Breast Cancer Prediction using Machine Learning

EFFORTS BY:

ARYAMAN AGARWAL	-	16BCB0025
APURV SHRIKANT TODKAR	-	16BCB0048
VAASU GUPTA	-	16BCB0062
KESAV REDDY	-	16BCB0100
SIDDHARTH SHAILENDRA	-	16BCB0129

Introduction

- Machine learning is branch of Data Science which fuses an expansive arrangement of statistical strategies.
- These methods empower data scientists to make a model which can learn from past information and recognize patterns from huge, complex and noisy datasets
- Researchers use machine learning for disease expectation and visualization.
- Machine learning permits inferences or choices that generally can't be made utilizing regular measurable approaches.
- With vigorously validated machine learning model, odds of right analysis improve.
- It specially helps in interpretation of results for marginal cases.

Breast Cancer: A brief introduction

- The most common cancer in women around the world.
- The standard reason for death from cancer among women around the world.
- Early discovery is the best method to diminish deaths related to breast cancer.
- Early analysis requires an exact and dependable system to recognize between benign breast tumors from malignant ones
- Breast Cancer Types - three types of breast tumors:
  - Benign breast tumors
  - In-situ cancers
  - Invasive cancers
- most of breast tumors detected by mammography are benign.

Breast Cancer: A brief introduction (continued)

- They are non-cancerous growths and can't spread outside of the breast to different organs.
- At times, it is hard to recognize certain benign masses from malignant lesions with mammography.
- If the malignant cells have not gone through the basal membrane but is completely contained in the lobule or the ducts, the cancer is called in-situ or noninvasive.
- If the cancer has gotten through the basal membrane and spread into the encompassing tissue, it is called invasive.
- This analysis assists in differentiating between benign and malignant tumors.

Data Source

- https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
- The data used is from University of Wisconsin.
- This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg.

Literature Review

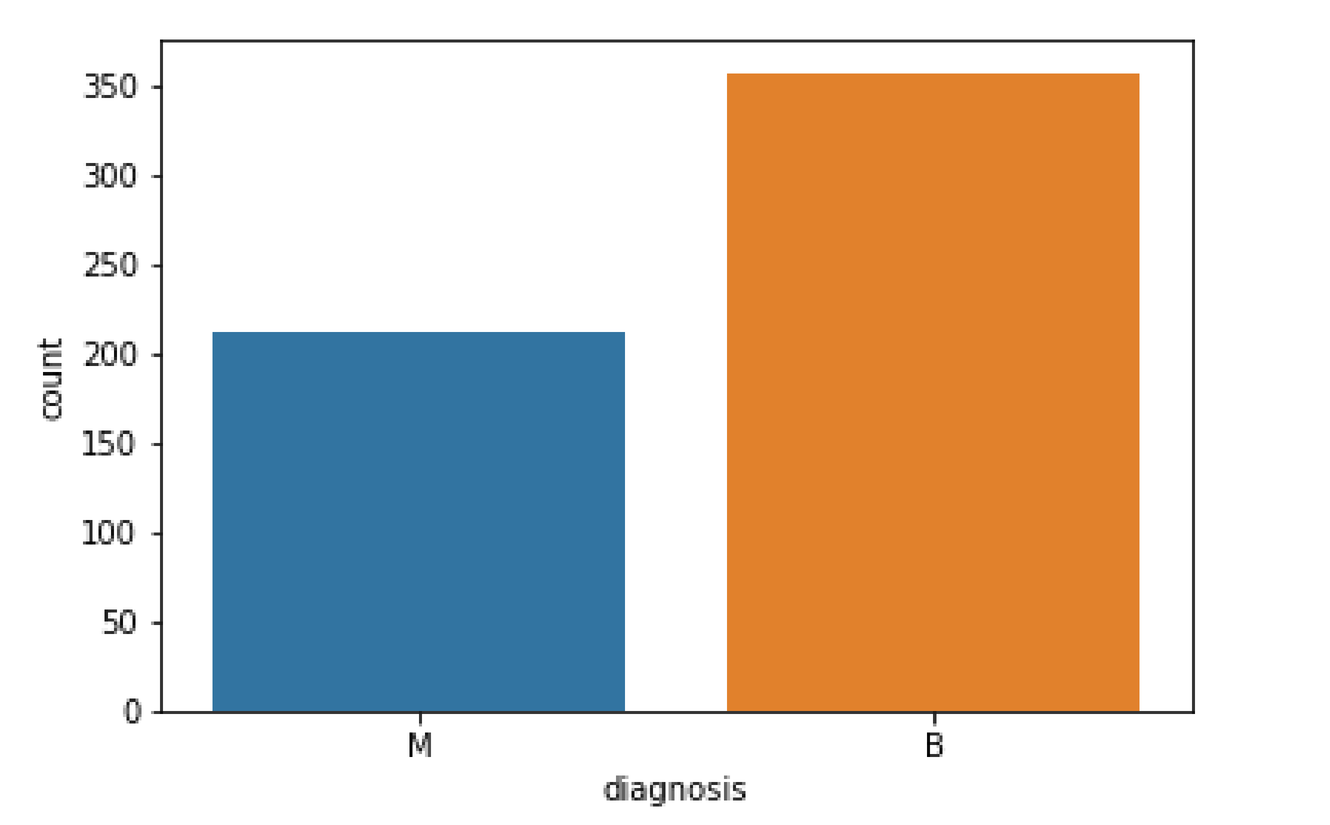
- O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.
- William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.
- O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.
- K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).

Flow of data

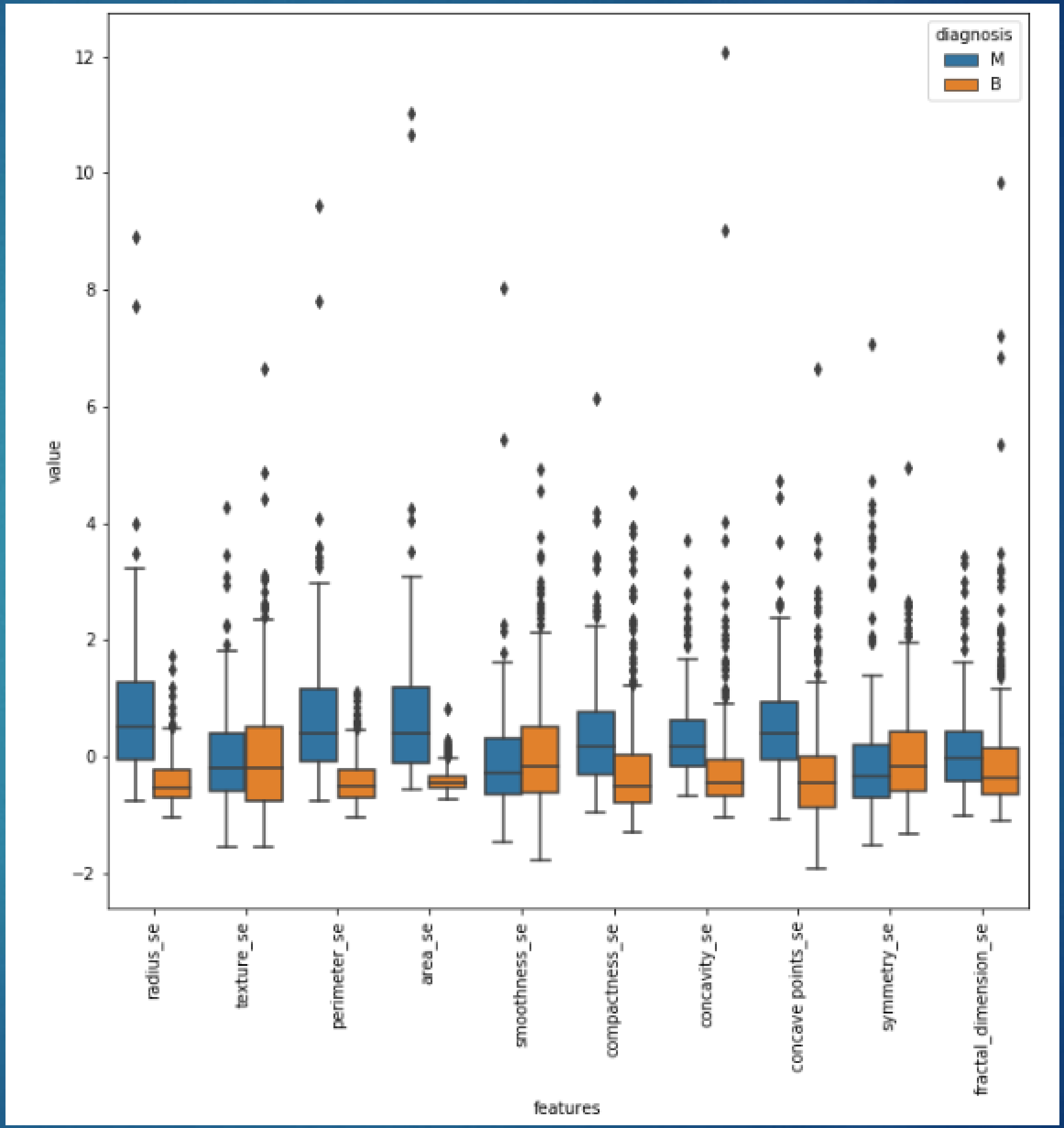

Biopsy Procedure
Measurements
Reports
Evaluation
Diagnosis
Analysis of measurements
Preparation of ML Models
Predictions and validation

Initial distribution of Data before split

- Number of Benign: 357
- Number of Malignant : 212



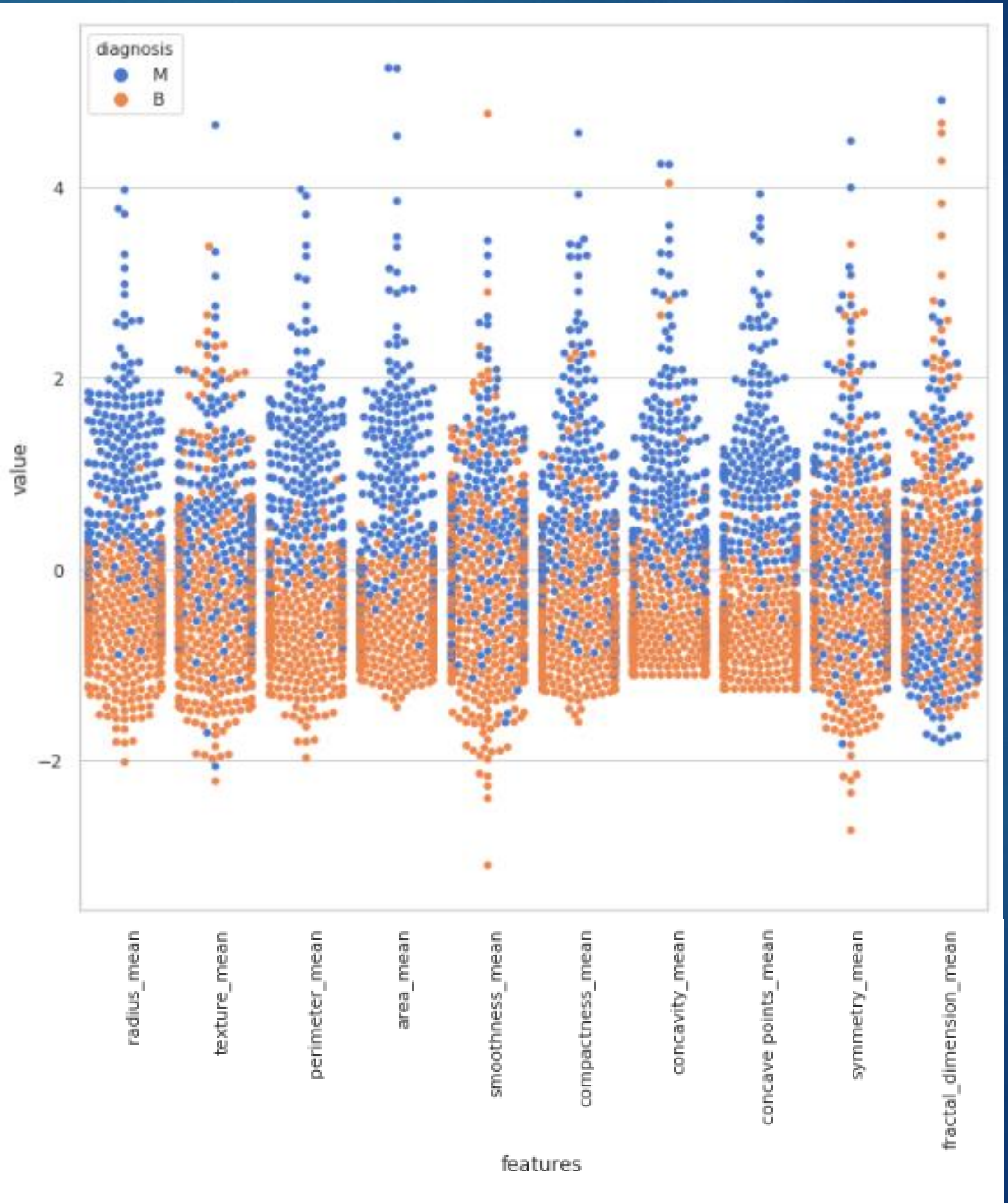
A boxplot of all the features



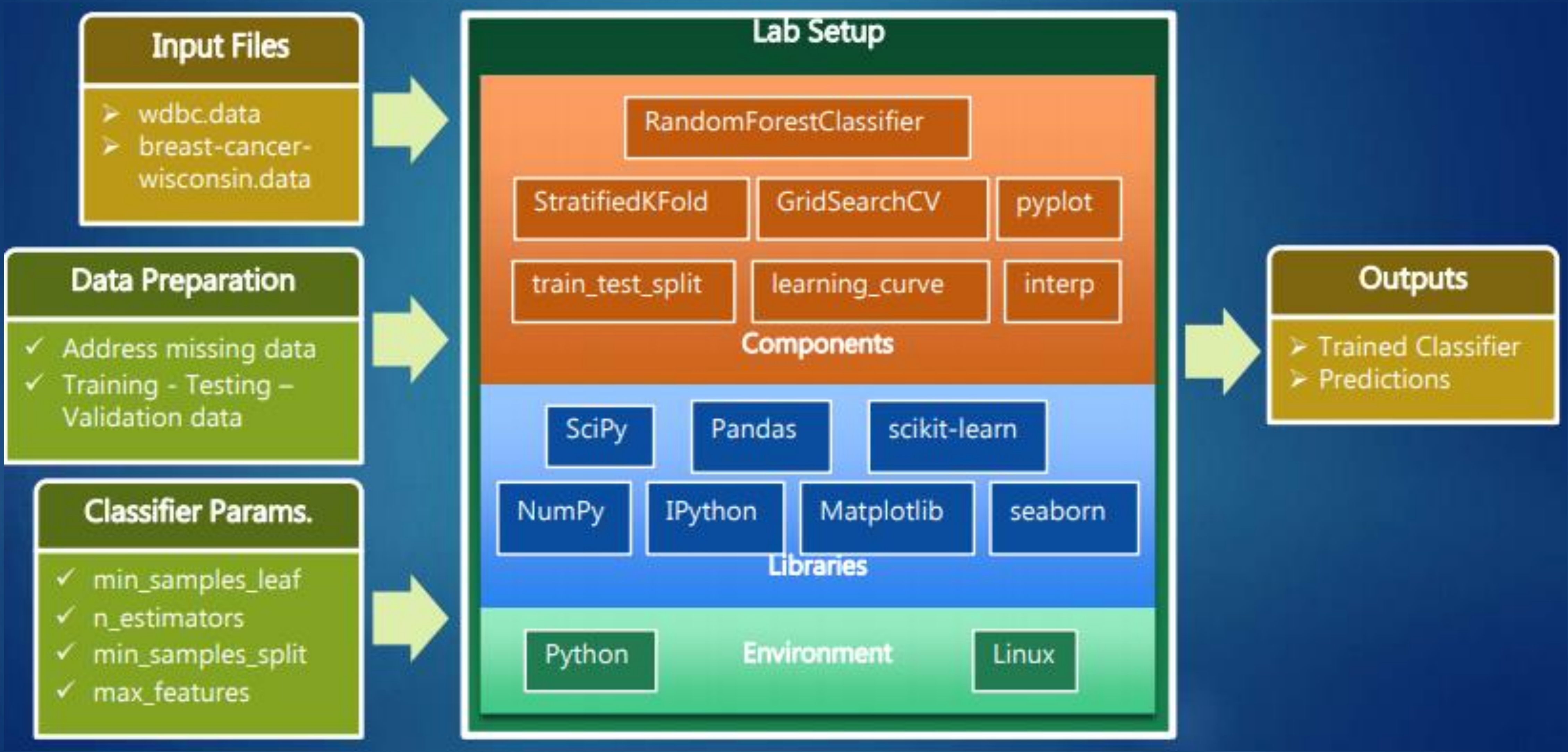
- ➢ Here the Blue colour denotes Malignant.
- ➢ Orange colour denotes Benign

Swarm plot of the data

- It is useful for getting a categorical scatterplot with non-overlapping points.
- This gives a better representation of the distribution of values.



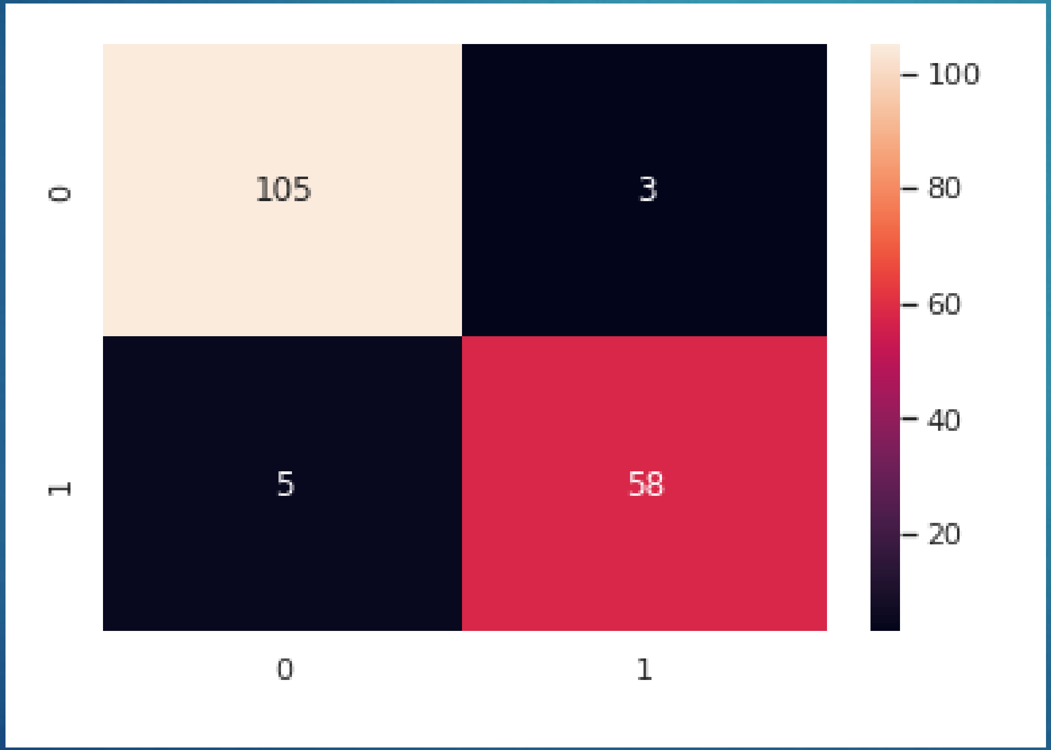
Method



Input Files
wdbc.data
breast-cancer-wisconsin.data
Data Preparation
Address missing data
Training - Testing – Validation data
Classifier Params.
min_samples_leaf
n_estimators
min_samples_split
max_features
Lab Setup
RandomForestClassifier
StratifiedKFold
GridSearchCV
pyplot
train_test_split
learning_curve
interp
Components
SciPy
Pandas
scikit-learn
NumPy
IPython
Matplotlib
seaborn
Libraries
Python
Environment
Linux
Outputs
Trained Classifier
Predictions

Results

- Data is divided into training and test set.
- Test data has a total of 171 records.
- Accuracy score is : 0.9532163742690059
- Confusion matrix is as follows :



Conclusion

- We were successfully able to train a model on Wisconsin dataset for breast cancer.
- We used logistic regression in our model.
- We used K-fold cross validation.
- Random forest classifier was also used.
- Decision Tree classifier was also used.
- Using all of the above tools we built a pipeline which helped us in achieving a high accuracy level of roughly 95%.
- For future work we could extend the classification using more concurrent features thus achieving even greater accuracy in our predictions.

Thank You