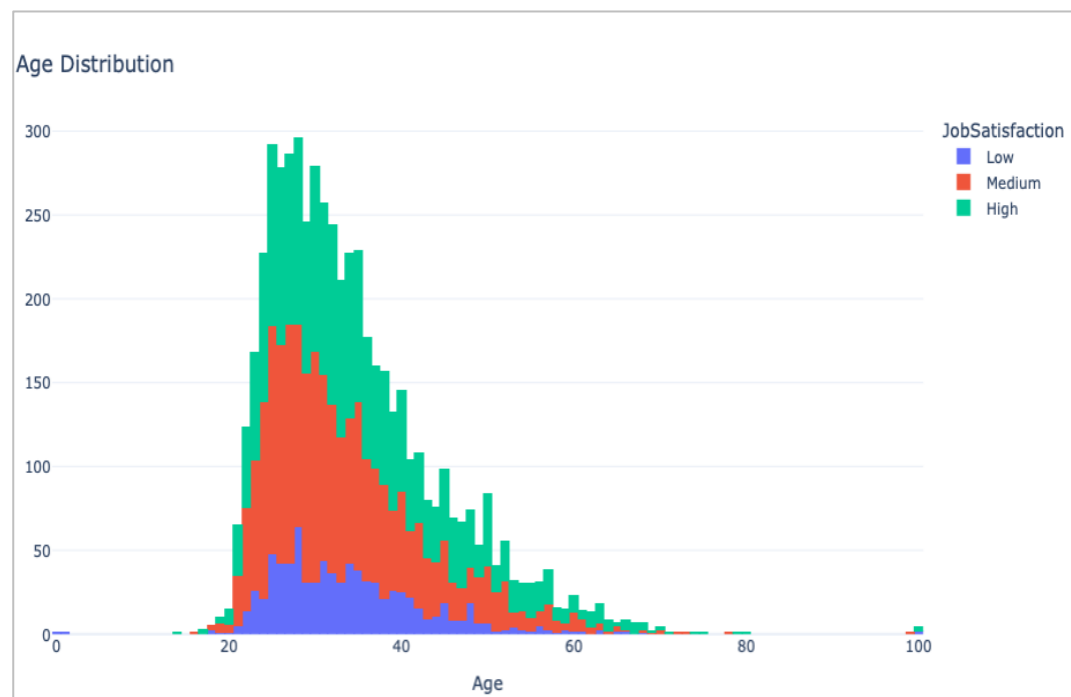


*Project data set is an imbalanced data set consist of 43.48% Medium, 41.38% High and 15.14% Low labels*

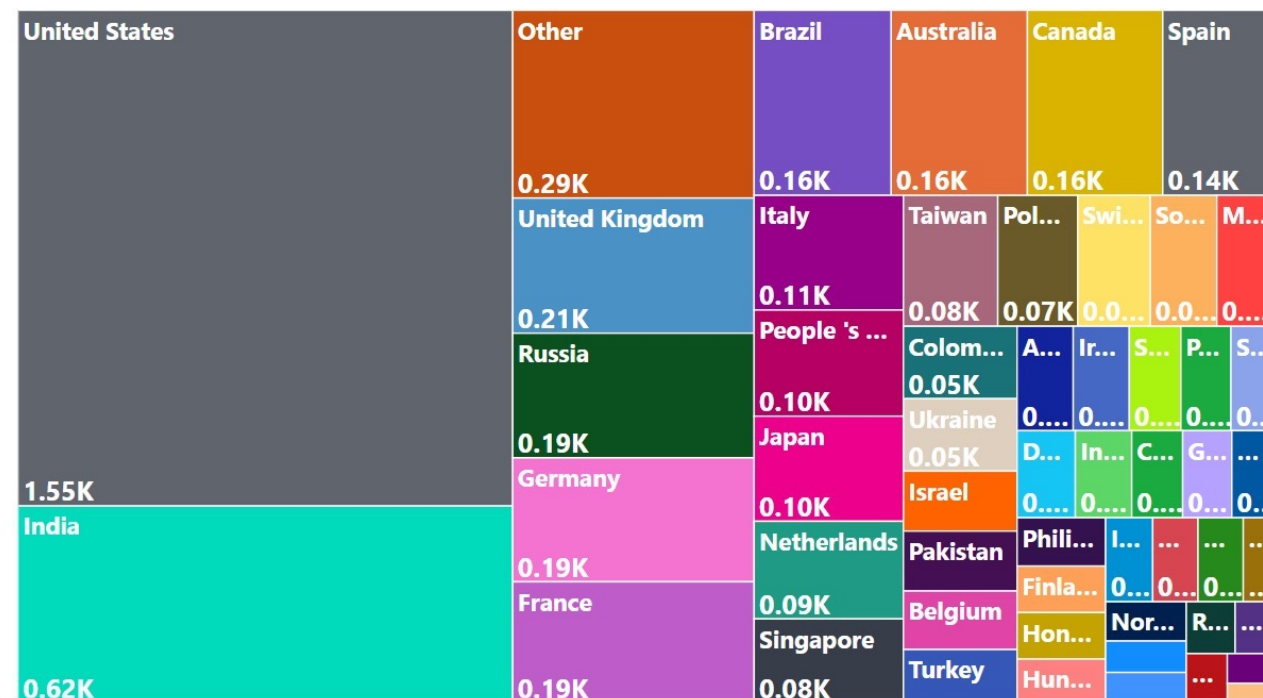


*Exploratory data analysis findings show that 85% of Kagglers are male and 14% are female. Female Kagglers have low job satisfaction with 16% which is 2 point greater than Male Kagglers' low job satisfaction ratio*

# Statistics About Kagglers

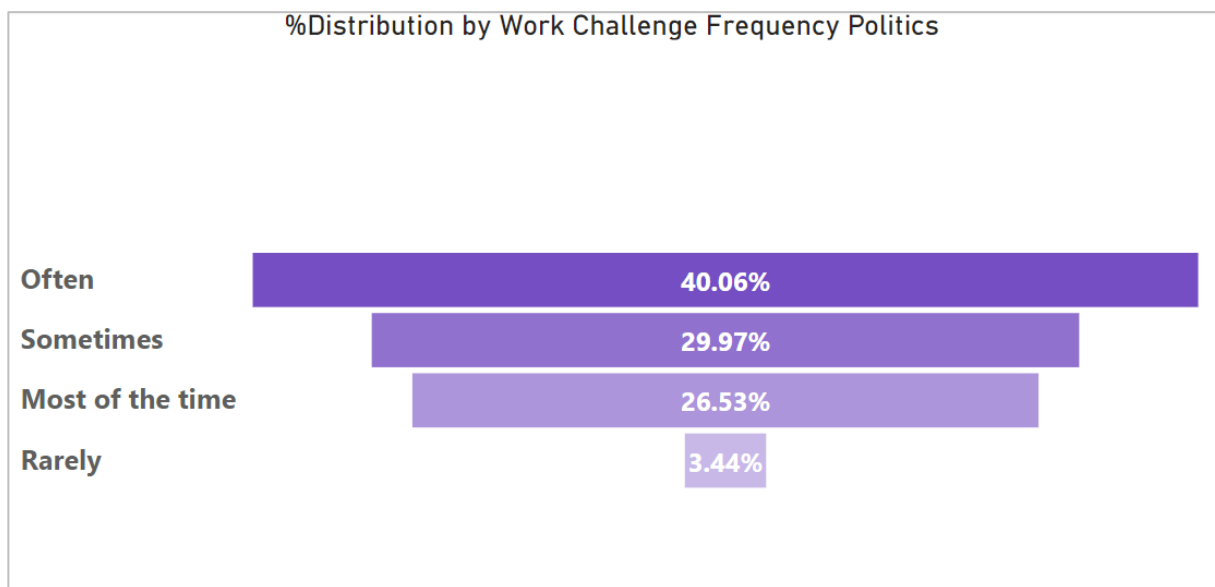


*Kagglers' age has a skewed distribution with the mean age of 35 approximately*

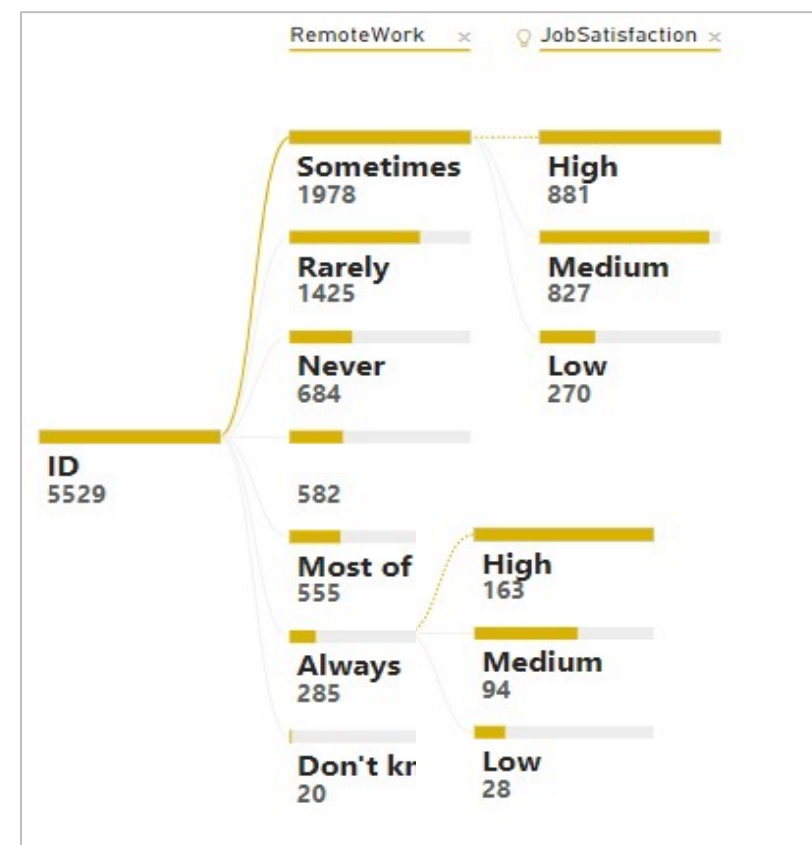


*Kagglers are mostly located in the USA, Asia and West Europe while the rest is located in other regions*

# Statistics About Kagglers

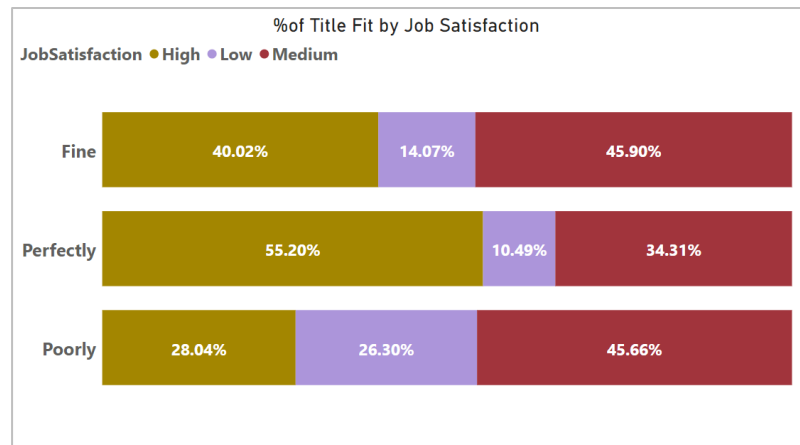


*Kagglers who face challenges like company politics, lack of management and financial support for a data science team most of the time in their working environment have the lowest job satisfaction level.*

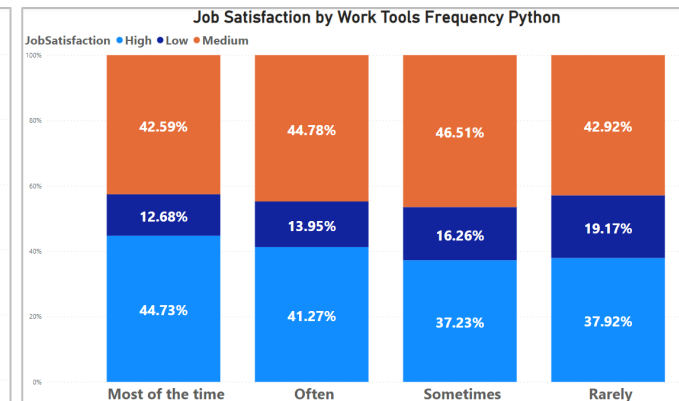
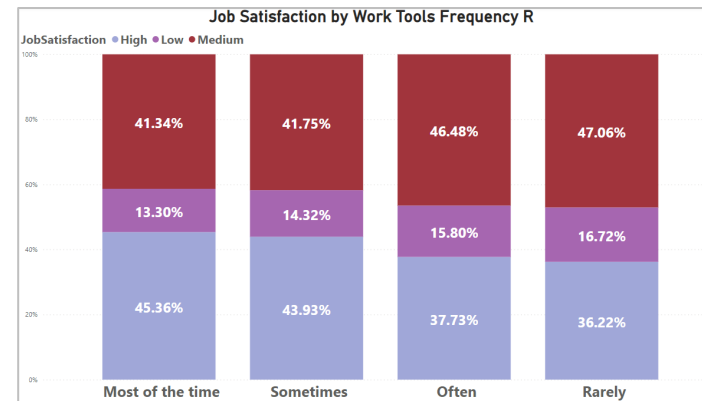


*Kagglers who always work remotely have higher job satisfaction.*

# Statistics About Kagglers



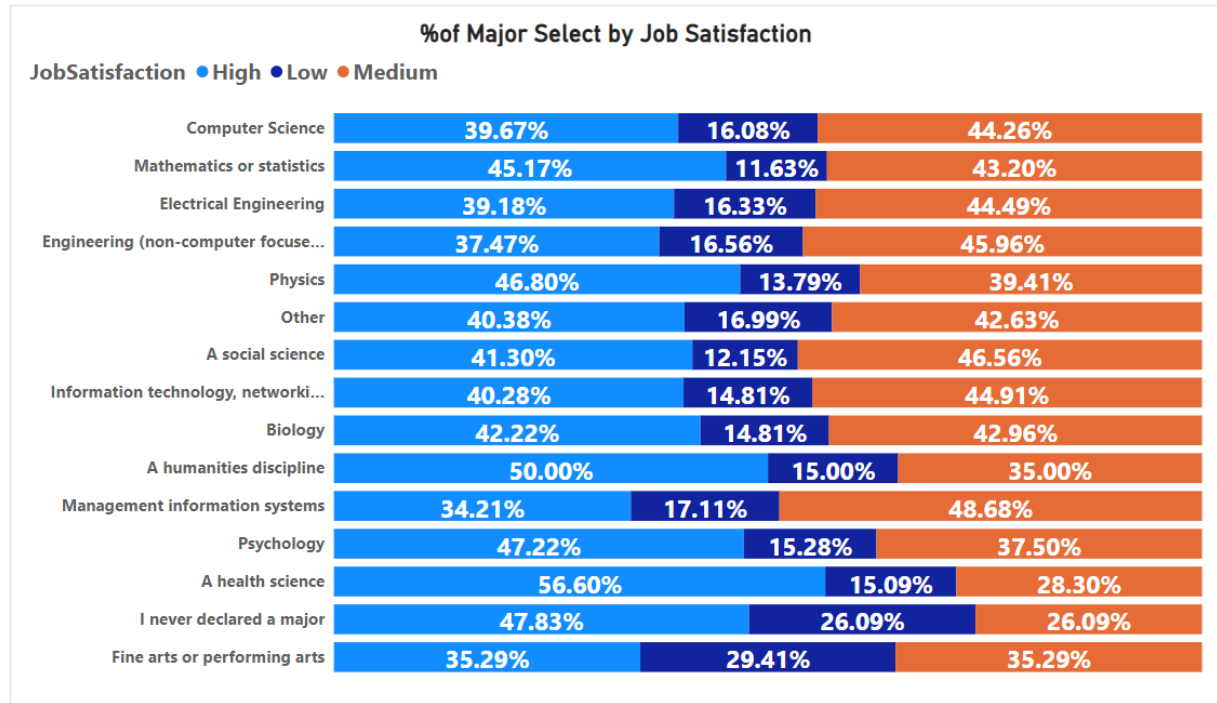
*Kagglers who have claimed that they are perfectly fit to their title are more satisfied (55%) with their jobs than others who claim that they are fine or poorly fit to their title.*



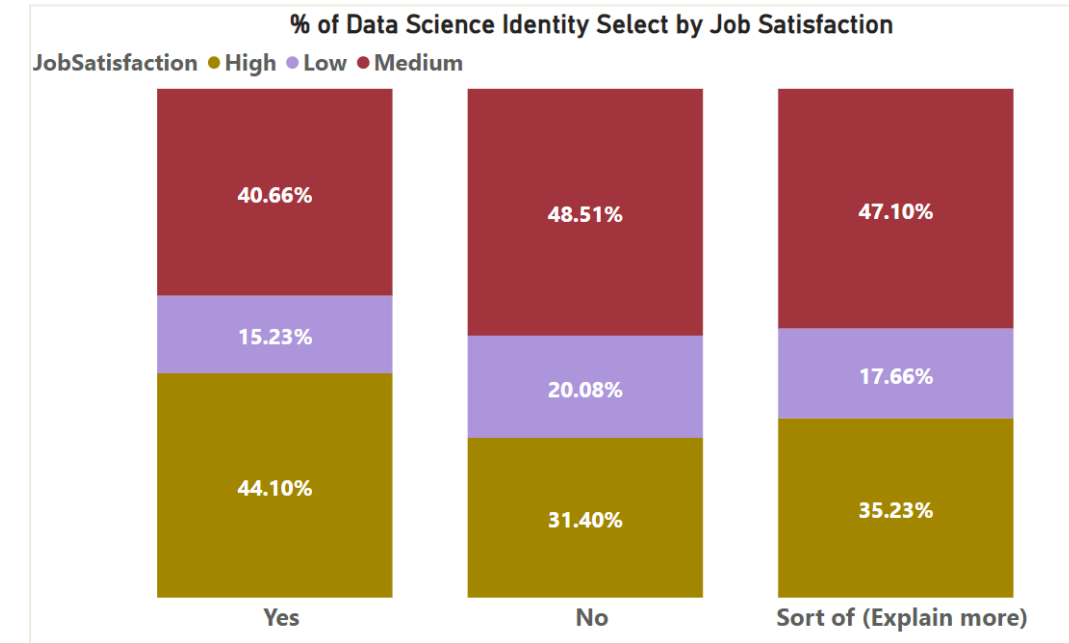
*Python and R users indicate higher satisfaction levels if they use these tools most of the time or often.*



# Statistics About Kagglers

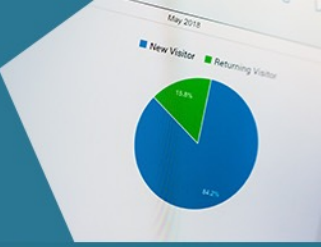


Kagglers mostly have Bachelor Degrees from Computer Science, Mathematics-Statistics and Engineering (Non-computer Focus) majors and their mostly employed in the academic and technology fields



Kagglers specified several current job titles and Data Scientist title is the most frequent one. Also, Kagglers who identified themselves as Data Scientist are dominant in the data set.

# Model Selection



FEN:

Originally an Arabic Word that means **Art**, but in Turkish it means **Science**  
Like Machine Learning...



# Model Building



Score Table

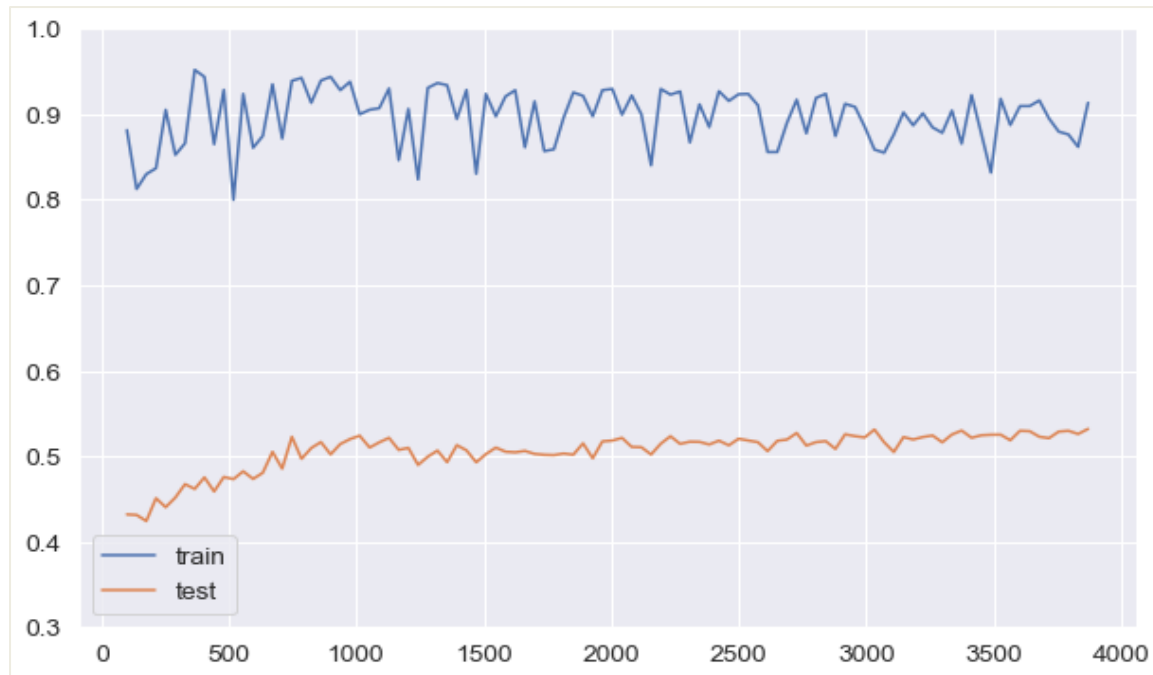
method	TP	FP	TN	FN	f1
Baseline LR	104	55	460	67	0.501698
GB Classifier	0	0	380	58	0.471919
GB Classifier with new parameters	17	8	420	62	0.496330
XGBoost	15	8	424	61	0.505600
XGBoost with new parameters	29	14	420	61	0.508666
Random Forest	102	50	440	55	0.524855
RF with DecisionTreeRegressor iterative imputer	91	44	426	58	0.533324
RF with ExtraTreesRegressor iterative imputer	92	42	442	59	0.524585
RF with KNeighborsRegressor iterative imputer	98	52	435	62	0.530858
RF with BayesianRidge iterative imputer	88	44	420	60	0.529858
Random forest with SMOTENC	69	54	409	64	0.502873

- In this part of the project, there have been built **11 models** based on different approaches including **Logistic Regression**, Random Forest, Gradient Boosting Classification and **XGBoost** algorithms.
- As this dataset consists of **imbalanced multiclass labels**, it was expected to have better F1 scores with algorithms including **weighted class distribution**
- ★ The best performing model was found as **RF with Decision Tree Regressor Iterative Imputer** with 53.3% F1 score
- F1 score improvement 50% → 53.3%

# Model Results



Learning Curve for RF model with Decision Tree Regressor Iterative Imputer



*It can be observed that there is overfitting with this model since there is a gap between train and test trend. However, we may say that they can converge closer if more data added.*

Classification Report

	Precision	Recall	F1-Score	Support
High	0.57	0.62	0.59	687
Low	0.41	0.36	0.39	251
Medium	0.54	0.52	0.53	721
accuracy			0.54	1659
macro avg	0.51	0.50	0.50	1659
weighted avg	0.53	0.54	0.53	1659

*When the class related F1-score is checked, it was found that Low class has the worst performance meaning that, the model has poorly performed classifying Low class compared to High and Medium classes.*



THANK YOU