Breast Cancer detection using random forest

Leela Sekhar Chowdary Mannava
700740979
University of Central Missouri
Lee Summit

Sravya Mannava
700747750
University of Central Missouri
Lee Summit

Phani Datta Pabisetty
700741652
University of Central Missouri
Lee Summit

Venkata Naga Nikhil Reddy
Sanikommu
700746372
University of Central Missouri
Lee Summit

Abstract- A hazardous form of cancer is breast cancer. Breast cancer can cause women to die suddenly. Using the automatic disease detection system, doctors may diagnose and evaluate patients more quickly, with more effectiveness and a lower chance of death. The likelihood of survival may be considerably decreased if breast cancer is detected late. A normal tumour has a clear border, is spherical, and moves fluidly.. Malignant tumors, on the other hand, are often thin in nature, uneven in shape, and devoid of clear borders. Therapy, chemotherapy, and the use of hormone and immune therapies for detection all have different rules. Essentially a silent disease, this cancer won't show any outward symptoms.

Many patients could recognize the signs and symptoms of the disease once the tumor had progressed to a more advanced stage. Applying machine learning to the method and accuracy of diagnosis in medicine is a significant development and foreseeable course of the future medical model. Breast cancer can take one of two main forms.

Benign breast cancer is one form that, when found in its early stages, can be treated with medication. Malignant breast cancer, the other form, can be deadly and exhibits severe symptoms. Many algorithmic techniques, such as the ensemble machine learning algorithm and Naive Bayes, are utilized to detect breast cancer in its early stages.

This study utilized one machine learning method to show how breast cancer might be predicted using the Random Forest Classifier. Using a random forest classifier, the results from several small classifiers can be combined to produce reliable classification results. By segmenting the data into different trees, the classifier assesses whether a person is at risk of acquiring breast cancer or not. The accuracy of this model is 98%.

I. INTRODUCTION

In the past years, there has been a considerable rise in the high incidence of breast cancer among women. Only lung cancer causes more cancer-related deaths in women than in men. By World Health Organization's statistics, there are roughly 1.38 million new breast cancer cases and 458000 fatalities yearly.

In 2020, breast cancer diagnosis rates surpassed those of lung cancer to become the most prevalent type of cancer. Many women cannot receive the treatment they require due to a lack of qualified specialists and the high expenses of consultations. There is a critical issue with doctors in developing nations like India. Therefore, adopting automated clinical decision systems could help solve this issue. The deployment of such systems is hampered by the need for a general understanding of how machine learning models work in medicine.

A disease detected on time must be treated with a certain amount of human effort Most people are unaware of their ailment until it becomes chronic. Based on the performance all around the world, the death rate rises. Breast cancer is one of the illnesses that may be cured if it is found at an early stage and hasn't already spread to every part of the body.

It is difficult for doctors to establish a therapy plan that might lengthen patient survival time because there are no prognosis models available. It takes time to develop a method that produces the fewest errors possible while increasing accuracy. The existing methods to detect breast cancer, such as mammography, ultrasound, and biopsy, took a lot of time, thus there was a need for a computerized diagnostic system that used machine learning techniques. This approach makes use of algorithms to be more exact and quickly identify cells, classify tumors, and classify tumors.

Breast cancer is serious worldwide, including in industrialized and developing countries. Breast cancer can be identified medically by reviewing a patient's medical history and using several diagnostic procedures. Some of the terminology and methods used to predict breast cancer

- Mammography is an X-ray technique used to find breast cancer in its early stages. Changes in breast tissue that might represent signs of cancer can be found with the help of Mammography.
- A small portion of breast tissue is extracted during a biopsy and examined under a microscope to identify whether or not it is malignant.
- Ultrasound is a diagnostic process that produces images of the breast tissue using high-frequency sound waves. Breast tissue alterations that could be cancerous can be found with ultrasound.
- Magnetic resonance imaging (MRI) creates images of the breast tissue by using strong magnets and radio waves. Breast tissue alterations that can indicate signs of cancer can be found using MRI.

Ensemble classification involves learning to increase classification accuracy. Machine learning is frequently used to determine whether a person has a particular disease. As an example, consider Cancer Classification Using Fuzzy C-Means with Feature Selection Application. High-dimensional breast cancer database categorization using the Normed kernel function-based fuzzy probabilistic C-means algorithm selection is based on Laplacian Score, an Application of machine learning on brain cancer multiclass classification, and Kernel Modified Fuzzy C-Means for Gliomatosis Cerebri are a few instances.

The certainty of the diagnosis is critical for building a rule-based diagnosis support system. It is common practice to extract rules from a single learning model, such as a decision tree. However, the precision of these extracted rules is constrained. One of the most used learning models, ensemble learning, can make a great candidate for precise, understandable machine learning. However, because of the data, most ensemble learning models have not been interpretable or widely used for medical diagnosis. In the past few decades, some work has been done to comprehend the choice principles in ensemble learning by deriving decision rules from them. However, the bulk of ensemble learning models that have been written about in the past need to be clarified, and as a result, they are only sometimes used for identifying health issues. By deriving decision rules from decision principles in ensemble learning, some work has been done in recent years to understand these concepts better.

The Random Forest approach is our primary model because it performs better than other algorithms at classifying breast cancers as benign or malignant. It is trained using two distinct dataset subsets with 16 and 8 features that were selected using various feature selection techniques.

KEYWORDS: Decision trees, Random forest classifier, Machine learning algorithms, Training datasets, testing dataset, Python Interface, NumPy, and mathplot functions().

II. MOTIVATION

Breast cancer is the cause of a staggering number of fatalities worldwide. It is brought on by several health-related, lifestyle-related, cultural, and economic problems. By following clinical methods, this cancer can either be predicted in the early stages or the later stages. In the latter stages, there is no suitable treatment for the patients. Breast cancer is categorified into two types, likely benign and malignant. Benign can be treated with medications but malignant cannot be treated with proper medication. Mainly malignant is of improper shape and painful.

The medical procedure takes Mammography, an x-ray that includes ultrasounds with many side effects. Machine learning techniques are helpful for early cancer prediction. Machine learning can forecast breast cancer based on traits concealed in data. The random forest technique uses classification and regression to train large datasets and combine numerous decision trees to give accurate results that help patients determine if they have cancer in its initial phases.

Ensemble learning helps to bag and vote on the datasets using the Python interface machine learning algorithms. The datasets are divided into training and test data using the random forest classifier. The random forest With the aid of the random forest classifier, data will be predicted. multiple decision trees; by using various decision trees, the predicted data which provide accurate data with efficient results. So, machine learning techniques can help patients know about breast cancer in its early stages without medical reports.

III. OBJECTIVES

- Collect the data and Analyse it from the repository and identify the missing values and clean the data.
- The aim is to collect long-time datasets for training and testing data for the model. Because the dataset is smaller and limited, ensemble models efficiently predict the results.
- This Project is used to predict breast cancer using ensemble models like boosting, voting, and bagging techniques.
- The results of implementing random forest to identify breast cancer will be accurate.
- Implement models and mention the updated framework.
- The random forest algorithms are used and implemented in the Python interface.

IV. Main contribution

- Requirement analysis - Sravya
- Data collection and cleaning- Naga Nikhil
- Exploratory data analysis -Leela Shekar
- Choosing Model - naga Nikhil + Phani Data
- Performance evaluation - Phani Datta
- Summarizing the results- Sravya

- Python code -Leela Shekar

V RELATED WORK

Generally, Breast cancer mainly occurs in the breast tissue and causes Irritation on the breast skin, changes in the breast size, and pain in the breast area. Traditionally, breast cancer is detected using a mammogram x-ray, which is very costly, and sometimes, this test will not show accurate results. Breast cancer can be identified and predicted using X-ray imaging. However, it might be difficult to conclude when combining imaging and clinical data. These cancers mainly occur in young and adult women. The patient's disease cannot be identified in the early stages by following the medical report. So we use some machine learning techniques to overcome this problem.

The data is identified from the Wisconsin data repository. The data has to be cleaned in the first step, and it should be analyzed and should replace the missing values with the mean value and then sent to data pre-processing. It has three types of data cleaning, transforming, and reduction, that help clean the noisy data. Data reduction will eliminate the columns which have so many missing values. The exploratory data analysis is used to understand all the variables and understand them and to visualize and analyze the results. It helps process the raw data from the Wisconsin data set and prepare it for other data pre-processing processes.

Ensemble model To improve the accuracy of predictive analytics, ensemble modeling combines the outputs of two or more related but distinct analytical models into a single score or spread. After all models have been fitted to different subsets of a training dataset, the predictions from all models are combined in a process known as bagging.

Boosting method reduces errors in predictive data analysis of the tanning data sets. It is a machine learning estimating fact that generates various basic models or

estimators and makes predictions based on averaging their results. A voting classifier helps machine learning evaluate that it generates multiple basic models or estimators and makes predictions based on averaging their results.

A decision tree is a hierarchical structure in which the nodes express precise requirements on a set of features, and the branches allocate options in favor of the child nodes. The students choose class names. A decision tree can be constructed using either a conditional inference tree or recursive partitioning. Recursive partitioning is used to build a Decision Tree incrementally by splitting or not splitting each node. In our scenario, the tree is trained by dividing the source set into subgroups based on an attribute value test. The recursion is considered to be complete when each node's subset only contains occurrences of the target variable's same value. A statistical method called a conditional inference tree corrects for multiple testing and uses non-parametric tests as splitting criteria to prevent overfitting.

Since it can handle missing values and continuous, categorical, and binary data, Random Forest is appropriate for high dimensional data modeling. However, for vast data sets, the size of the trees can consume a lot of memory. Tuning the hyper-parameters is necessary because it has a propensity to over-fit.

One of the supervised machine learning models is random forest.. A randomly selected data set is initially chosen by the random forest classifier, producing many decision trees. The final classification of the test object is then decided after summarizing the votes from various decision trees. A single decision tree has a greater chance of producing an error. Still, when several decision trees are used in the classification process, we see a decrease in error and an increase in accuracy. When evaluating the impact of each output or decision from any decision tree, this algorithm leverages the concept of weights. Low weight is assigned to a tree with a high error, and increased weight is given to a tree with a common misconception.

Randomforestclassifier has three parameters they are The number of decision trees utilized in the model is n, and the purpose of creating two methods is to employ them consistently throughout the model. The minimal working set size at a node required for splitting is denoted by S, which is utilized to minimize the sample split defaults to 2 in this case.

The machine learning models use Python as an interface to implement their models. To implement this Project, firstly download and install the Python application. Python is user-friendly and can easily understand by beginners. Python has some predefined packages, and we can access them using the import keyword. Packages like Mathplot, sklearn, seaborn, os, and pandas are imported to implement the random forest model. The Python package will compare different types of datasets and helps to predict accurate results.

VI Proposed Framework

This Project has four phases. Data pre-processing (data cleaning), exploratory data analysis, ensemble modeling, and web deployment. In the first stage, unclean data is removed, unneeded columns are renamed, and the target variables are label-encoded. The data structure and observations are derived from the visualizations in the second stage. Cleansed data is sent to the ensemble models in the third stage. Among the three forms of ensemble models, the best model is chosen. Predictions for a given data set are displayed using the Python online interface. Using the scikit-learn train test split method, pre-processed data is divided into train and test segments in the proportion of 80:20. In terms of statistics, an ensemble is a collection of related systems or a collection of the same designs with various states.

Ensemble approaches combine several different models to display the output of machine learning models. Ensemble models come in three flavors: bagging, boosting, and voting.

A meta-estimator called an ensemble classifier fits individual weak learners to random subsets of the original dataset. Then it averages or votes on each learner's predictions to get a final prediction. Samples drawn using replacement are known as "bagging" samples. By instructing the weaker learners, the boosting method in the ensemble technique enhances the outcomes. AdaBoost and GradientBoosting are two of its primary algorithms. The list of invalid students in each ensemble type is displayed in the table below. The Project's ensemble algorithms are all implemented with the scikit-learn library. We considered the accuracy parameter throughout the performance evaluation step to determine the optimal model for classifying breast cancer.

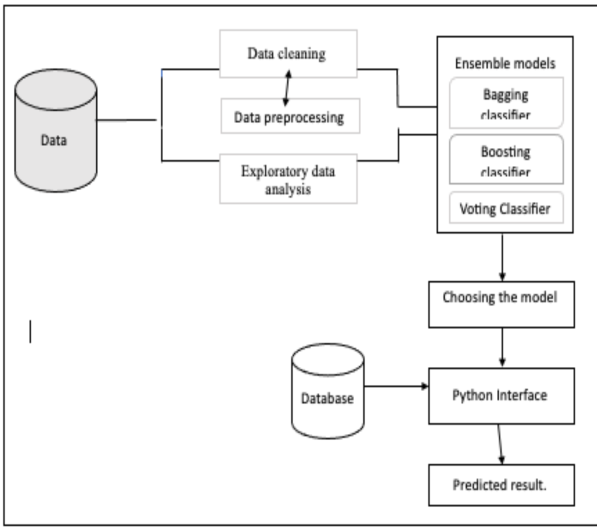


Fig 1

Data pre-processing: In data, pre-processing is divided into three types: cleaning, transformation, and reduction.

- Data cleaning helps to clean the data by removing noisy data from the Wisconsin data repository.
- Data transforming helps to replace the missing values with the mean value or minimum or maximum values based on program needs.
- If a column has a lot of missing values, data reduction can help to Traduce it. However, missing values are never counted as an attribute while collecting data from cancer patients.

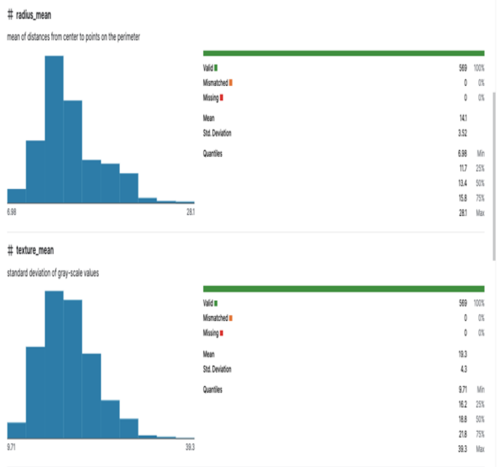


Fig-2

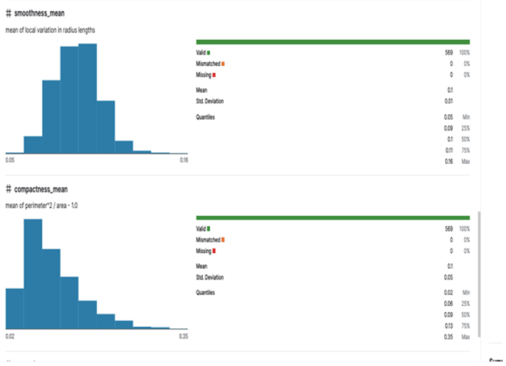


Fig-3

Exploratory data analysis: We noticed the data's underlying structure during the exploratory data analysis. The widest distribution occurs between the experimenter mean. We need to comprehend the distribution of each category to move forward with the models. We have the most samples in both types, demented and non-demented, for the two forms of cancer. Because using the algorithms depends on the data, this is a crucial observation that should be considered.

Ensemble methods use a combination of different models to improve the results of machine learning models. There are three ensemble model types: bagging, boosting, and voting.

- Bagging is an ensemble approach that combines the predictions from all models after fitting them to various subsets of a training dataset.
- Boosting method reduces errors in predictive data analysis of the tanning data sets. It is a machine learning estimating fact that generates various basic models or estimators and makes predictions based on averaging their results.
- A voting classifier helps machine learning evaluate that it generates multiple basic models or estimators and makes predictions based on averaging their results.

In a random forest, the classifier will take numerous decision trees to get an accurate result from that considerable amount of data. Bootstrapping is one method used in random forests. While training the data sets, each tree in the random forest learns from the samples of the datasets. Some of the datasets are taken with replacements. All the ensemble methods in the Project are implemented using the scikit-learn library. In the performance evaluation step, we have considered the accuracy parameter to decide the best model that classifies breast Cancer. Gradient boosting, Extra tree classifier, and Random forest are the best-performing models.

The Python packages are imported to implement these ensemble models in this Project. Python is an interface used to implement and predict accurate results for the random forest approach. The random forest combines multiple decision trees to get accurate results. In the Python interface, we import packages to predict the precise outcome.

In Python, we import the packages like pandas, seaborn, NumPy, sklearn, math,train_test_split, matplotlib, and os to implement the Project using the random forest classifier () algorithm.

Os: provide the ability to set up the user's interface with the operating system. It gives various practical OS features that may be utilized to carry out OS-based tasks and obtain pertinent OS-related data.

Numpy: a widely used open-source Python library in practically all branches of science and engineering.

Pandas: provides quick, adaptable, and expressive data structures that make using "relational" or "labeled" data simple and intuitive.

Seaborn: The Seaborn library is a popular visualizing data toolkit for Python that is frequently used for machine learning activities.

Matplotlib: Python may be used to create static, animated interactive infographics. Matplotlib makes both tough problems and simple ones doable.. Produce plots fit for publication.

Sklearn: a machine learning (ML) package that is free and open source that is considered to be the industry standard in the Python ecosystem

Pre-processing: provides a number of general utility functions and classes to transform unrepresentative raw feature vectors into more useful representations for downstream estimators.

Train_test_split(): merges calls to check_arrays, next(iter(ShuffleSplit(n_samples)), and the framework to input data into a function call for splitting (and possibly subsampling) data.

Making use of the random forest and breast cancer. Based on the patient's symptoms regarding their condition, the random forest employs several decision trees using historical data to get reliable results.

VII DATA DESCRIPTION

s.no	Type	keywords	Description
1	Clinical Parameter	Diagnosis	Checking the diagnosis of Breast tissue
2	Clinical Parameter	Malignant	early stage of breast cancer
3	Clinical Parameter	benign	Next stage of breast cancer
4	Clinical Parameter	texture	This will calculate using gray-scale values.
5	Clinical Parameter	radius	distances from the centre to the entire

			breast tissue.
6	Clinical Parameter	perimeter	size of the core tumor.
7	Clinical Parameter	area	Calculate the area of the breast tissue.
8	Clinical parameter	smoothness	Calculates local variation in radius lengths

One of the biggest challenges to treatment is the struggle to identify symptoms in the early stages of the disease. One typical sign is the development of breast tissue lumps, which are also related to various skin diseases. Wisconsin provided the datasets for this Project. The 13 columns in the 14-column dataset used in this model are considered significant parameters, and the final column comprises the prediction values.

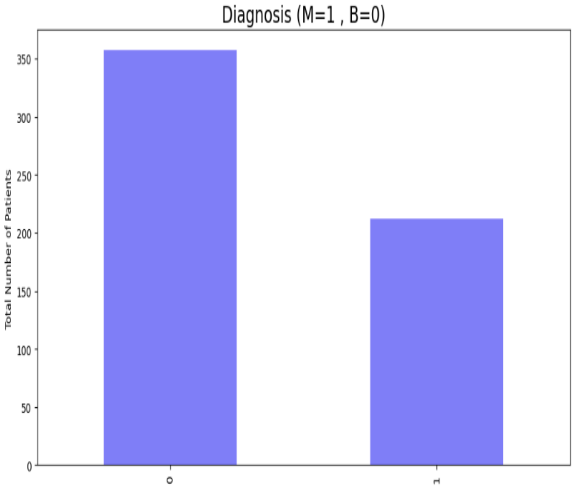
The parameters radius, texture, perimeter, area, smoothness, compactness, concavity, concave point, symmetry, and fractal dimension are all represented in the dataset along with their mean, best case, and worst case values. The diagnosis will also be one of the crucial variables in the dataset. Unique components like malignant and benign tissue are included in the diagnosis of breast tissue. Malignant illnesses are ones in which abnormal cells multiply uncontrollably and can invade neighboring tissues.

A benign tumor cannot spread to other parts of the body, despite the possibility of growth. Developing a model that can accurately anticipate the current illness without using the expensive monogram test is crucial.

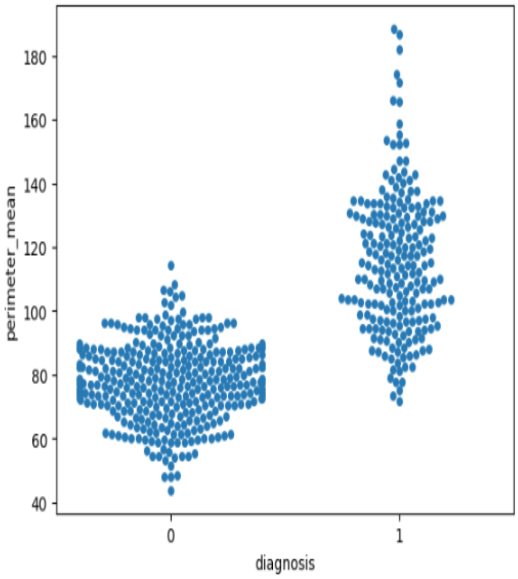
VIII Results/ Experimentation & Comparison/Analysis

Comparing the data and showing how many have cancer. i.e., two types (Malignant, benign)

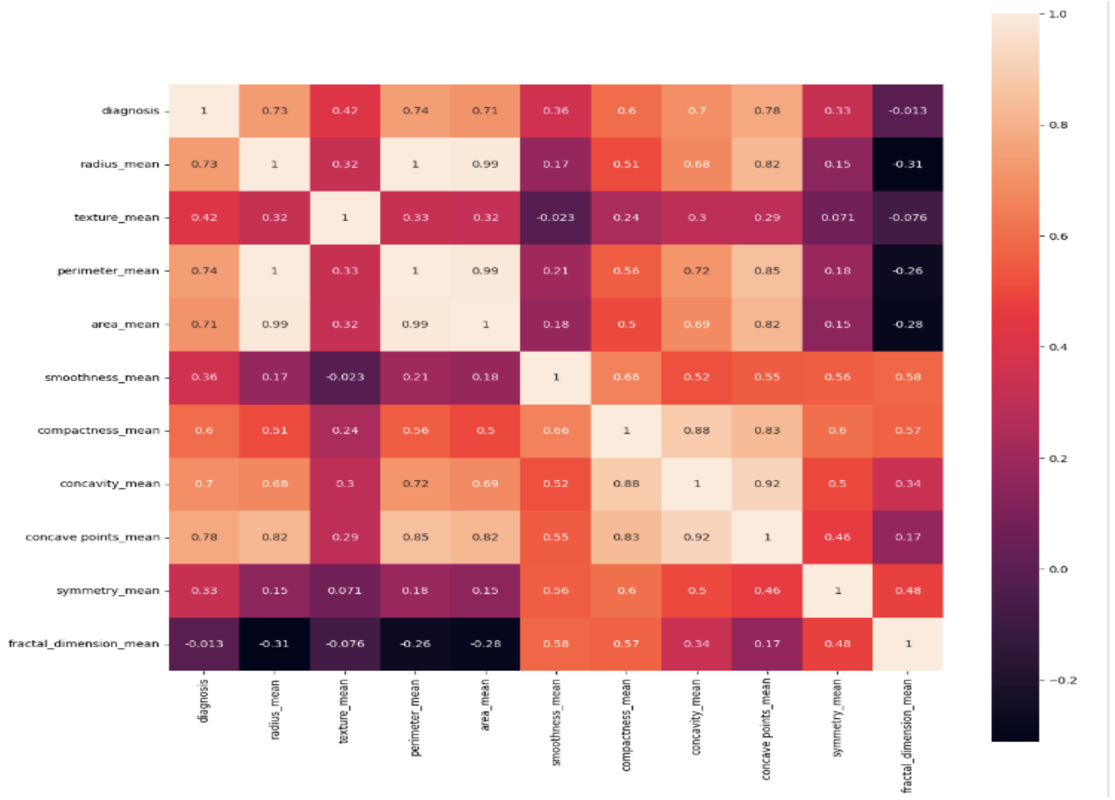
```
: Text(0, 0.5, 'Total Number of Patients')
```



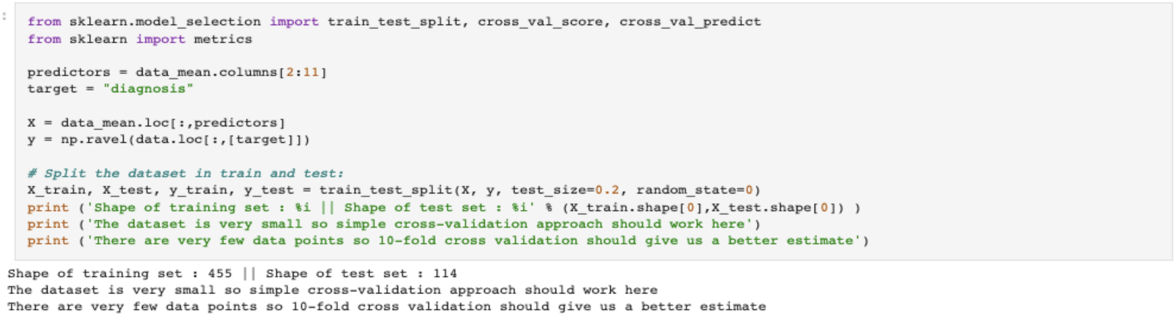
Only 65% of the patients had benign tumors, and the remaining had malignant tumors.



Shows the categorical values representing the values of both Malignant and Benign values.



The diagnosis and the means for radius, perimeter, area, compactness, concavity, and concave points are highly related.
The other parameters don't seem to have a significant effect on diagnoses.



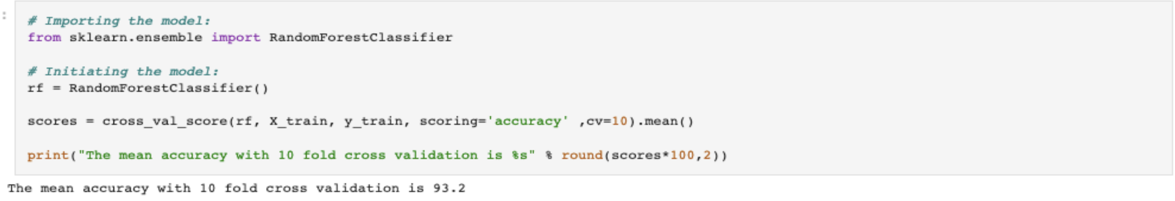
```
from sklearn.model_selection import train_test_split, cross_val_score, cross_val_predict
from sklearn import metrics

predictors = data_mean.columns[2:11]
target = "diagnosis"

X = data_mean.loc[:,predictors]
y = np.ravel(data.loc[:,[target]])

# Split the dataset in train and test:
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
print ('Shape of training set : %i || Shape of test set : %i' % (X_train.shape[0],X_test.shape[0]) )
print ('The dataset is very small so simple cross-validation approach should work here')
print ('There are very few data points so 10-fold cross validation should give us a better estimate')
```

```
Shape of training set : 455 || Shape of test set : 114
The dataset is very small so simple cross-validation approach should work here
There are very few data points so 10-fold cross validation should give us a better estimate
```



```
# Importing the model:
from sklearn.ensemble import RandomForestClassifier

# Initiating the model:
rf = RandomForestClassifier()

scores = cross_val_score(rf, X_train, y_train, scoring='accuracy' ,cv=10).mean()

print("The mean accuracy with 10 fold cross validation is %s" % round(scores*100,2))
```

```
The mean accuracy with 10 fold cross validation is 93.2
```

IX REFERENCES

[1] https://pandas.pydata.org/docs/getting_started/overview.html#:~:text=pandas%20is%20a%20Python%20package,world%20data%20analysis%20in%20Python.

[2] https://matplotlib.org/

[3] https://www.activestate.com/resources/quick-reads/what-is-scikit-learn-in-python/#:~:text=Scikit%2Dlearn%20is%20an%20open,categorizing%20data%20based%20on%20patterns

[4] https://scikit-learn.org/0.15/modules/generated/sklearn.cross_validation.train_test_split.html#:~:text=Quick%20utility%20that%20wraps%20calls,subsampling)%20data%20in%20a%20oneliner.&text=Python%20lists%20or%20tuples%20occurring,converted%20to%201D%20numpy%20arrays.

[5] https://www.simplilearn.com/tutorials/python-tutorial/python-seaborn#:~:text=Box%20plot-,What%20Is%20Seaborn%20in%20Python%3F,and%20can%20perform%20exploratory%20analysis

[6] https://aws.amazon.com/what-is/boosting/

[7] Ram, Vallam Sudhakar Sai, Namrata Kayastha, and Kewei Sha. "OFES: Optimal feature evaluation and selection for multi-class classification." *Data & Knowledge Engineering* 139 (2022): 102007.

[8] Anisha, P. R., et al. "Early Diagnosis of Breast Cancer Prediction using Random Forest Classifier." *IOP Conference Series: Materials Science and Engineering*. Vol. 1116. No. 1. IOP Publishing, 2021.

[9] https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm#:~:text=Step%201%3A%20Select%20random%20samples,as%20the%20final%20prediction%20result.

[10] https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[11] https://www.javatpoint.com/machine-learning-random-forest-algorithm

[12] https://www.researchgate.net/publication/351893325_Early_Diagnosis_of_Breast_Cancer_Prediction_using_Random_Forest_Classifier

[13] Bhise, Sweta, et al. "Breast cancer detection using machine learning techniques." *Int. J. Eng. Res. Technol* 10.7 (2021).

[14] https://www.javatpoint.com/python-os-module#:~:text=Python%20OS%20module%20provides%20the,under%20Python's%20standard%20utility%20modules

[15] https://www.cancerresearchuk.org/about-cancer/breast-cancer/symptom

[16] echtarget.com/searchdatamanagement/definition/data-preprocessing#:~:text=Data%20pre

processing%20transforms%20the%20data,pipeline%20to%20ensure%20accurate%20results

[17] https://seaborn.pydata.org/generated/seaborn.swarmplot.html.

[18] https://stackoverflow.com/questions/36153410/how-to-create-a-swarm-plot-with-matplotlib

[19] https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/

[20] https://towardsdatascience.com/random-forest-classification-678e551462f5#:~:text=Using%20Random%20Forest%20classification%20yielded,and%20without%20much%20parameter%20tuning.

[21] https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing#:~:text=What%20is%20data%20preprocessing%3F,for%20the%20data%20mining%20process