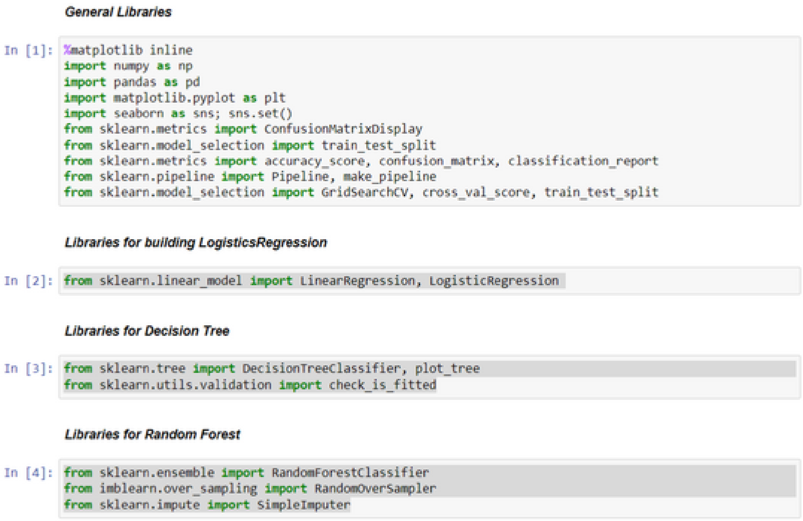
FIRST ASSIGNMENT - BINARY CLASSIFICATION

An introduction Section explaining the development process

The libraries used.

The libraries used for a this data classification project include;

1. NumPy and Pandas for data manipulation,
2. Matplotlib and Seaborn for data visualization,
3. Scikit-learn for machine learning tasks such as data preprocessing, model selection, and evaluation, Imblearn for handling imbalanced data, GridSearchCV is used for hyperparameter tuning. ConfusionMatrixDisplay is used for visualizing the confusion matrix and SimpleImputer for imputing missing values.



General Libraries

```
In [1]: %matplotlib inline
        import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns; sns.set()
        from sklearn.metrics import ConfusionMatrixDisplay
        from sklearn.model_selection import train_test_split
        from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
        from sklearn.pipeline import Pipeline, make_pipeline
        from sklearn.model_selection import GridSearchCV, cross_val_score, train_test_split
```

Libraries for building LogisticsRegression

```
In [2]: from sklearn.linear_model import LinearRegression, LogisticRegression
```

Libraries for Decision Tree

```
In [3]: from sklearn.tree import DecisionTreeClassifier, plot_tree
        from sklearn.utils.validation import check_is_fitted
```

Libraries for Random Forest

```
In [4]: from sklearn.ensemble import RandomForestClassifier
        from imblearn.over_sampling import RandomOverSampler
        from sklearn.impute import SimpleImputer
```

The classification methods used, and how hyperparameters for each classification method selected

The models used include Logistic Regression, Decision Tree Classifier, and Random Forest Classifier,

1. For Logistics Regression, a GridSearchCV method is used to perform K-Fold Cross Validation with k=5 to select the best hyperparameters for the model. The parameter grid includes values for the regularization strength C and the penalty function.
2. The Decision Tree model uses a range of values for the maximum depth hyperparameter, and the training and validation accuracy scores are plotted to select the best value.
3. For the Random Forest model, the dataset is first resampled using the RandomOverSampler method, and GridSearchCV is used to select the best hyperparameters for the SimpleImputer and RandomForestClassifier. The best hyperparameters are then used to train the model.

The training and testing process

The training and testing process varied for each of the classification methods used.

- For Logistic Regression, the hyperparameters C and penalty were selected using GridSearchCV with a parameter grid of possible values. The dataset was split into train and test sets for this model.
- For Decision Tree, the maximum depth hyperparameter was selected using a line chart that plotted the training and validation accuracy scores against different max depth values. The dataset was split into train, validation, and test sets in a ratio of 60:20:20 for this model.

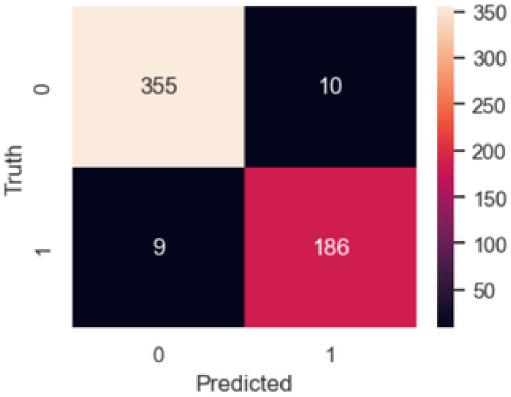
- For Random Forest, the hyperparameters were selected using GridSearchCV with a parameter grid of possible values. Random oversampling was also used to resample the dataset. The dataset was splited into train and test sets, and the accuracy scores were obtained using cross-validation.

Overall, the models were trained and tested using a variety of techniques to select the best hyperparameters and improve accuracy.

An evaluation section describing and explaining your results

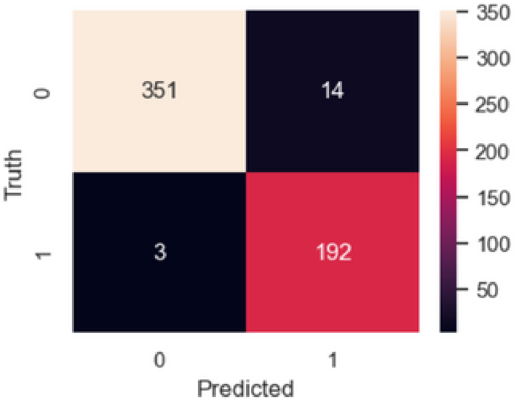
Confusion matrix of the best version of each classification method

- Logistics Regression



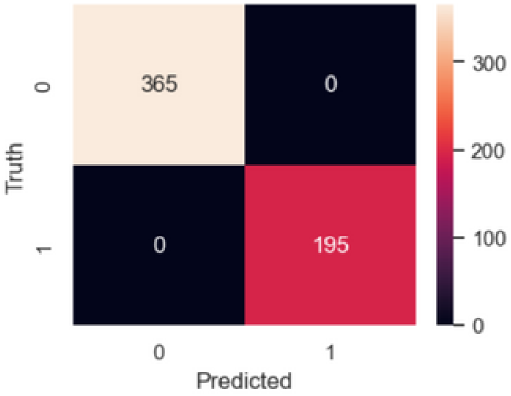
The confusion matrix shows the performance of a best Logistics Regression classifier, where 355 instances were correctly classified as negative and 186 instances were correctly classified as positive, while 10 negative instance was wrongly classified as positive and 9 positive instances were wrongly classified as negative.

- Decision Three



The confusion matrix shows the performance of a best Decision Tree classifier, where 351 instances were correctly classified as negative and 192 instances were correctly classified as positive, while 14 negative instance was wrongly classified as positive and 3 positive instances were wrongly classified as negative.

- Random Forest



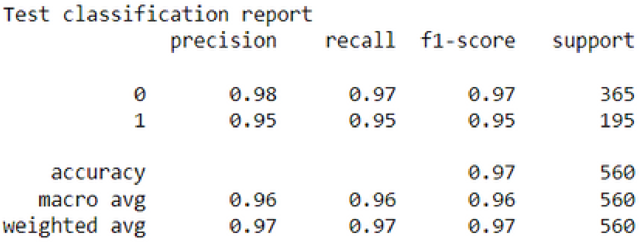
The confusion matrix shows the performance of a best Random Forest classifier, where 365 instances were correctly classified as negative and 195 instances were correctly classified as positive, while 0 negative instance was wrongly classified as positive and 0 positive instances were wrongly classified as negative.

It appears that the Random Forest classifier performed the best with no instances being wrongly classified. However, it's important to consider other metrics such as precision, recall, and F1-score

Classification Report (precision, recall, and F1-score)

- **Logistics Regression**

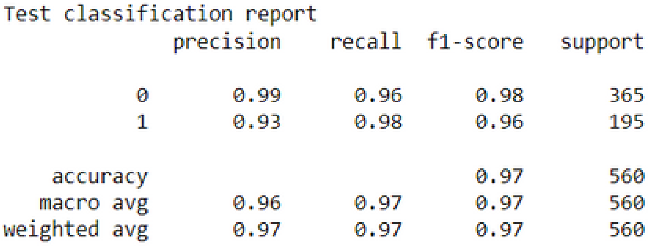
  The Logistics Regression model has an overall accuracy of 0.97, and the weighted average precision, recall, and f1-score are 0.97. The macro-averaged precision, recall, and f1-score are 0.96, and the support for class 0 and class 1 is 365 and 195, respectively.



```
Test classification report
              precision    recall  f1-score   support

           0       0.98      0.97      0.97       365
           1       0.95      0.95      0.95       195

    accuracy                           0.97       560
   macro avg       0.96      0.96      0.96       560
weighted avg       0.97      0.97      0.97       560
```

- **Decision Three**

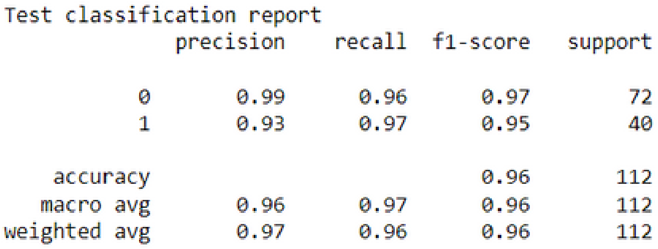
  The accuracy of the Decision Tree model on is 0.97. Class 0 has a precision of 0.99 and recall of 0.96, while class 1 has a precision of 0.93 and recall of 0.98. The weighted average F1-score is 0.97, indicating good overall performance. The macro-average F1-score is 0.97, which takes into account class imbalance.



```
Test classification report
              precision    recall  f1-score   support

           0       0.99      0.96      0.98       365
           1       0.93      0.98      0.96       195

    accuracy                           0.97       560
   macro avg       0.96      0.97      0.97       560
weighted avg       0.97      0.97      0.97       560
```

- **Random Forest**

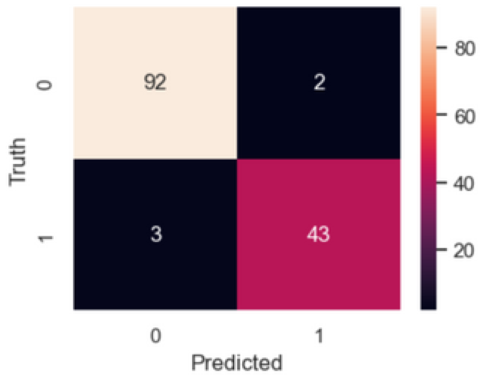
  The Random Forest model achieved a perfect accuracy of 1.00 on the training set with precision, recall, and f1-score of 1.00 for both the positive and negative classes.



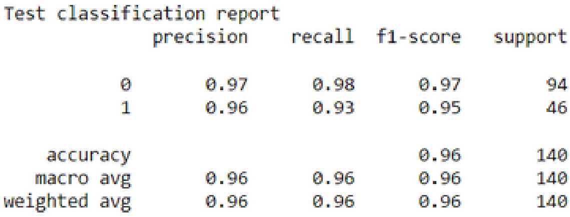
```
Test classification report
              precision    recall  f1-score   support

           0       0.99      0.96      0.97        72
           1       0.93      0.97      0.95        40

    accuracy                           0.96       112
   macro avg       0.96      0.97      0.96       112
weighted avg       0.97      0.96      0.96       112
```

Final conclusions

Based on the evaluation metrics, the Random Forest model appears to be the best model to adopt as it achieved perfect accuracy and F1-scores on both classes in the test classification report. This indicates that the model correctly classified all instances in the test set. Therefore, it can be concluded that the Random Forest model has the highest predictive power among the models evaluated. We then fit the random forest modet to yjr test dataset



The random forest classifier achieved an accuracy of 0.96 on the test set, with precision and recall of 0.97 and 0.98 for class 0, and 0.96 and 0.93 for class 1, respectively. The f1-score was 0.97 for class 0 and 0.95 for class 1.



```
Test classification report
              precision    recall  f1-score   support

           0       0.97      0.98      0.97        94
           1       0.96      0.93      0.95        46

    accuracy                           0.96       140
   macro avg       0.96      0.96      0.96       140
weighted avg       0.96      0.96      0.96       140
```

The Random Forest model achieved an accuracy of 97% on the test set, correctly classifying 92 negative instances and 43 positive instances, with 2 false negatives and 3 false positives.