Exploratory Data Analysis

Data Science Persistency of a Drug Project

DS HealthCare Group Members:

- Yuchi Chen
- Bolutife Akinlawon
- Alexis Collier
- Han-Fu Lin
- Runtian Wang

Agenda

- Data Analysis Approach
- Problem Understanding
- Business Understanding
- Data Preparation
- Exploratory Data Analysis
- EDA Summary
- Model Building
- Model Selection
- Final Recommendation
- Conclusion

Data Analysis Approach

One of the challenges for all Pharmaceutical companies is understanding the persistency of drugs as per the physician's prescription.

To solve this problem, ABC pharma company is seeking to automate this process of identification by:

- Explore and Understand the data
- Prepare and clean the data
- Analyze the data and find the features/variables that affect drug persistency
- Give recommendations for the classification model that is to be built to automate the process of drug persistency identification

Problem Understanding

ABC Pharma is a pharmaceutical company that aims to automate the identification of the persistence of a drug.

In this data analysis, the various factors of a patient are evaluated to determine if they affect persistency. The data is collected for analysis to determine the durability of the drug.

At the end of this project, we will suggest a model follow for deployment.

BUSINESS UNDERSTANDING

In summary, the task can be represented as follows:

DATA PREPARATION

- Python was utilized for data preparation, as well as pandas library specifically.
- Data was cleaned and prepared for analyzation.
  - Method of Approach:
    - Look for null or missing values.
    - Identify values that are improbable.
    - Create visualizations of the data.

EXPLORATORY DATA ANALYSIS

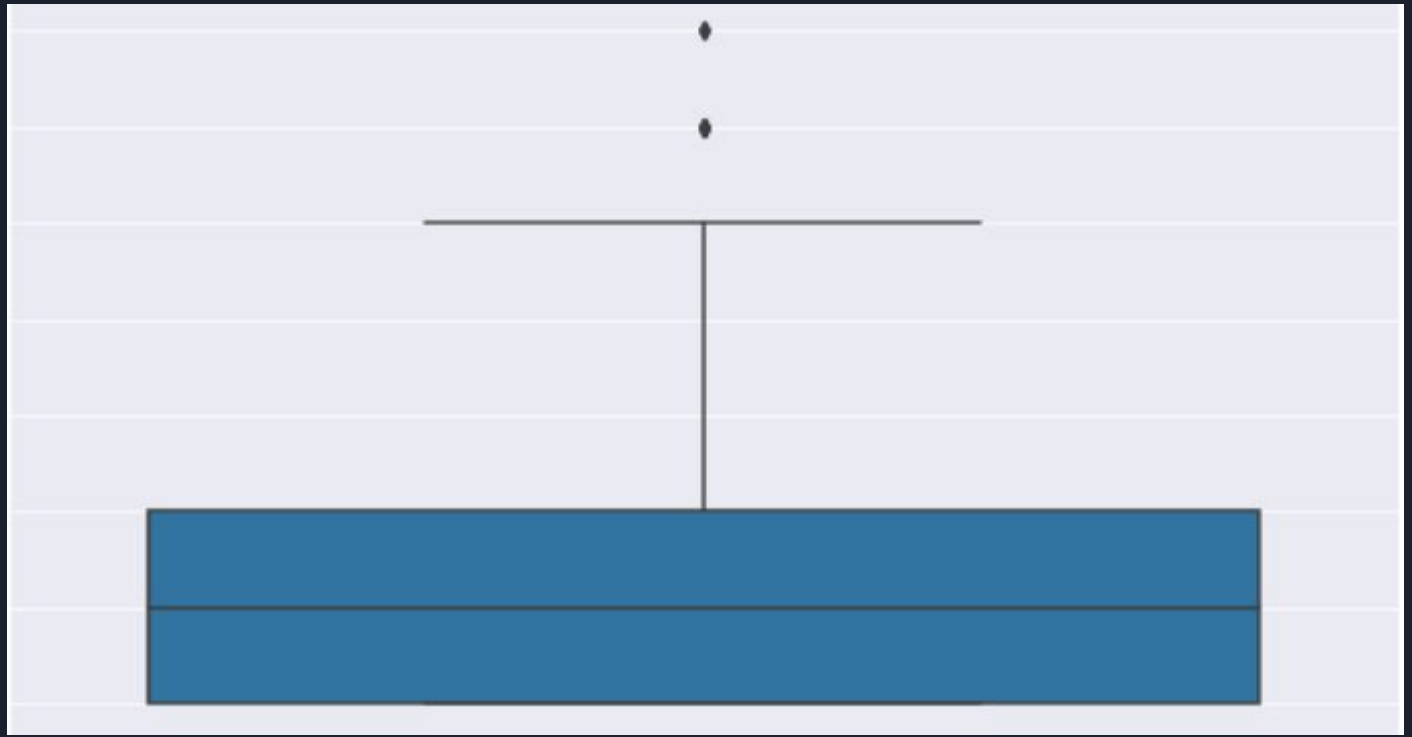
- We analyzed the demographic data characteristics, including gender, race, ethnicity, region, age bracket, and iodine indicator.

- The data variables are represented via stacked bars and correlation heatmaps.

EXPLORATORY DATA ANALYSIS

- Some demographic observations include the following:
  - More significant number of female participants.
  - There are more persistent than non-persistent participants, indicating an imbalanced dataset.
  - The primary racial demographic of the dataset is caucasian (non-Hispanic).
  - Most of the participants are 75 and older.
  - ‘Below 55’ age bracket is the least represented age demographic.

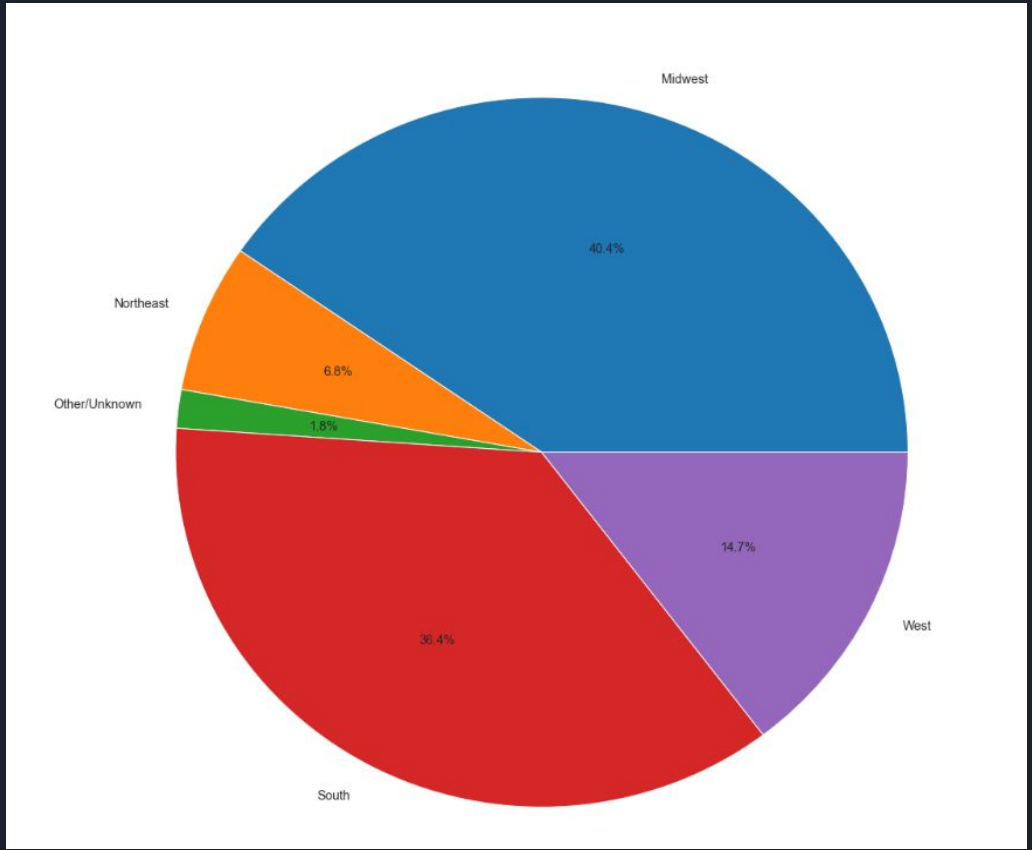
Outlier Test



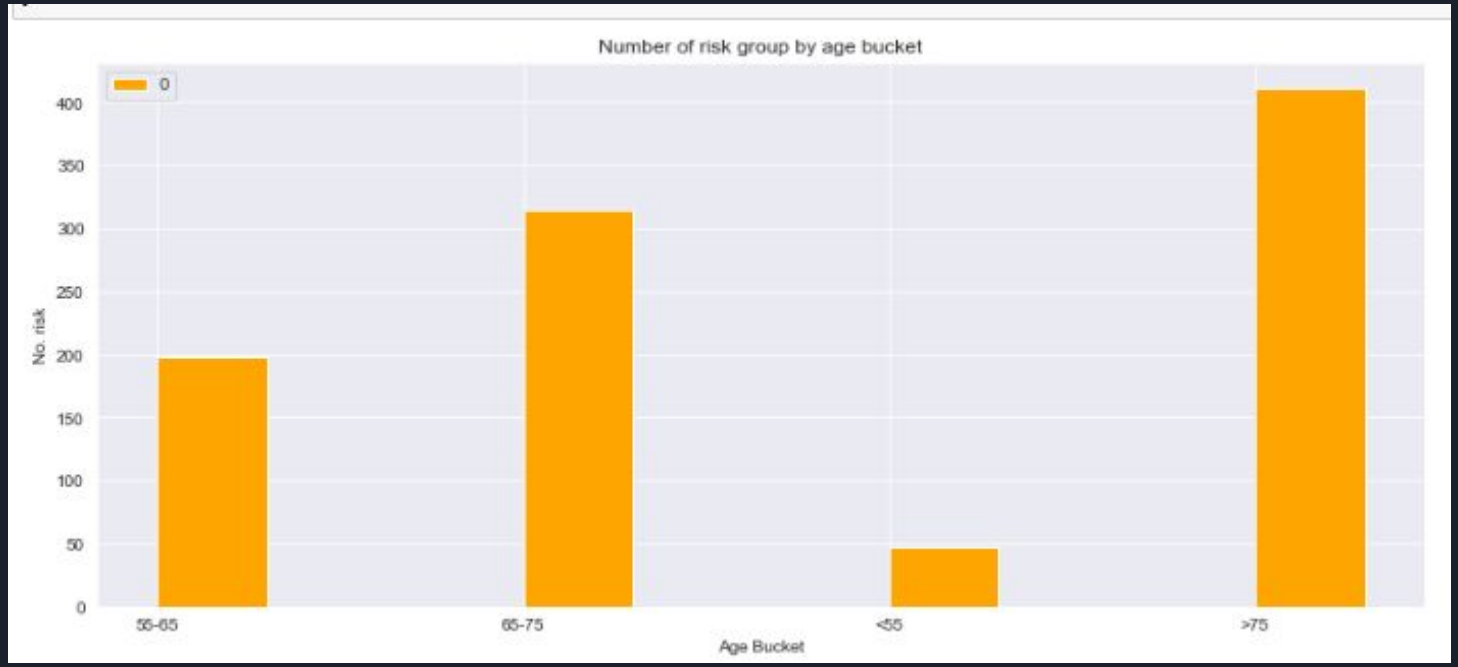
Heatmap

```
plt.figure(figsize=(20,10))
sns.heatmap(df.corr(),annot=True,vmin=-1,vmax=1,cmap='magma')
plt.show()
```

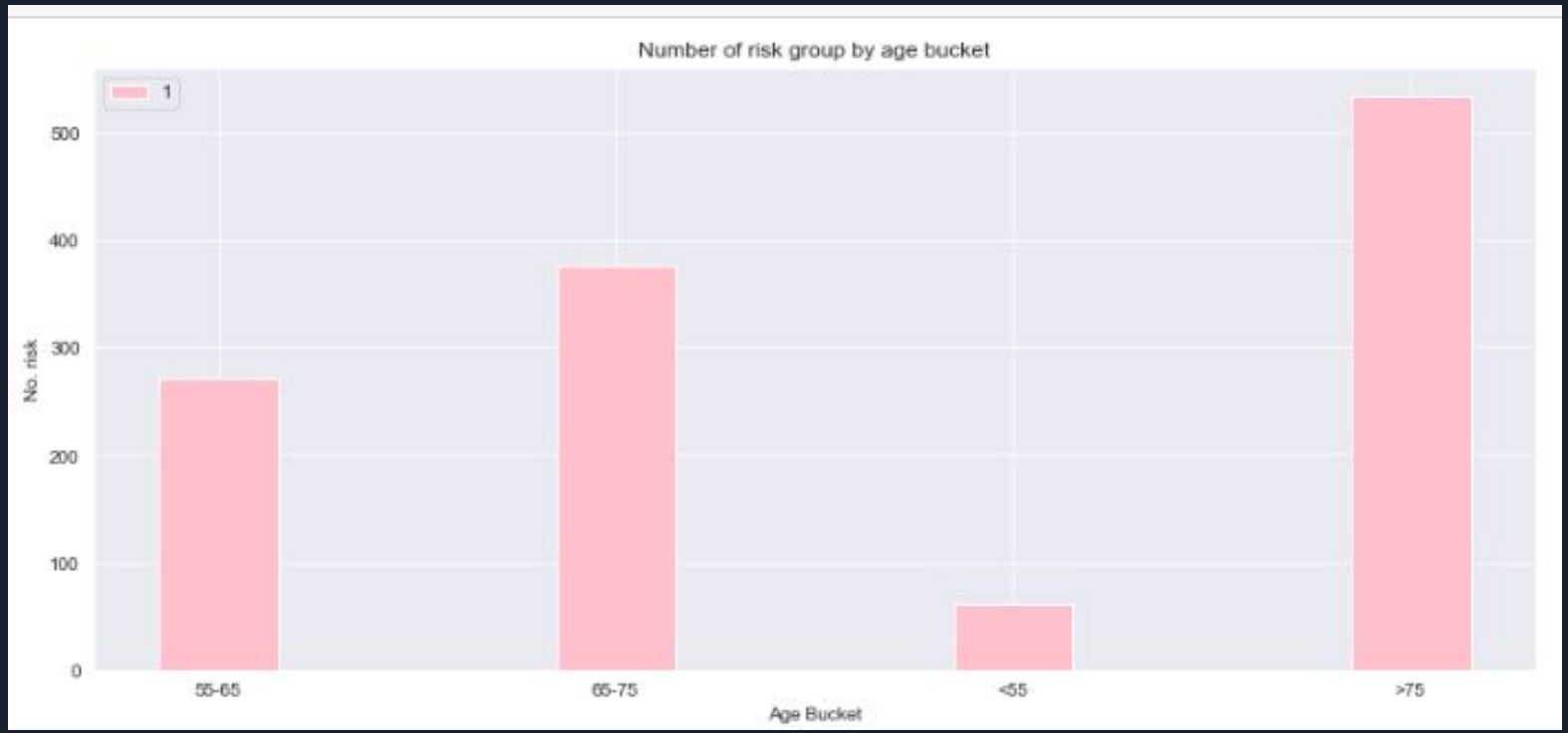
Data Amount by Region



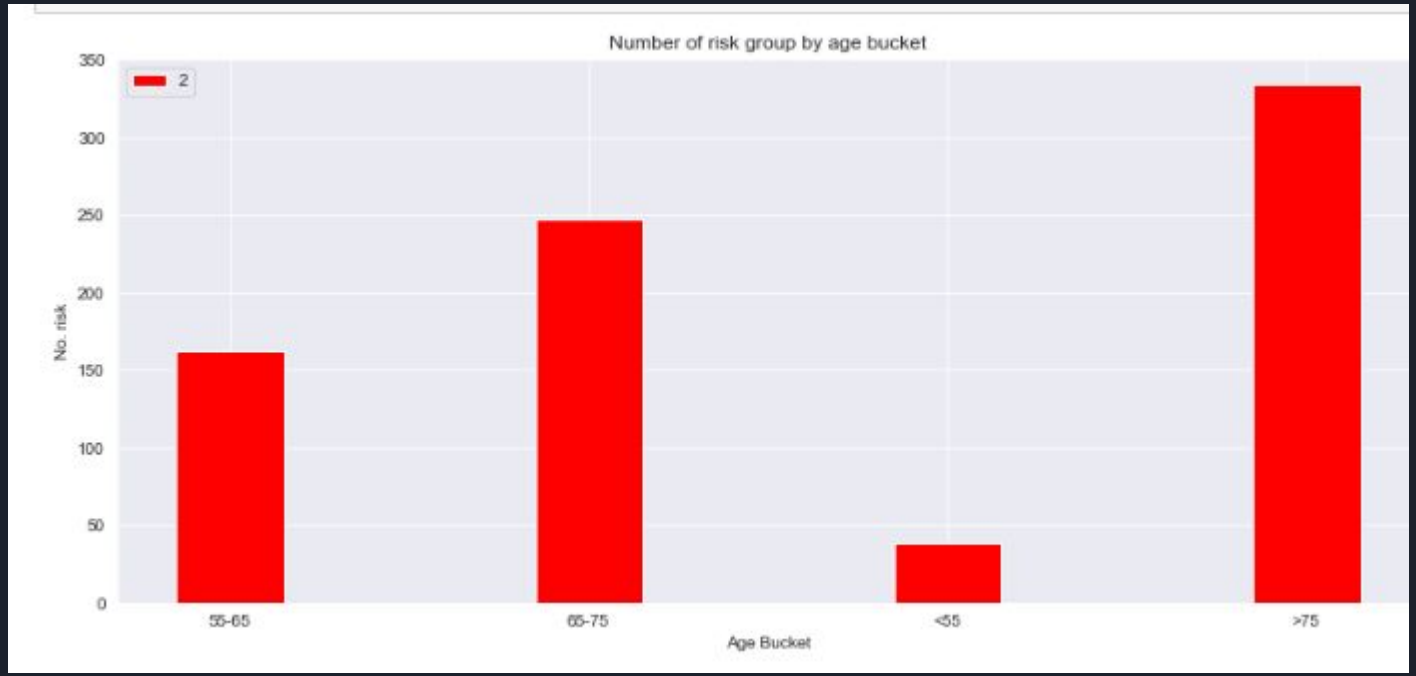
Risk Count by Different Age Group



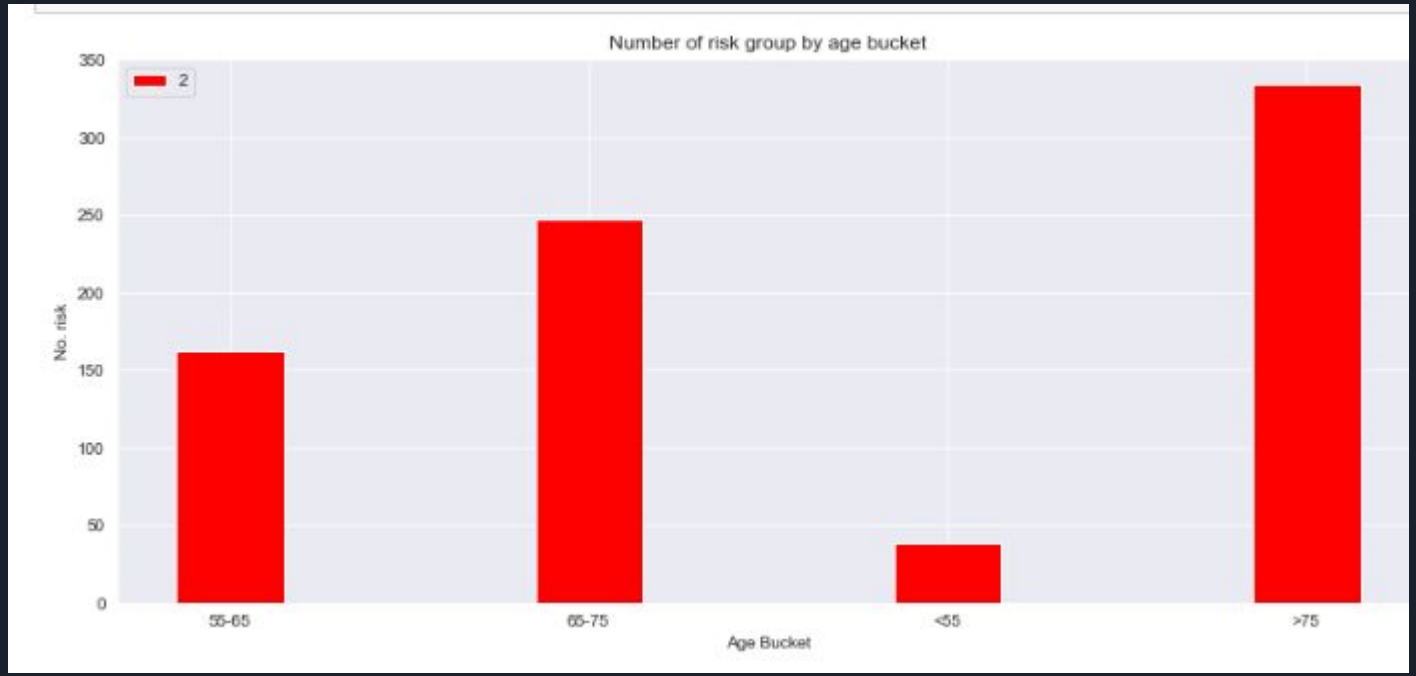
Number of Risk Group by Age Bucket



As a result, the count of risk doesn’t show much difference in different age group.



As a result, the count of risk doesn’t show much difference in different age group.



EDA Summary

Out of the many factors that may affect a patient's persistency to a drug, having risks, being of a certain race, ethnicity, age group, and the specialty of the HCP has the strongest effect on a patient's persistency towards a drug.

Now that the factors have been identified, we can use a model to create an automated process that predicts whether a drug will be persistent on a patient or not. The random forest classifier model would be the best one to use in this scenario for predicting whether or not a patient would have persistency with a drug.

Model Building

In the model-building stage, various regression techniques were used to classify the Drug Persistency of subjects based on the predictor variables. After preparing the features of the machine learning model to predict data by transforming categorical data into numbers, it is then time for the model-building step.

We used two different regression techniques:

- Decision Tree Classifier
- Random Forest Classifier

Model Selection

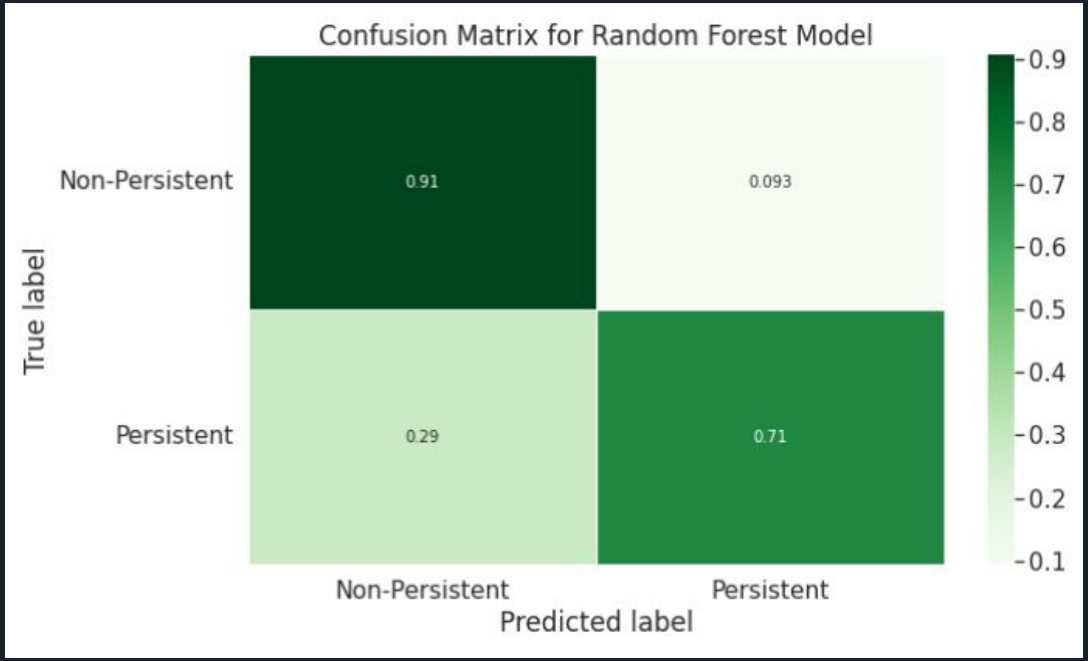
To select the best model, we looked at the performance metrics used, which are: Accuracy, Precision, Recall, F1-Score, Support, and AUC.

The best model is Random Forest Classifier.

```
Confusion matrix:
[[379  39]
 [ 77 190]]
```

	precision	recall	f1-score	support
Non-Persistent	0.83	0.91	0.87	418
Persistent	0.83	0.71	0.77	267
accuracy			0.83	685
macro avg	0.83	0.81	0.82	685
weighted avg	0.83	0.83	0.83	685

Model Selection



Final Recommendation

Random Forest Classifier is the best fit model for our dataset, with an accuracy score of 0.83.

From various Classification models, Random Forest Classifier was chosen based on the different parameter values:

- It has an accuracy of 83
- This could be the efficient model for the automation of the prediction of persistent or non-persistent drugs

Thank You