CLARK UNIVERSITY

STAT 4650: Final Project

Breast Cancer Diagnosis and Prediction by Machine Learning

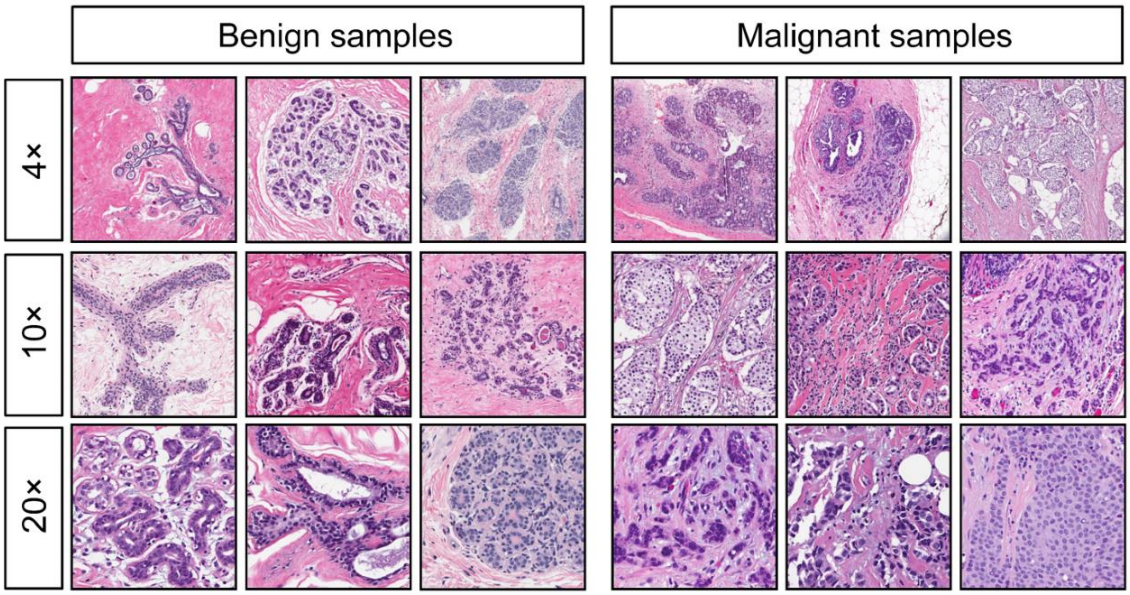
Supervised by:
Prof Yue Gao

Group Members:
Aashiya Aryal
Anuj Ursal
Rosy Shrestha

May 2022

Table of Contents

Breast Cancer Diagnosis and Prediction by Machine Learning



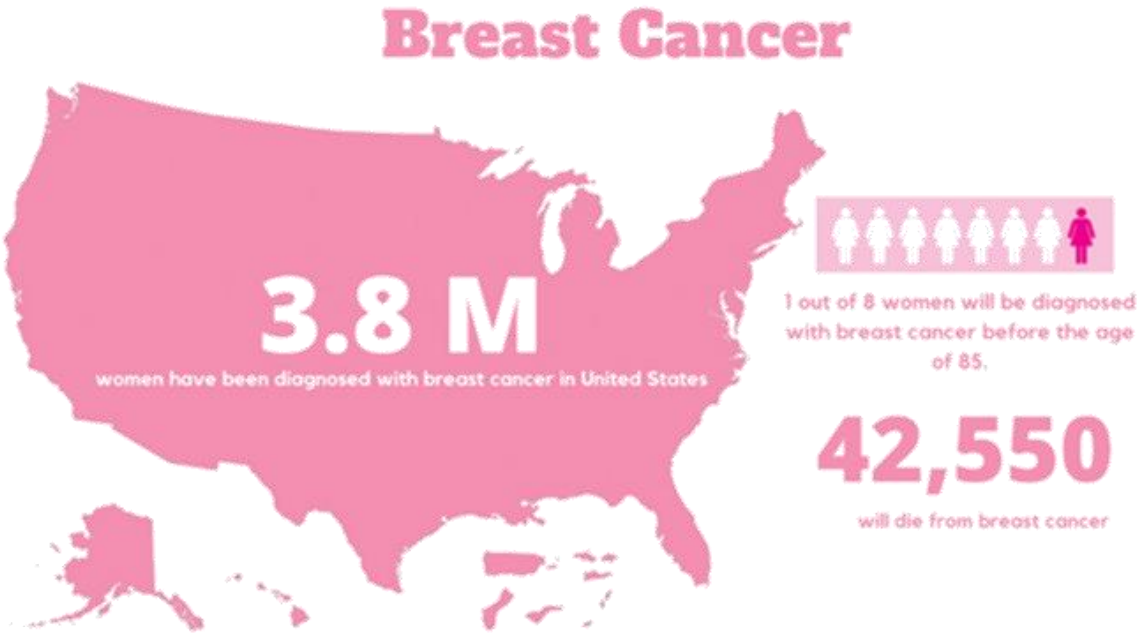
Abstract

Breast cancer is the most common type of cancer in women worldwide, and it has the second-highest fatality rate. Diagnosis of breast cancer is performed when an abnormal lump is found. After a tumor is detected, the doctor will conduct a diagnosis to determine whether it is malignant or benign. Malignant means cancerous and benign means non-cancerous. The severity of the disease can be reduced if detected and diagnosed early.

Machine learning could be of great help to detect if the cancer is benign or malignant using measures of area, smoothness, texture, etc., from a digitized image of cell nuclei. Machine learning can also contribute to the process of early diagnosis of breast cancer. Breast cancer detection is a time-consuming and difficult process, given the fact that a pathologist must scan large areas of benign tissue to locate spots of malignancy. Using machine learning to detect cancer can help to reduce the dependence on the pathologist which is more useful in regions where pathologists are not available.

In this study, we applied four machine learning algorithms such as K-nearest neighbors, Random Forest, Logistic Regression, and Support Vector Machine (SVM) in the Breast Cancer dataset. After getting the results, a performance evaluation and comparison is done.

1. Introduction



Breast cancer is a malignant tumor that is in or around breast tissue which usually begins as a lump/calcium deposit that develops from abnormal cell growth. Most breast lumps are benign, but few of them can be precancerous or cancerous. Breast cancer can be localized which means initially appearing within the breast or metastatic that spreads to another part of the body. It is classified as primary or metastatic. The malignant tumor which develops within the breast tissue is called primary breast cancer which sometimes can also be found when it is spread to lymph nodes located nearby in the armpit. When cancer cells located in the breast break away and then travel to another organ or part of the body metastatic breast cancer is formed.

Breast Cancer is the second major cause of women's death. The disease accounts for 1 in 3 of new female cancers annually. Each year in the United States, about 255,000 cases of breast cancer are diagnosed in women and about 2,300 in men. About 42,000 women and 500 men in the U.S. die each year from breast cancer. Currently, there are more than 3.8 million women who have been diagnosed with breast cancer in the United States. Since 2007, the count of women aged 50 and over who have died of breast cancer has decreased while the number of women under age 50 who have died of breast cancer is steady and from 2013 to 2018, the death rate for women with breast cancer dropped by 1% each year. (Cancer.net, 2022).

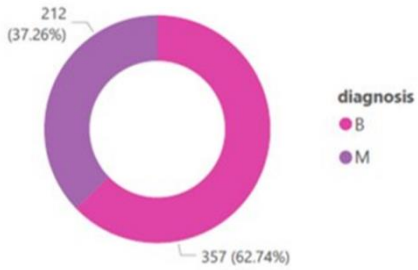
1.1 Objective

The objective of this project is to develop a machine learning model to determine whether a breast cancer is benign or malignant. We will achieve this by fitting a model that can predict the discrete class of new input using machine learning classification algorithms.

Research Question – Predict whether the tumor cell is Malignant or Benign

1.2 Data Description

The dataset is obtained from Kaggle. The data set contains 569 rows and 32 columns. These 33 features are extracted using information from a digitized image of the fine-needle aspiration test (FNA) of a breast mass.



We will create a model that can categorize a breast cancer tumor using two types of training classification:

1= Malignant (Cancer) - Present
0= Benign (non-cancerous) -Not present

Link - https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/code

1.2.1 Attribute Information

Dependent Variable	Independent Variable
Diagnosis	Id, radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave_points_mean, symmetry_mean, factual_dimension_mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave_points_se, symmetry_se, fracal_Dimension_Se, radius_worst, compactness_worth, concavity_worst, concave_points_worst, symmetry_worst, fractual_dimension_worth, Unnamed: 32

1.3 Data Preparation

Data preparation involves multiple activities such as handling the missing values, assigning numerical values to categorical data, selecting attributes and feature extraction.

1.3.1 Checking Null and Missing Values

:t	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst	Unnamed: 32
e	False	False	False	False	False	False	False	False	True
e	False	False	False	False	False	False	False	False	True
e	False	False	False	False	False	False	False	False	True
e	False	False	False	False	False	False	False	False	True
e	False	False	False	False	False	False	False	False	True
..	...	...	...	...	...	...	...	...	...
e	False	False	False	False	False	False	False	False	True
e	False	False	False	False	False	False	False	False	True
e	False	False	False	False	False	False	False	False	True
e	False	False	False	False	False	False	False	False	True
e	False	False	False	False	False	False	False	False	True

It has one column with missing values - Unnamed: 32. We can see from the data that id and Unnamed: 32 are unnecessary columns, so we will drop them.

1.3.2 Statistical Description of Data

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	f
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	
mean	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919	0.181162	
std	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803	0.027414	
min	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	0.106000	
25%	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	0.161900	
50%	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500	0.179200	
75%	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	0.195700	
max	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	0.304000	

8 rows × 30 columns

Statistic characteristic such as "mean," "median," "variance," etc. is calculated. These characteristics are required for future investigation and the development of a clear prediction model.

2. Methodology

1. **Data Analysis**: The dataset is obtained from Kaggle. The dataset features – shape, missing values, and data types will be observed.

2. **Data Processing**: After analysis, data cleaning, selecting attributes, and features extraction is performed. Variables will be analyzed to find the patterns and correlations

3. **Build Model** - Modelling will be done using Logistic Regression, Random Forest, SVC, and KNN, with fit performance measured using metrics like Accuracy, Precision, and Recall.

4. **Result**: Based on the performance and evaluation metrics, we will choose the best model.

All of the experiments on the machine learning algorithms presented in this work were carried out with Python programming language.

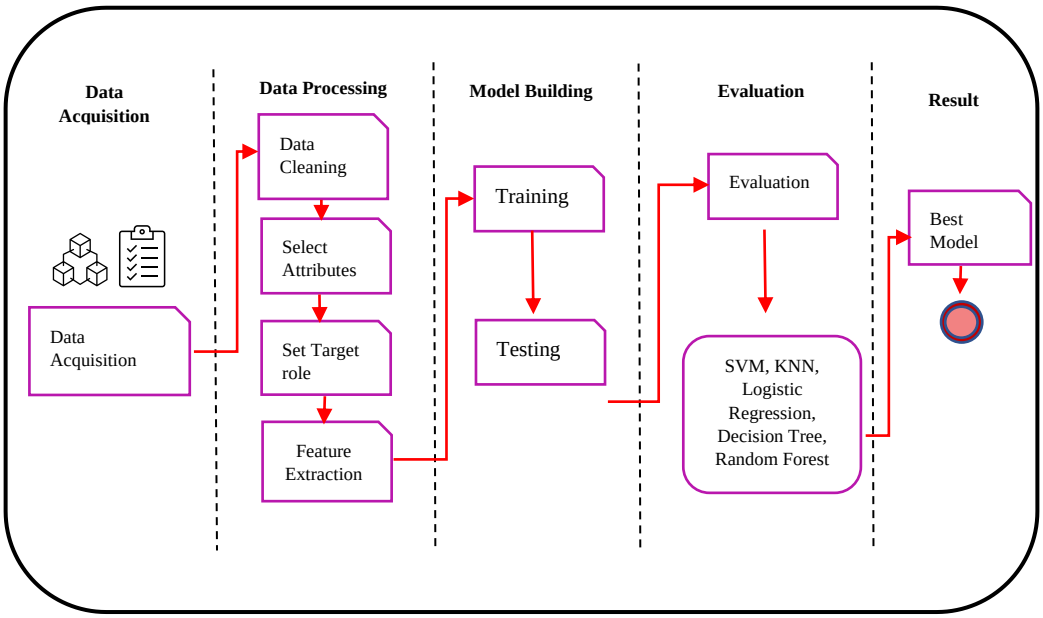


Figure 1: Project Methodology

2.1.1 Feature Selection using Correlation Heatmap

When the number of features with similar characteristics is high, there is a decrease in accuracy in machine learning models. Therefore, we used feature selection on the data to increase the accuracy of the models and minimize overfitting. To choose the characteristics, we employed a correlation matrix with heatmap visualization. We chose only one feature to represent all the characteristics in each set of strongly associated features to avoid overfitting and reserve most of the information.

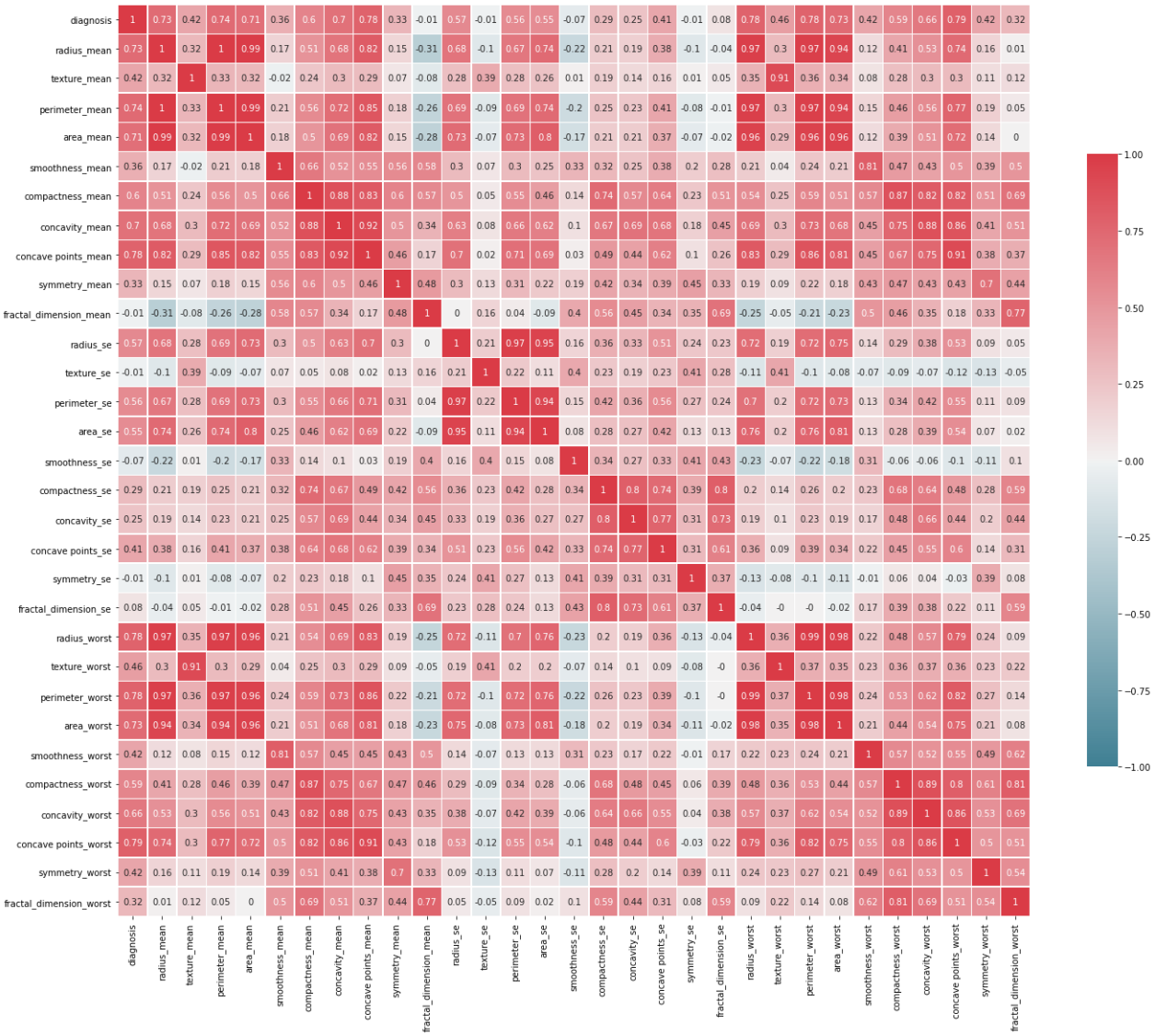


Figure 2 Correlation Heatmap

Observation: We can see that some of the features like perimeter_mean and radius_mean is very highly corelated. To observe if there is multicollinearity or not, we further generate scatter plot matrix with mean columns.

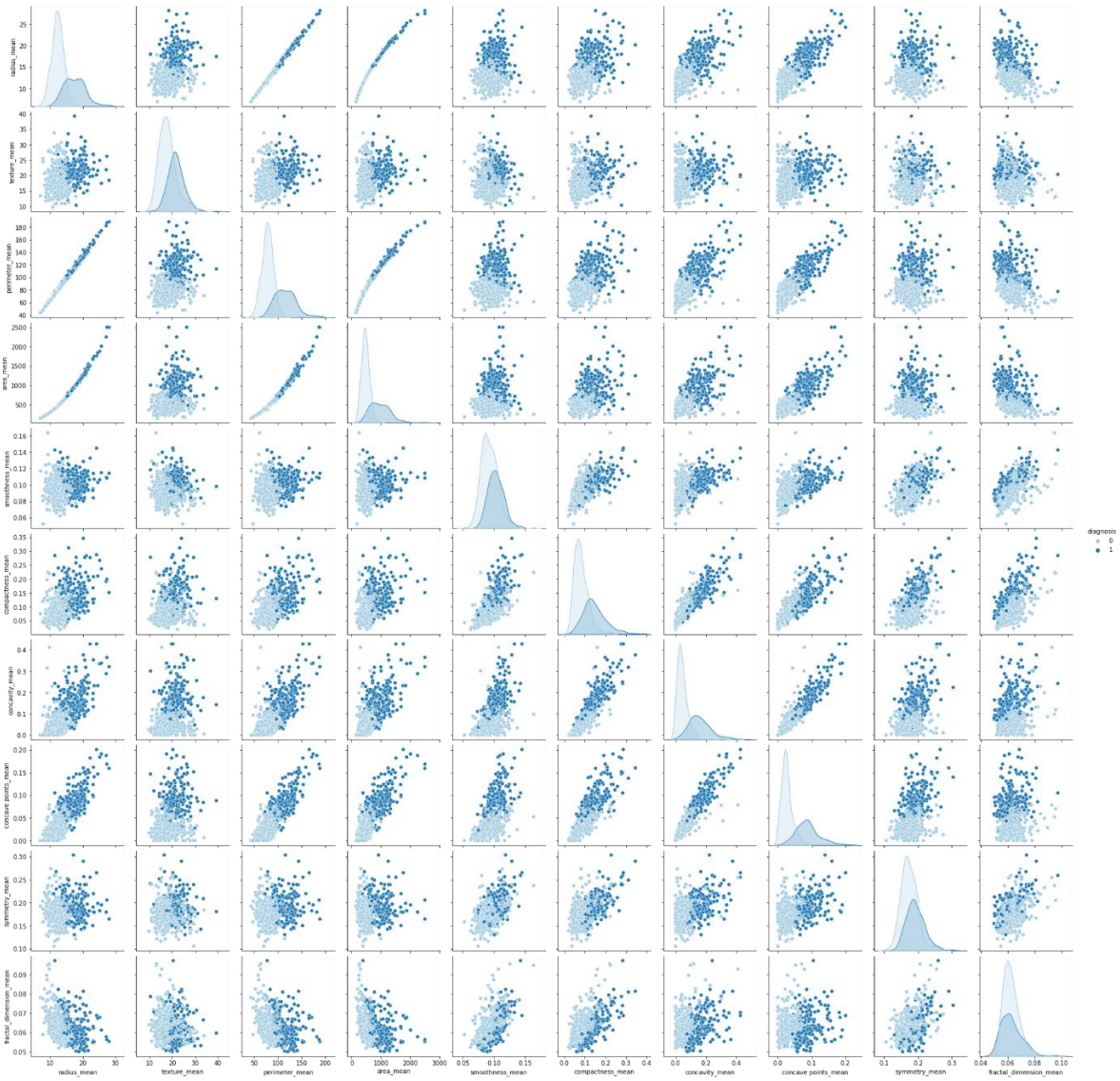


Figure 3 Scatter Plot Matrix

We can confirm that some variables are multicollinear. For example, given the perimeter mean and area mean columns, the radius mean column has a correlation of 1 and 0.99, respectively. This implies that the features consist of similar information. So, for further analysis, we can select only one features from the three features above.

The difference between the "mean" and "worst" columns is another example of multicollinearity. The radius_mean and radius_worst columns, for example, have a 0.97 correlation which is very high.

Lastly, we can observe linear relation between three other features which are concave point, compactness and concavity. Any one features can be selected among this three.

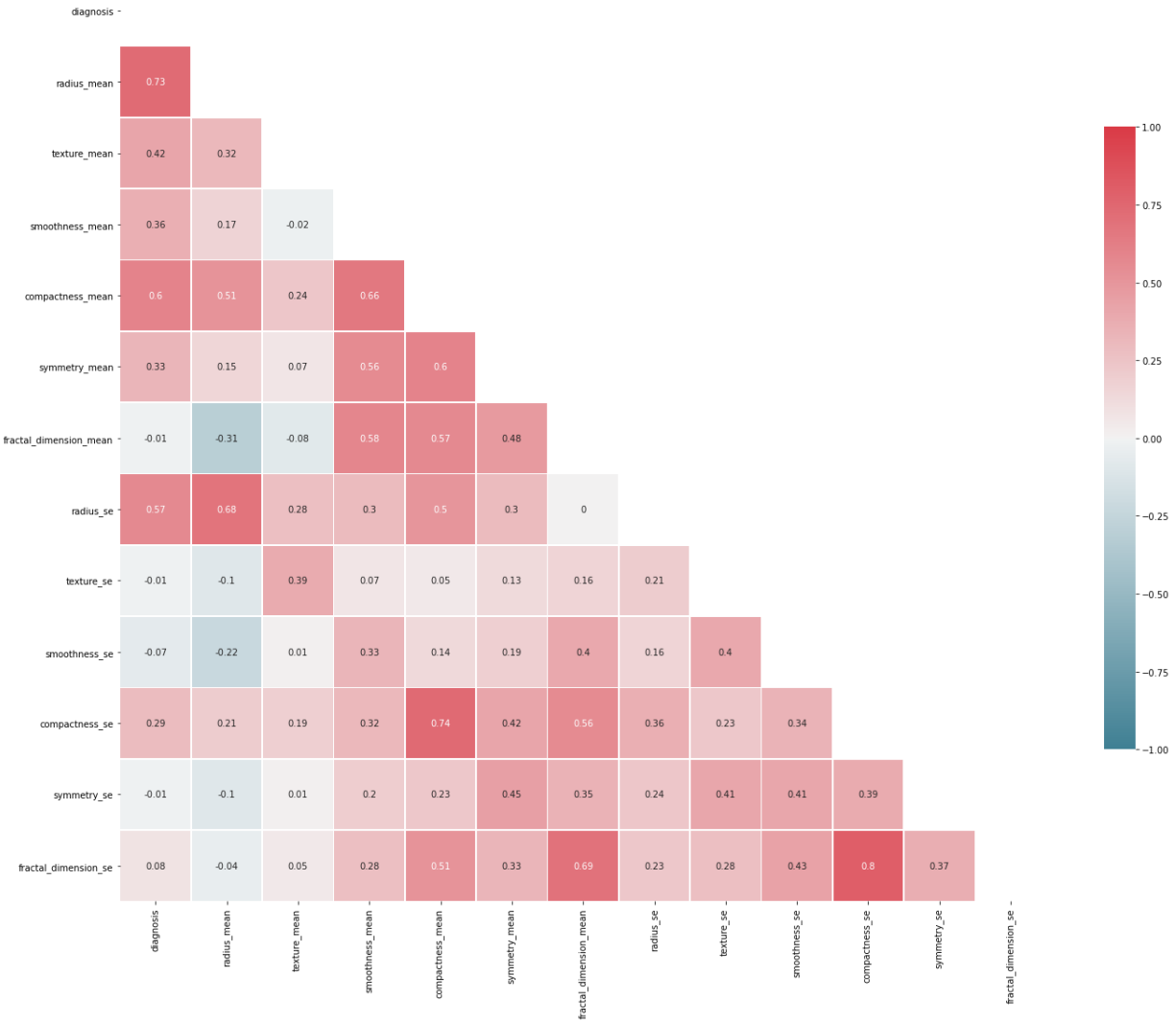
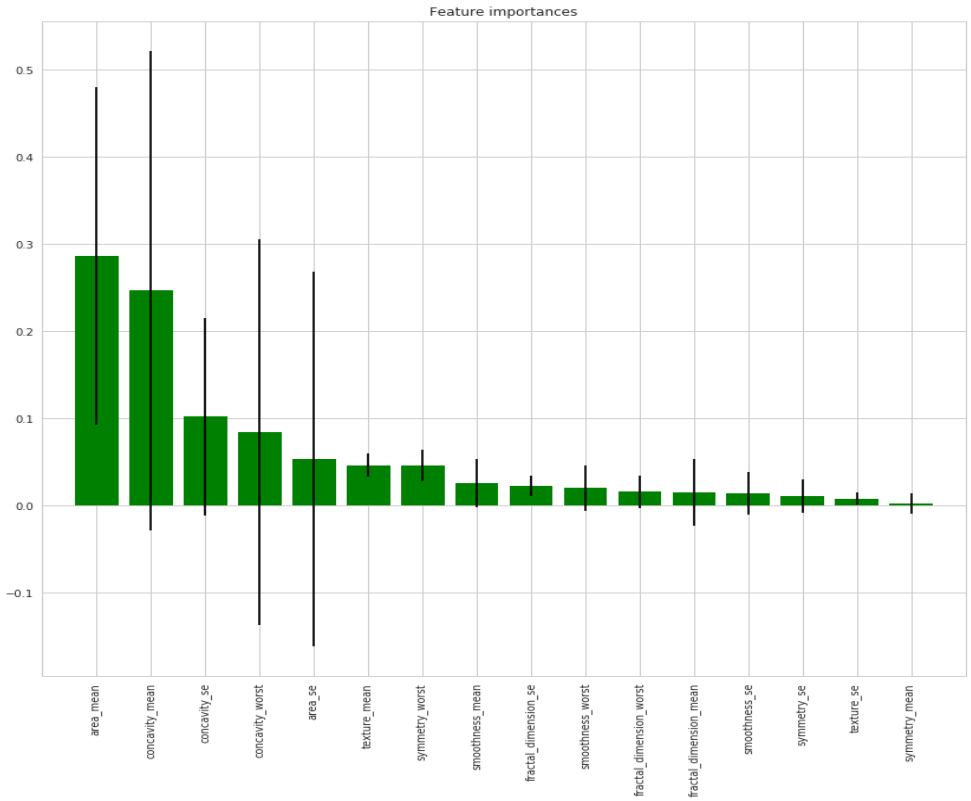


Figure 4 Updated Heatmap

After dropping the correlated features, we plot the heatmap with 12 features. We can observe that these features are not highly corelated with each other. Hence, we will use these features to further perform machine learning algorithm. The 11 features selected as independent variables against dependent variable diagnosis are:

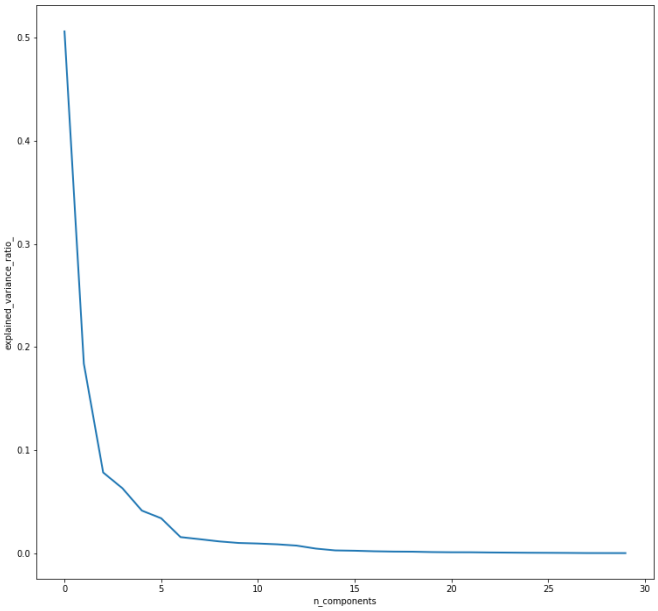
- radius_mean
- texture_mean
- smoothness_mean
- compactness_mean
- symmetry_mean
- fratal_dimension_mean
- radius_mean
- texture_mean
- smoothness_mean
- compactness_mean
- symmetry_se
- fractual_dimension_se

2.1.2 Tree based feature selection



We found 16 most important features. They are area_mean, concavity_mean, concavity_se, concavity_worst, area_se, texture_mean, symmetry_worst, smoothness_mean, fractal_dimension_worst , smoothness_worst, fractal_dimension_mean, smooothness_se, symmetry_se, texture_se and symmetry_mean.

2.1.3 Feature Selection with PCA



Observation - Three components can be chosen based on the variance ratio.

3 Building Machine Learning Algorithms

Firstly, we will divide the data into two sets: a training set and a testing set. We will use the 12 features that was extracted from correlation heatmap for further analysis. Then, the algorithm is trained in the first part, predictions are made in the second part, and the predictions are evaluated and compared in the third section.

3.1 Machine Learning Algorithms

We will employ four classifiers for machine learning algorithms:

3.1.1 KNN

The k-nearest neighbors (KNN) algorithm is a powerful machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity.

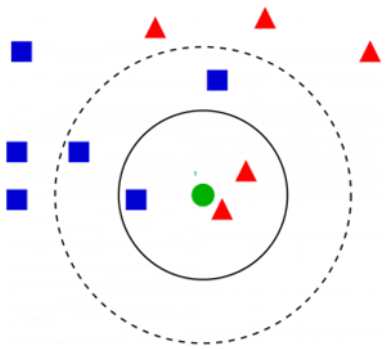


Figure 5 K-nearest neighbors

	precision	recall	f1-score	support
0	0.93	0.97	0.95	115
1	0.94	0.86	0.90	56
accuracy			0.94	171
macro avg	0.94	0.92	0.93	171
weighted avg	0.94	0.94	0.93	171

Observation: KNN provides accuracy of 94%, precision of 94%, recall of 92% and f1 score of 93%.

3.1.2 Random Forest

Random decision forests are an ensemble method for classification, regression, and other tasks, that operate by constructing a multitude of decision trees.

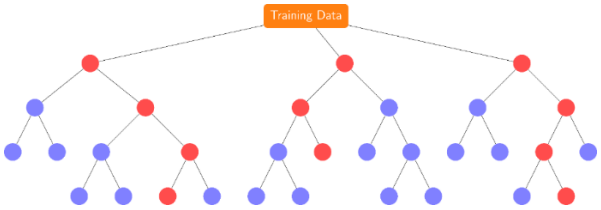


Figure 6: Random Forest

	precision	recall	f1-score	support
0	0.94	0.95	0.94	115
1	0.89	0.88	0.88	56
accuracy			0.92	171
macro avg	0.92	0.91	0.91	171
weighted avg	0.92	0.92	0.92	171

Observation: Random Forest provides accuracy of 92%, precision of 92%, recall of 92% and f1 score of 91%.

3.1.3 Logistic regression

Logistic regression is a very powerful modeling tool, is a generalization of linear regression. It is used primarily for predicting binary or multiclass dependent variables.

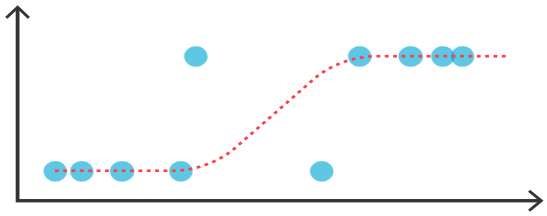


Figure 7: Logistic Regression

	precision	recall	f1-score	support
0	0.98	0.96	0.97	115
1	0.92	0.96	0.94	56
accuracy			0.96	171
macro avg	0.95	0.96	0.95	171
weighted avg	0.96	0.96	0.96	171

Observation: Logistic Regression provides accuracy of 96%, precision of 95%, recall of 96% and f1 score of 95%.

3.1.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a classifier which divides the datasets into classes to find a maximum marginal hyper plane (MMH) via the nearest data points.

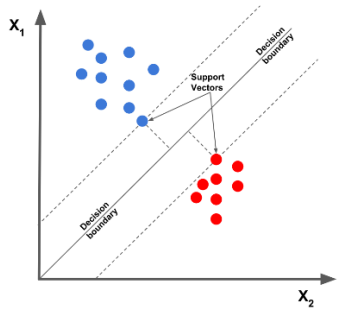


Figure 8: Support Vector Machine

	precision	recall	f1-score	support
0	0.97	0.97	0.97	115
1	0.95	0.95	0.95	56
accuracy			0.96	171
macro avg	0.96	0.96	0.96	171
weighted avg	0.96	0.96	0.96	171

Observation: SVM provides accuracy of 96%, precision of 96%, recall of 96% and f1 score of 96%.

4. Result

	Logistic Regression	Random Forest	K-Nearest Neighbor	Support Vector Machine
Accuracy	95%	92%	93%	96%
Precision	0.96	0.92	0.94	0.96
Recall	0.96	0.91	0.92	0.96
F1 Score	0.95	0.92	0.93	0.96

Support Vector Machine demonstrate the highest accuracy, precision, recall and F1 score. After SVM, logistic regression was second best performing machine learning algorithm.

5. Conclusion

In conclusion, for our dataset, the Support Vector Machine (SVM) model produced the best results. We experiment with feature selection utilizing a correlation matrix and feature importance to increase performance. Overall, we provided four machine learning models that may be used to increase the accuracy of breast cancer diagnosis and hence help with early detection.

5.1 Business Implication

Using machine learning can help in early diagnosis of breast cancer. We can minimize false negatives and false positives with better modeling, which will help pathologists make more accurate diagnoses.

5.2 Limitation

The study is limited to Wisconsin dataset obtained from Kaggle. It is necessary to reflect and apply the same methods to a larger dataset to confirm the results obtained and achieve a higher accuracy.

References

Hastie, T., Robert, T., & Friedman, J. (2009). *The Elements of Statistical Learning.* Springer Nature Switzerland.

Larose, D. T. (n.d.). *Discovering Knowledge in Data.* Retrieved from https://doc.lagout.org/Others/Data%20Mining/Discovering%20Knowledge%20in%20Data_%20An%20Introduction%20to%20Data%20Mining%20%282nd%20ed.%29%20%5BLarose%20%26%20Larose%202014-06-30%5D.pdf

Noble, W. S. (n.d.). *Nature.com*. Retrieved from What is a support vector machine?: https://www.nature.com/articles/nbt1206-1565

Services, U. D. (n.d.). *Breast Cancer Basic information*. Retrieved from Centers for Disease Control and Prevention: https://www.cdc.gov/cancer/breast/basic_info/index.htm#:~:text=Each%20year%20in%20the%20United,each%20year%20from%20breast%20cancer.

(RadiologyInfo.org, n.d.): https://www.radiologyinf so.org/en/info/breast-cancer

(Cancer.net, 2022): https://www.cancer.net/cancer-types/breast-cancer/statistics

(ScienceDirect, n.d.)https://www.sciencedirect.com/science/article/pii/S1877050921014629

https://www.kaggle.com/code/priyanka841/breast-cancer-diagnostics-prediction/notebook?scriptVersionId=32115512

https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

https://tex.stackexchange.com/questions/503883/illustrating-the-random-forest-algorithm-in-tikz

https://www.tibco.com/reference-center/what-is-logistic-regression

https://learnopencv.com/support-vector-machines-svm/

https://www.kaggle.com/code/kanncaa1/feature-selection-and-data-visualization