HURU SCHOOL

DIANA NABATURU MUTEKHELE

STUDENT ID: DS2021048867

DATA SCIENCE PROJECT REPORT

BREAST CANCER PREDICTION USING MACHINE LEARNING

1.Background

The common type of cancer and the second highest in terms of mortality rate worldwide is breast cancer with 1.67 million new cases diagnosed in 2012 and approximately 500000 annual deaths. In global statistics it represents majority of new cancers and cancers related deaths which makes it a significant public health problem in todays' society. It occurs mostly in women, and it starts when cells begin to grow out of control. This cancer begins in the ducts that carry milk to the nipple (ductal cancers). When the cancer cells get into the blood or lymph system, they are carried to other body parts.

Diagnosis is performed by self-examination or x-ray. It can also be diagnosed using Fine Needle Aspiration Test. This is a type of biopsy whereby cells are obtained from the lump or mass using a needle. Doctors then examine them under a microscope to determine if it is malignant(cancerous) or benign(non-cancerous). Not all lumps found in the body are cancerous. According to a recent study, in women younger than 40, 80 to 85 percent of breast lumps are benign and non-cancerous. The best cancer screening test is mammography. It detects cancer up to 2 years before a tumour is felt.

Risk factors for breast cancer

Age-breast cancer is mostly found in women over the age of 50 which means as women age their risk of getting cancer increases.
Breast cancer history- one is likely to get cancer if she once had cancer.
Family history of breast cancer- there is a high risk of getting breast cancer if you have relatives with breast cancer especially before the age of 40.
Genetic factors- genetic mutations in BRCA1 and BRCA2 genes can lead to breast cancer.
Childbearing and menstrual history- there is a high risk of getting breast cancer when a woman bears a child at an older age. Other factors include early menses (before 12) and late menopause (after 55) in women. Not having children also increases chances of getting breast cancer.

2.Problem Statement

In high income countries, more than 70% of breast cancer patients are diagnosed in stages I and II. In low- and middle-income countries 20%-50%of patients are diagnosed in these stages. Mortality rates are higher in developing countries. Studies show there is an association between the advanced and delays in getting treatment. To improve prognosis and chances of survival, early diagnosis and timely clinical treatment to patients is important. Accurate classification of tumours to either benign or malignant is important to avoid undergoing unnecessary treatment. Early detection will also help one to know the type of cancer one has(***invasive lobular carcinoma or Invasive ductal carcinoma***).This subject has led to a lot of research. **Machine learning** is a widely used methodology in breast cancer classification and forecast modelling because of critical feature detection in complex datasets. I will use machine learning classification methods to fit a function that can predict the discrete class of new input.

2.1 Dataset Description

Dataset is collected from *https://www.kaggle.com/uciml/breast-cancer-wisconsin-data*

This breast cancer dataset was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. It contains 569 samples of malignant and benign tumor cells. The first two columns in the dataset store the unique ID numbers of the samples and the corresponding diagnosis (M=malignant, B=benign), respectively. The columns 3-32 contain 30 real-value features that have been computed from digitized images of the cell nuclei, which can be used to build a model to predict whether a tumor is benign or malignant.
Column names description:
1) ID number 2) Diagnosis (M = malignant, B = benign) 3-32)
Ten real-valued features are computed for each cell nucleus:
a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)
The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features.
All feature values are recoded with four significant digits.
Missing attribute values: none
Class distribution: 357 benign, 212 malignant

2.2 Goals for the Analysis

To create a predictive model to determine if cancer is benign(non-cancerous) or malignant(cancerous) and view unique trends that may help us in model selection and hyper parameter selection.

3.Methodology

Python programming is used to clean, analyse, visualise data and create predictive model.
Analysis is done in 4 Phases

3.1 Phase 1- Exploratory Data Analysis

Data-pre-processing-This involves loading dataset to python using pandas and creating a folder called Breast_cancer where the dataset is stored. Inspection of the data is done by observing the columns and the head. It also involves checking for null values, missing values and outliers.

The findings were, the dataset has 569 rows and 32 columns. All dataset types are floats except 2 which were integer for id column and object for diagnosis column. The diagnosis count is 357 for “B” and 212 for “M”. There are no missing nor null values.

Dataframe.describe() is used summarise measures of central tendency and dispersion for the columns. Data is visualised using a boxplot with the library seaborn. It indicated that the data had outliers which majority were above the upper limit.

Below is a graph(boxplot) showing the distribution of various features against the values.

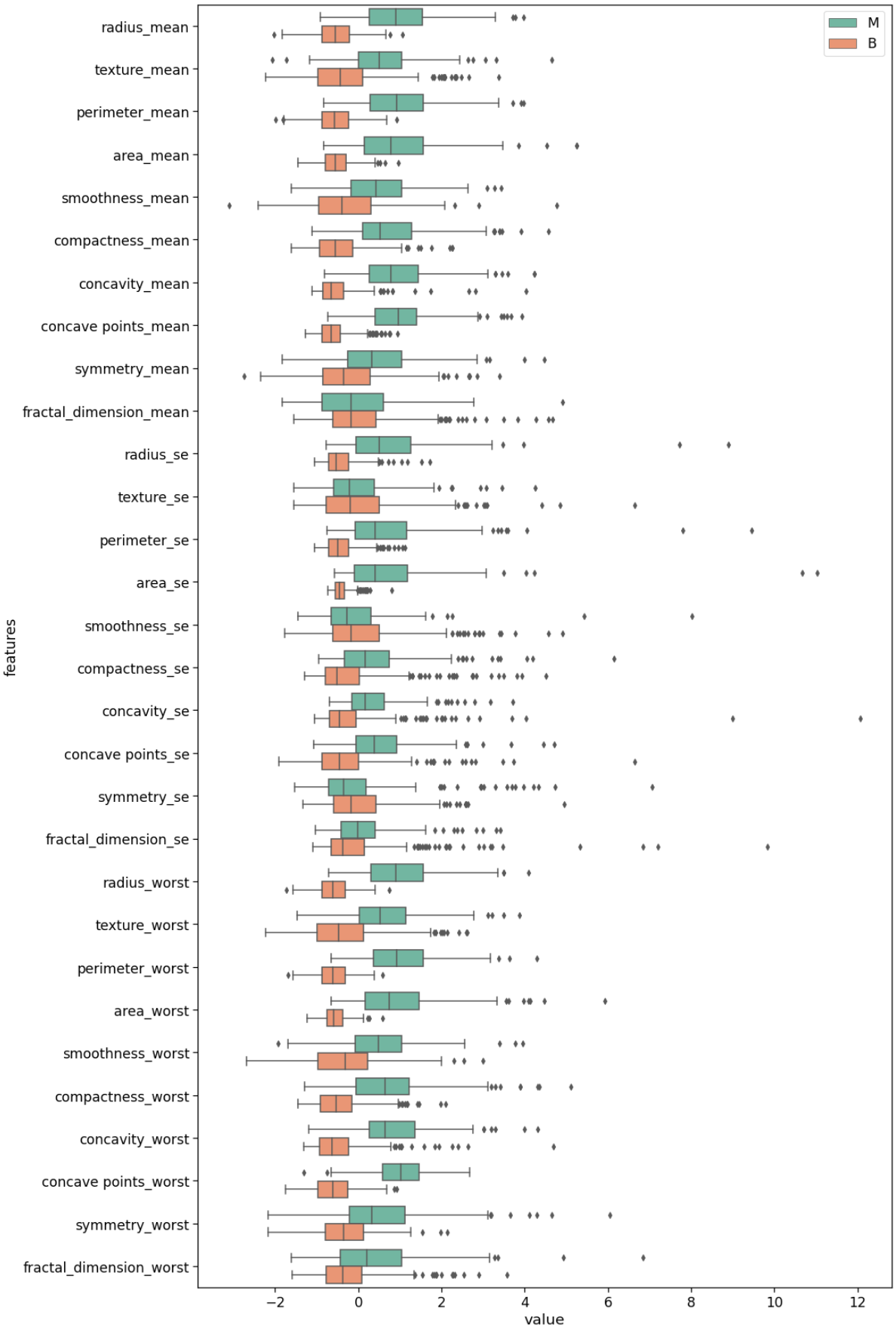


Figure 1 Boxplot of the features against the values for the breast cancer dataset

Data is also visualised using violin plots which are like density plots but unlike the bar graphs with means and error bars, they have all data points which makes them an excellent tool to visualise samples of small size.

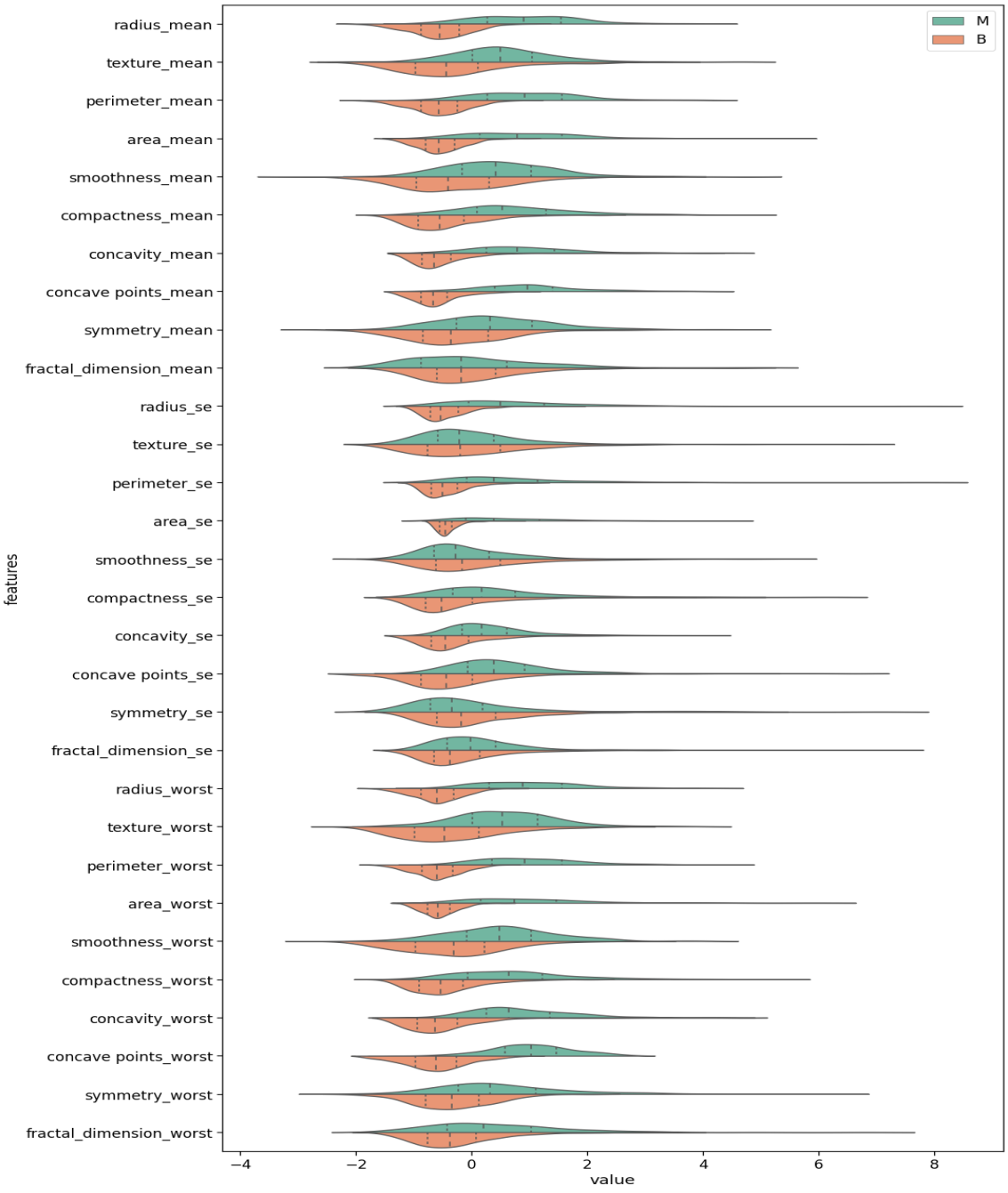


Figure 3Graph showing the violin plots of features against the value

From the above graphs (boxplot and violin plots), it is evident that the fractal dimension mean, texture se mean and smoothness se, the medians and means of the malignant and benign groups are very close to each other. therefore, they are not good features for classification. The shape of the violin plot for area se and the distribution of data points for benign and malignant are different from the others.

Area_worst and perimeter_worst looks well separated therefore good for classification.
To check the correlation between the features, I plotted a correlation matrix. It is effective in summarizing a large amount of data where the goal is to see patterns and find relationship between features.
The means, std errors and worst dimension lengths of compactness, concavity and concave points of tumors are highly correlated amongst each other (correlation > 0.4) Peimeter_mea n and area_mean have a correlation of 1. Perimeter_worst and area_worst, radius_mean,perimetre_mean,area_mean are highly corelated.

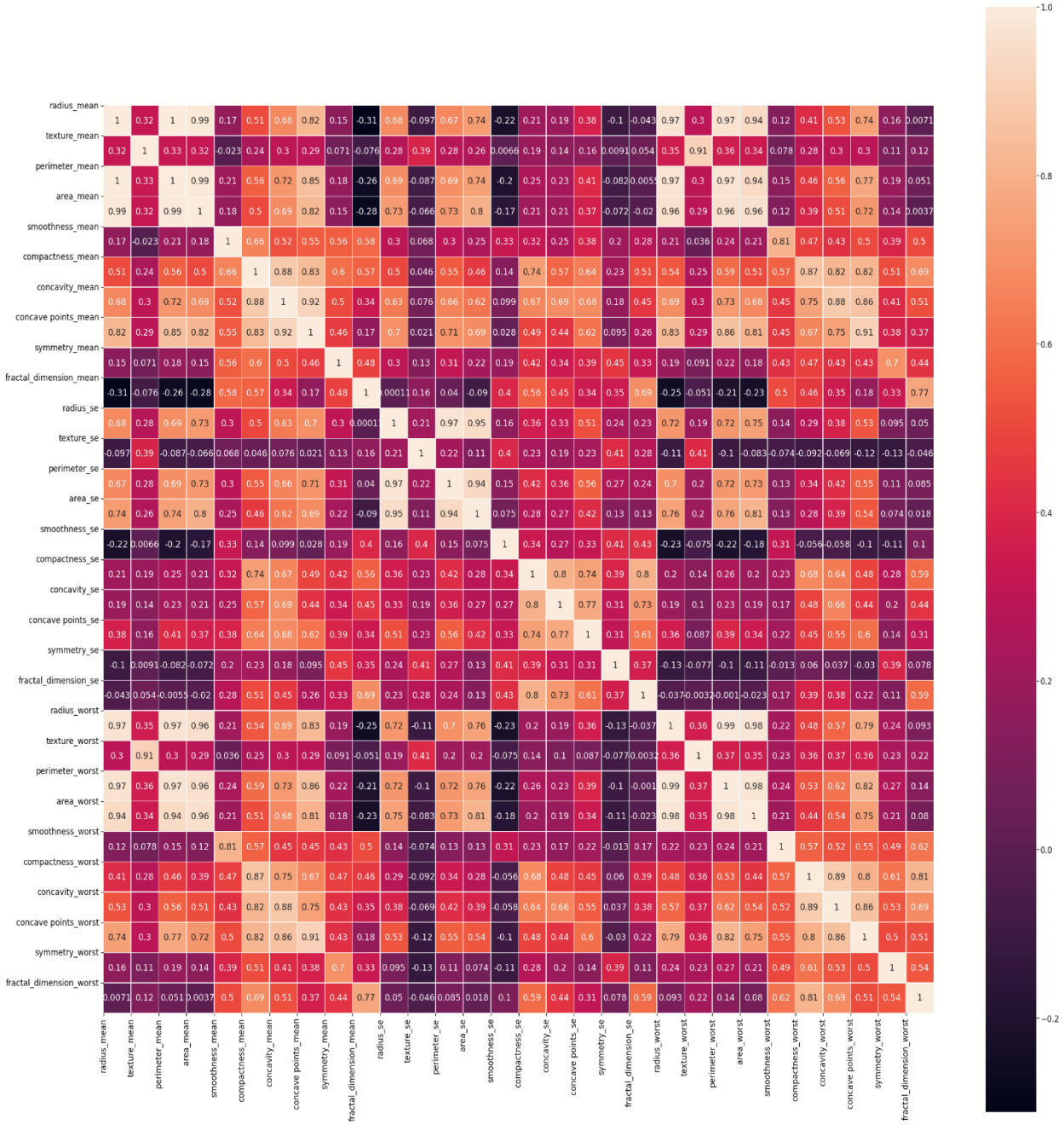
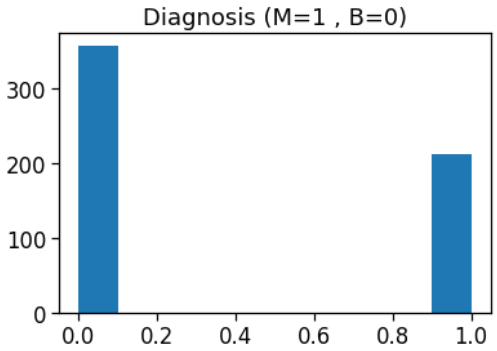


Figure 4 correlation matrix for all the features

3.2 Phase 2- Categorical Data

The categorical data in the column 'diagnosis' is converted to numerical data by mapping diagnosis column(object) to integer value :0,1. The head was the observed to confirm the mapping. The diagnosis count is also visualised using the histogram as shown below



3.3 Phase 3 Splitting the data

The dataset is split using Scikit learn library which splits the dataset using the train_test_split method. Dataset is split into training set (70%) which contains a known output, and the model learns on this data in order to be generalized to other data later on. The test set (30%) tests our model prediction. The features for building the model are selected by removing some columns before splitting the data.

3.3 Phase 4- Model Selection

Machine learning types include

Supervised learning- it uses known labelled data as input and has feedback mechanism

Unsupervised learning- No labels are given to the learning algorithm. It learns on its own to find structure in its input.

Our dataset has the outcome variable or Dependent variable i.e Y having only two set of values, either M (Malign) or B(Benign). So, I used Classification algorithm of supervised learning.

Classification algorithms include *logistic Regression, Nearest Neighbor, Support Vector Machines, Naïve Bayes, Decision Tree Algorithm, Random Forest Classification*

The features selected to build the model are *radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean,compactness_mean,concavity_mean'concave points_mean, symmetry_mean, radius_se, perimeter_se, area_se, compactness_se, concavity_se, concave points_se, fractal_dimension_se, texture_worst, perimeter_worst, area_worst, smoothness_worst.* 10 features were eliminated.

Sklearn library is used to import all the methods of classification algorithm and applied the different classification models to compare accuracies with different models. The one with highest accuracy is selected to predict the test set results and check the accuracy with each of the model.

The model with the highest accuracy is *Support Vector Machines with 96% Accuracy which formed the best model*

	precision	recall	f1-score	support
0	0.96	0.98	0.97	112
1	0.96	0.93	0.95	59

accuracy			0.96	171
macro avg	0.96	0.96	0.96	171
weighted avg	0.96	0.96	0.96	171

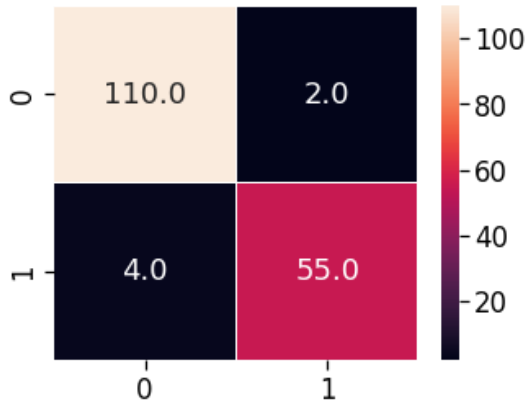


Figure 5 Confusion matrix

Out of 112 Benign cases 110 were predicted correctly and 2 were misclassified as shown above. Out of 59 Malignant cases, 55 were correctly predicted and 4 were misclassified.
I imported confusion matrix method of metrics class. The confusion matrix is a way of tabulating the number of misclassifications or evaluating predictions.

4.Conclusion

Cancer consists of very many different subtypes(heterogenous). To facilitate clinical management of patients, early diagnosis and prognosis is necessary for cancer research. Classifying cancer into benign or malignant has led many research groups in biomedical field to study application of machine learning method. Machine learning has been utilized as an aim to model the progression and treatment of cancerous conditions. It can detect key features from complex datasets. Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Decision Trees (DTs) have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making. ML techniques have been used to predict cancer susceptibility, cancer recurrence and cancer survival.

References

[1] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144: 646–74.
[3] Mei-Yin C. Polley, Boris Freidlin, Edward L. Korn, Barbara A. Conley, Jeffrey S. Abrams, Lisa M McShane20. November 2013.Machine learning applications in cancer prognosis and prediction. *JNCI: Journal of the National Cancer Institute*, Volume 105, Issue 22, Pages 1677–1683, https://doi.org/10.1093/jnci/djt282
[4] Mugdha Paithankar,(Nov 9, 2020). Breast Cancer Classification Using Python. https://medium.com/swlh/breast-cancer-classification-using-python-e83719e5f97d
[5] Dan Hoicowitz,(Jan 9, 2019). Breast Cancer Detection— A Classification Problem in Python. https://medium.com/@dandatascienceblog/breast-cancer-detection-a-classification-problem-in-python-ae2b0c9579ba