ANALYTICS

Predicting App Ratings in the Google Play Store

An in-depth analytical approach
using machine learning

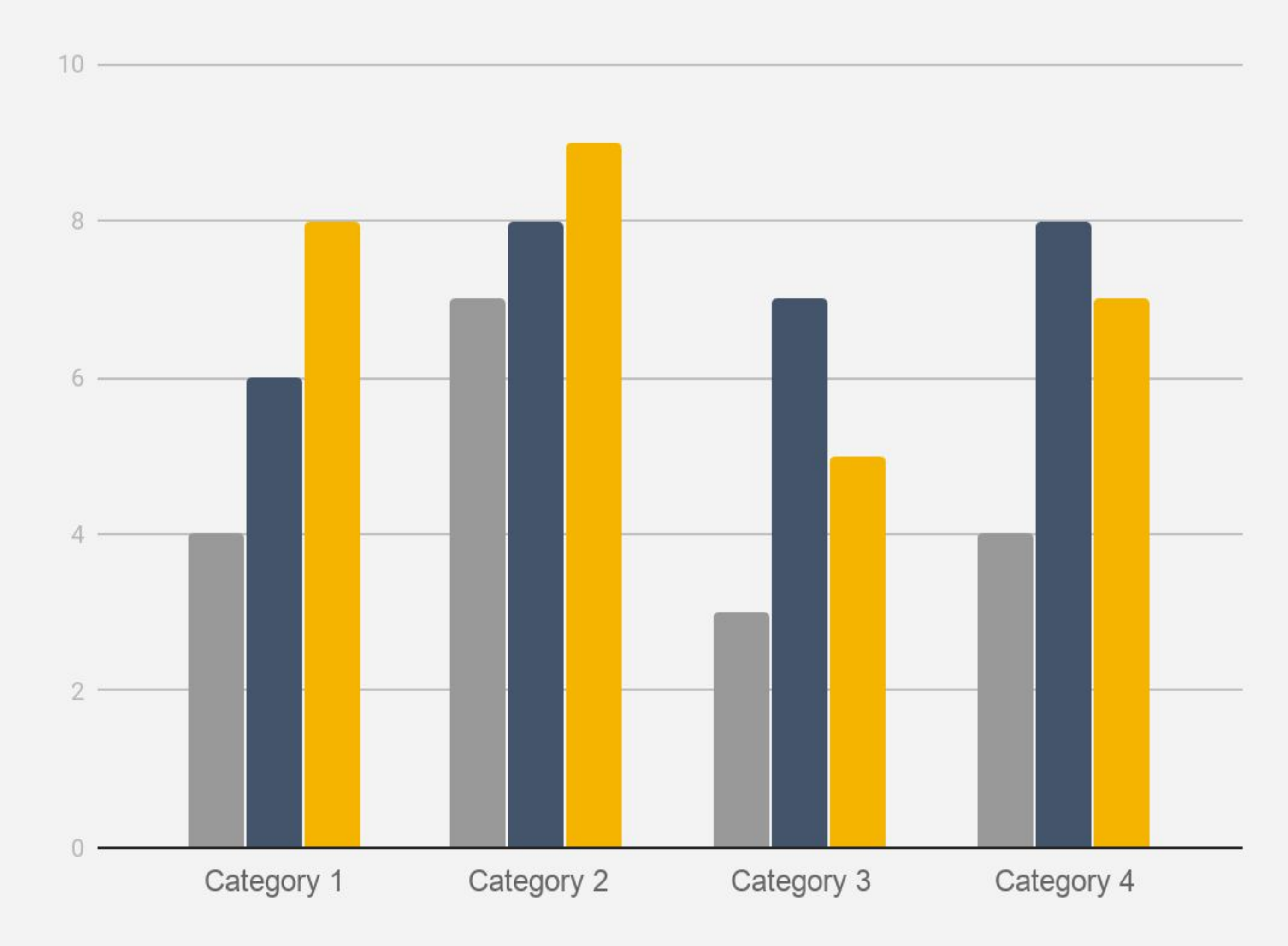
Group 8
Ryhan, Sharmi, Shalini
Humber IGS

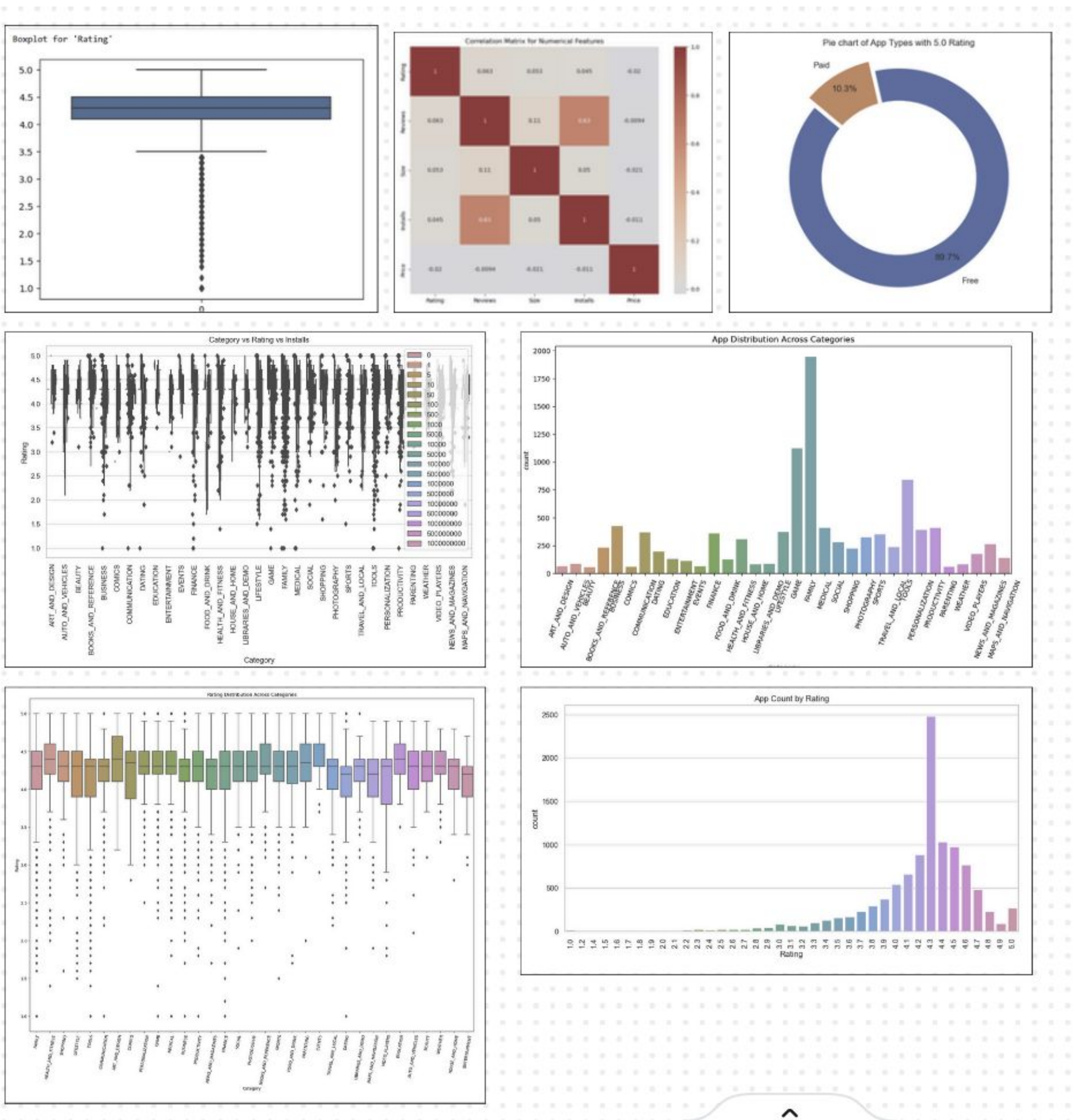
Problem Statement and Dataset Discussion

Google Play

PROBLEM STATEMENT:
How well will an app perform in the Google Play Store based on its reviews, size, and user demographics?

DATASET
Google Play Store Apps Data: Over 10,000 apps with features such as Category, Reviews, Size, Installs, Type, and Price.



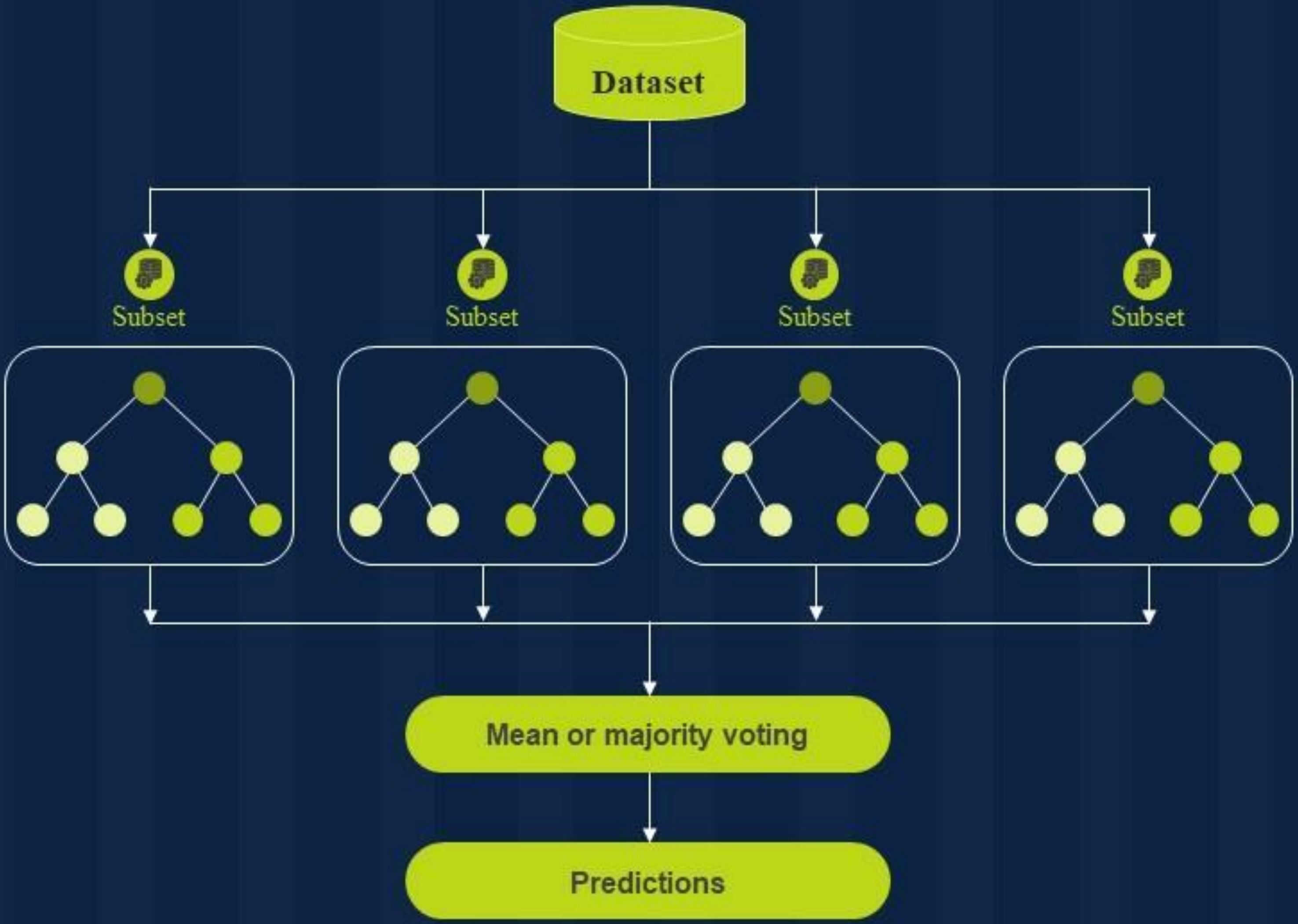


Statistical plots of the dataset

- **Data Preprocessing**: Addressed missing values, duplicates, and outliers to refine the dataset for accurate analysis.
- **Category Impact**: 'Games' lead in app categories, possibly influencing user ratings.
- **Rating Dynamics**: High ratings prevail, indicating user satisfaction is often reflected in their ratings.
- **Review-Rating Link**: Detected a positive correlation between review quantity and higher ratings, highlighting the importance of user engagement.
- **App Type Preference**: A striking 90% of apps rated 5 stars are free, suggesting a user bias towards free apps.
- **Significance of Updates**: Frequent app updates seem to be associated with better ratings, suggesting that consistent improvement may lead to higher user appreciation.

Random forest technique for classification model

This slide represents the random forest technique to implement a classification model that simultaneously works on individual subsets of sample data. It also includes its working and benefits, including multiple input handling, overfitting resistance, and so on.



Dataset

Subset

Subset

Subset

Subset

Mean or majority voting

Predictions

Working

- Properly categorize enormous amounts of data
- Employs bagging method that makes subsets of information from training data sets picked randomly with substitute
- Users can train on multiple subsets simultaneously by choosing them from extensive sample data
- Add text here

Benefits

- Manages many input parameters without removing any of them
- Offers effective strategies for predicting missing information
- Overfitting resistance
- Provides accuracy even when a significant data percentage is absent
- Determines beneficial traits for categorization
- Add text here

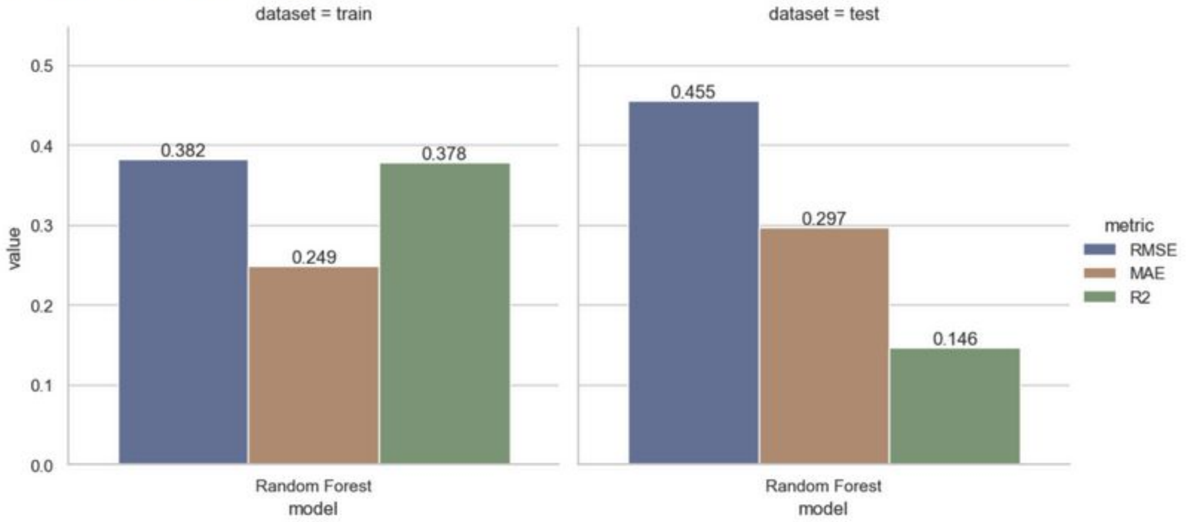
Proposed Predictive Model Justification

Why Random Forest for Our Prediction?

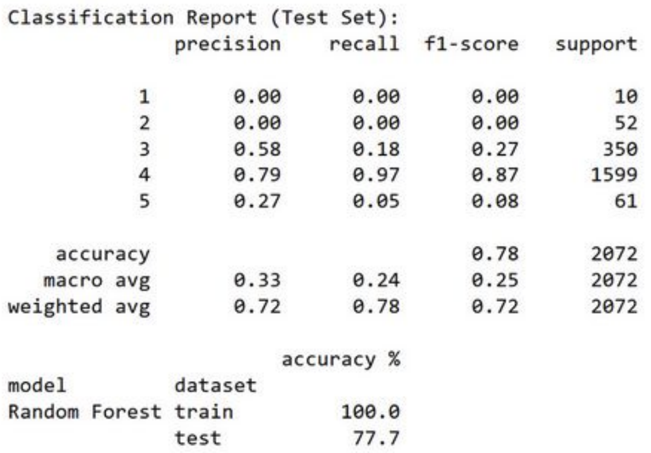
- Handles non-linear relationships in data, essential for diverse app attributes.
- Effectively reduces overfitting by averaging multiple decision trees.
- Random Forest's feature importance helps identify which app characteristics most influence ratings.

Model Performance

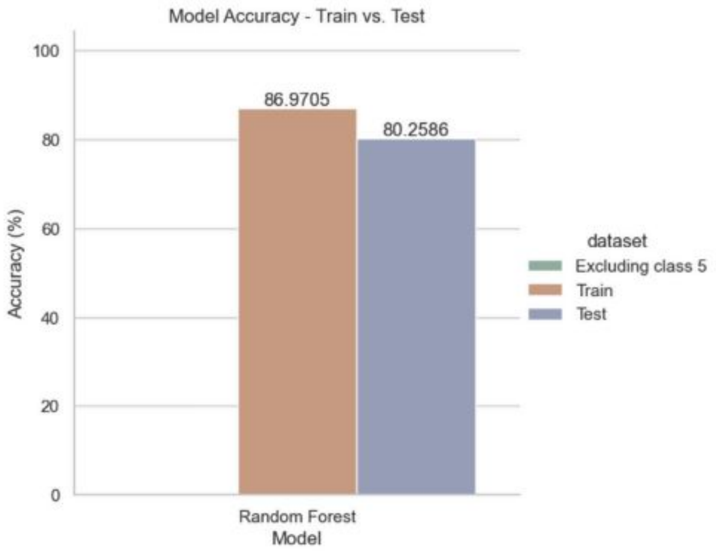
Random Forest Regressor:



Random Forest Classifier:

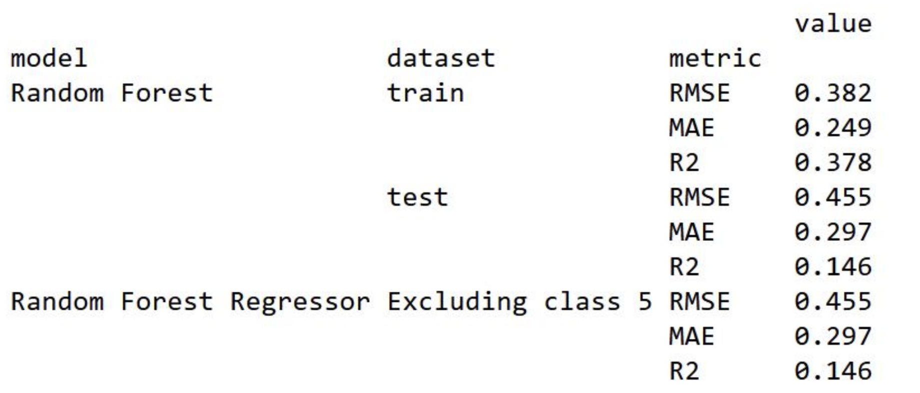


```
Classification Report (Test Set):
              precision    recall  f1-score   support

           1       0.00      0.00      0.00        10
           2       0.00      0.00      0.00        52
           3       0.58      0.18      0.27       350
           4       0.79      0.97      0.87      1599
           5       0.27      0.05      0.08        61

    accuracy                           0.78      2072
   macro avg       0.33      0.24      0.25      2072
weighted avg       0.72      0.78      0.72      2072

                      accuracy %
model         dataset
Random Forest train        100.0
              test          77.7
```



We tried both Random Forest Regressor and Random Forest Classifier to compare which model fits the best (regression vs classification).

Our conclusion was that the Random Forest Classifier performed far better than Random Forest Regressor, achieving an 80.3% accuracy



```
                                                    value
model                   dataset           metric
Random Forest           train             RMSE      0.382
                                          MAE       0.249
                                          R2        0.378
                        test              RMSE      0.455
                                          MAE       0.297
                                          R2        0.146
Random Forest Regressor Excluding class 5 RMSE      0.455
                                          MAE       0.297
                                          R2        0.146
```

Discussion and Takeaways

100%

ACCURACY ON TRAINING SET

77.7%

ACCURACY ON TESTING SET WITHOUT HYPERPARAMETER TUNING

80.3%

ACCURACY ON TESTING SET WITH HYPERPARAMETER TUNING

- **Interpretation**: Reviews are a moderate predictor of ratings; other features like Size and Price are less influential.
- **Implications**: Developers should focus on garnering reviews to improve app ratings.

THANK YOU