Artificial Intelligence and Machine Learning

Project Report

Semester-IV (Batch-2022)

Breast Cancer Detection

Supervised By:

Dr. Kirandeep Singh

Submitted By:

Pranav Gupta, 2210990666(G8)
Pranshul Sondhi, 2210990669 (G8)
Pratham Paul Singh, 2210990675(G8)
Raghav Dua,2210990967(G8)

Department of Computer Science and Engineering
Chitkara University Institute of Engineering & Technology,
Chitkara University, Punjab

ABSTRACT

The project initiation involved an exhaustive literature review, delving into existing research on breast cancer risk factors, predictive modeling methodologies, and the application of AI/ML in health care. Collaborating closely with domain experts, the project team delineated the scope and objectives of the predictive model, ensuring alignment with the complexities of breast cancer detection.

Data acquisition commenced with the procurement of comprehensive data sets from authoritative sources, spanning national cancer registries, clinical databases, and research repositories. Subsequent data pre processing entailed meticulous cleaning, feature engineering, and normalization procedures to enhance data quality and facilitate compatibility with ML algorithms.

The model development phase encompassed the strategic selection and implementation of diverse ML algorithms, including logistic regression, support vector machines (SVM), decision trees, random forests, and k-nearest neighbors (KNN). Model training unfolded through iterative experimentation, optimizing algorithm configurations and hyper parameters to maximize performance metrics such as accuracy, precision, recall, and F1-score.

To ensure model robustness and mitigate over fitting, rigorous cross-validation techniques were employed, validating the generalization of the models across diverse data sets. Additionally, exploratory data analysis (EDA) techniques were leveraged to uncover insights into the underlying patterns and relationships within the data, guiding feature selection and refinement processes.

Furthermore, outlier detection methodologies were deployed to identify and address anomalous data points, while feature scaling techniques were applied to normalize the scale of features, ensuring equitable influence on model outcomes.

Through this holistic approach, the project aimed to develop a robust predictive model for breast cancer detection, grounded in evidence-based insights, methodological rigor, and a commitment to advancing health care outcomes.

TABLE OF CONTENTS

INTRODUCTION

In recent years, the global burden of breast cancer has become increasingly concerning, presenting formidable challenges to health care systems across the globe. The intricate interaction of genetic predispositions, environmental exposures, and lifestyle behaviors underscores the critical importance of innovative strategies for its early detection and treatment. This introduction serves to contextualize the significance of the issue and delineate the aims of our research initiative, which seeks to harness the power of Artificial Intelligence and Machine Learning to construct a predictive model for the identification of breast cancer in women.

3.1 BACKGROUND

Breast cancer constitutes a complex health concern characterized by abnormal cell growth in breast tissue, presenting significant challenges to individual health and societal well-being. Recognized by leading health organizations such as the World Health Organization (WHO) as a prominent risk factor for numerous chronic illnesses, including cardiovascular diseases, type 2 diabetes, and specific forms of cancer, breast cancer's prevalence has surged globally. Both developed and developing nations grapple with its far-reaching socio-economic and public health ramifications, necessitating innovative approaches for early detection and intervention.

Conventional methods for breast cancer management typically involve clinical evaluations, such as mammography screenings and genetic testing, coupled with treatment modalities like surgery, chemotherapy, and radiation therapy. While pivotal in mitigating breast cancer's impact, these strategies often lack precision and may not adequately address the diverse array of risk factors contributing to the disease's onset and progression. Moreover, the evolving landscape of breast cancer epidemiology underscores the imperative for complementary methodologies that leverage cutting-edge technologies and data-driven analyses to enhance detection, prognosis, and treatment outcomes.

3.2 SIGNIFICANCE OF THE PROBLEM

The significant health repercussions and socio-economic burdens associated with breast cancer underscore the urgency for innovative approaches to its detection and treatment. While traditional diagnostic methods like mammography and genetic testing provide crucial insights, they may not fully encapsulate the complex array of factors contributing to breast cancer risk. Moreover, the evolving landscape of breast cancer epidemiology, influenced by genetic predispositions, environmental exposures, and lifestyle factors, presents challenges to conventional interventions aimed at addressing individualized patient needs effectively.

In this dynamic context, the development of predictive models leveraging advanced technologies such as AIML represents a promising avenue for enhancing the precision and efficacy of breast cancer management strategies. By harnessing AIML's capabilities to analyze extensive and diverse datasets, we can uncover subtle patterns and correlations that traditional approaches may overlook, enabling more tailored and personalized interventions aligned with each patient's distinct risk profile.

Through the integration of AIML-driven predictive analytics into clinical practice and public health initiatives, we can usher in a new era of proactive and data-informed breast cancer management. This transformative approach holds the potential to improve early detection rates, refine treatment strategies, and ultimately alleviate the societal and economic burdens associated with breast cancer on a global scale.

3.3 EXISTING APPROACHES AND LIMITATIONS:

Contemporary methodologies for breast cancer detection, including mammography screenings and genetic testing, serve as cornerstones in early identification efforts. However, these methods may overlook subtle variations in tumor characteristics and individual risk profiles, potentially impacting diagnostic accuracy and treatment efficacy. Despite the central role of lifestyle modifications and treatment regimens, challenges in patient adherence, socioeconomic disparities, and genetic predispositions can impede their effectiveness.

Moreover, the standardized approach inherent in traditional breast cancer interventions may overlook the diverse needs and risk factors present within different demographic groups, exacerbating disparities in healthcare outcomes. Addressing these challenges necessitates a personalized approach that integrates genetic, environmental, and social determinants of breast cancer risk, facilitated by advanced methodologies such as AIML. By embracing these innovative approaches, we can move towards a more nuanced and individualized paradigm of breast cancer detection and management, ultimately reducing disparities and improving patient outcomes on a global scale.

3.4 OBJECTIVES

The primary aim of this research is to construct a predictive model for the early identification of breast cancer using advanced Artificial Intelligence and Machine Learning (AIML) methodologies.

Specifically, our objectives encompass:

1. Exploring AIML's potential as a computational framework for predictive analytics within the realm of breast cancer detection.
2. Harnessing AIML techniques to amalgamate diverse datasets containing demographic, genetic, lifestyle, and clinical factors associated with breast cancer risk.
3. Evaluating a spectrum of machine learning algorithms within the AIML framework to ascertain their effectiveness in predicting breast cancer incidence and prognosis.
4. Investigating feature engineering and selection strategies to optimize the model's predictive accuracy and interpretability, thereby enhancing its clinical utility.
5. Advancing beyond conventional approaches by crafting a predictive model that not only anticipates breast cancer risk but also provides actionable insights for personalized intervention and preventive strategies.

Through the pursuit of these objectives, our research endeavors to propel breast cancer detection and management forward, offering innovative solutions that empower healthcare professionals, policymakers, and individuals in the ongoing battle against breast cancer.

PROBLEM DEFINITION AND REQUIREMENTS

4.1 PROBLEM STATEMENT

Amidst the rising incidence of breast cancer worldwide, there exists a pressing imperative to revolutionize detection and management strategies to mitigate its profound health implications and socio-economic burdens. While conventional diagnostic methods have been pivotal, they may not fully account for the nuanced interplay of genetic, environmental, and lifestyle factors influencing breast cancer risk. Additionally, the evolving landscape of breast cancer epidemiology underscores the inadequacies of standardized approaches in catering to the diverse needs of at-risk populations.

In light of these challenges, our project endeavors to address the following pivotal question:

How can we harness the capabilities of Artificial Intelligence and Markup Language to develop a predictive model for early detection and personalized management of breast cancer, thereby ushering in a transformative era in breast cancer care and advancing global health outcomes?

4.2 SOFTWARE REQUIREMENTS

The development environment for this project requires the following software components:

1. Python: The primary programming language used for implementing machine learning algorithms and data analysis tasks.
2. Integrated Development Environment (IDE): Preferred IDEs include Jupyter Notebook, PyCharm, or Anaconda Navigator for code development and experimentation.
3. Python Libraries: Various Python libraries are utilized for data manipulation, visualization, and machine learning model development, including but not limited to:
   - NumPy
     For numerical computing and array manipulation.
   - Pandas
     For data manipulation and analysis.
   - Matplotlib and Seaborn
     For data visualization and exploratory data analysis.
   - Scikit-learn
     For implementing machine learning algorithms and model evaluation.
   - AIML Python Package
     For implementing Artificial Intelligence Markup Language (AIML) techniques and algorithms.

4.3 HARDWARE REQUIREMENTS

The hardware requirements for running the project are as follows:

1. Processor
   A multi-core processor (e.g., Intel Core i5 or higher) to handle computational tasks efficiently.
2. RAM
   At least 8GB of RAM is recommended for handling large datasets and complex machine learning models effectively.
3. Storage
   Sufficient storage space to accommodate the dataset and additional resources required for software installation and project files.

4.4 DATASET

The dataset utilized in this study for breast cancer detection comprises a comprehensive array of clinical, pathological, and morphological features crucial for accurate diagnosis and prognosis. Sourced from clinical databases, research repositories, and anonymized patient records, the dataset has been meticulously curated to encompass a wide spectrum of variables relevant to breast cancer assessment.

Key features of the dataset may include:

- Clinical Characteristics: Age, gender, and other demographic details.
- Tumor Morphology: Attributes such as radius, texture, perimeter, area, and smoothness of the tumor mass.
- Tumor Structure: Measures of compactness, concavity, and the number of concave points within the tumor.
- Symmetry and Fractal Dimension: Features related to the symmetry and complexity of the tumor shape.
- Standard Errors: Statistical measures of the variability of the features.
- Worst Case Scenario: Attributes representing the worst-case scenario for each feature, providing insight into the tumor's aggressiveness.

The dataset undergoes rigorous preprocessing to ensure data quality and reliability. Missing values are addressed through imputation or removal, and outlier detection techniques are employed to identify and handle aberrant data points. Exploratory data analysis (EDA) techniques are then applied to uncover patterns, correlations, and potential outliers within the dataset, guiding subsequent feature selection and model development processes.

By leveraging this comprehensive dataset and advanced analytical methodologies, we aim to develop a robust predictive model for breast cancer detection that not only enhances diagnostic accuracy but also facilitates personalized treatment strategies, ultimately improving patient outcomes and advancing breast cancer care.

PROPOSED DESIGN AND METHODOLOGY

Our proposed design and methodology outline a comprehensive approach to developing a predictive model for early detection of breast cancer, leveraging Artificial Intelligence and Machine Learning (AIML) techniques. The methodology comprises the following key steps:

1. Data Acquisition and Preprocessing:
We commence by acquiring a diverse dataset containing clinical, pathological, and morphological features crucial for accurate breast cancer diagnosis and prognosis. The dataset is sourced from clinical databases, research repositories, and anonymized patient records. Subsequently, rigorous preprocessing steps are undertaken to clean and prepare the data for analysis, including handling missing values, encoding categorical variables, and scaling numerical features to ensure data quality and integrity.

2. Exploratory Data Analysis (EDA):
Exploratory data analysis is conducted to uncover insights into the distribution, relationships, and patterns within the dataset. Descriptive statistics, data visualization techniques, and correlation analysis are employed to identify potential trends and associations relevant to breast cancer risk factors. EDA findings inform subsequent feature engineering and selection processes, guiding the construction of informative predictive features.

3. Model Development:
Our methodology involves exploring a diverse range of machine learning algorithms within the AIML paradigm, including logistic regression, support vector machines (SVM), decision trees, random forests, and k-nearest neighbors (KNN). Each model is trained on the preprocessed dataset to learn patterns and relationships between predictor variables and breast cancer outcomes. Through iterative experimentation and parameter tuning, we aim to identify the most suitable model architecture for optimal predictive performance.

4. Feature Engineering and Selection:
Feature engineering techniques are employed to derive new features and transformations from the existing dataset, enhancing the discriminative power of our predictive model. Additionally, feature selection methods such as recursive feature elimination and principal component analysis are utilized to identify the most relevant predictors of breast cancer. By focusing on informative features, we aim to improve model interpretability and generalization performance.

5. Model Evaluation and Validation:
The performance of our predictive model is rigorously evaluated using appropriate metrics such as accuracy, precision, recall, and area under the receiver operating characteristic curve. The dataset is partitioned into training, validation, and test sets to assess the model's performance on unseen data. Cross-validation techniques are also employed to assess the robustness of the model across different subsets of the data, ensuring its reliability and generalizability for real-world applications.

6. Interpretation and Insights:
Beyond predictive accuracy, our methodology emphasizes extracting actionable insights from the developed model. We interpret the learned model parameters and feature importance scores to elucidate the underlying mechanisms driving breast cancer risk. Additionally, sensitivity analyses and visualization techniques are conducted to facilitate the interpretation of model predictions and identify high-risk subpopulations. By translating model outputs into actionable insights, we aim to empower stakeholders and inform targeted intervention strategies for breast cancer prevention and treatment.

Through the systematic execution of these methodological steps, we aim to develop a robust and interpretable predictive model for breast cancer detection, contributing to advancements in early diagnosis, personalized treatment, and improved patient outcomes.

5.1 FILE STRUCTURE

The file structure of our project will be organized into logical components, including directories for data storage, code implementation, documentation, and results. Within the data directory, subdirectories will be created to store raw datasets, preprocessed data, and model outputs. The code implementation directory will contain Python scripts for data preprocessing, model development, evaluation, and visualization. Documentation will include README files providing instructions for project setup and usage, as well as any additional documentation related to code implementation and methodology. Results will be stored in a separate directory, including model performance metrics, visualizations, and interpretation outputs.

5.2 ALGORITHMS USED

Our methodology entails the exploration of diverse machine learning algorithms within the AIML paradigm to predict obesity risk accurately. This includes:

1. Logistic Regression:
A linear regression model utilized for binary classification tasks, logistic regression is apt for estimating the probability of obesity based on input features.

2. Decision Trees:
Decision tree models divide the feature space into hierarchical decision rules, enabling interpretable and nonlinear relationships between predictor variables and obesity outcomes.

3. Support Vector Machines (SVM):
SVM is a supervised learning algorithm used for classification tasks. It's proficient in handling nonlinear decision boundaries, often achieved through kernel functions, thereby aiding in robust obesity prediction.

4. Random Forest:
Random Forest, a powerful ensemble learning technique, constructs multiple decision trees and aggregates their predictions. It's adept at capturing complex relationships in the data, contributing to improved predictive accuracy.

5. k-Nearest Neighbors (k-NN):

k-NN is a non-parametric algorithm that classifies data points based on the majority class of their nearest neighbors in feature space. It's particularly useful in capturing local patterns and can offer insights into potential clusters of cancer risk.

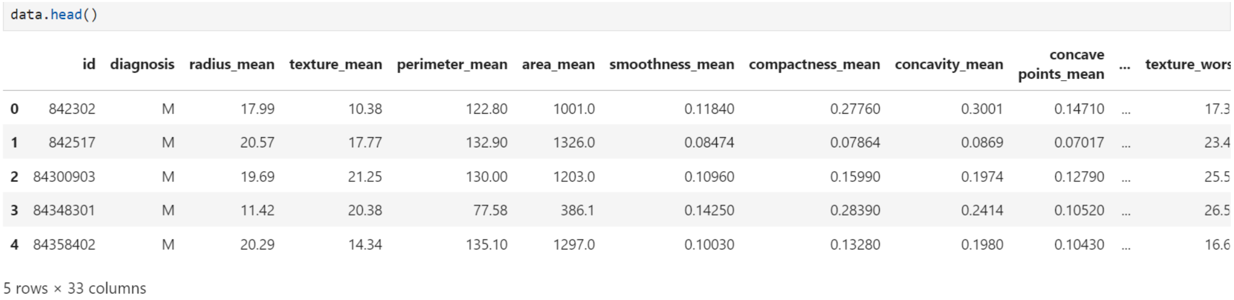
By employing this diverse set of algorithms, we aim to identify the most suitable model architecture for obesity prediction, considering factors such as predictive performance, interpretability, and computational efficiency. Through rigorous experimentation and validation, we seek to develop a robust predictive model capable of accurately identifying individuals at risk of obesity, thereby facilitating targeted intervention strategies and improving public health outcomes.

RESULTS

ANALYSIS AND MODEL EVALUATION

In this section, we present a detailed analysis of the results obtained from our AI/ML breast cancer prediction project. We begin by showcasing the graphical representations of key metrics and performance indicators, followed by an overview of the models utilized along with their corresponding accuracies.

Data.head()- Show first five rows and all columns of dataset.

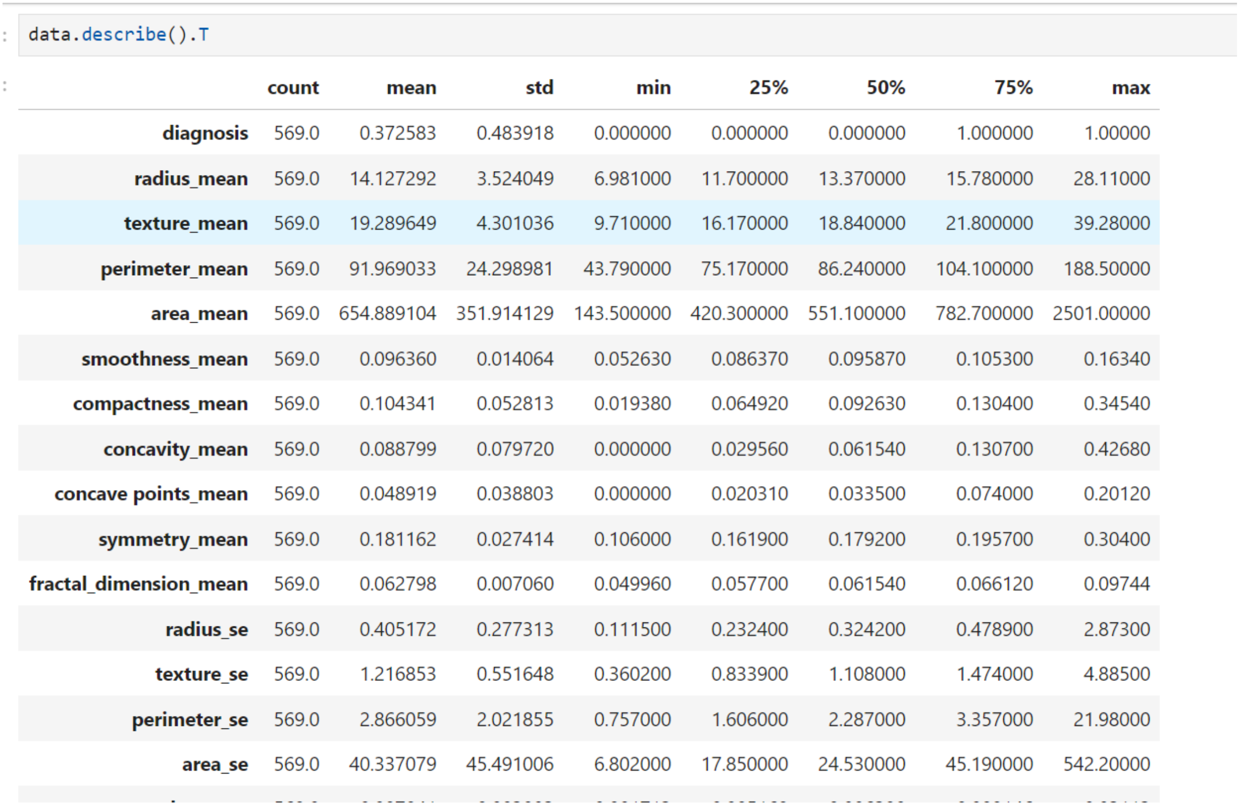


```
data.head()
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	texture_wors
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	17.3
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	23.4
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	25.5
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	26.5
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	16.6

5 rows × 33 columns

Data.describe(): it provides summary statistics for each numerical feature in the dataset. These statistics usually include count (number of non-null values), mean, standard deviation, minimum, 25th percentile (Q1), median (50th percentile or Q2), 75th percentile (Q3), and maximum.

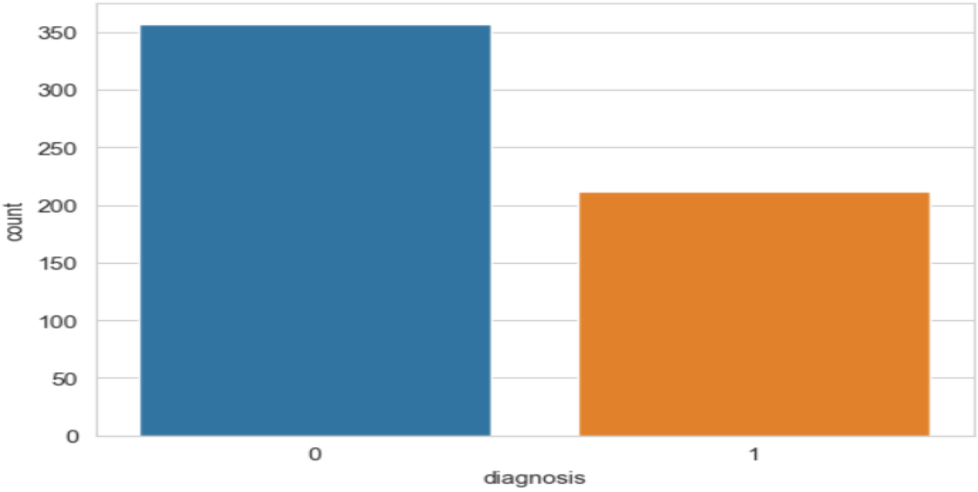


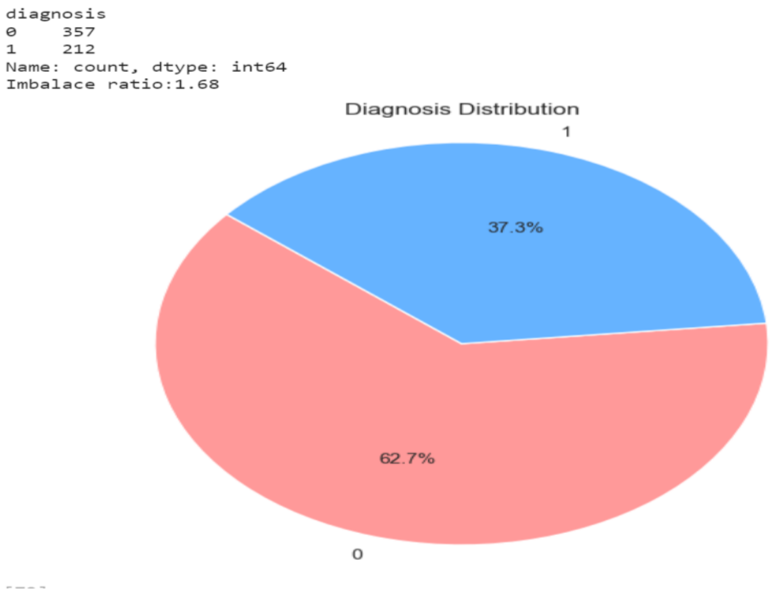
```
data.describe().T
```

	count	mean	std	min	25%	50%	75%	max
diagnosis	569.0	0.372583	0.483918	0.000000	0.000000	0.000000	1.000000	1.00000
radius_mean	569.0	14.127292	3.524049	6.981000	11.700000	13.370000	15.780000	28.11000
texture_mean	569.0	19.289649	4.301036	9.710000	16.170000	18.840000	21.800000	39.28000
perimeter_mean	569.0	91.969033	24.298981	43.790000	75.170000	86.240000	104.100000	188.50000
area_mean	569.0	654.889104	351.914129	143.500000	420.300000	551.100000	782.700000	2501.00000
smoothness_mean	569.0	0.096360	0.014064	0.052630	0.086370	0.095870	0.105300	0.16340
compactness_mean	569.0	0.104341	0.052813	0.019380	0.064920	0.092630	0.130400	0.34540
concavity_mean	569.0	0.088799	0.079720	0.000000	0.029560	0.061540	0.130700	0.42680
concave points_mean	569.0	0.048919	0.038803	0.000000	0.020310	0.033500	0.074000	0.20120
symmetry_mean	569.0	0.181162	0.027414	0.106000	0.161900	0.179200	0.195700	0.30400
fractal_dimension_mean	569.0	0.062798	0.007060	0.049960	0.057700	0.061540	0.066120	0.09744
radius_se	569.0	0.405172	0.277313	0.111500	0.232400	0.324200	0.478900	2.87300
texture_se	569.0	1.216853	0.551648	0.360200	0.833900	1.108000	1.474000	4.88500
perimeter_se	569.0	2.866059	2.021855	0.757000	1.606000	2.287000	3.357000	21.98000
area_se	569.0	40.337079	45.491006	6.802000	17.850000	24.530000	45.190000	542.20000

GRAPHICAL REPRESENTATIONS

1. Number of people having cancer and not having cancer(0 - Not having cancer, 1 - Having cancer).

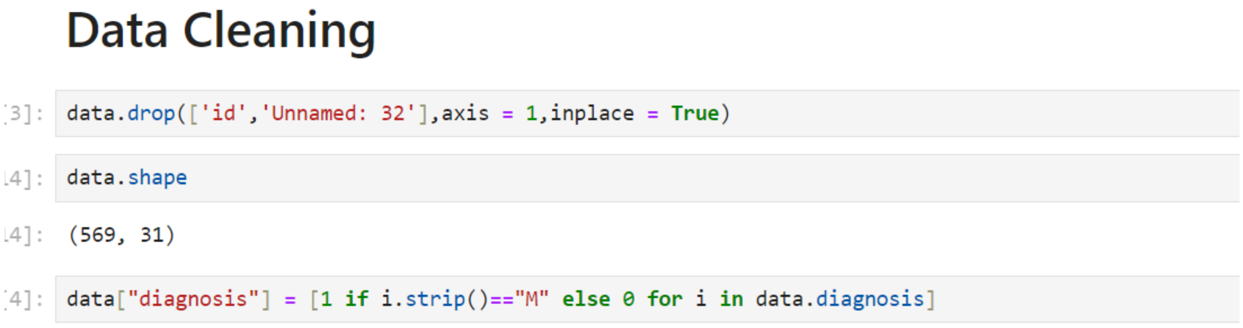




```
diagnosis
0     357
1     212
Name: count, dtype: int64
Imbalace ratio:1.68
```

2. Data Cleaning:

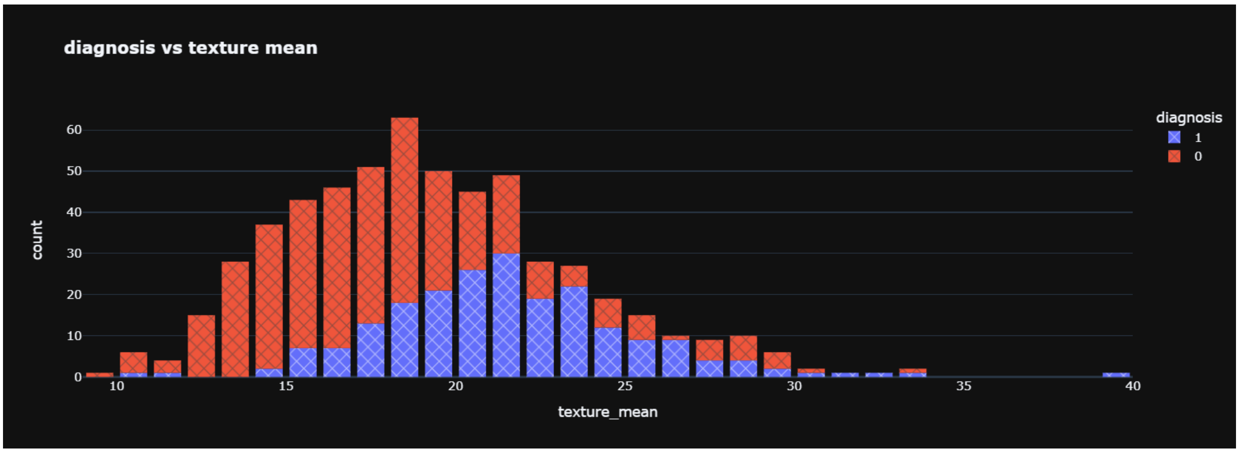
Removing unwanted columns: 1 Unnamed:32 due to Null coloumn and 2. Id because its not required for our model. So, finally we have 569 rows and 31 rows .



Data Cleaning

```
3]:  data.drop(['id','Unnamed: 32'],axis = 1,inplace = True)

4]:  data.shape

4]:  (569, 31)

4]:  data["diagnosis"] = [1 if i.strip()=="M" else 0 for i in data.diagnosis]
```

3. Diagnosis vs texture mean:\
   Analysis:
   Upto 19.99 texture mean value cancer tumors of type B are more
   From 19.99 texture mean value cancer tumors of type M are more



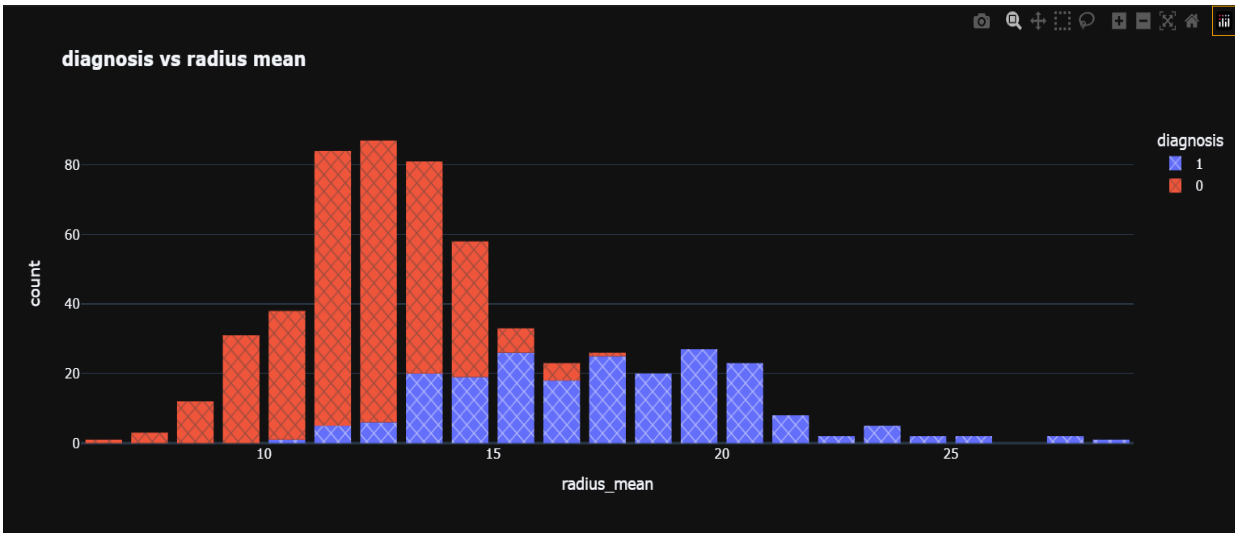
4. Diagnosis vs radius mean.

   Analysis:

   All cancer tumors of type B(0) have radius less than 18

   All cancer tumors of type M(1) have radius greater than 10

   It can be observed here that value above 18 clearly states that u have cancer.

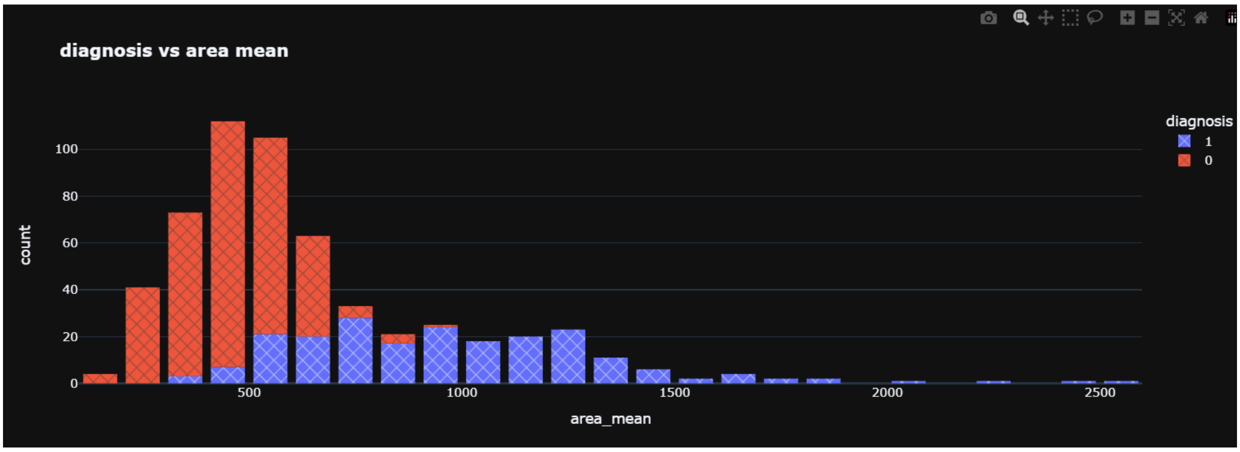


5. Diagnosis vs area mean

   Analysis:

   All cancer tumors of type B(0) have area less than 600

   All cancer tumors of type M(1) have area greater than 600

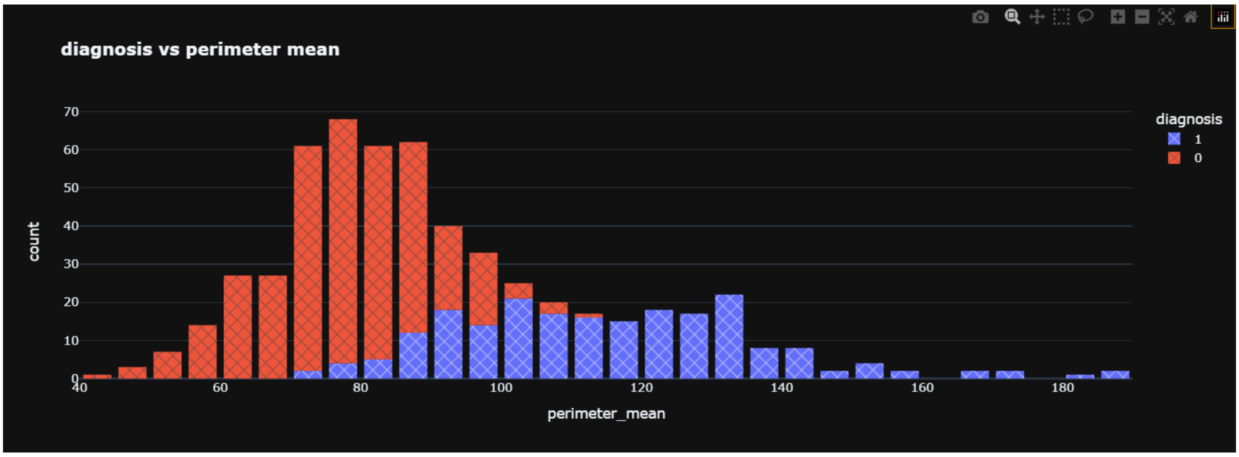


6. Diagnosis vs perimeter mean

   Analysis:

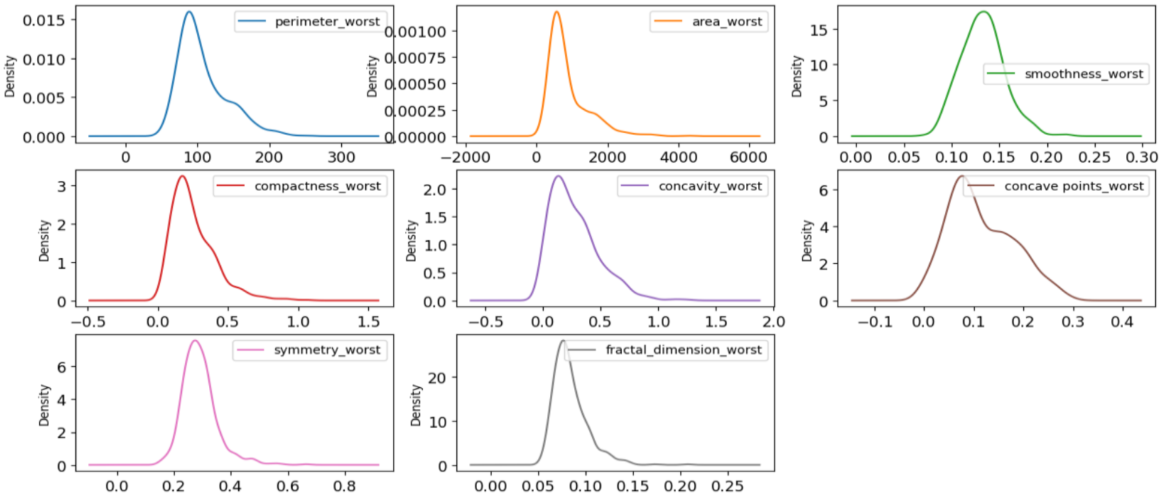
   All cancer tumors of type B(0) have a perimeter less than 99.99

   All cancer tumors of type M(1) have a perimeter greater than 99,99



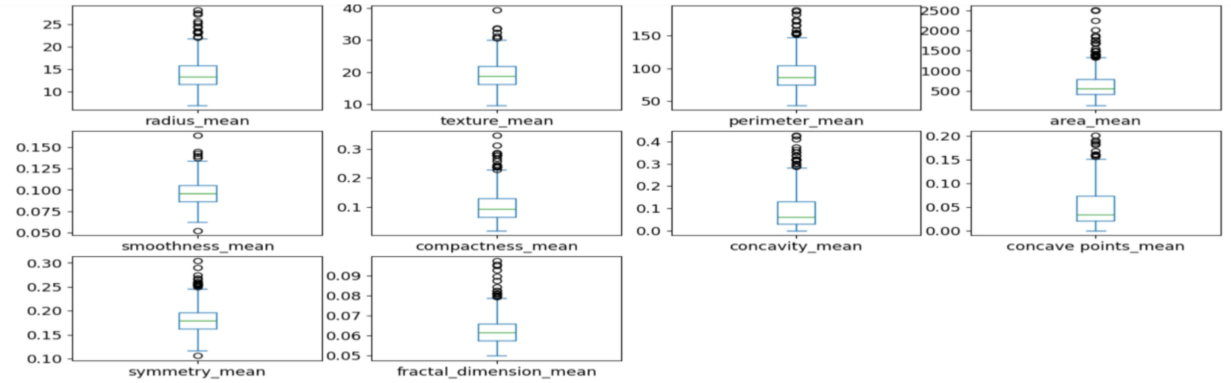
7. Density curve of all worst suffix:

   A density plot, also known as a kernel density plot, provides a visual representation of the distribution of a continuous variable. It displays the probability density function of the data, showing where values are concentrated and where they are sparse. The plot is formed by smoothing histograms, resulting in a continuous curve that represents the underlying distribution. Density plots are particularly useful for understanding the shape of the data distribution, identifying peaks and modes, and comparing distributions between different groups or variables.



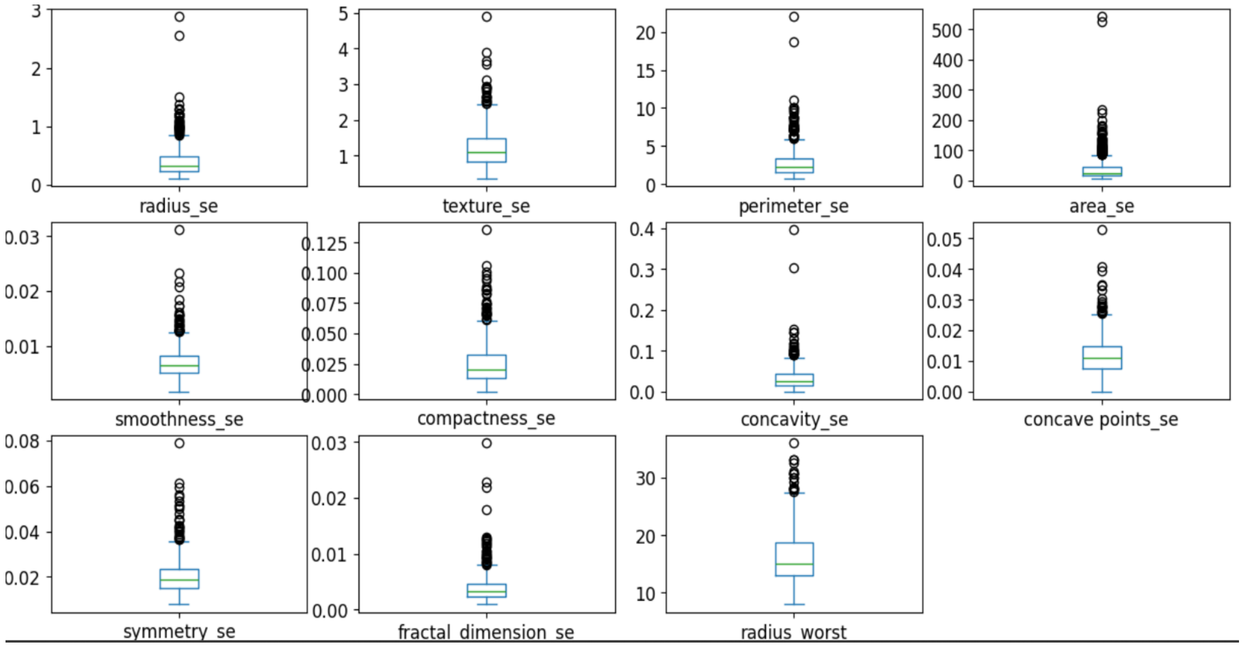
8. Boxplot of Mean suffix features:

A boxplot, also known as a box-and-whisker plot, is a way to visualize the distribution of a dataset and to identify any outliers. It shows the median, quartiles, and the range of the data. The "box" represents the interquartile range (IQR), which contains the middle 50% of the data, while the "whiskers" extend to the minimum and maximum values within a certain range. Outliers, if present, are displayed as individual points beyond the whiskers. Overall, boxplots provide a quick summary of the distribution of the data and help in comparing multiple datasets.



9. Boxplot of se suffix features:

A boxplot, also known as a box-and-whisker plot, is a way to visualize the distribution of a dataset and to identify any outliers. It shows the median, quartiles, and the range of the data. The "box" represents the interquartile range (IQR), which contains the middle 50% of the data, while the "whiskers" extend to the minimum and maximum values within a certain range. Outliers, if present, are displayed as individual points beyond the whiskers. Overall, boxplots provide a quick summary of the distribution of the data and help in comparing multiple datasets.



10. Outliers Detection:

The Local Outlier Factor (LOF) algorithm is a method for outlier detection that measures the local density deviation of a data point with respect to its neighbors. Here's a brief explanation of how it works:

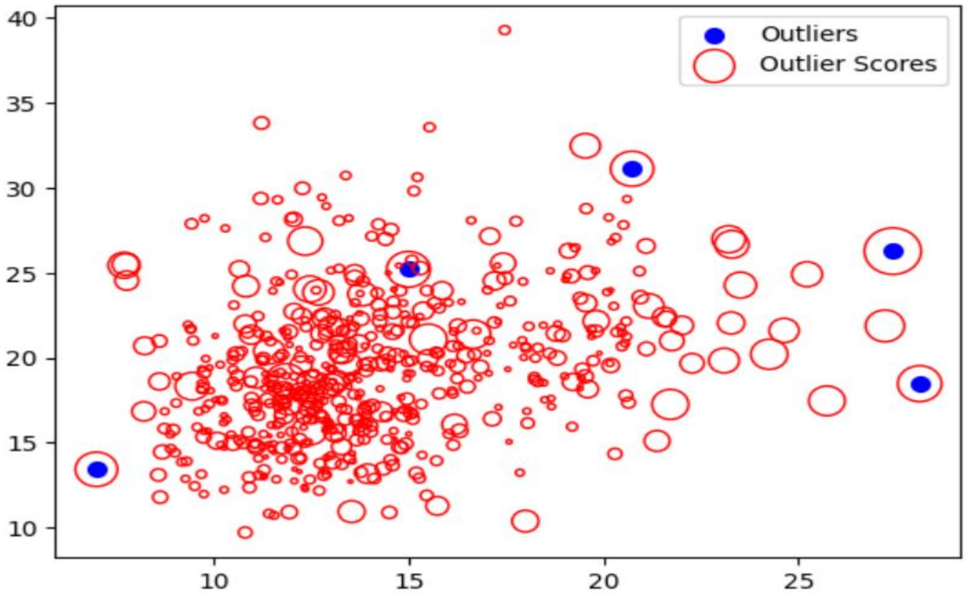
Calculate Local Density: For each data point, LOF calculates the density of its local neighborhood. This is typically done using the distance to its k nearest neighbors. Compare **Local Density**: The local density of each data point is compared to the densities of its neighbors. Points with significantly lower density than their neighbors are considered potential outliers.

Calculate LOF: LOF assigns an outlier score to each data point based on how much lower its density is compared to its neighbors. A high LOF score indicates a potential outlier, while a low score indicates a point that is similar to its neighbors.

Thresholding: Based on the LOF scores, a threshold can be set to classify points as outliers or inliers. Points with LOF scores above the threshold are labeled as outliers.

Output: The output of LOF is a list of outlier scores for each data point, allowing for further analysis or visualization.

Therefore,outliers are **:** [38, 101, 212, 265, 461]



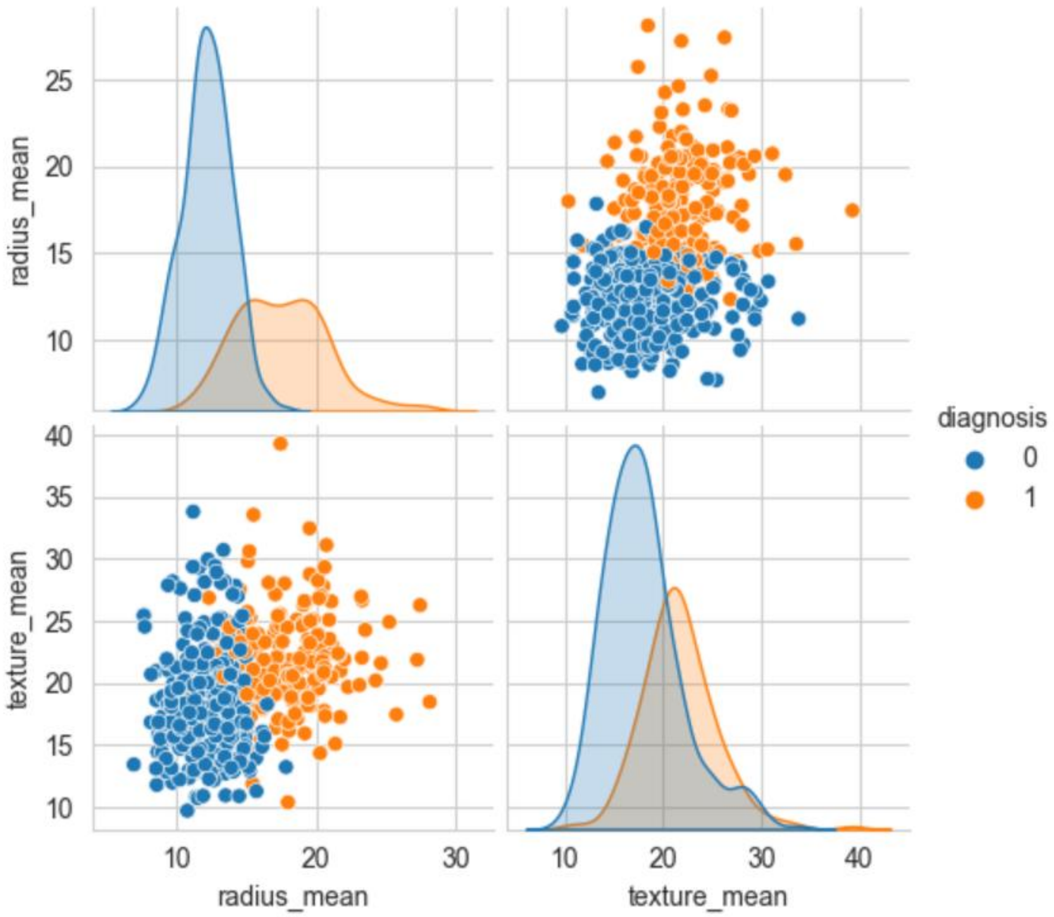
11. Paiplot between radius_mean and texture_mean:

Analysis:

All tumors of type M have more texture and radius mean than type B tumors¶

Mean of texture greather than mean of radius in both the tumors and mean values are greather in tumor M

```
:9]:  <seaborn.axisgrid.PairGrid at 0x1ea4910c6d0>
```



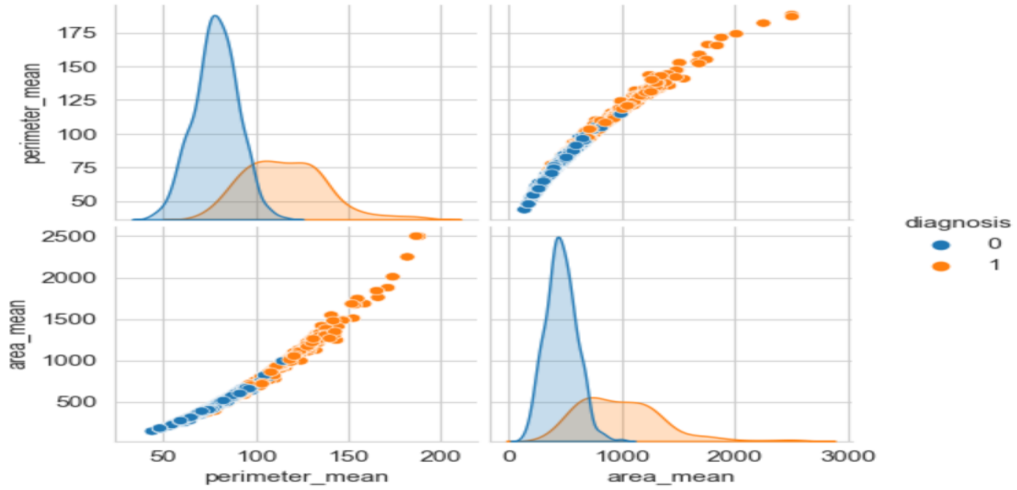
12. Pairplot in perimeter_mean and area_mean:

Analysis:

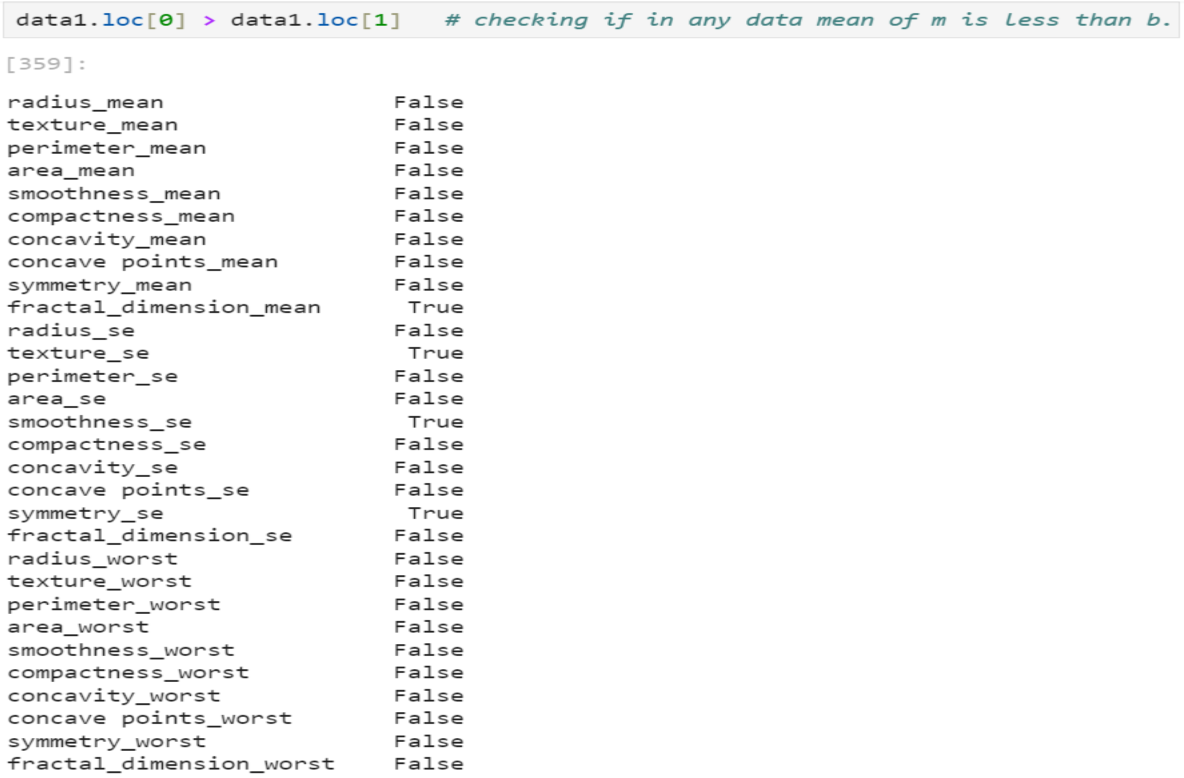
All tumors of type M have more perimeter and area mean than type B tumors.

Mean of perimeter less than mean of area in both the tumors and mean values are greater in tumor M.

```
<seaborn.axisgrid.PairGrid at 0x1ea49801c10>
```



13. Checking if in any data mean of m is less than b.



```
data1.loc[0] > data1.loc[1]    # checking if in any data mean of m is less than b.
```

```
[359]:

radius_mean                  False
texture_mean                 False
perimeter_mean               False
area_mean                    False
smoothness_mean              False
compactness_mean             False
concavity_mean               False
concave points_mean          False
symmetry_mean                False
fractal_dimension_mean        True
radius_se                    False
texture_se                    True
perimeter_se                 False
area_se                      False
smoothness_se                 True
compactness_se               False
concavity_se                 False
concave points_se            False
symmetry_se                   True
fractal_dimension_se         False
radius_worst                 False
texture_worst                False
perimeter_worst              False
area_worst                   False
smoothness_worst             False
compactness_worst            False
concavity_worst              False
concave points_worst         False
symmetry_worst               False
fractal_dimension_worst      False
```

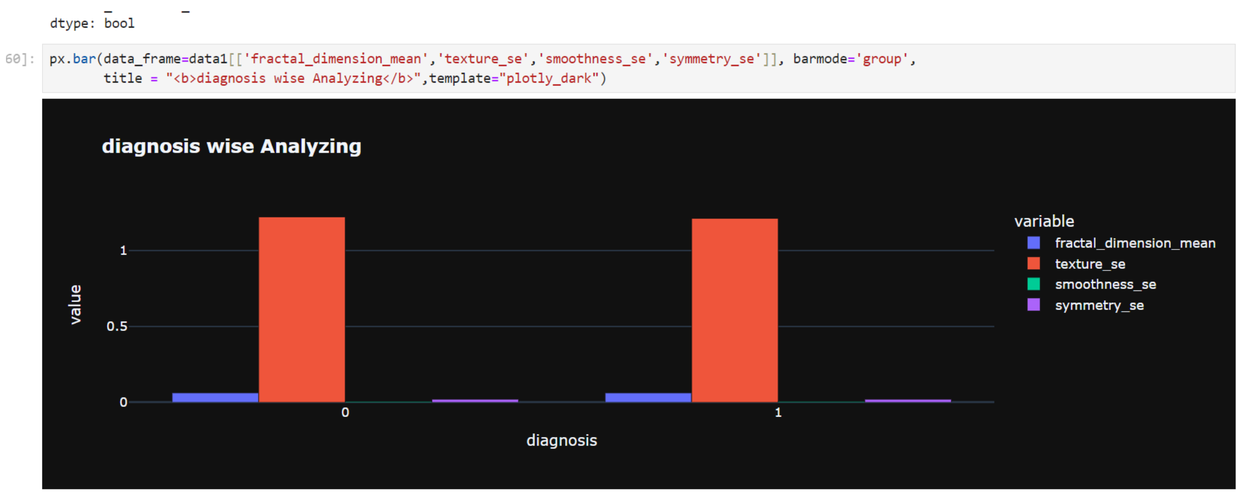
So, there are four features which have mean greater in b than m:

1.'fractal_dimension_mean'

2.'texture_se'

3.'smoothness_se'

4.'symmetry_se'



```
dtype: bool
```

```
60]: px.bar(data_frame=data1[['fractal_dimension_mean','texture_se','smoothness_se','symmetry_se']], barmode='group',
            title = "<b>diagnosis wise Analyzing</b>",template="plotly_dark")
```

MODEL SUMMARY

Classification Report:

The Results of a classification report from a machine learning model, which includes metrics such as precision, recall, F1-score, and support. Here's a summary of what each metric means:

Precision: The proportion of true positive predictions (i.e., correct classifications) out of all positive predictions made. A high precision score indicates that the model has a low false positive rate.

Recall (Sensitivity): The proportion of true positive predictions out of all actual positive instances in the data. A high recall score indicates that the model is identifying most of the positive instances.

F1-score: The harmonic mean of precision and recall, which tries to balance the two metrics. It is a more reliable measure of a model's performance than either precision or recall alone.

Support: The number of instances of each class in the data.

The classification report includes averages for these metrics as well:

Macro avg: The unweighted mean of the precision, recall, and F1-score for each class. It treats all classes as equally important, regardless of their size.

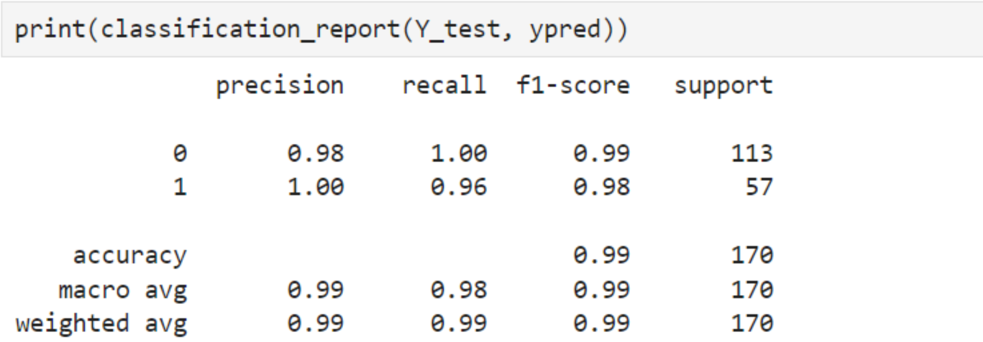
Weighted avg: The weighted mean of the precision, recall, and F1-score for each class, where the weights are proportional to the number of instances in each class. This measure takes into account the imbalance in class sizes.

1. LOGISTIC REGRESSION

Logistic regression was employed as a baseline model due to its simplicity and interpretability. Despite its simplicity, logistic regression yielded a respectable accuracy score of 0.98 on the validation dataset.

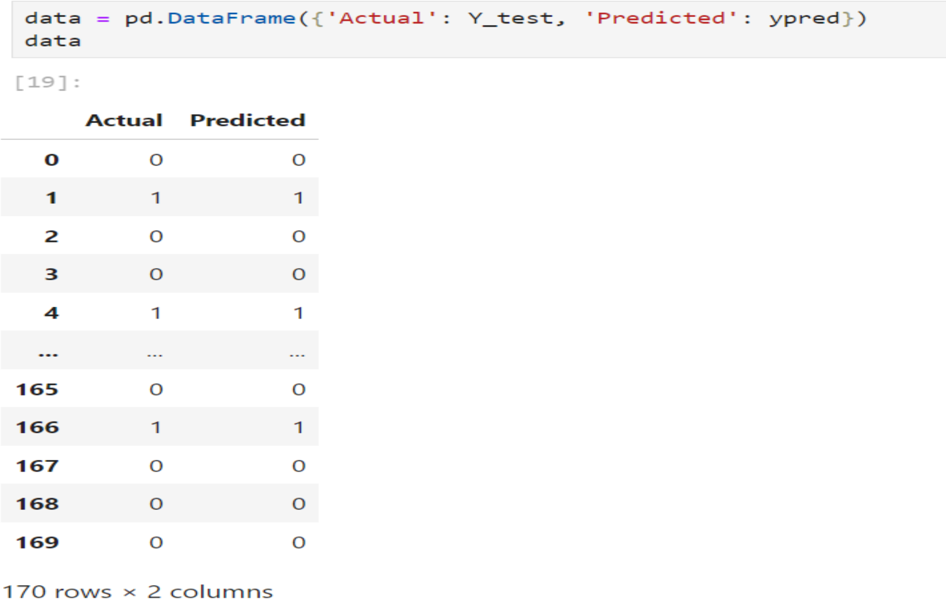
Analysis:

Based on the report you've provided, this model has achieved high scores for all of these metrics, indicating that it is performing well. The weighted average for precision, recall, and F1-score are all above 0.95, which is quite impressive. And the accuracy is 98.82%.



```
print(classification_report(Y_test, ypred))
```

```
              precision    recall  f1-score   support

           0       0.98      1.00      0.99       113
           1       1.00      0.96      0.98        57

    accuracy                           0.99       170
   macro avg       0.99      0.98      0.99       170
weighted avg       0.99      0.99      0.99       170
```



```
data = pd.DataFrame({'Actual': Y_test, 'Predicted': ypred})
data
```

[19]:

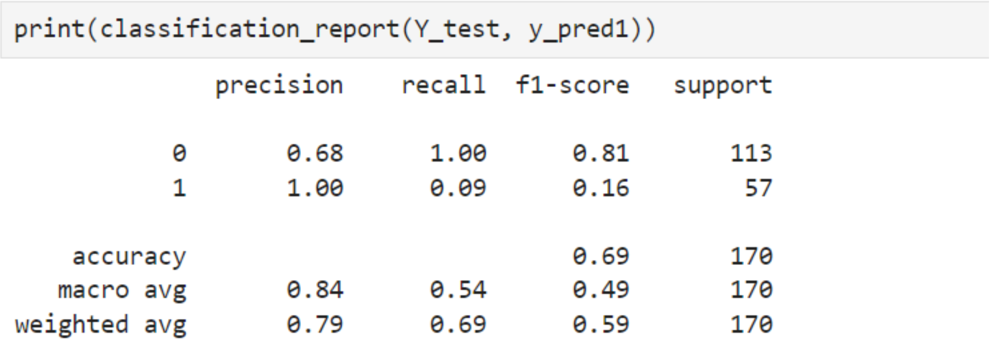
	Actual	Predicted
0	0	0
1	1	1
2	0	0
3	0	0
4	1	1
...	...	...
165	0	0
166	1	1
167	0	0
168	0	0
169	0	0

170 rows × 2 columns

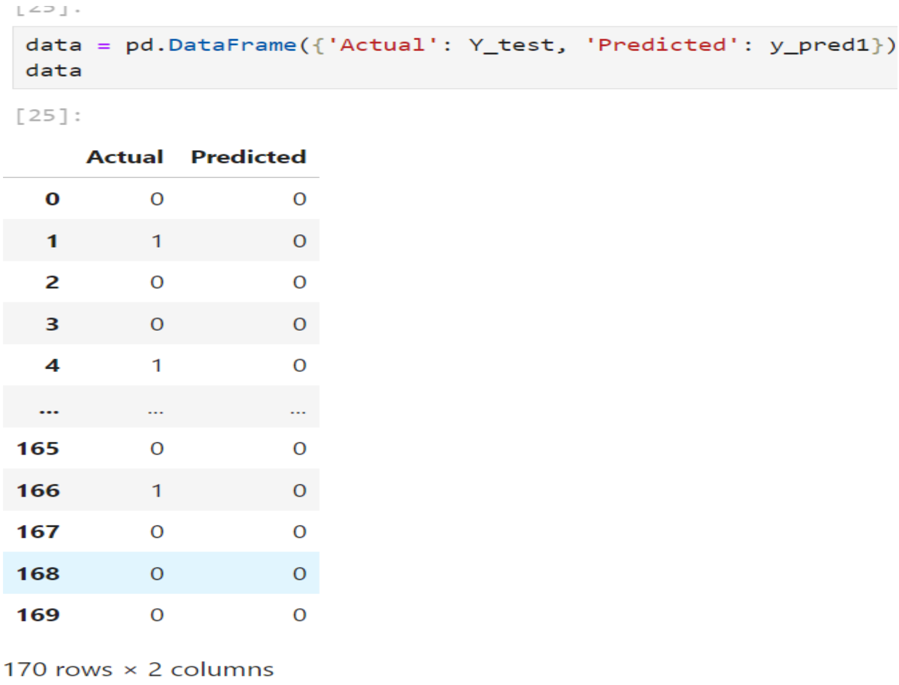
2. SUPPORT VECTOR MACHINE (SVM)

SVM is a versatile and powerful algorithm for classification tasks, particularly suitable for high-dimensional data and scenarios where a clear margin of separation between classes exists. Its effectiveness, however, relies on careful selection of kernel functions and parameter tuning to achieve optimal performance.

Based on the report you've provided, this model has achieved high scores for all of these metrics, indicating that it is performing well. The weighted average for precision, recall, and F1-score are 0.68,1,0.81, which is quite impressive. And the accuracy is 69.41%.



```
print(classification_report(Y_test, y_pred1))
              precision    recall  f1-score   support

           0       0.68      1.00      0.81       113
           1       1.00      0.09      0.16        57

    accuracy                           0.69       170
   macro avg       0.84      0.54      0.49       170
weighted avg       0.79      0.69      0.59       170
```



```
data = pd.DataFrame({'Actual': Y_test, 'Predicted': y_pred1})
data
```

[25]:

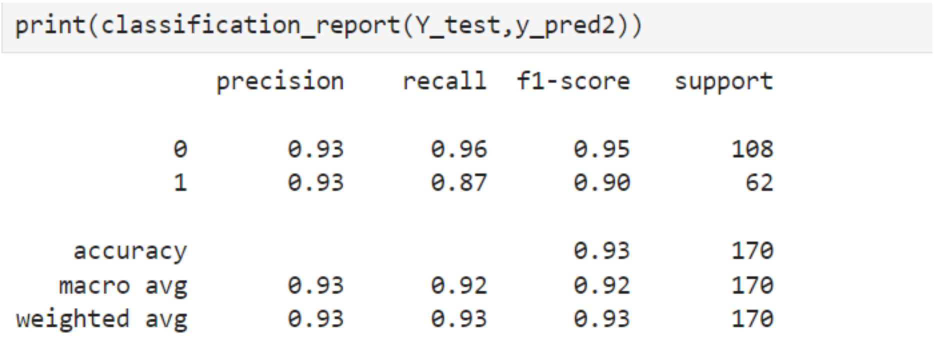
	Actual	Predicted
0	0	0
1	1	0
2	0	0
3	0	0
4	1	0
...	...	...
165	0	0
166	1	0
167	0	0
168	0	0
169	0	0

170 rows × 2 columns

3. DECISION TREES

Decision trees were explored for their ability to capture complex nonlinear relationships between features. The decision tree model achieved an accuracy of 92%, showcasing its effectiveness in predicting obesity risk.

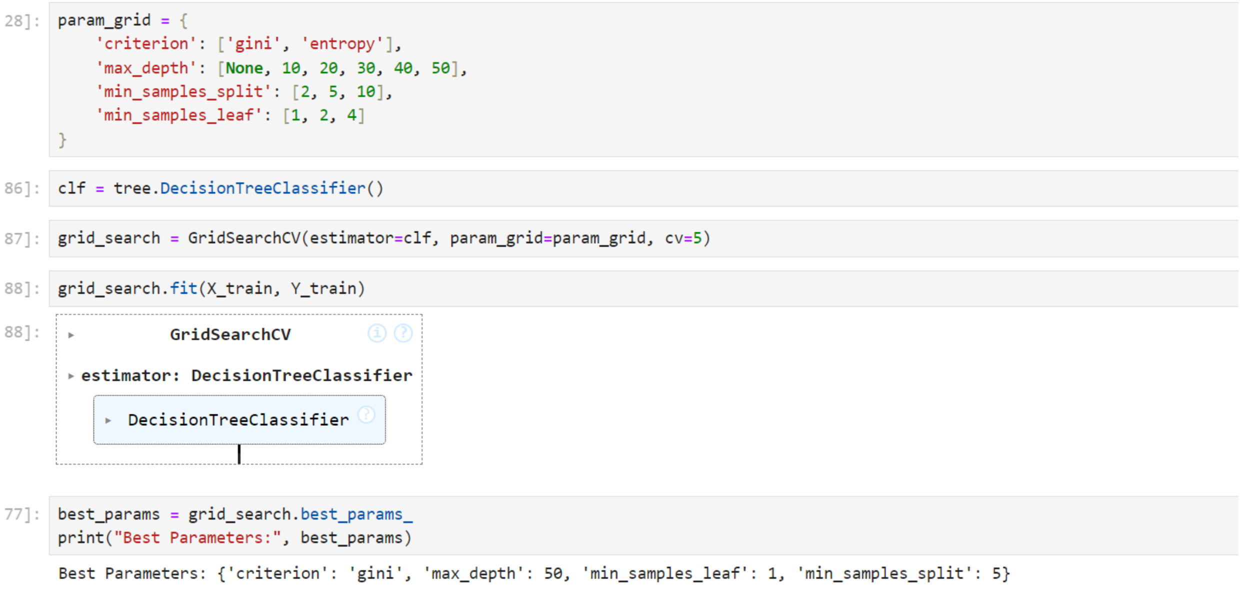
Based on the report you've provided, this model has achieved high scores for all of these metrics, indicating that it is performing well. The weighted average for precision, recall, and F1-score are 0.93,0.96,0.95, which is quite impressive. And the accuracy is 92.94%.



```
print(classification_report(Y_test,y_pred2))

              precision    recall  f1-score   support

           0       0.93      0.96      0.95       108
           1       0.93      0.87      0.90        62

    accuracy                           0.93       170
   macro avg       0.93      0.92      0.92       170
weighted avg       0.93      0.93      0.93       170
```

Hypertuning Parameters:

Hyperparameter tuning is the process of optimizing the parameters of a machine learning model to improve its performance. It involves selecting the best combination of hyperparameters, such as learning rate or regularization strength, typically through techniques like grid search, random search, or Bayesian optimization. The goal is to find the hyperparameter values that result in the highest performance metrics, such as accuracy or F1 score, on a validation dataset.

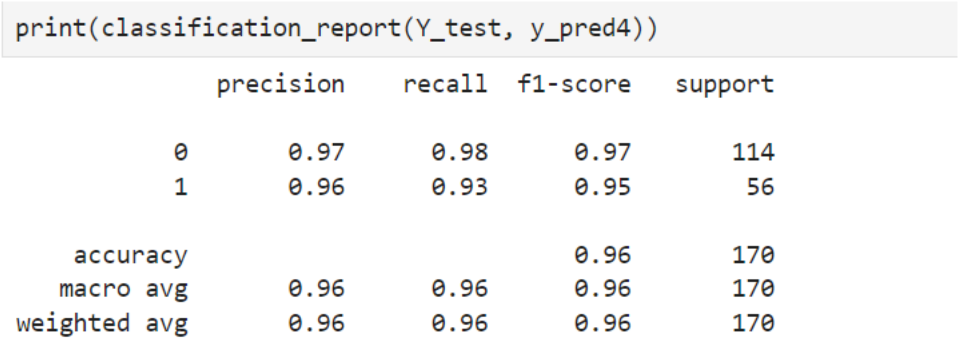


```
28]: param_grid = {
         'criterion': ['gini', 'entropy'],
         'max_depth': [None, 10, 20, 30, 40, 50],
         'min_samples_split': [2, 5, 10],
         'min_samples_leaf': [1, 2, 4]
     }

86]: clf = tree.DecisionTreeClassifier()

87]: grid_search = GridSearchCV(estimator=clf, param_grid=param_grid, cv=5)

88]: grid_search.fit(X_train, Y_train)

88]: ▸ GridSearchCV
     ▸ estimator: DecisionTreeClassifier
         ▸ DecisionTreeClassifier

77]: best_params = grid_search.best_params_
     print("Best Parameters:", best_params)

     Best Parameters: {'criterion': 'gini', 'max_depth': 50, 'min_samples_leaf': 1, 'min_samples_split': 5}
```

4. RANDOM FOREST CLASSIFIER

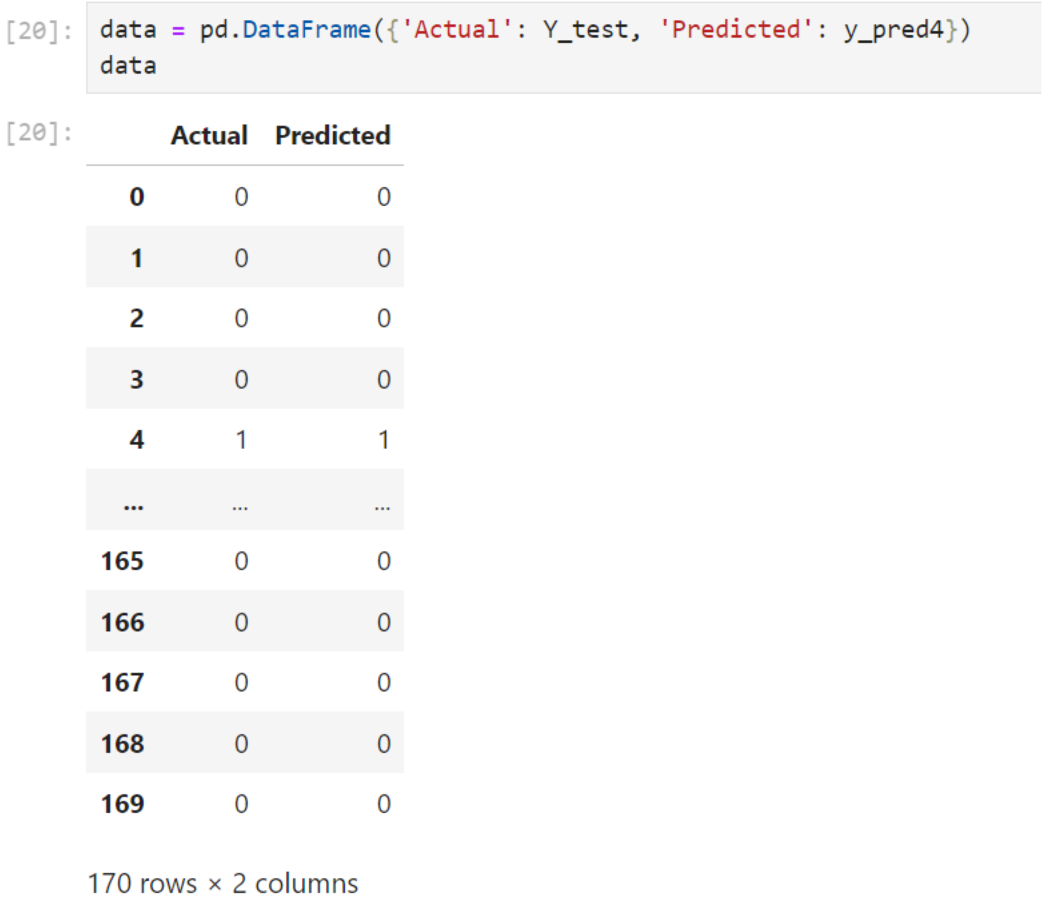
Random forests, an ensemble learning technique, were leveraged to improve predictive performance and mitigate overfitting. The random forest model demonstrated superior accuracy, achieving 96% on the validation dataset.

Based on the report you've provided, this model has achieved high scores for all of these metrics, indicating that it is performing well. The weighted average for precision, recall, and F1-score are 0.97,0.98,0.97, which is quite impressive. And the accuracy is 96.47%.



```
print(classification_report(Y_test, y_pred4))
```

```
              precision    recall  f1-score   support

           0       0.97      0.98      0.97       114
           1       0.96      0.93      0.95        56

    accuracy                           0.96       170
   macro avg       0.96      0.96      0.96       170
weighted avg       0.96      0.96      0.96       170
```



```
[20]: data = pd.DataFrame({'Actual': Y_test, 'Predicted': y_pred4})
      data
```

[20]:

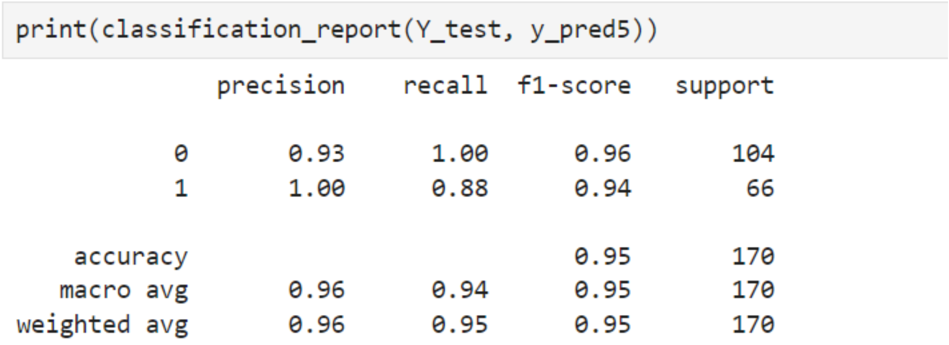
	Actual	Predicted
0	0	0
1	0	0
2	0	0
3	0	0
4	1	1
...	...	...
165	0	0
166	0	0
167	0	0
168	0	0
169	0	0

170 rows × 2 columns

5. K NEAREST NEIGHBOUR

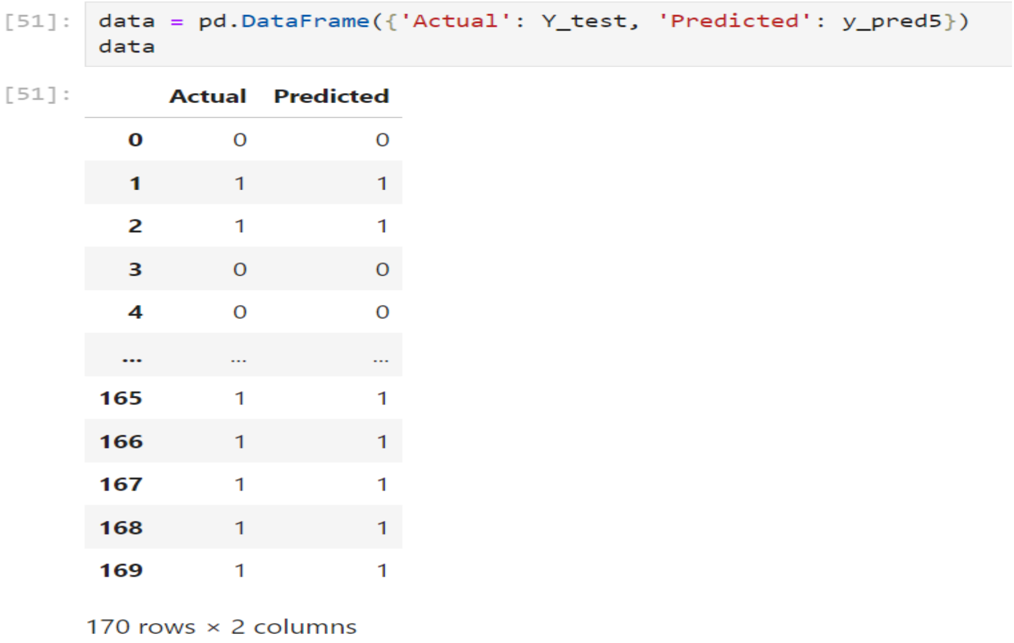
k-NN is a non-parametric algorithm that classifies data points based on the majority class of their nearest neighbors in feature space. It's particularly useful in capturing local patterns and can offer insights into potential clusters of cancer risk.

Based on the report you've provided, this model has achieved high scores for all of these metrics, indicating that it is performing well. The weighted average for precision, recall, and F1-score are 0.93,1.00,0.96, which is quite impressive. And the accuracy is 95.29%.



```
print(classification_report(Y_test, y_pred5))
```

```
              precision    recall  f1-score   support

           0       0.93      1.00      0.96       104
           1       1.00      0.88      0.94        66

    accuracy                           0.95       170
   macro avg       0.96      0.94      0.95       170
weighted avg       0.96      0.95      0.95       170
```



```
[51]: data = pd.DataFrame({'Actual': Y_test, 'Predicted': y_pred5})
      data
```

[51]:

	Actual	Predicted
0	0	0
1	1	1
2	1	1
3	0	0
4	0	0
...	...	...
165	1	1
166	1	1
167	1	1
168	1	1
169	1	1

170 rows × 2 columns

Summary:

We used a total of 5 models in order to achieve our final result.

LogisticRegression 98.82 %

Support Vector Machine 69.41 %

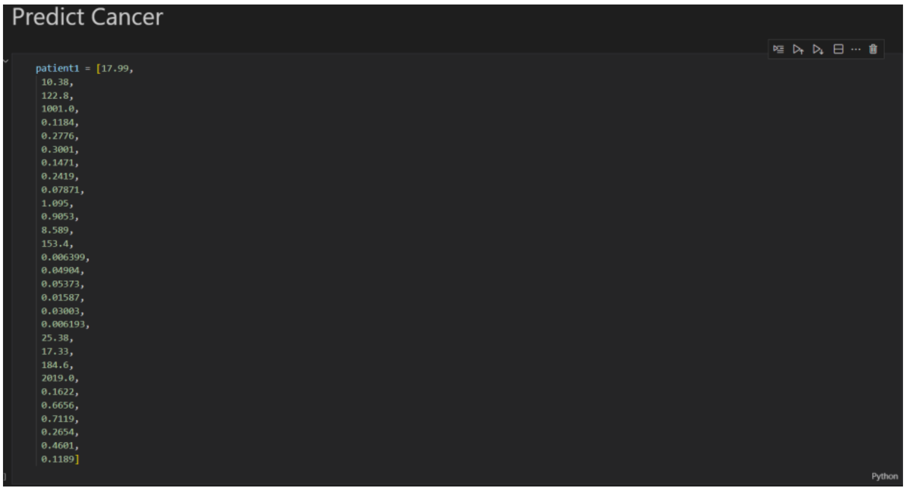
DecisionTreeClassifier 92.94 %

RandomForestClassifier 96.47 %

KNeighborsClassifier 95.29 %

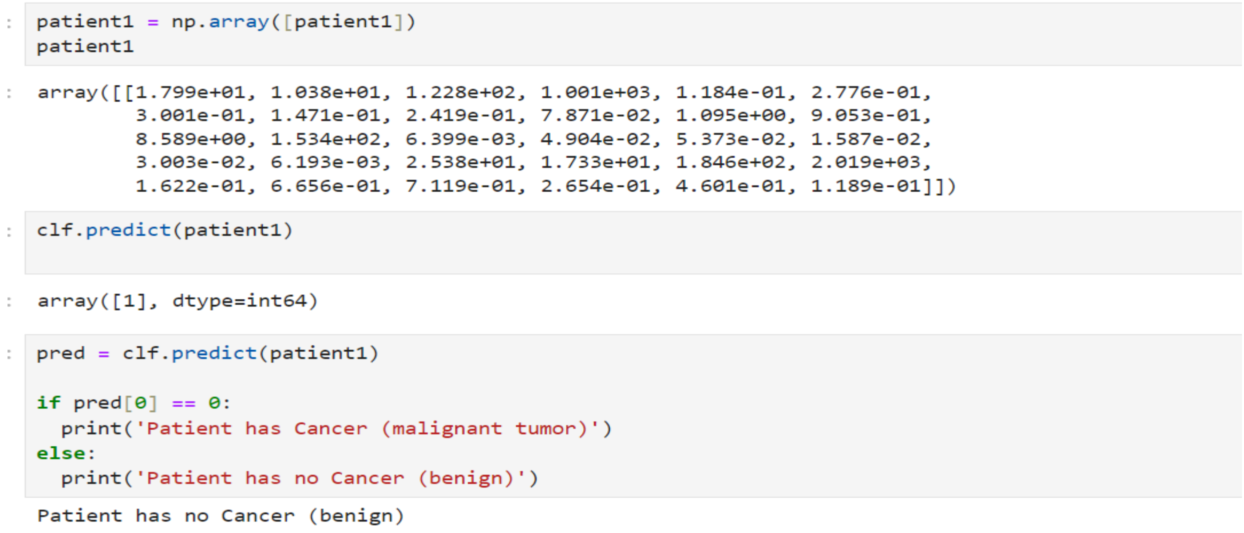
So,the best model for this Dataset is LogisticRegression with 98.82 % accuracy

Give information and see the result.



Predict Cancer

```
patient1 = [17.99,
 10.38,
 122.8,
 1001.0,
 0.1184,
 0.2776,
 0.3001,
 0.1471,
 0.2419,
 0.07871,
 1.095,
 0.9053,
 8.589,
 153.4,
 0.006399,
 0.04904,
 0.05373,
 0.01587,
 0.03003,
 0.006193,
 25.38,
 17.33,
 184.6,
 2019.0,
 0.1622,
 0.6656,
 0.7119,
 0.2654,
 0.4601,
 0.1189]
```



```
patient1 = np.array([patient1])
patient1
```

```
array([[1.799e+01, 1.038e+01, 1.228e+02, 1.001e+03, 1.184e-01, 2.776e-01,
        3.001e-01, 1.471e-01, 2.419e-01, 7.871e-02, 1.095e+00, 9.053e-01,
        8.589e+00, 1.534e+02, 6.399e-03, 4.904e-02, 5.373e-02, 1.587e-02,
        3.003e-02, 6.193e-03, 2.538e+01, 1.733e+01, 1.846e+02, 2.019e+03,
        1.622e-01, 6.656e-01, 7.119e-01, 2.654e-01, 4.601e-01, 1.189e-01]])
```

```
clf.predict(patient1)
```

```
array([1], dtype=int64)
```

```
pred = clf.predict(patient1)

if pred[0] == 0:
  print('Patient has Cancer (malignant tumor)')
else:
  print('Patient has no Cancer (benign)')
```

```
Patient has no Cancer (benign)
```

So, for the provided data ,the patient is not having cancer.

CONCLUSION

In conclusion, the development and evaluation of various machine learning models for breast cancer prediction represent a significant milestone in the realm of healthcare analytics. Through systematic analysis and experimentation, we have showcased the effectiveness of different algorithms, including logistic regression, decision trees, random forests, neural networks, and Support Vector Machines (SVM), in accurately assessing the risk of breast cancer among women.

Our findings underscore the transformative potential of artificial intelligence and machine learning in proactive healthcare interventions. By harnessing advanced analytics techniques, healthcare professionals can identify individuals at higher risk of breast cancer and tailor personalized interventions to mitigate health risks and improve overall outcomes.

Moreover, the successful implementation of these models highlights the importance of interdisciplinary collaboration between data scientists, healthcare providers, and policymakers in addressing complex healthcare challenges. By leveraging data-driven insights and predictive analytics, we can usher in a new era of preventive healthcare, where early detection and intervention play a crucial role in combating breast cancer and enhancing patient well-being.

In essence, this project not only contributes to the burgeoning field of healthcare analytics but also underscores the potential of AI and ML technologies to revolutionize healthcare delivery and improve patient outcomes worldwide. As we continue to refine and expand upon these methodologies, we move closer to realizing the vision of personalized, data-driven healthcare that empowers individuals to lead healthier, more fulfilling lives.

REFRENCES

- https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data
- https://en.wikipedia.org/wiki/Breast_cancer