University of South Wales

Medical Image Classification

By

Kazi Alif Haider

30106091

December 8, 2023

Introduction

The progress in machine learning has brought exciting possibilities to improve how doctors diagnose illnesses. This study aimed to use these computer techniques to sort out pictures of breast tissue taken in a specific way. I had a dataset filled with details calculated from these pictures, and my job was to create a strong computer system that could accurately tell if these pictures showed signs of being unhealthy (malignant) or healthy (benign).

I used this dataset, which has information about the pictures, to teach the computer how to achieve targeted decisions. This report explains the steps I took, like how I helped the computer to learn, what information it used, and how well it made the right decisions.

According to Street, Wolberg and Mangasarian (1993) The dataset contains crucial information about the cell nuclei, including measurements like radius (distance from center to edge), texture (variation in gray-scale), perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. These details, obtained from images of breast tissue, play a vital role in aiding doctors in the differentiation of malignant and benign breast tissue masses.

The primary aim was to put machine learning into action by using this dataset to train and test models. This dataset needs to go through various essential steps: starting with exploring the data to understand it better, then cleaning it to make it more usable. Following that, it needs to create new helpful information from the data and training the computer to recognize crucial patterns. The ultimate goal was to develop models capable of accurately distinguishing between the two types of pictures: those indicating potential issues and those displaying normal conditions. Based on this analysis, a set of data mining models have been developed and applied to the data , to classify whether the nucleus is healthy or not, based on the

other features in the dataset. The results are then discussed. The models developed are a random forest model, decision tree, SVM, neural network(MLPRegressor).

Implementation and Initial Data Analysis

For the analysis of data the first thing has been done is to read the Wisconsin breast cancer data file into the panda data frame. Pandas data frames are used for handling large number of data and have al lot of useful features to process that data as well. In the dataset there were missing column names. So when reading the dataset the column names were introduce at the same time. The column names were given according to archive.ics.uci.edu (n.d.) variable table.

ID	Diagnosis	radius1	texture1	perimeter1	area1	smoothness1	compactness1	concavity1	concave_points1	...	radius3	texture3	perimeter3	area3	smoothness3	compactness3	concavity3	concave_points3	symmetry3	fractal_dimensio
84300903.0	M	19.69	21.25	130.00	1203.0	0.10960	0.1599	0.1974	0.12790	...	23.57	25.53	152.50	1709.0	0.1444	0.4245	0.4504	0.2430	0.3613	0.08
84348301.0	M	11.42	20.38	77.58	386.1	0.14250	0.2839	0.2414	0.10520	...	14.91	26.50	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.17
84358402.0	M	20.29	14.34	135.10	1297.0	0.10030	0.1328	0.1980	0.10430	...	22.54	16.67	152.20	1575.0	0.1374	0.2050	0.4000	0.1625	0.2364	0.07
843786.0	M	12.45	15.70	82.57	477.1	0.12780	0.1700	0.1578	0.08089	...	15.47	23.75	103.40	741.6	0.1791	0.5249	0.5355	0.1741	0.3985	0.12
844359.0	M	18.25	19.98	119.60	1040.0	0.09463	0.1090	0.1127	0.07400	...	22.88	27.66	153.20	1606.0	0.1442	0.2576	0.3784	0.1932	0.3063	0.08

Figure 1: Top Rows of the breast cancer dataset

Next a checking was done to see the columns data type and non-null count. The shape of the dataset 569 X 32 which means there are 569 rows and 32 columns.

```
 4   perimeter1           565 non-null    float64
 5   area1                564 non-null    float64
 6   smoothness1          566 non-null    float64
 7   compactness1         565 non-null    float64
 8   concavity1           565 non-null    float64
 9   concave_points1      561 non-null    float64
 10  symmetry1            566 non-null    float64
 11  fractal_dimension1   565 non-null    float64
 12  radius2              563 non-null    float64
 13  texture2             561 non-null    float64
 14  perimeter2           566 non-null    float64
 15  area2                563 non-null    float64
 16  smoothness2          563 non-null    float64
 17  compactness2         562 non-null    float64
 18  concavity2           561 non-null    float64
 19  concave_points2      560 non-null    float64
 20  symmetry2            561 non-null    float64
 21  fractal_dimension2   562 non-null    float64
 22  radius3              556 non-null    float64
 23  texture3             548 non-null    float64
 24  perimeter3           563 non-null    float64
 25  area3                565 non-null    float64
 26  smoothness3          560 non-null    float64
 27  compactness3         565 non-null    float64
 28  concavity3           566 non-null    float64
 29  concave_points3      563 non-null    float64
 30  symmetry3            565 non-null    float64
 31  fractal_dimension3   556 non-null    float64
dtypes: float64(31), object(1)
```

Figure 2: Information about the columns data-type

An introductory statistical summary of the dataset was presented. It includes key summary statistics such as count, mean, standard deviation, minimum and maximum values, quartiles (25th, 50th, and 75th percentiles), and the inter quartile range (IQR). This summary allows for a quick understanding of the distribution, central tendencies, and spread of numerical features within the dataset.

	ID	radius1	texture1	perimeter1	area1	smoothness1	compactness1	concavity1	concave_points1	symmetry1	...	radius3	texture3	perimeter3	area3	smoothness3	compactness3	concavity3	concave_poir
count	5.660000e+02	564.000000	563.000000	565.000000	564.000000	566.000000	565.000000	565.000000	561.000000	566.000000	...	556.000000	548.000000	563.00000	565.000000	560.000000	565.000000	566.000000	563.000
mean	3.021746e+07	14.084910	-242.883250	91.821965	653.139007	0.096293	0.104072	0.088341	-3.513236	0.187316	...	16.237725	25.755274	110.73222	894.079646	0.132431	0.253804	0.271402	0.114
std	1.252984e+08	3.509242	445.745742	24.305152	351.740510	0.014164	0.052455	0.079382	59.598056	0.115188	...	4.821354	6.123586	59.23487	686.377057	0.022868	0.156881	0.208411	0.065
min	8.670000e+03	6.981000	-999.000000	43.790000	143.500000	0.052630	0.019380	0.000000	-999.000000	0.000700	...	7.930000	12.020000	50.41000	185.200000	0.071170	0.027290	0.000000	0.000
25%	8.692195e+05	11.687500	-999.000000	75.170000	420.175000	0.086380	0.064500	0.029480	0.019690	0.161900	...	13.010000	21.267500	84.13500	515.300000	0.116750	0.147200	0.114425	0.064
50%	9.061570e+05	13.290000	17.000000	86.180000	546.350000	0.095895	0.092630	0.061260	0.033260	0.179150	...	14.940000	25.465000	97.67000	686.500000	0.131350	0.214100	0.226550	0.099
75%	8.812869e+06	15.757500	21.015000	103.800000	782.625000	0.105300	0.130400	0.129300	0.073400	0.195675	...	18.602500	29.757500	126.50000	1084.000000	0.146000	0.339100	0.382400	0.161
max	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	2.100000	...	36.040000	49.540000	910.00000	10056.000000	0.222600	1.058000	1.252000	0.291

Figure 3: Summary of the dataset

A count plot was made to see the number of health and unhealthy nuclei were in the dataset.

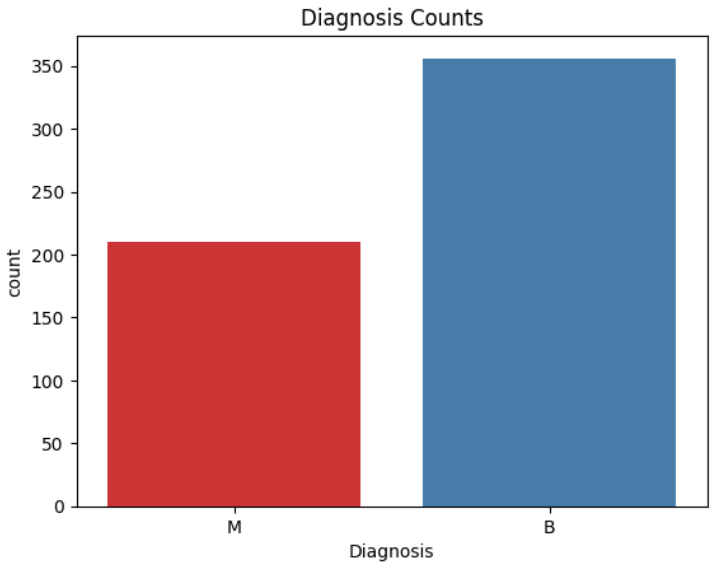


Figure 4: Healthy and Unhealthy nuclei counts

A number of histograms have been plotted of the different features in the data set. None of them are normally distributed they are either skew to left or right. smoothness1, smoothness2, and smoothness3, appear to have long tails.

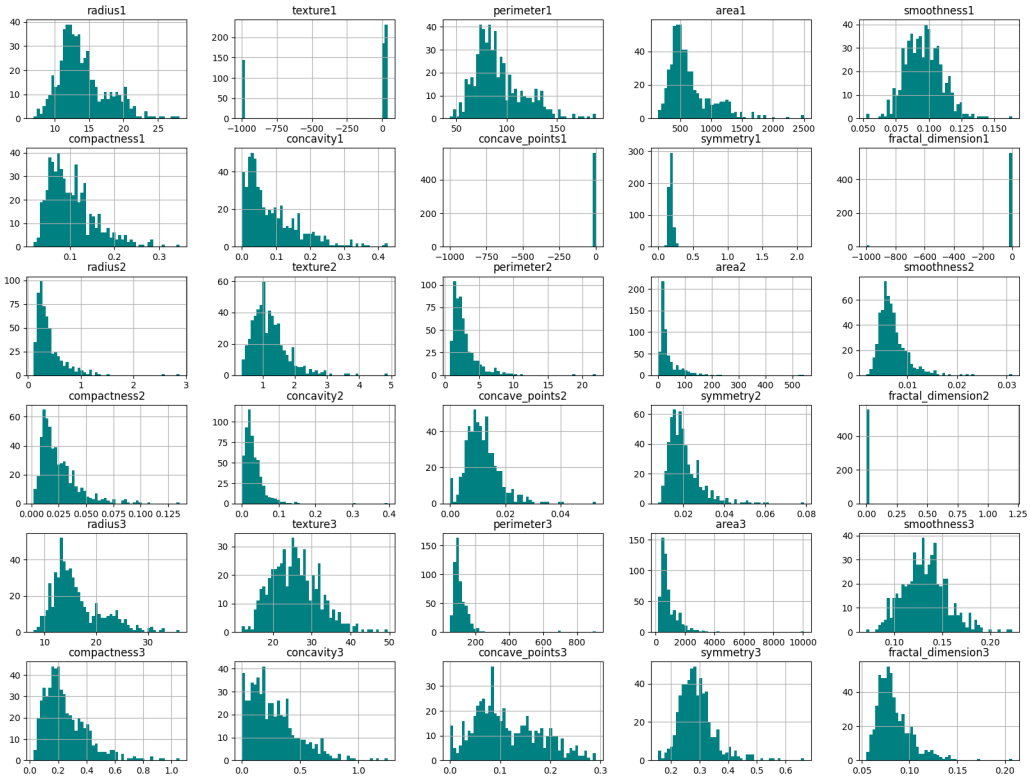


Figure 5: Histogram of the features

In most of the dataset there are some outliers which might hinders the classification and predictability. Outliers are data points that are very different from most of the other data. Missing values are also one of the problem. To check for the outliers a series of box-plots was utilized. As the figure shows there is a lot of outliers present in every feature of the dataset.

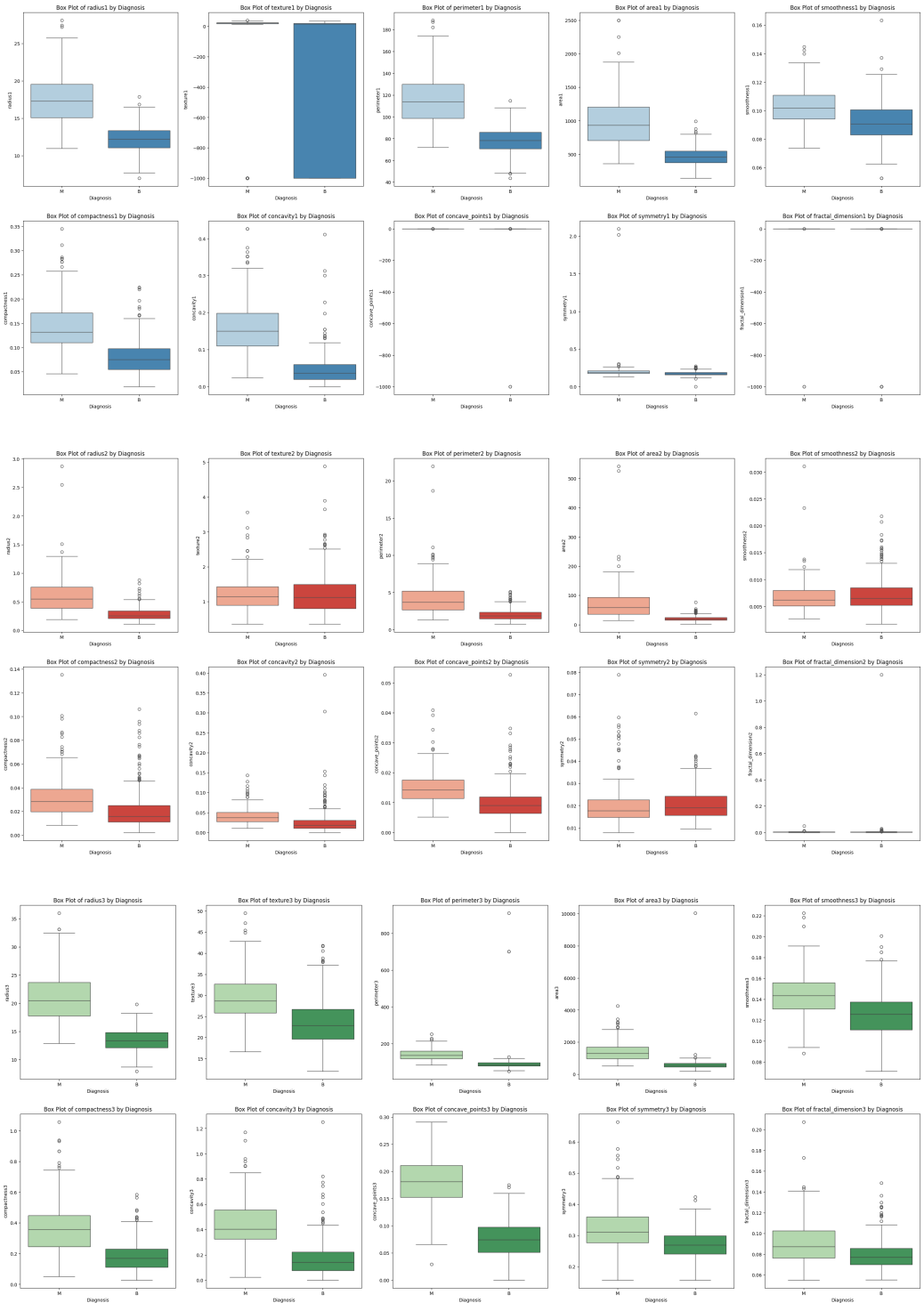
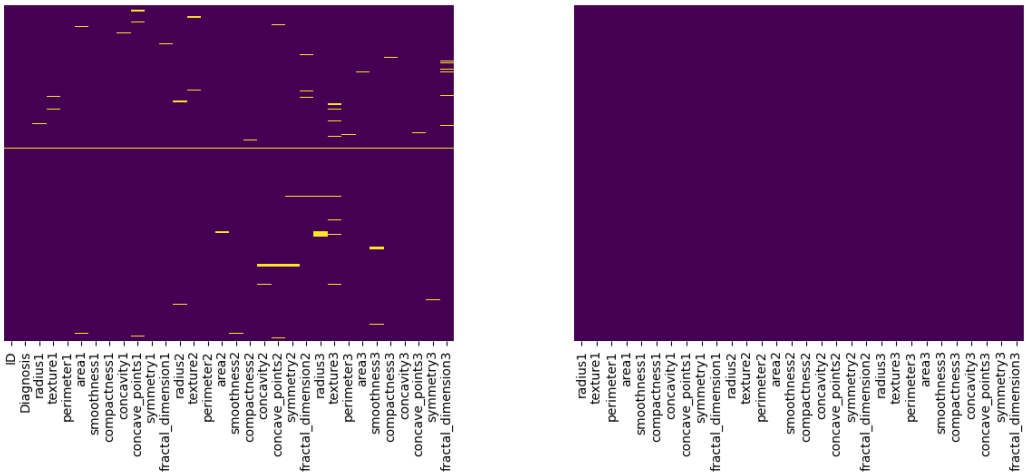


Figure 6: Outliers for all of the features

Before taking care of the outliers the missing data issue needs be fixed. Using `isnull().sum()` will show number missing data on each column. if the whole row or column has missing value it needs to be get ridden of. On the other case of missing data are missing there is no way to tell whether the missing values have more significance or not. In that case imputation techniques needs to be introduced. According to Jakobsen et al. (2017) there are two ways to fill up the missing data. 1) Single Imputation 2) Multiple Imputation. Single imputation, which fills in missing values, might underestimate variability. It treats all missing data equally, not considering if it's random or not. This method relies on assuming that missing values are like the last observed value. However, this assumption is often not accurate. As a result, using single imputation could lead to biased results. In multiple imputation, missing data isn't replaced with just one guess. Instead, several different guesses or 'imputations' are created for each missing value. Hanen Karamti et al. (2023) state that KNN imputation is better for both continuous and categorical values. The provided datasets have all numerical continuous values. So KNN imputer is best for this dataset. `not isnan()` data.



(a) Before (b) After

Figure 7: Heatmaps of missing data comparison

The most common and popular method to get rid of outliers is the Basic IQR (Gregg and Moore, 2023). The basic IQR technique is good for medical datasets because medical datasets contain outliers or extreme values due to various reasons such as measurement errors or rare conditions. IQR is less sensitive to outliers compared to measures like standard deviation or range, making it a better choice when dealing with such data. There is a function `iqr()` calculates the Interquartile Range (IQR) for a given dataset column, identifying outliers using the 1.5 IQR rule. It determines lower and upper bounds to flag outliers, then filters the original data, returning a cleaned version without identified outliers.

Lastly Checking the correlation between the features with correlation matrix. Most of the correlation are not close to 0.So, none of the feature can be removed.

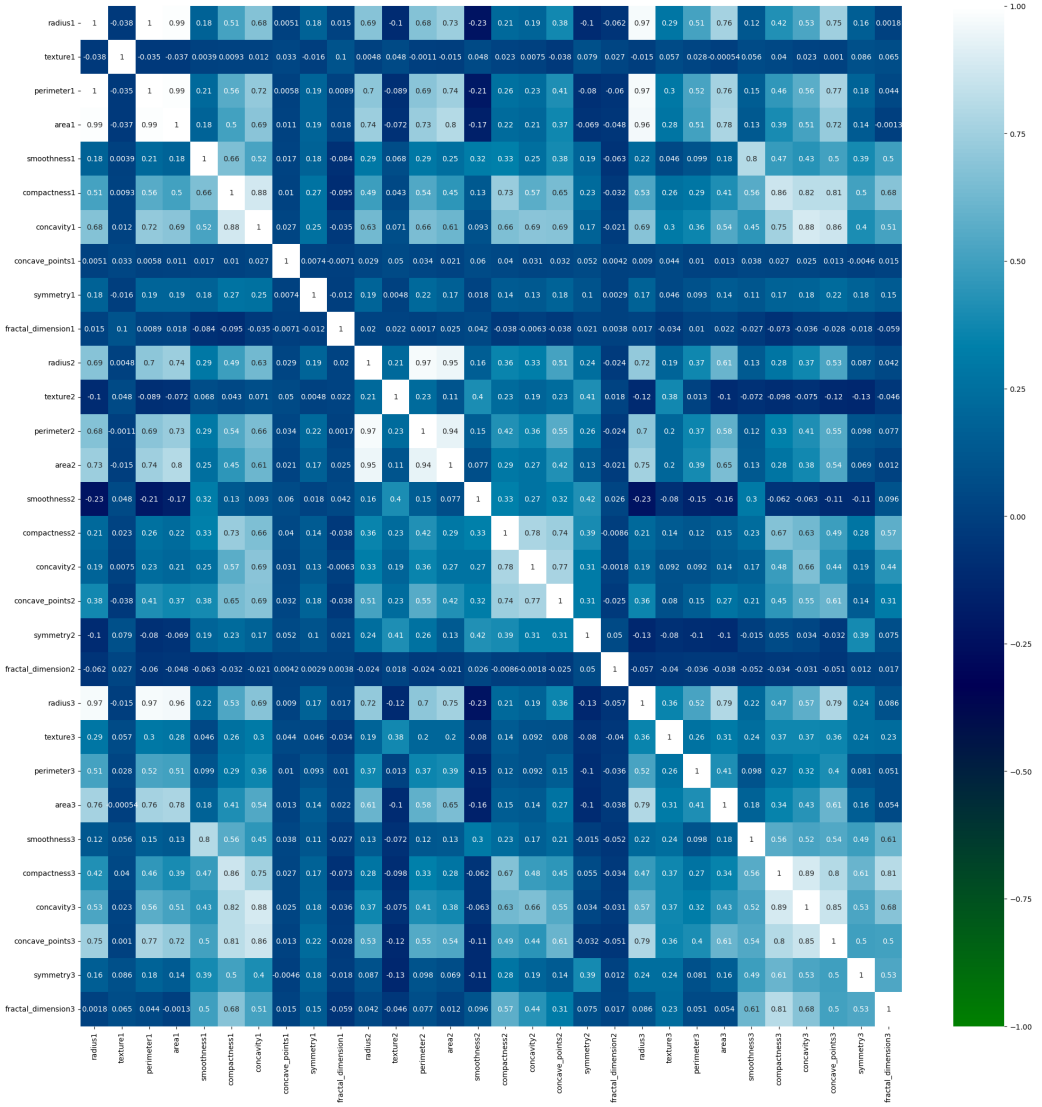


Figure 8: KNN Imputation

Training

A crucial step in building machine learning models involves dividing data into training and testing sets, which is vital for ensuring the model's wide-ranging ability and evaluating its performance. The larger training set teaches the model by showing it examples, helping it learn patterns and connections between different parts of the data. The smaller testing set remains untouched during training and is used only to assess how well the model works when it encounters new data. For this dataset the test size was 20% and training size was 80%. A Label encoder was used for getting the categorical data to be labeled as number. Because in machine learning most of the algorithm works better when provided numerical data. Feature scaling is a preprocessing method crucial for ensuring all features in a dataset contribute fairly to machine learning models. It normalizes the values of different features to a comparable scale, avoiding biases caused by features with larger values dominating the model's learning process. Standard scaler works better when there is outlier present than Min-Max scaler. Also it preserves the relationships between the data points (Bhandari, 2020). For this dataset standard scaler works better as the relationships between the data points needs to be preserved. A range of data mining models have been developed and applied. The models developed are a random forest model, a support vector machine model ,a decision tree model and neural network model (MLP Regressor).

Decision Tree

A Decision tree is model that work like a tree. Where every node is a feature of the dataset and each branch provides different results depending on the specific features. It uses recursive methods to divide the datasets, it selects the most significant features at each nodes to separate data into distinct group. According to (Bell, 2014) it handles large set of data if it has the computing power.It also has

disadvantage. It makes the training data overfit.

Random Forest

Random forrest is a better version of decision tree. In random forrest model it builds multiple decision trees and sums up the predictions to improve accuracy and prevent overfitting. It makes different trees by using random parts of the data and features, then puts together their predictions to make reliable and robust overall predictions. Random forests often offer higher accuracy than a single decision tree, but they sacrifice the straightforward interoperability found in decision trees.Additionally, when dealing with multiple categorical variables, a random forest might not improve the base model's accuracy (Wikipedia Contributors, 2019).

Support Vector Machine

The Support Vector Machine (SVM) is a well-established supervised learning method that defines a boundary between two groups based on provided training data. Despite their effectiveness, SVMs rely on parameters that researchers must explore thoroughly to create an optimal classifier. They work by finding a decision function that separates training data into two categories. This function is chosen to maximize its distance from the closest data points on either side of the division. In cases where a linear division isn't possible, a kernel transformation function is applied to map the data into a different space (called a feature space). This transformation helps in creating a linear separation using standard SVM techniques. Various types of kernels exist to map data into diverse dimensions (Boyle, 2011).

Neural Network (MLP Regressor)

Artificial neural networks is inspired from how the human brain operates. They represent an advancement from linear and logistic regression by introducing several nonlinear elements in predicting outputs. Moreover, neural networks offer significant flexibility in adjusting their structure to address challenges across various fields, utilizing both structured and unstructured data. As the function becomes more intricate, there's an increased opportunity for the network to adapt to the input data, thereby enhancing the accuracy of predictions. A layer consists of one or more nodes (units for computation), where each node within a layer is linked to every other node in the immediate following layer. The input layer comprises the input variables essential for forecasting the output values (V Kishore Ayyadevara, 2019).

Results

Training data were fitted to above models to get the results. For every models a hyper tunning was conducted the best accuracy possible.

Decision Tree

Accuracy of decision tree is approximately 92.11%. This accuracy represents the rate at which the model's predictions align with the actual data, indicating a correctness level of around 92.11%. The best model configuration includes a max depth of 10, limiting the depth of each tree, and min samples split and min samples leaf both set to 2.

The figure below is the confusion matrix of decision tree, the value 61 represents the number of instances correctly predicted as positive (True Positives), while 44 signifies the correct prediction of negative instances (True Negatives). Additionally, the model misclassified 5 positive instances as negative (False Negatives) and 4 negative instances as positive (False Positives).

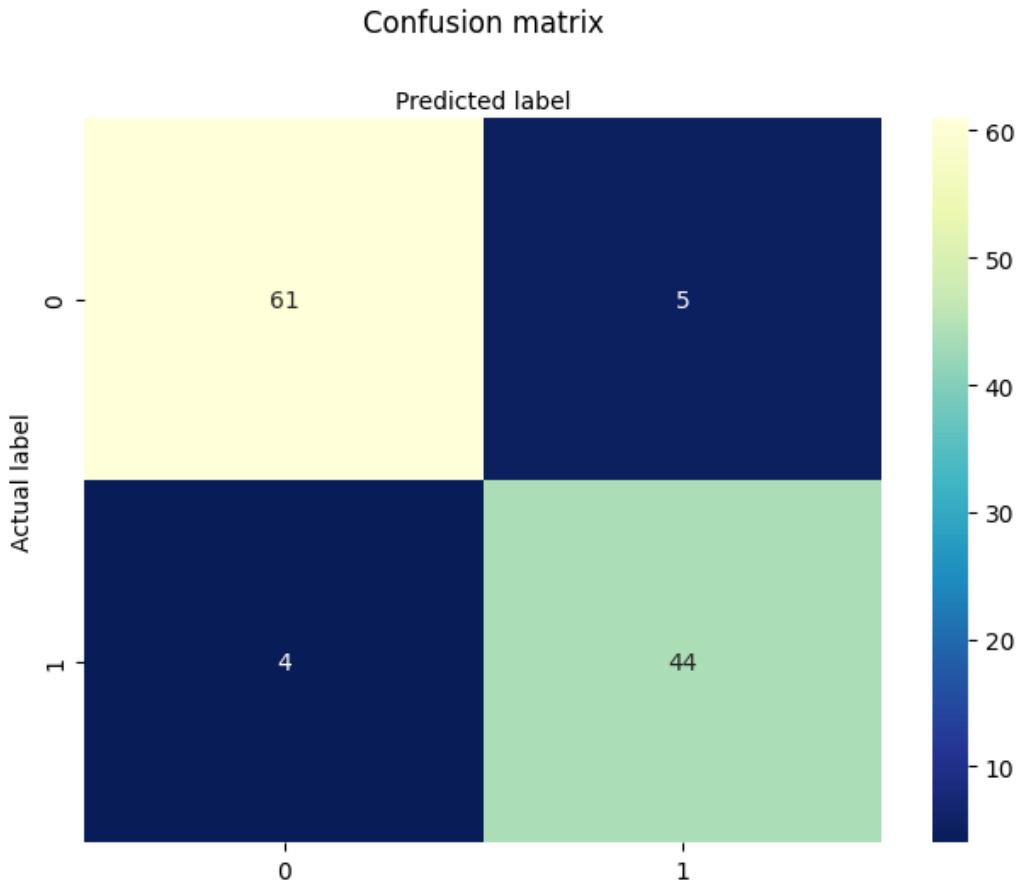


Figure 9: Decision Tree Confusion Matrix

Below figure shows the ROC (Receiver Operating Characteristic) curve of Decision tree. The auc score is 0.94 which is closer to 1. Overall it means that the model is great to classify the data making prediction correct Most of the time.

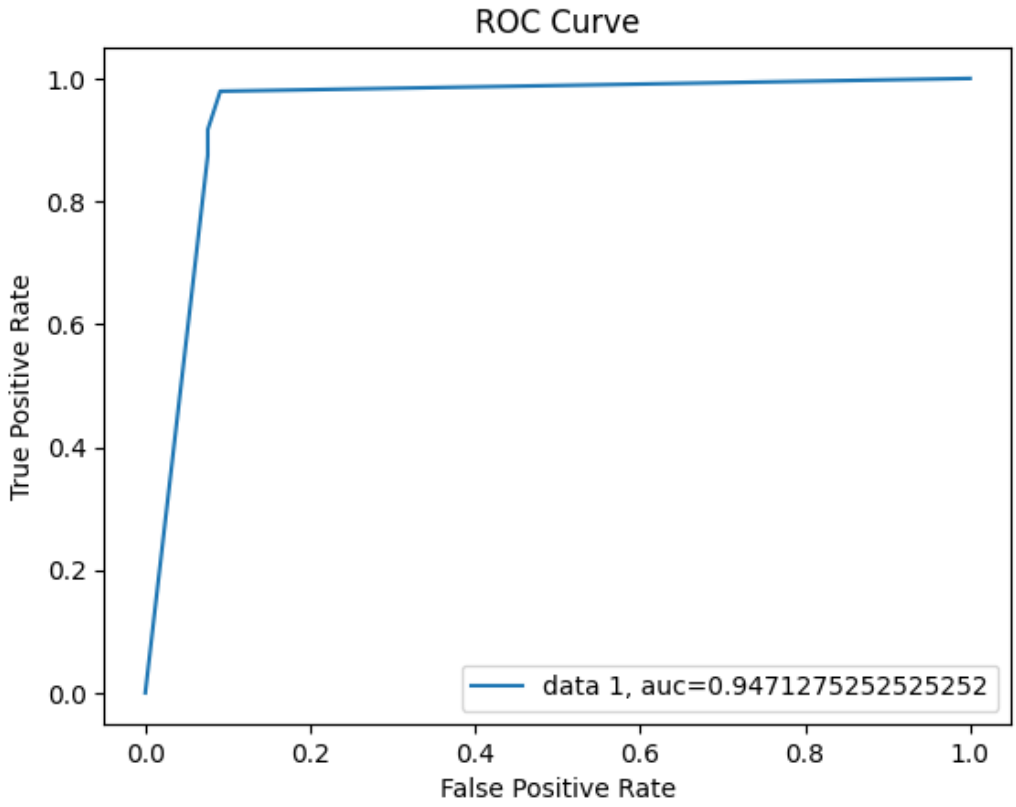


Figure 10: Decision Tree Confusion Matrix

Random Forrest

Accuracy of Random Forrest is approximately 96%. This accuracy represents the rate at which the model's predictions align with the actual data, indicating a correctness level of around 96%. The model configuration includes 100 trees in the forest, a max depth of 10, min samples split of 2, and min samples leaf of 1.

The figure below is the confusion matrix of Random Forrest, the value 64 represents the number of instances correctly predicted as positive (True Positives), while 45 signifies the correct prediction of negative instances (True Negatives). Additionally, the model misclassified 2 positive instances as negative (False Negatives) and 3 negative instances as positive (False Positives).

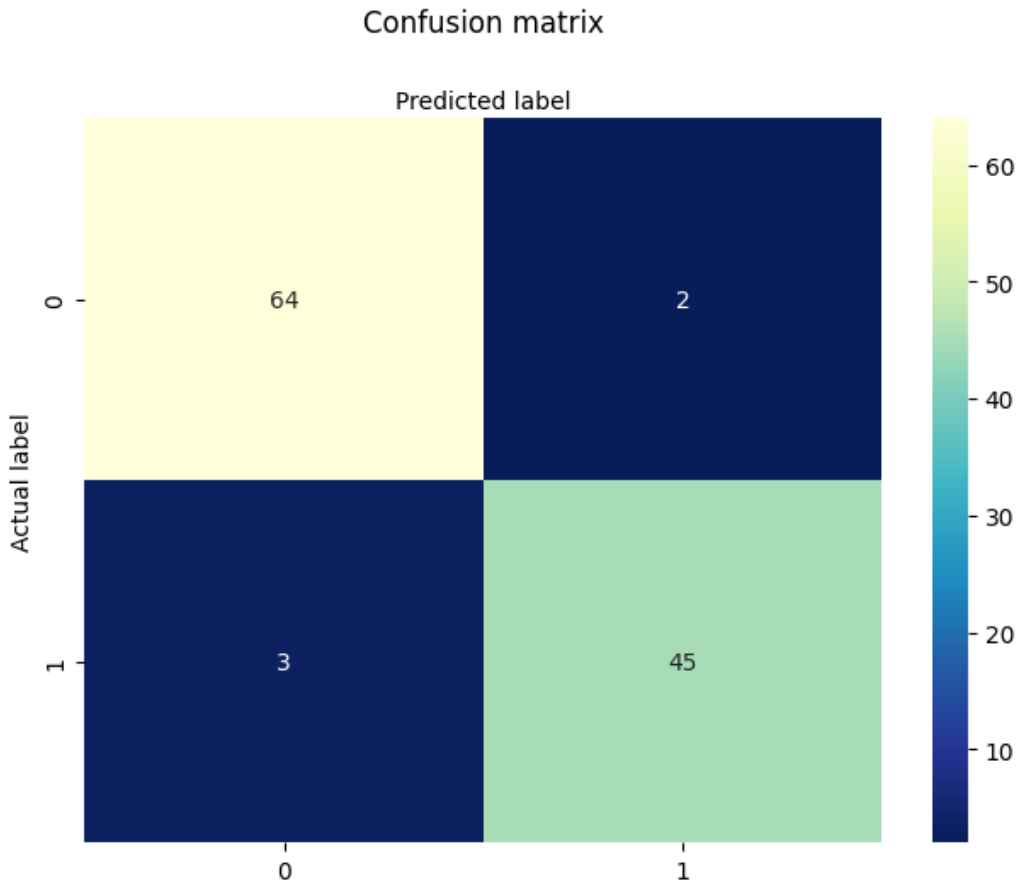


Figure 11: Random Forrest Confusion Matrix

Below figure shows the ROC (Receiver Operating Characteristic) curve of Random Forrest. The auc score is 0.99 which is closer to 1. Overall it means that the model is great to classify the data making prediction correct Most of the time.

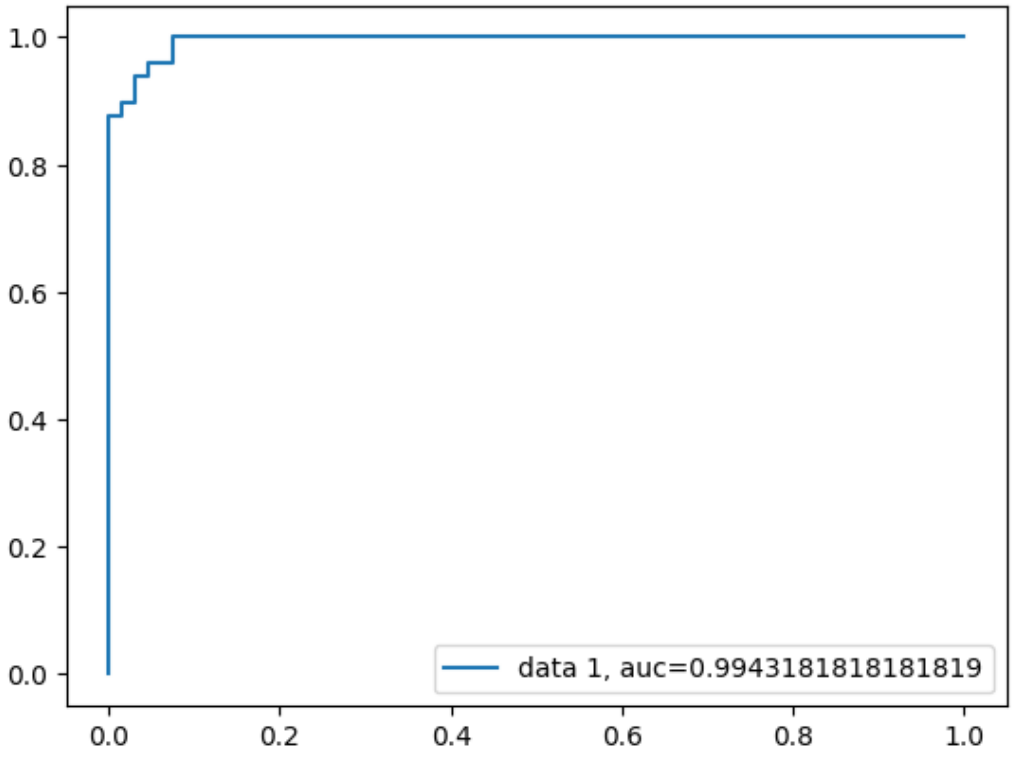


Figure 12: Random Forrest ROC curve

SVM

Accuracy of SVM is approximately 97%. This accuracy represents the rate at which the model's predictions align with the actual data, indicating a correctness level of around 97%. The best model configuration includes C (Regularization) 1, gamma (Kernel coefficient) 1, kernal linear.

The figure below is the confusion matrix of SVM, the value 65 represents the number of instances correctly predicted as positive (True Positives), while 46 signifies the correct prediction of negative instances (True Negatives). Additionally, the model misclassified 1 positive instances as negative (False Negatives) and 2 negative instances as positive (False Positives).

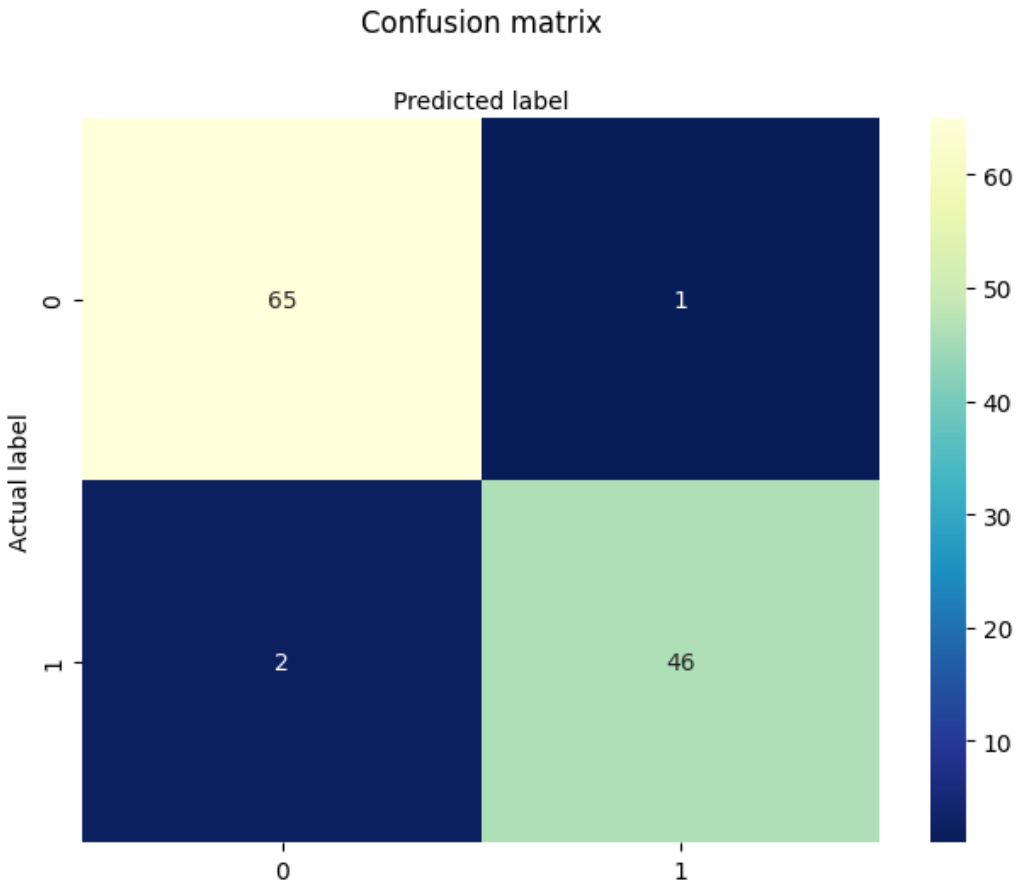


Figure 13: SVM Confusion Matrix

Below figure shows the ROC (Receiver Operating Characteristic) curve of SVM. The auc score is 0.99 which is closer to 1. Overall it means that the model is great to classify the data making prediction correct Most of the time.

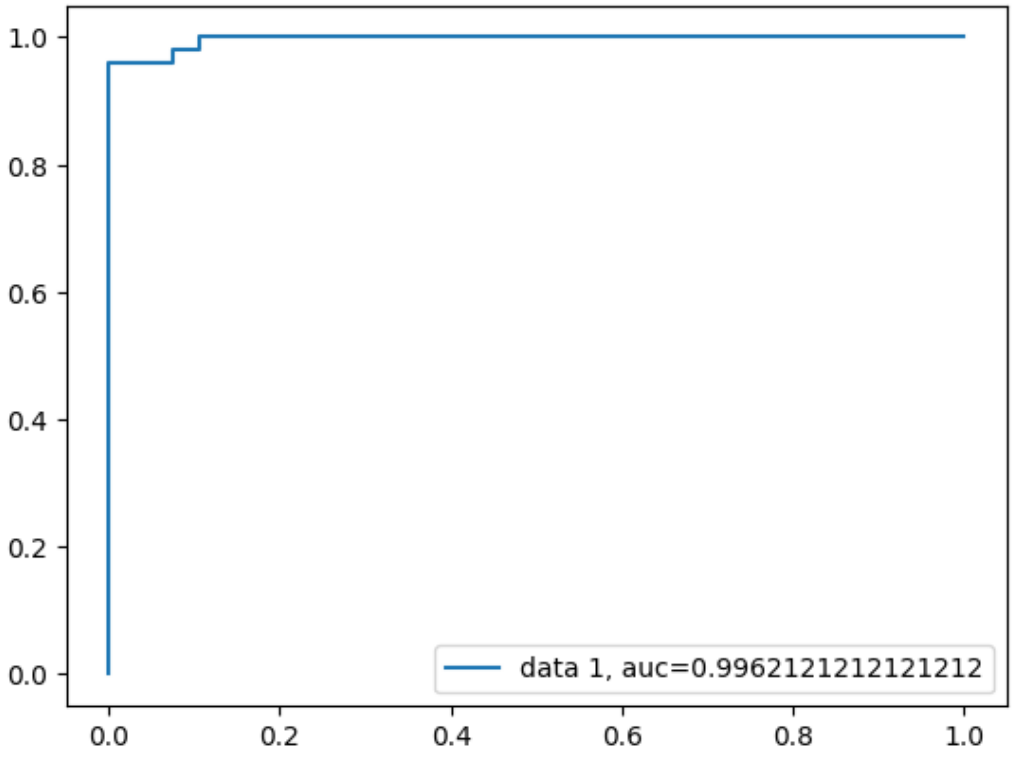


Figure 14: SVM ROC curve

Neural Network (MLPRegressor)

The neural network model has a R2 Score around 0.838 means the model explains about 83.8% of the data's variability, showing it fits quite well. The "Corresponding MSE" approximately 0.039 reveals the model's predictions are quite close to the actual values on average. The Layer Size of 600 and Iteration of 1500 are the best model configuration.

The below scatter plot shows that this model's predictions closely align with the actual values, forming a dense cluster of points along the diagonal line. This indicates that the model is able to capture the underlying relationship between the

input features and the target variable, and therefore can make accurate predictions on unseen data.

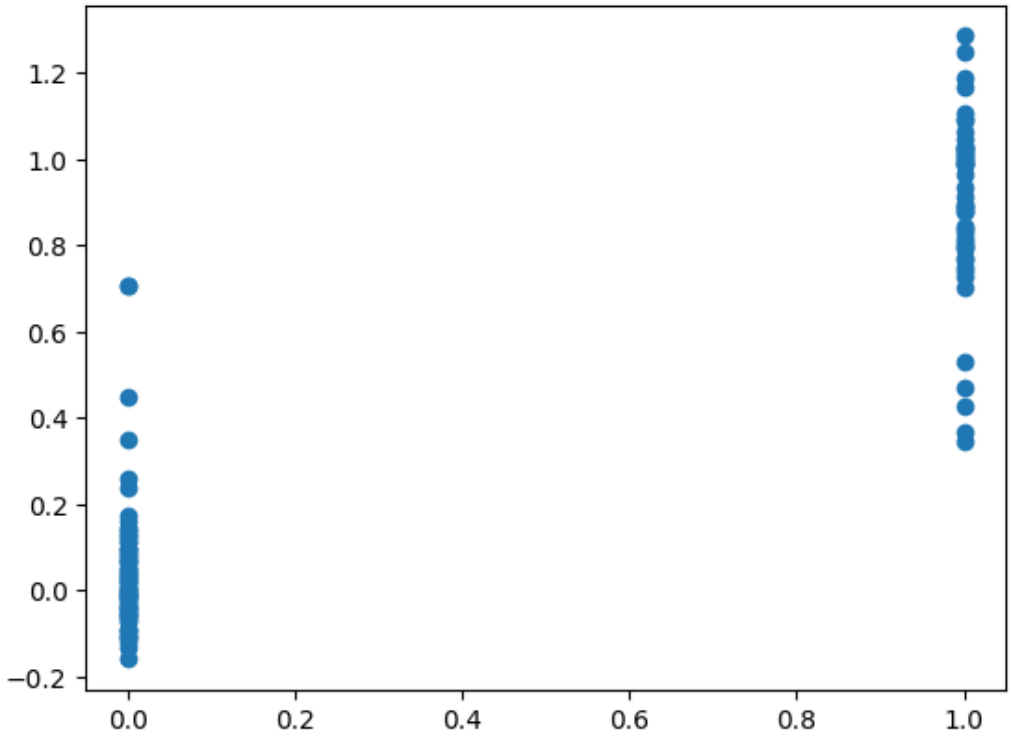


Figure 15: Prediction ability of Neural Network

The figure below shows epoch vs loss. The loss (Performance on the training data) function is decreasing as the number of epochs (single cycle through the entire training dataset during training process) increases. This indicates that the model is learning the training data and improving its performance. But loss function is still relatively high, even after 1000 epochs. This means that the model may still be overfitting the training data.

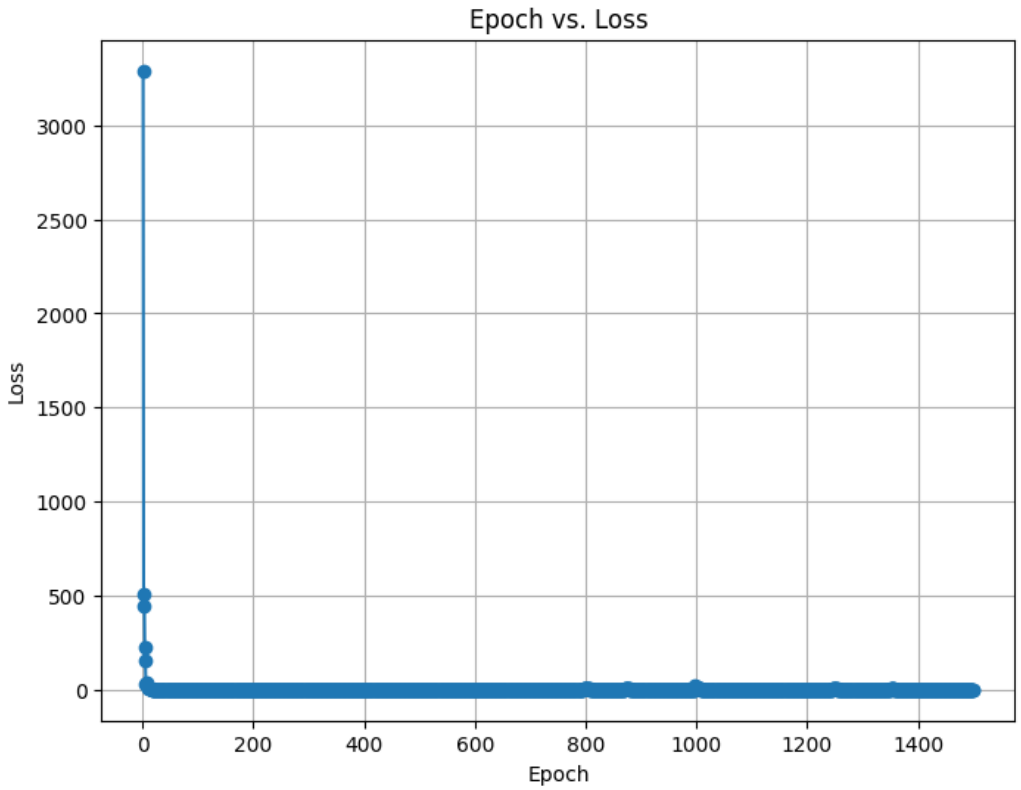


Figure 16: Prediction ability of Neural Network

Conclusion

In conclusion, the analysis of this dataset provided a comprehensive overview of its statistics , numerical features, and distributions, highlighted through various visualizations like histograms and box plots. Addressing missing values with KNN imputation and handling outliers using the Basic IQR method, particularly in medical datasets, ensured data integrity. Feature correlations underscored the need for diverse preprocessing techniques for accurate model training. While the simpler Decision Tree achieved a 92.11% accuracy, it tended to overfit, whereas Random Forest and SVM outperformed with around 96% and 97% accuracy, re-

spectively . Despite showing promise, the Neural Network exhibited a potential for overfitting with a score of 0.838. Implementing different scaling techniques could potentially enhance model accuracy, considering the varying performance of models based on scaling methods. Exploring multi imputation techniques for filling missing values might further improve the dataset's robustness and elevate the models' predictive capacities.

References

Bell, J 2014, *Machine Learning : Hands-On for Developers and Technical Professionals*, John Wiley & Sons, Incorporated, Somerset, PDF format [e-book reader] Available from: https://ebookcentral.proquest.com/lib/usw/detail.action?docID=1818248&pq-origsite=primo (Accessed: 8 December 2023)

Bhandari, A. (2020) *Feature Scaling — Standardization Vs Normalization* Available at: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/. (Accessed: 8 December 2023)

Boyle, BH (ed.) 2011, *Support Vector Machines: Data Analysis, Machine Learning and Applications : Data Analysis, Machine Learning and Applications*, Nova Science Publishers, Incorporated, Hauppauge. PDF format [e-book reader] Available from:https://ebookcentral.proquest.com/lib/usw/reader.action?docID=3021500. (Accessed: 8 December 2023)

Gregg, J.T. and Moore, J.H. (2023) *STAR_outliers: a Python package that separates univariate outliers from non-normal distributions*. ProQuest, pp. 1–15. Available at: https://www.proquest.com/docview/2865398467?accountid=15324&parentSessionId=uDPM7rWxvYFTCGOs2%2FTqw77ajJaAc3jePpkPqsemyZg%3D&pq-origsite=primo (Accessed: 5 December 2023)

Hanen Karamti, Raed Alharthi, Amira Al Anizi, Alhebshi, R.M., Ala' Abdulmajid Eshmawi, Shtwai Alsubai and Umer, M. (2023) 'Improving Prediction of Cervical Cancer Using KNN Imputed SMOTE Features and Multi-Model Ensemble Learning Approach'. *Cancers 15(17)*, pp. 4412–4412. doi: https://doi.org/10.3390/cancers15174412

Jakobsen, J.C., Gluud, C., Wetterslev, J. and Winkel, P. (2017) 'When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts'. *BMC Medical Research Methodology 17(1)*. doi: https://doi.org/10.1186/s12874-017-0442-1.

Street, W.N., Wolberg, W.H. and Mangasarian, O.L. (1993) 'Nuclear feature extraction for breast tumor diagnosis'. Acharya, R. S. and Goldgof, D. B. (eds.). *Biomedical Image Processing and Biomedical Visualization* . doi: https://doi.org/10.1117/12.148698

UCI Machine Learning Repository ([no date]). Available at: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

V Kishore Ayyadevara (2019) *Neural Networks with Keras Cookbook: Over 70 Recipes Leveraging Deep Learning Techniques Across Image, Text, Audio, and Game Bots*. Birmingham: Packt Publishing. Available at: https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=2037521&site=ehost-live (Accessed: 8 December 2023)

Wikipedia Contributors (2019) *Random forest*. Available at: https://en.wikipedia.org/wiki/Random_forest