Comparative Study of Medical Data Recognition

Takshshila Rawat (1219489817), Rahul Naik (1225612271), Shahil Mohammed (1225369223),
Gagandeep Konana Nagaraj (1225293134), Mihir K Gandrakota (1225376100)

December 8, 2022

Abstract

This project's goal is to determine which classification algorithm provides the best estimate for accurately predicting health problems for the following datasets: breast cancer detection, heart failure prediction, and stroke prediction. To process the data and optimize the output, we employ appropriate normalization and regularization techniques. As a result of the proposed project, a comparative analysis of the algorithms will be performed.

1 Introduction and Motivation

This project focuses on implementing various classifiers based on input features to predict or classify a disease. In this project, we used three different health datasets and implemented three different classifiers. The primary goal of the project is to evaluate and compare different classifiers' performance. The proper application of Machine Learning in the health industry will have a positive impact by allowing early disease detection and treatment by doctors. As a result, performance analysis and comparison will provide insight into which classier can be used to solve similar problems. This is the driving force behind the proposed topic. The topic of the project is chosen based on its relevance to the course. The algorithms are chosen in accordance with the course syllabus. Additionally, Normalization and Regularization are being implemented.

2 Problem Description

Our study shows that one of the most significant causes of death worldwide today is heart disease. We also studied that due to unhealthy lifestyle of people, most of them in their middle age are more likely to suffer from strokes. It is linked to high mortality and morbidity rates. One of the most prevalent diseases affecting women in today's world is breast cancer. In 2020 there were lakhs of deaths from breast cancer worldwide which is something to be worried about. One of the crucial step in rehabilitation and treatment is obtaining an accurate and prompt diagnosis. There might be more fixes or longer endurance rates assuming illness is recognized before. A significant reduction in disease-related mortality can be achieved through early diagnosis and treatment. Due to the digitization of this world, machine learning plays an important role in the diagnosis of these diseases. During the COVID-19 pandemic, we observed that many of us were confined to our homes, increasing the demand for digitized medical records. In this project, we use machine learning models on which distinct methods of feature extraction and data preprocessing will be done. Furthermore, by analyzing the performance of each disease and predicting it more accurately, our project aims to identify the best algorithms for each disease so that it would help people to diagnose the disease much earlier.

3 Dataset

All of the datasets were obtained from Kaggle and then spilt into train and test datasets before modeling them.

3.1 Heart Failure

The heart failure dataset has a total of 917 data points and 11 features. In 918 data points, 507 had heart disease and 410 did not have heart disease.

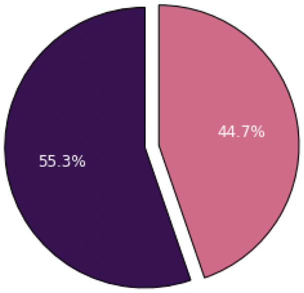


Figure 1: Heart Failure label distribution

3.2 Stroke

The stroke dataset has a total of 5110 data points with 10 features. In 5110 data points, 249 are suffering from stroke and 4861 do not have a stroke.

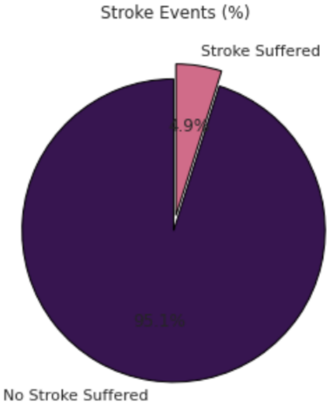


Figure 2: Stroke label distribution

3.3 Breast Cancer

The breast cancer dataset has total of 569 data points with 31 features. In 569 data points 212 are malignant and 357 are benign.

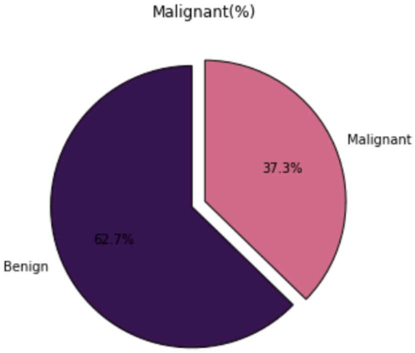


Figure 3: Breast Cancer label distribution

4 Methodology

4.1 EDA

Exploratory Data Analysis was done on each dataset to get insights into the feature-diagnosis co-relation.

4.1.1 Breast Cancer Data-set

Fig 4 maps the direct co-relation of features on the Breast Cancer diagnosis. Lighter-coloured mappings have high co-relation.

The below graphs represent the individual features vs Breast cancer diagnosis.

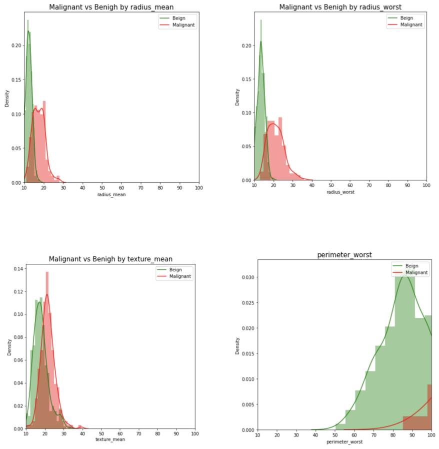


Figure 4:

EDA conclusion for Breast cancer Dataset: Radius and density: higher radius and lower density can mean a higher chance of breast cancer

Texture means: higher texture means a higher chance of positive diagnosis.
Smoothness and Concavity have a higher influence on the diagnosis output.

4.1.2 Heart Failure Data-set

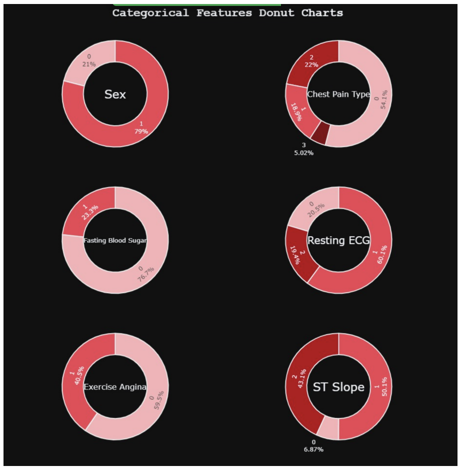


Figure 5: Features Analysis of Heart Failure Dataset

Fig 12 maps the direct co-relation of features on Heart Disease diagnosis. Lighter-colored mappings have high co-relation.
Age, Fasting, and Oldpeak features have the highest influence on the heart disease dataset.

4.1.3 Stroke Data-set

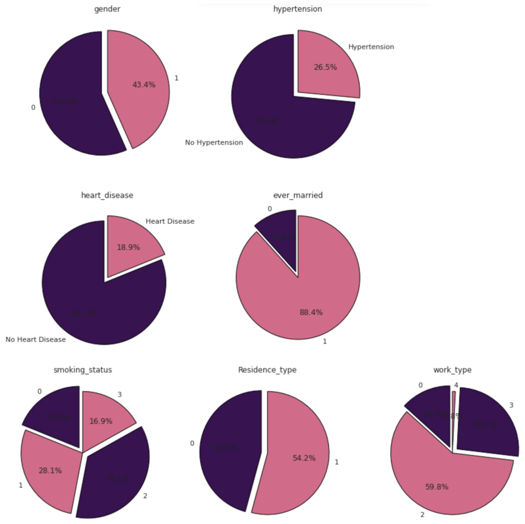


Figure 6: Features Analysis of Stroke Dataset

6 maps the direct co-relation of features on Stroke diagnosis. Lighter-colored mappings have high co-relation.

EDA conclusion for Stroke: From the Fig 6, samples in the dataset with female gender is more than that of male, 26.5% of of samples have hypertension. 11.6% of the samples were never married. 28.1% of the data formerly smoked, 16.9% still smoke and the remaining never smoked. 59.8% of the data are privately employed, 26% are self employed, 13.3% have govt jobs and remaining are children. 54% of the data are urban population. In total, 18.9% of the samples have heart disease. The age of the data range from 55-80 with avg glucose level of 80-2000 and bmi 20-40.

4.2 Algorithms

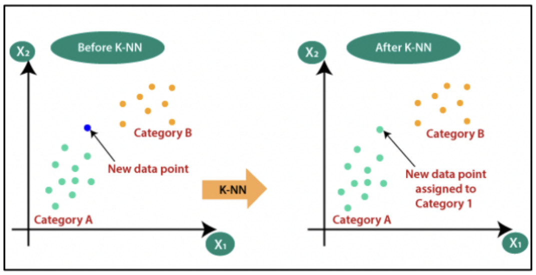
I. 

Figure 7: K-Nearest Neighbors

Regression and classification both make use of the k-nearest neighbors (k-NN) method of supervised learning. k-NN attempts to predict the appropriate class for the test data by calculating the distance between the test data and all of the training points[1]. The *k* number of points that are closest to the test data will then be chosen. The *k* training data class with the highest probability is selected after the k-NN algorithm examines the likelihood of test data belonging to that class. Figure 1 shows the working of k-NN.

II. ***SVM:*** A support vector machine (SVM) is a supervised ML model that is used as a classification algorithm for group classification problems. Each observation is plotted as a point in n-dimensional space(n is the number of features in the given dataset)[2]. An optimal hyperplane is built for classifying the data points into their appropriate classes. Figure 8 shows the working of SVM where A and B are support vectors and C is the hyperplane.

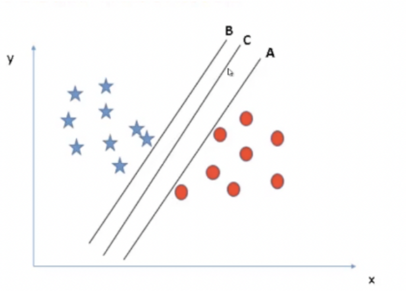


Figure 8: Support Vector Machine

III. ***Logistic Regression:*** A classification algorithm from Machine Learning called logistic regression is used to predict the likelihood of particular classes based on some dependent variables. In a nutshell, the logistic regression model computes the logistic of the result by adding up the input features (most of the time with a bias term).Reference image is shown

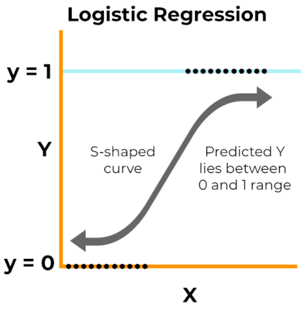


Figure 9: Logistic regression

IV. ***Random Forest:*** It is a classification system that comprises many decision trees. A "Forest" is created using many trees and out of an ensemble of decision trees, which are commonly trained using the "bagging" approach. The main notion behind the bagging approach is that combining several learning models improves the final outcome and performance of the model.

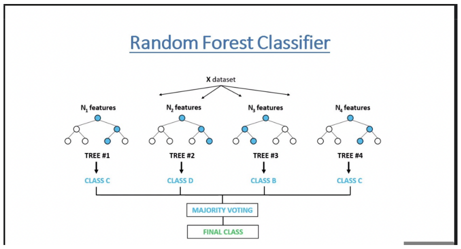


Figure 10: Random Forest

V. ***XGBoost:*** Extreme Gradient Boosting (XGBoost) is a distributed, scalable gradient-boosted decision tree (GBDT) machine learning library. In this calculation, decision trees are made in sequential order. The weights of the variables play an important role in this algorithm. The weight of variables that are predicted wrong by the tree is increased and these variables are then put into the next tree that is being built and the iteration continuous. These individual classifiers/predictors then ensemble to the best model[3].

VI. ***Gaussian Naive Bayes:*** The Bayes algorithm is implemented using the Naive Bayes theorem[4]. Each feature pair is assumed to be conditionally independent. While working on continuous data it is assumed that each class's value has a gaussian normal distribution. The formula is given as below,

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

VII. ***Decision Tree***: Decision Trees constantly separate data based on a parameter.Data is divided into divisions and further divided on each branch using this iterative method. The categorization model is created with the help of a decision tree. The attribute is represented by each node in the tree, and the attribute's potential value is represented by each branch that descends from that node.

4.3 Tuning and Evaluation of Models

4.3.1 Tuning techniques

- *GridSearchCV*: It is a method for determining the best values for a grid's parameters from a given set. It is essentially a method of cross-validation. It is necessary to enter both the model and the parameters.In our project,the best hyperparameters of the model was found using the GridSearchCV.

- *Smote*: SMOTE is a data augmentation algorithm that creates synthetic data points from the original data points.As we encountered imbalanced dataset in the project which is done we used SMOTE to handle the problem.

- *K-fold*: The holdout method is repeated k times, with each of the k subsets serving as the test set and the other k-1 subsets serving as the training set, in the K-Fold validation technique. The project uses K-fold cross-validation for the resampling procedure to evaluate machine learning models as we had a limited data sample.

4.3.2 Evaluation Metrics

The evaluation metrics used in our project to evaluate the best-built models are as follows. The metrics we used are Accuracy, Precision, Recall, F-1 Score, and AUC.

5 Results

All the models were built successfully and their results in percentages are tabulated below;

5.1 Heart Failure

Table 1: Heart Failure

Metric Model	**Acc**	**F-1**	**Rec**	**Pre**	**K-fold**
XGBoost	86.4	89	91	87	88.04
Random Forest	84.78	87	91	84	87.92
Log Reg	82	85	85	84	86.79
SVM	69.02	71	64	78	70.82
k-NN	67.39	72	73	72	71.45
NB (Baseline)	85.86	88	88	88	86.03
Decision Tree	78.26	80	75	86	88.75

5.2 Stroke

Table 2: Stroke Results

Metric Model	**Acc**	**F-1**	**Rec**	**Pre**	**K-fold**
XGBoost	94.12	4	4	20	96.02
Log Reg	77.37	24	69	15	77.95
SVM	71.72	23	78	13	78.40
k-NN	81.7	22	48	14	88.94
NB (Baseline)	74.75	23	72	14	78.78
Decision Tree	90.90	18	19	17	94.39

5.3 Breast Cancer

Table 3: Breast Cancer Results

Metric Model	**Acc**	**F-1**	**Rec**	**Pre**	**K-fold**
XGBoost	97.37	97	96	98	96.38
Rand Forest	94.74	94	98	90	96.38
Log Reg	93.86	93	98	88	94.83
SVM	93	91	91	91	89.5
k-NN	92.10	91	94	88	93.10
NB (Baseline)	93	91	91	91	93.97
Decision Tree	97.37	97	98	96	92.93

We considered Naive Bayes as our baseline as it strongly assumes that features are independent of each other given the output. In real-world features influences, each other to a certain extent as Fig 12 shows in what capacity features influence each other.

From the above results in Table1and Table 3, we observed that our baseline (Naive Bayes) outperforms KNN. The KNN is a supervised lazy classifier having complex decision boundaries that need a lot of data to learn and generalize better. Both heart failure and breast cancer have fewer data points, so KNN fails to learn complex features.

we observed that our baseline (Naive Bayes) outperforms SVM. Our data for stroke prediction was highly imbalanced, although we created more data points through oversampling, the newly created data sampled failed to mimic real-world data scenarios.

XGboost is our winning algorithm, performing best in all the datasets having less data, imbalance features, and complex features. The boosting method incorporates the results of the previous classifiers.

6 Related Work

As a background study, we referred to some of the papers in a similar area to get familiar with the implementation of classifiers on health-related datasets. The summary of the references is as follows:

The Wisconsin Breast Cancer (original) datasets are used in the research [5] to compare the performance of four different machine learning algorithms: Support Vector Machine (SVM), Decision Tree, Naïve Bayes (NB), and k Nearest Neighbors (k-NN). The major goal of the work is to assess the accuracy of data classification in terms of each algorithm's efficiency and effectiveness in terms of accuracy, precision, sensitivity, and specificity. According to experimental data in this study, SVM provides the best accuracy and lowest error rate.

In a study, [6], four distinct models for prediction are trained using a variety of machine learning methods, including Logistic Regression (LR), Decision Tree (DT) Classification, Random Forest (RF) Classification, and Voting Classifier. The algorithm with the best results for this task was Random Forest. The open-access Stroke Prediction dataset was utilized in the method's development.

The various machine learning techniques based on a quick examination of heart disease diagnosis are presented in [7]. First, a weighted version of Nave Bayes is employed to forecast cardiac disease. The second one is automated and analyzes the localization and identification of ischemic heart disease by the frequency domain, temporal domain, and information theory characteristics. In this strategy, the two most effective support vector machines (SVM) using XGBoost classifiers are chosen for the classification. The third technique is an enhanced SVM-based duality optimization methodology for the automated detection of heart failure.

7 Conclusion

With the exponential increase in technological advancements in the modern era, people have begun to recognize the importance of technology and have begun to use it for their specific needs. Machine learning is one such technology that has been widely used, particularly in the medical field, which is the most important requirement of any individual, using Machine learning techniques to predict the possibility of having a fatal disease can help us dodge the bullet early on and take necessary precautions. This project examines various ML algorithms modeled with three different patient disease datasets, namely Breast Cancer, Heart Failure, and Stroke. This aids in determining the best classifier for accurately predicting health problems. Initial data screening using evaluation metrics such as Accuracy, Precision, Recall, and F1 score resulted in poor model test accuracy due to the limited data samples available, however, the use of tuning techniques involving GridSearchCV, SMOTE and K-fold helped in achieving higher accuracy. Oversampling the dataset helps in eliminating the label bias and leads to more accurate predictions.

References

[1] G. Guo, H. Wang, D. Bell, and Y. Bi, "Knn model-based approach in classification," 08 2004.

[2] T. Evgeniou and M. Pontil, "Support vector machines: Theory and applications," vol. 2049, pp. 249–257, 01 2001.

[3] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” *CoRR*, vol. abs/1603.02754, 2016.

[4] I. Rish, “An empirical study of the naïve bayes classifier,” *IJCAI 2001 Work Empir Methods Artif Intell*, vol. 3, 01 2001.

[5] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, “Using machine learning algorithms for breast cancer risk prediction and diagnosis,” *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016. The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016) / Affiliated Workshops.

[6] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. Monirujjaman Khan, “Stroke disease detection and prediction using robust learning approaches,” *Journal of Healthcare Engineering*, vol. 2021, p. 7633381, Nov 2021.

[7] U. Nagavelli, D. Samanta, and P. Chakraborty, “Machine learning Technology-Based heart disease detection models,” *J Healthc Eng*, vol. 2022, p. 7351061, Feb. 2022.

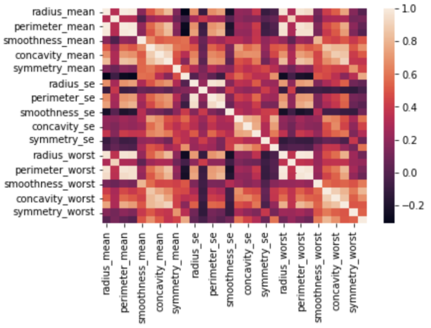


Figure 11: Co-relation between features and Positive Diagnosis of Breast Cancer

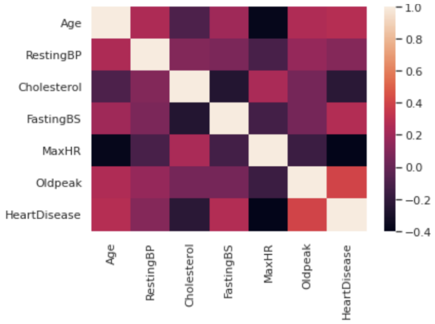
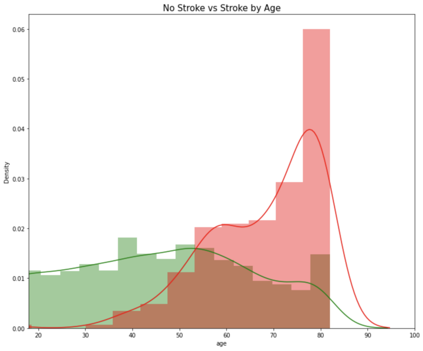
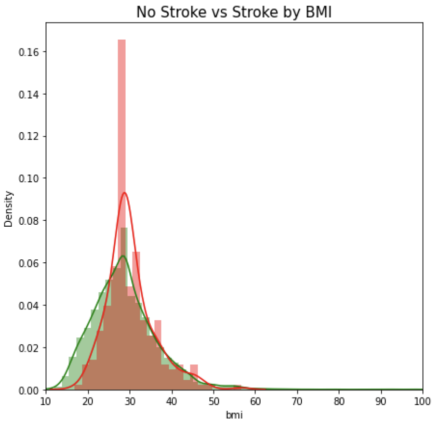


Figure 12: Co-relation between features and Positive Diagnosis of Heart Disease





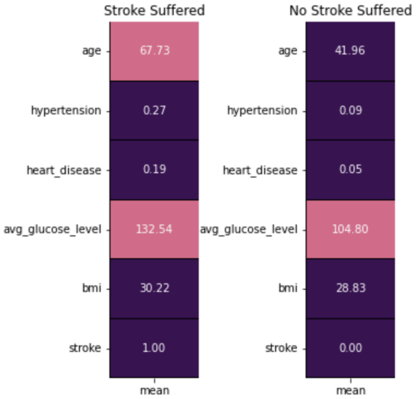
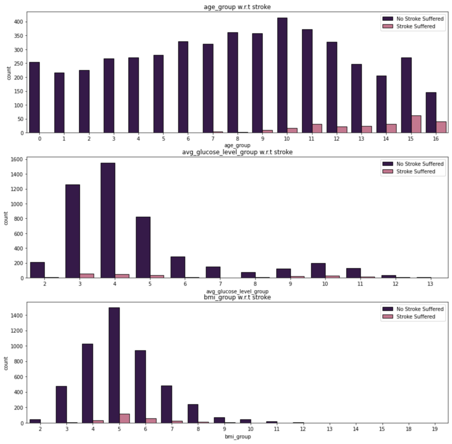


Figure 13: label statistics for positive and negative diagnosis of Stroke.



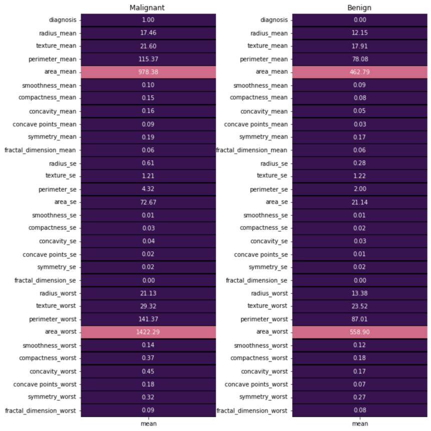


Figure 14: label statistics for positive and negative diagnosis of Breast Cancer.