COMPARITIVE ANALYSIS OF AGING DISEASES USING MACHINE LEARNING MODELS

[1]Dr. T. PanduRanga Vital, [2]Pingali Sai Phani Kumar, [3]Telli Sravani, [4]Bommali Anjali [5]K. Nikhil Sai Patnaik

[1, 2, 3, 4, 5] Dept of CSE, Aditya Institute of Technology and Management, Tekkali, Srikakulam, Andhra Pradesh-India

Abstract: *Aging diseases are often called as the age-associated diseases. Age plays a vital role in these type of diseases. Generally these diseases are occurred at the very elderly stage of the humans. Cancer, Diabetes, Kidney, Lung, Liver etc., are comes under aging diseases. In this project, we collect and prepare 5 different types of datasets that comes under the Aging Diseases (AD). The five different types of aging diseases datasets that we collected are Breast Cancer, Diabetes, Heart, Liver, Kidney datasets. For the mentioned datasets we perform Classification, Prediction, Model fitting using the Various Machine Learning Classifiers (Algorithms) and prepare a Ranking Table which include the Datasets and the Accuracies of ML classifiers that obtained. In this research we calculate the average of all the accuracies of all the ML classifiers of each aging disease dataset and provide the highest average accuracy. The various Machine Learning classifiers that we used for our analyses are Naïve-Bayes (NB), K-Star, J48(C4.5), Random Forest(RF) and Random Tree(RT). These five Machine Learning Classifiers are applied on each dataset that we collected which are related to the aging diseases. We also perform statistical Analyses for each individual dataset and also we perform Attribute Ranking to determine which attribute is considered to be highly prioritized. Our main intention of this project is to say which Machine Learning Model is the best classifier among the other classifiers in giving the best and highest Accuracy and taken the least amount of time to construct the model. As per our Machine Learning Analyses Random Forest (RF) ML classifier gives the highest Average Accuracy of 83.0494% and ranks 1 among all the Machine Learning Classifiers.*

Key Terms: Aging Diseases, Machine Learning (ML), Random Forest (RF),Ranking.

1. INTRODUCTION

Age-related diseases are illnesses and conditions that occur more frequently in people as they grow old, which mean that age is considered to be a significant risk factor. Age plays a vital role in these types of diseases. Aging diseases comprises a group of biological, physiological, and psychosocial processes that may result in the wide variations among individuals. Adult aging diseases often involve some common changes which will not be harmful like diabetes. For many of the people, however, aging is a progressive and inexorable loss of function resulting in increased vulnerability for any disease or a disability. Many of the hypotheses and theories are explaining that this decline have been offered through from the years, but none by itself can explain the array of biological, physical, and psychological changes take place as people age. Many older people also suffer from diabetes, alzheimers, chronic lung disease, liver, heart, weight loss, dizziness etc. Research into underlying causes for these and some other geriatric syndromes are required, for developing the new prevention strategies and treatment approaches. The major symptoms of these aging diseases include Greater susceptibility to the infection, High risk of heat stroke, A Slight decrease in the height of the bones of our spines so that they become thinner and also loses some height, Bones may break

more easily, Joints in the body may changes, starts from normal to severe arthritis, Urinary incontinence, Slight slowing of the thoughts and memory, and it also impacts in thinking, Weight loss after a age of 55 in the men and after a age of 65 in the women etc., which comes under the symptoms of the aging diseases.
In this analysis the 'age' is considered to be as the common and primary attribute. Along with the age attribute there are some more other attributes that are used in the various datasets such as Gender, Total Proteins, Glucose, Hemoglobin, Insulin, Blood Pressure, BMI, Mitosis, Cholesterol, Albumin, Bilirubin etc., are the other important attributes.
In each dataset there is a target attribute which represents that there is disease to a person or not. All the datasets that are used in this analysis are two class problems, which states either yes (person may have disease) or no (person may not have any disease).
For this work, we use the ML models such as the Naïve Bayes, K-Star, J-48 (C4.5), Random Forest (RF), Random Tree (RT) to predict the unknown values from the datasets and provides the classification accuracies for each Dataset. This is one of the valuable work to determine the different stages of aging diseases and to determine which ML model will give faster predictions among other ML classifiers.

Organization of the paper: This paper is a well-organized research about predicting that how many have diseases and don't have diseases among the total number of the instances. The prediction of the accuracies can be done by using different ML classifiers that are applied on each aging disease dataset. The research can be organized as

- Collecting the five different aging disease datasets such as Breast Cancer, Diabetes, Heart, Liver, Kidney.
- Remove the noise values and make the dataset clean by using various preprocessing techniques. Save them as the new CSV (Comma Separated value) files. Convert from CSV to ARFF files.
- Applying the classification mechanism. Identify the target attribute to classify diseased and non diseased.
- Apply various machine learning classifiers to each dataset. Predict the accuracies of each dataset. Perform attribute ranking for each dataset. Finally, preparing a ranking table that gives ranking to the various ML classifiers of all datasets.
- We reviewed lots of papers of each aging disease that considers age as a primary factor. The model diagram represents our overall view of this research. Our research references are discussed in the contribution of authors tabular representation. In the results and discussions section, it represents the overall output containing accuracies and ranking among the ML classifiers.

2. LITERATURE REVIEW

For this research work, we have reviewed 30 research papers from the reputed journals like IEEE, ELSVEIR, IJSCE, SPRINGER and other papers that are related to our research work. Aging Diseases are represented as the highest number of diseases when compared with the other diseases. Any disease that represents age as a key factor is comes under the aging

diseases. For our research work, we prefer and collect the major aging diseases that include Breast Cancer, Diabetes, Heart, Liver, and Kidney. We perform statistical analysis for each individual dataset.

There are different statistics that are taken into consideration for our research work. The various statistics that are used in this research are Kappa statistic, Mean Absolute Error, Root Mean Squared Error, Relative absolute Error, Root relative squared error. Evaluation metrics are the most important elements in classifying the accuracies by each class for each individual dataset. The various Evaluation Metrics followed in this work are TP Rate, FP Rate, Precision, Recall, F-Measure, Class. The class is the major evaluation metric that defines whether the dataset contains diseased instance or not by classifying the dataset into yes or no classes. After that we generate confusion matrix. A confusion matrix is a tabular representation of both true and false values or correctly classified and incorrectly classified instances for the dataset. In this work we also prepare a ranking table that contains the accuracies of different datasets for each ML classifier. For each dataset there are 5 ML classifiers are applied in this work. Each ML classifier produces some accuracies on each individual dataset. We represented these accuracies of each dataset in a tabular format in order to increase the correctness and predictions of the outputs. Average accuracy is calculated for each ML classifier on each dataset. Average Accuracy is the sum of all the accuracies obtained for each ML classifier that applied on each dataset. This average accuracy work is also included in the ranking table. Ranking Table increases the smoothness of our work and also decreases the confuseness in representing the accuracies. Later on that we perform Attribute Ranking to each dataset. The main idea behind this Attribute Ranking is to rank the effectiveness of each individual attribute and choose the best and top attributes that are used for the classification purpose. This analysis is used in removing the irrelevant attributes that are present inside the dataset. This work is based on the Feature Selection. Feature Selection is the process of selecting the number of features or attributes that are included in the dataset to model the problem. Here features are nothing but the individual attributes of a dataset. Generally Feature Selection is classified into two types. They are: Attribute Evaluator and Search Method. The Attribute Evaluator is a technique in which each attribute or feature in a dataset is evaluated in the context of a output (or) target variable. For example, class is a target (or) output variable for a dataset. The Search Method is another technique in which it tries or navigates the different combinations of attributes or features that are present in the dataset in order to arrive on a short list of chosen attributes or features. In this work we perform Feature Selection that is based on the correlation. Pearson's correlation coefficient in statistics is the key coefficient factor that is used to perform correlations to each attribute or feature of a dataset. Ranking Table and Attribute Ranking are the principal mechanisms of this project and is used to predict the unknown values from the known datasets.

Most of the researchers and authors work on predicting the different types of aging diseases. But in our project we consider mainly five different aging diseases based on the two class problem.

The following are some of the authors that contribute their work in predicting the different aging diseases.

Tablle1: Contribution of Authors In Predicting Different Aging Diseases

Ref.no.	Author	Year	Technique used	Highlighted contributions
[1]	José m. Jerez et al.,	2010	Multi-layer perceptron (mlp), self-organisation maps (som) and k-nearest neighbour (knn)	The methods based on machine learning algorithms gives the statistical analysis in the prediction and comparision of patient data.
[2]	Thomas noel et al.,	2016	Support vector machine (svm), decision tree (c4.5), naive bayes (nb) and k nearest neighbors (k-nn)	The objective is to assess the check and correct the classified data with respect to efficiency and effectiveness of each algorithm with the statistical measures as accuracy, precision, sensitivity and specificity.
[3]	M. Kumar et al.,	2016	C4.5, random forest (93%), support vector machines (svm), logistic, nn and naive bayes	To improve the prediction accuracy and finally identify the exact cause for getting the specific disease with the basis of machine learning algorithms.
[4]	T. Panduranga vital et al.,	2015	Knn and svm	A new decision support system is generated for prediction of ckd datasets and gives the accuracy result.
[5]	Abid sarwar et al.,	2013	Naïve bayes, artificial neural networks (ann), and *k*-nearest neighbors (knn)	To calculate the efficiency, time, ROC curves with the results of prediction system and they are compared with the actual medical diagnosis of the specific patients.
[6]	Ioannis kavakiotis et al.,	2017	Support vector machines (svm)	A systematic effort was made to identify and give the review machine learning algorithms and data mining approaches which are applied on dm research.
[7]	Jie luo et al.,	2013	Artificial neural network (ann) and support vector machine (svm)	This method can not only achieve an accurate normal and abnormal classification but also estimate the proceeding stage of severe cases.
[8]	Yojiro sakiyama et al.,	2008	Random forest, support vector machine (svm), logistic regression, and recursive partitioning.	Classification and prediction done by random forest as well as svm yielded kappa values is slightly higher than other machine learning classification tools.
[9]	Nidhi bhatla et al.	2012	Naive bayes, decision tree	The analysis shows that neural network has shown the highest accuracy. On the other hand, decision tree has also performed well with good accuracy

Nidhi Batla et al., applied the basic naïve bayes and the decision tree classifiers for performing the prediction of classification Accuracies. On the other hand decision tree is used to represent the possible decisions and the occurrences of reactions.

Yojiro et al., work of research is mainly depends upon the support vector machine (svm) and Random Forest (RF) classifiers. The entire dataset used in this research is classified into two different classes separated through an optimal Hyper plane.

Marjia et al., applied k-star and C4.5 ml classifiers for the prediction of the classification accuracies. K-star is used to find the k-shortest paths. C4.5 will generate a decision tree that is used for the classification purpose.

3. PROPOSAL MODEL

In this research work, we collect, analyze and predict the different aging disease datasets in order to justify the number of diseased instances and the number of non diseased instances using the various ML classifiers. Each dataset is again classified into training and testing sets. Statistical Analyses is done to all the datasets. Attribute Ranking is also prepared to each dataset in order to know the attributes that are used in a dataset and which attribute is mostly preferable and prioritized. After all the preprocessing work, we prepare a ranking table that defines the datasets and the applied ML classifier to predict the highest average accuracy. This research work may contain other representations included such as the visualization of the decision tree for a particular ML classifier, visualization of the cost curve, Marginal curve, ROC curves, Threshold curves etc., As a result of this research work we concluded in finding the best machine learning classifier in predicting the classification accuracy among the other ML classifiers. The detailed description of each dataset can be seen in the next further steps. In this proposal model we primarily collect the five different aging disease datasets. We choose and apply five basic machine learning classifiers for each dataset to predict the classification accuracies and determine the best ML classifier among the other classifiers and apply rankings to the datasets by making a ranking table. All the datasets contain the data in the form of a CSV file. CSV stands for the Comma Separated Value. The CSV file is a file that stores the data in the form of instances. Each instance in the CSV file represents a diseases or non diseased. All the CSV files contain age as a primary attribute and also contain some other attributes that are common in different aging disease datasets. CSV file is a combination of both the correctly classified instances and incorrectly classified instances. The data in the CSV file is arranged in rows and columns. The rows represent the Instances and the columns represent the Attributes or Features. In this research work we give more priority to the model representations. We built a basic machine learning classifier model and another model that represents our research work in detail.

Basic Machine Learning Model:

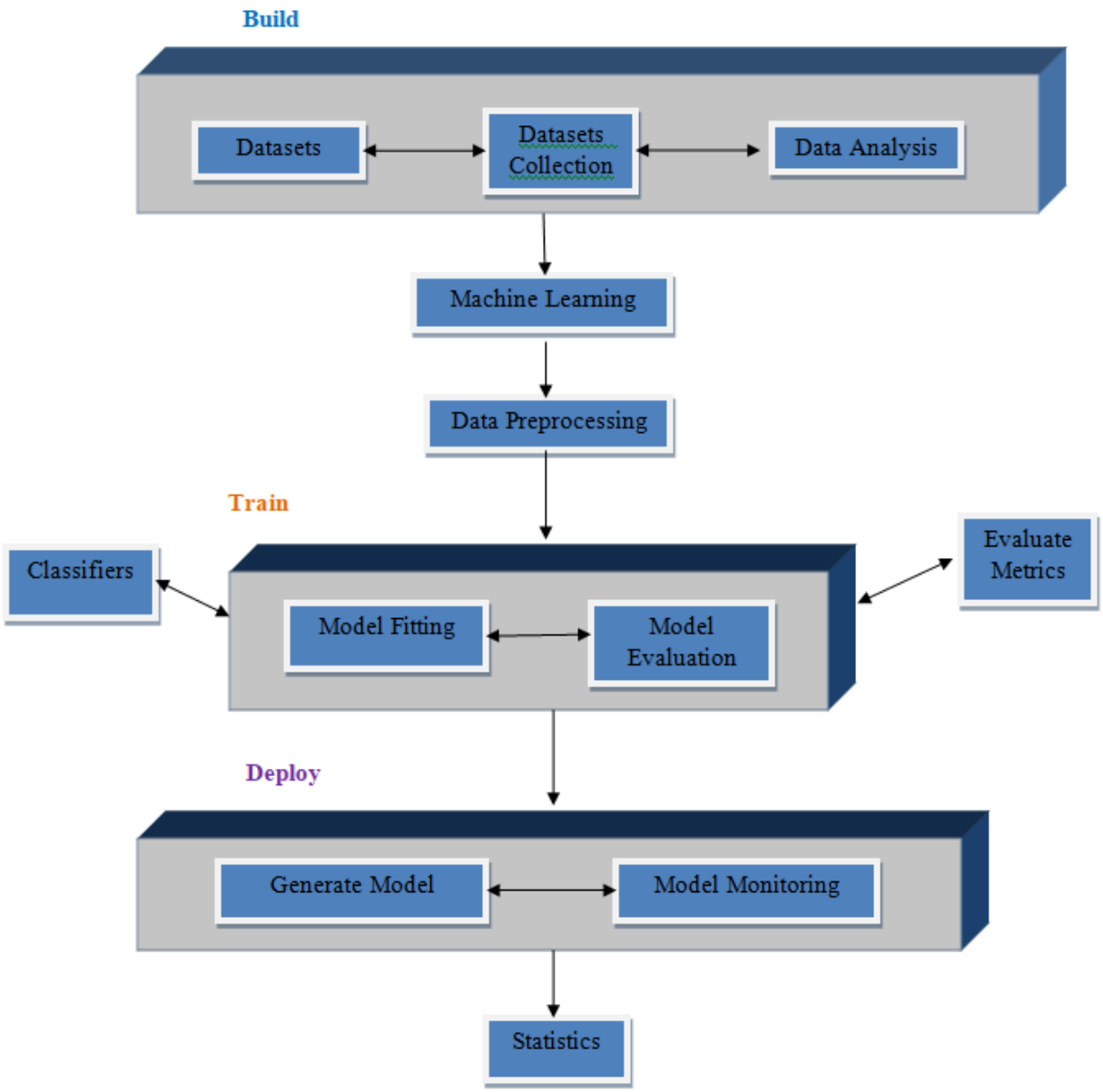


Figure 1: Basic Machine Learning Model

The ML model is basically classified into three sections. They are Build, Train and Deploy sections. In the Build section Datasets are being collected. This mechanism is known as the dataset collection. In this phase datasets are being collected and analyzed. This is known as data analysis. Later on applying machine learning to the collected datasets. After that data preprocessing work is being started. Converting the raw data into cleaned data using various preprocessing techniques is known as the Data Preprocessing. After that training phase will starts. In this section dataset is classified into training and testing sets. Model Fitting is done to

datasets in this phase. Model fitting is the process of applying or fitting an ML classifier or model to a particular dataset. Model Evaluation can also be done by using various evaluation metrics. The last phase of the entire ML model is the deployment section. In this section various statistics are used to represent the class accuracies such as mean, correlation, root mean squared error. These statistics are applied to the final model that is generated by the training phase.

Proposal Model Description

In the above section the basic machine learning can be briefly explained. Now, in this section a detailed overview will be given about how this project works and organized by using this project model. The project model is also having some similar tasks when compared with the Basic ML model. The primary step in this model is also collecting the datasets that are related to aging diseases. Apply some data preprocessing techniques to collected raw data to remove the noise data from the datasets. After that store the preprocessed or cleaned data in the local storage. Apply various Machine Learning Models to each aging disease preprocessed dataset. Perform Attribute Ranking to each dataset by using the feature selection process. After that apply the various classification and regression evaluation metrics to all the datasets. Now calculate the average accuracy obtained by the accuracies of all the machine learning classifiers applied on all the datasets. Prepare a ranking table based on the accuracies obtained and assign individual ranks to the machine learning classifiers. And finally filter the best machine learning classifier among the other machine learning classifiers. This section is just an overview about the project and in the further sections all the techniques used in this project can be briefly explained.

4. DATASETS DESCRIPTION

In this work, we use five different datasets. They are Breast Cancer, Diabetes, Heart, Liver and Kidney datasets. Each dataset can be described as follows:

4.1 Breast Cancer: The abnormal growing of cells in the breast causing Breast Cancer. The cells contain cancer in the breast are said to be Benign cells and the non cancerous cells are said to be malignant cells. The symptoms of breast cancer include a lump in the breast, bloody discharge from the nipple and changes in the shape of the nipple. The dataset contains 10 different attributes such as classes, mitosis, cell size and cell shape etc.,

4.2 Diabetes: Diabetes is a aging disease that occurs when the blood containing glucose is very high. Diabetes can be represented in different types which are named as Type 1 Diabetes, Type 2 Diabetes, Prediabetes, Gestational Diabetes. The major symptoms of diabetes are increased hunger, increased thirst, weight loss, frequent urination. In this diabetes dataset it contains majorly 9 attributes such as Age, glucose, insulin, hemoglobin, pregnancy, outcome etc.,

4.3 Heart: Heart is the major organ in the human body. Any disease caused to heart can be named as a heart disorder. Heart diseases are of several types. It may contain numerous complications such as heart attack, heart failure. Heart diseases are comes under cardiovascular diseases. The major symptoms for heart diseases are Mild pain in the chest, swelling in hands or legs, Irregular heartbeats. This heart dataset determines the target value has the heart disease or not. This dataset contains 14 attributes such as gender, age, cholesterol, Blood pressure etc.,

4.4 Liver: The liver is the principal organ in the human body and the diseases occurs to liver are said to be liver disorders. Liver is used for the digestion and respiration purpose. The major symptoms of liver diseases are skin and eyes that appear yellowish, Itchy skin, Dark urine color. In this liver dataset it contains 11 attributes such as age, gender, alkaline, albumin etc.,

4.5 Kidney: The kidneys are the two bean shaped organs in the human body. Kidneys are used to filter the extra water and outs the waste from the blood and finally filters the blood from waste. Any damage related to kidney leads to the kidney diseases. The major symptoms of the kidney diseases are high blood pressure, severe unintentional weight loss etc., There are around 15 attributes used in this kidney dataset. Some of the major attributes are age, gender, hemoglobin, blood pressure, ph level, sugar etc.,

For all the above datasets 'Age' is the common attribute that is to be used. So age is considered to be as the key factor in each dataset.

5. MACHINE LEARNING CLASSIFIERS DESCRIPTION

In this work, the major ML classifiers that are to be applied on each dataset are can be briefly explained as follows:

5.1 Naïve-Bayes:

The Naïve-Bayes classifier is a machine learning classifier that is used for the classification problems. It involves in high dimensional training datasets. Bayes theorem can be stated as the probability of the event B given A is equal to the probability of the event A given B multiplied by the probability of A upon probability of B. Naïve bayes can be of many types such as Gaussian NB, Multinomial, Bernoulli. Naïve Bayes can be represented as follows:

$$P(A/B) = \frac{P(B/A)\,P(A)}{P(B)}$$

In the above formula, A is called Proposition, B is called Evidence.
P(A) is the prior probability of Proposition
P(B) is the prior probability of Evidence

P(A/B) is the Posterior
P(B/A) is the Likelihood

5.2 K-Star:

K-Star is a heuristic search algorithm that is used to find the k-shortest paths between the designated pair of the vertices in a given directed weighted graph. K* uses the entropic distance measure. It uses the values returned by the distance measure to give a prediction.

$$k^*(b/a) = -\log 2\, p^*(b/a)$$

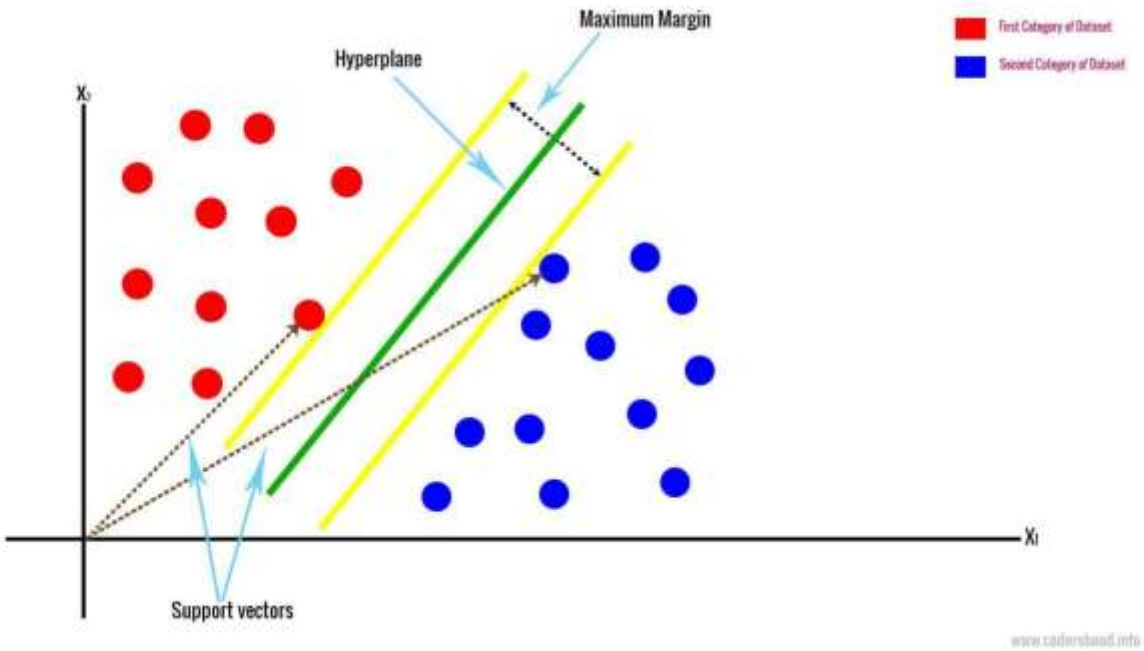
In the above formula,
P is a probability function on T*
a is the starting instance
b is the ending instance

5.3 C4.5:

The C4.5 machine learning classifier is used to develop the decision tree. It is an extension for the ID3 machine learning classifier. These decision trees are generated by the C4.5 that can be used for the classification. C4.5 is referred as the statistical classifier. This classifier builds the decision trees from a set of training data by using the concept Information Entropy. Training data is a set of already classified samples. Each sample consists of a P-dimensional vector.

5.4 Support Vector Machine (SVM):

Support Vector Machine which is simply termed as SVM, is a supervised machine learning classifier which is used for the classification and regression problems. The main objective of the SVM is to define a hyperplane in a N-dimensional space that distinctly classifies two classes. The Hyperplane in SVM is a decision boundary that is used to define and classify the classes or the data points. Each point in a N-dimensional space is called as Vector. The points that are close to the optimal hyperplane on both sides are called as the support vectors.



5.5 Random Forest (RF):

Random Forest, simply termed as RF, is also a supervised machine learning classifier that is used for the classification and regression. Random Forest classifier simply creates the decision trees on the samples of data and gets the prediction from each data sample and selects the best feasible solution by means of voting.

6. EVALUATION METRICS

Metrics means Measurements. The Evaluation Metrics are the measurements that are used to measure the accuracy of each machine learning classifier belongs to a particular dataset. The Evaluation Metrics can be used in two types of problems. They are Evaluation Metrics for Classification Problems and the Evaluation Metrics for Regression problems. For this work we consider some of the major Evaluation Metrics from both the Classification and Regression Problems.

Classification Accuracy: The number of correct predictions to the total number of the input samples is defined as the classification Accuracy.

$$Accuracy = \frac{Number\ of\ correct\ Predictions\ (CP)}{Total\ Number\ of\ Predictions\ made\ (TNP)}$$

TP Rate: TP Rate stands for True Positive Rate. True Positive rate is defined as,

TP/(FN+TP).

FP Rate:FP Rate stands for false positives rate for the given class values.

FP/(TN+FP)

Precision:Precision is classified as the proportion of true instances of a class divided by the total instances classified in a class. It is defined as

TP/(TP+FP)

Recall:Recall is defined as the proportion of instances classified as a given class divided by the actual total in that class. It is equivalent to TP rate. The formula is defined as

TP/(TP+FN)

F-Measure:F-Measure is a combined measure for precision and recall and it is calculated as

2 * Precision * Recall / (Precision + Recall)

MCC: MCC(Matthews correlation coefficient) is used to measure of the quality of binary (two-class) classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure. It can be used in the classes of different sizes.

$$MCC = (TP.TN - FP.FN) / \sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$$

PRC(Precision Recall Curves) area :The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.

All the above metrics comes under classification. The below are the various evaluation metrics that comes under regression.

Kappa Statistic: Cohen's Kappa statistic is a very useful, but under-utilized, metric. Sometimes in machine learning we are faced with a Multi Class Classification problem.

The following tables determine all the evaluation metric values of regression applied on each dataset.

ROC (Receiver Operating Characteristics) Area : The ROC area is used for plotting the true positive rate (TPR) against the false positive rate.

Mean Absolute Error (MAE): The Mean Absolute Error (MAE) is the simplest regression error metric to understand that will be used to calculate the residual for every data point, taking only the absolute value of each so that negative and positive residuals do not cancel out.

Root Mean Squared Error (RMSE): The Root Mean Squared Error (RMSE) is a quadratic scoring rule and also a measurement that measures the average magnitude of the error.

Relative Absolute Error (RAE): The Relative Absolute Error (RAE) is a way to measure the performance of a predictive model. It is expressed as a ratio, comparing a mean error (residual) to errors produced by a trivial or naive model.

Root Relative Squared Error (RRSE): The Root Relative Squared Error (RRSE) is relative to what it would have been if a simple predictor had been used. This simple predictor is just the average of the actual values.

TABLE2: EVALUATION METRICS OF REGRESSION FOR BREAST CANCER

Dataset	Error Metrics	Machine Learning Classifiers				
		Naive Bayes	K-Star	J-48(C4.5)	Random Forest(Rf)	Random Tree(Rt)
Breast Cancer	Kappa Statistic	0.9127	0.5484	0.8799	0.934	0.8604
	Mean Absolute Error (MAE)	0.0403	0.1865	0.0691	0.0611	0.0635
	Root Mean Squared Error (RMSE)	0.1983	0.4213	0.2228	0.1636	0.251
	Relative Absolute Error (RAE)	8.926%	41.2667%	15.2992%	13.5192%	14.0599%
	Root Relative Squared Error (RRSE)	41.7119%	88.6307%	46.8739%	34.4268%	52.8039%

TABLE3: EVALUATION METRICS OF REGRESSION FOR DIABETES DATASET

Dataset	Error Metrics	Machine Learning Classifiers				
		Naive Bayes	K-Star	J-48(C4.5)	Random Forest (RF)	Random Tree(RT)
Diabetes	Kappa Statistic	0.4664	0.2895	0.4164	0.4576	0.3089
	Mean Absolute Error (MAE)	0.2841	0.3275	0.3158	0.3114	0.3151
	Root Mean Squared Error (RMSE)	0.4168	0.4969	0.4463	0.4033	0.5613
	Relative Absolute Error (RAE)	62.5028 %	72.055 %	69.4841 %	68.515 %	69.3289 %
	Root Relative Squared Error (RRSE)	87.4349 %	104.25 %	93.6293 %	84.603 %	117.7696 %

Table4: Evaluation Metrics of Regression for Heart Dataset

Dataset	Error Metrics	Machine Learning Classifiers				
		Naive Bayes	K-Star	J-48(C4.5)	Random Forest (RF)	Random Tree(RT)
	Kappa Statistic	0.6683	0.4516	0.5774	0.6051	0.3571
	Mean Absolute Error (MAE)	0.196	0.2849	0.232	0.2755	0.3208

Heart	**Root Mean Squared Error (RMSE)**	0.352	0.4698	0.4427	0.3644	0.5664
	Relative Absolute Error (RAE)	39.551 %	57.4706 %	46.8051 %	55.4809 %	64.7089 %
	Root Relative Squared Error (RRSE)	69.8355 %	93.2127 %	87.8271 %	73.1178 %	112.365 %

Table5: Evaluation Metrics of Regression for Liver Dataset

		Machine Learning Classifiers				
Dataset	**Error Metrics**	**Naive Bayes**	**K-Star**	**J-48(C4.5)**	**Random Forest (RF)**	**Random Tree(RT)**
Liver	**Kappa Statistic**	0.2468	1	0.6792	1	1
	Mean Absolute Error (MAE)	0.4436	0.0001	0.1921	0.1257	0.0011
	Root Mean Squared Error (RMSE)	0.6568	0.0005	0.3099	0.1592	0.0108
	Relative Absolute Error (RAE)	108.439 %	0.0207 %	46.9636 %	30.717 %	0.2597 %
	Root Relative Squared Error (RRSE)	145.28 %	0.1069 %	68.5359 %	35.203 %	2.3882 %

Table6: Evaluation Metrics of Regression for Kidney Dataset

		Machine Learning Classifiers				
Dataset	**Error Metrics**	**Naive Bayes**	**K-Star**	**J-48(C4.5)**	**Random Forest (RF)**	**Random Tree(RT)**
Kidney	**Kappa Statistic**	0.0949	0.5052	0.5404	0.4704	0.0629
	Mean Absolute Error (MAE)	0.4072	0.2371	0.2864	0.3425	0.4727
	Root Mean Squared Error (RMSE)	0.5899	0.4468	0.432	0.4331	0.6876
	Relative Absolute Error (RAE)	82.4805 %	48.0202 %	58.0028 %	69.3824 %	95.7507 %
	Root Relative Squared Error (RRSE)	119.129 %	90.2366 %	87.2497 %	87.4584 %	138.8538 %

7. RESULTS AND DISCUSSIONS

In this section as per our research work, statistical analysis is made for all the datasets based on the applied ML classifiers. In this statistical analysis, every dataset has to be applied by the ML classifiers. In the previous section which is Evaluation Metrics, we discussed that the classification evaluation metrics will be used for statistical analysis. Below is the following table (Table19) represents the various evaluation metrics including classification and regression. The various evaluation metrics that takes into consideration for this statistical analysis are TP rate, Precision, MCC, ROC area, Time to build the model, Accuracy of each ML classifier belongs to the dataset. Each ML classifier predicts their classification accuracy. The Breast Cancer dataset gives the accuracy of 96.99% with ranking 1 in Random Forest

Classifier. The dataset gives the accuracy of 76.30% with ranking 1 in Naive Bayes Classifier. The Kidney disease dataset gives the accuracy of 78.18% with ranking 1 in J48 Classifier. The Heart dataset gives the accuracy of 83.49% with ranking 1 in Naive Bayes Classifier. The Liver dataset gives the accuracy of 100% with ranking 1 in Random Forest, K-Star, and Random Tree Classifier. Among all the Machine Learning classifiers Random Forest (RF) is the best classifier which takes 0.02 seconds for the construction of the model performed on Breast Cancer Dataset. So according to the Machine Learning analyses, among all the aging diseases datasets, the Breast Cancer gives the best accuracy when compared with the other datasets and Random Forest(RF) classifier is the best classifier among the other classifiers in giving the highest accuracy and taken least amount of time to construct the model. In this Results and Discussions section there are two major tabular representations that contain the major data of the project. The two tabular representations are the Statistical Analysis of the Aging Disease Datasets and the other one is the Ranking Table.

Table7: Statistical Analysis of All the Aging Disease Datasets

Dataset	Metrics	Naïve Bayes	K-Star	J48	Random Forest	Random Tree
Breast Cancer	**TP Rate**	0.960	0.813	0.946	0.970	0.937
	Precision	0.962	0.818	0.946	0.970	0.937
	MCC	0.914	0.572	0.880	0.934	0.937
	ROC Area	0.986	0.840	0.955	0.989	0.930
	Time	0.01	0	0.06	0.08	0
	Accuracy	**95.9943%**	**81.259%**	**94.5637%**	**96.9957%**	**93.7053%**
Diabetes	**TP Rate**	0.763	0.691	0.738	0.758	0.685
	Precision	0.759	0.680	0.735	0.754	0.686
	MCC	0.468	0.293	0.417	0.458	0.309
	ROC Area	0.819	0.714	0.751	0.822	0.655
	Time	0	0	0.03	0.16	0.01
	Accuracy	**76.3021%**	**69.1406%**	**73.8281%**	**75.7813%**	**68.4896%**
Liver	**TP Rate**	0.557	1.000	0.873	1.000	1.000
	Precision	0.796	1.000	0.871	1.000	1.000
	MCC	0.352	1.000	0.681	1.000	1.000
	ROC Area	0.737	1.000	0.919	1.000	1.000
	Time	0.01	0.84	0	0.09	0.01
	Accuracy	**55.7461%**	**100%**	**87.307%**	**100%**	**100%**
Kidney	**TP Rate**	0.527	0.764	0.782	0.745	0.527
	Precision	0.569	0.762	0.783	0.744	0.546
	MCC	0.104	0.509	0.547	0.472	0.064
	ROC Area	0.754	0.817	0.764	0.774	0.533
	Time	0	0	0	0	0
	Accuracy	**52.7273%**	**76.3636%**	**78.1818%**	**74.5455%**	**52.7273%**
Heart	**TP Rate**	0.835	0.726	0.788	0.679	0.797
	Precision	0.835	0.726	0.795	0.679	0.798
	MCC	0.669	0.452	0.583	0.357	0.595
	ROC Area	0.898	0.793	0.802	0.679	0.878
	Time	0.02	0.13	0.01	0	0.02
	Accuracy	**83.4906%**	**72.6415%**	**78.7736%**	**67.9245%**	**79.717%**

Ranking Table for the Aging Disease Datasets

The Ranking Table in this project plays a vital role in examining the best machine learning classifier among the other machine learning classifiers. A rank is given to an individual

machine learning classifier on the basis of the accuracy obtained in each dataset. In this previous table of statistical analysis, the accuracy values are highlighted, because based upon the accuracy values obtained by each dataset the rank is determined to a particular machine learning classifier. Primarily, in this section, each dataset contains five different accuracies of the five machine learning classifiers. Once the individual ranking for each machine learning classifier of each dataset is done, then the average accuracy is to be calculated. The Average Accuracy is calculated on basis of the following computation:

Average Accuracy of each ML Classifier:

$$\frac{\text{Sum of all the accuracies of each dataset for an individual ML classifier}}{\text{Number of Datasets}}$$

According to our analysis, Random Forest is the best machine learning classifier among the other machine learning classifiers when compared with all the datasets.

Table8: Ranking Table For The Aging Disease Datasets

	Datasets	Naive Bayes	K-Star	J48	Random Forest	Random Tree
Accuracy	Breast Cancer	95.9943%(2)	81.259%(5)	94.5637%(3)	96.9957%(1)	93.7053%(4)
	Diabetes	76.3021%(1)	69.1406%(4)	73.8281%(3)	75.7813%2()	68.4896%(5)
	Liver	55.7461%(3)	100%(1)	87.307%(2)	100%(1)	100%(1)
	Kidney	52.7273%(4)	76.3636%(2)	78.1818%(1)	74.5455%(3)	52.7273%(4)
	Heart	83.4906%(1)	72.6415%(4)	78.7736%(3)	67.9245%(5)	79.717%(2)
	Average	72.85208%(5)	79.88094%(3)	82.53084%(2)	83.0494%(1)	78.92784%(4)

8. CONCLUSION

In this work, some of the major Machine Learning Classifiers or Models which are Naïve-Bayes, K-Star, J48, Random Forest and Random Tree have been applied on the five major

aging disease datasets such as Breast Cancer, Diabetes, Heart, Liver and Kidney. Each Machine Learning Classifier has their unique operations that are performed on the aging disease datasets. Also in this project, A Sample working of a machine learning model, A project model, Datasets and Attributes descriptions, Attribute Ranking and Visualizations, Evaluation Metrics, Calculating the average accuracies for each dataset, Statistical Analysis and Ranking of aging disease datasets have been explained in detail. This project allows us to use the different machine learning classifiers on different aging disease datasets and perform different visualizations.

REFERENCES

[1] Kashif, M., Malik, K. R., Jabbar, S., & Chaudhry, J. (2020). Application of machine learning and image processing for detection of breast cancer. In Innovation in Health Informatics (pp. 145-162). Academic Press.

[2] Sadhukhan, S., Upadhyay, N., & Chakraborty, P. (2020). Breast Cancer Diagnosis Using Image Processing and Machine Learning. In Emerging Technology in Modelling and Graphics (pp. 113-127). Springer, Singapore.

[3] Kumar, V., Mishra, B. K., Mazzara, M., Thanh, D. N., & Verma, A. (2020). Prediction of Malignant and Benign Breast Cancer: A Data Mining Approach in Healthcare Applications. In Advances in Data Science and Management (pp. 435-442). Springer, Singapore.

[4] Younus, M., Munna, M. T. A., Alam, M. M., Allayear, S. M., & Ara, S. J. F. (2020). Prediction Model for Prevalence of Type-2 Diabetes Mellitus Complications Using Machine Learning Approach. In Data Management and Analysis (pp. 103-116). Springer, Cham

[5] Baranwal, A., Bagwe, B. R., & Vanitha, M. (2020). Machine Learning in Python: Diabetes Prediction Using Machine Learning. In Handbook of Research on Applications and Implementations of Machine Learning Techniques (pp. 128-154). IGI Global.

[6] Maniruzzaman, M., Rahman, M. J., Ahammed, B., & Abedin, M. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems, 8(1), 7.

[7] Goyal, A., Hossain, G., Chatrati, S. P., Bhattacharya, S., Bhan, A., Gaurav, D., & Tiwari, S. M. (2020). Smart Home Health Monitoring System for Predicting Type 2 Diabetes and Hypertension. Journal of King Saud University-Computer and Information Sciences.

[8] Bhandary, A., Prabhu, G. A., Rajinikanth, V., Thanaraj, K. P., Satapathy, S. C., Robbins, D. E., ... & Raja, N. S. M. (2020). Deep-learning framework to detect lung abnormality–A study with chest X-Ray and lung CT scan images. Pattern Recognition Letters, 129, 271-278

[9] Chandra, T. B., & Verma, K. (2020). Pneumonia Detection on Chest X-Ray Using Machine Learning Paradigm. In Proceedings of 3rd International Conference on Computer Vision and Image Processing (pp. 21-33). Springer, Singapore.

[10] Xu, X., Wang, C., Guo, J., Yang, L., Bai, H., Li, W., & Yi, Z. (2020). DeepLN: A framework for automatic lung nodule detection using multi-resolution CT screening images. Knowledge-Based Systems, 189, 105128.

[11] Jena, L., Nayak, S., & Swain, R. (2020). Chronic Disease Risk (CDR) Prediction in Biomedical Data Using Machine Learning Approach. In Advances in Intelligent Computing and Communication (pp. 232-239). Springer, Singapore.

[12] Alloghani, M., Al-Jumeily, D., Hussain, A., Liatsis, P., & Aljaaf, A. J. (2020). Performance-Based Prediction of Chronic Kidney Disease Using Machine Learning for High-Risk Cardiovascular Disease Patients. In Nature-Inspired Computation in Data Mining and Machine Learning (pp. 187-206). Springer, Cham.

[13] Jena, L., Nayak, S., & Swain, R. (2020). Chronic Disease Risk (CDR) Prediction in Biomedical Data Using Machine Learning Approach. In Advances in Intelligent Computing and Communication (pp. 232-239). Springer, Singapore.

[14] Rubini, L. J., & Eswaran, P. (2015). Generating comparative analysis of early stage prediction of Chronic Kidney Disease. International Journal of Modern Engineering Research (IJMER), 5(7), 49-55.

[15] Amaral, J. L., Lopes, A. J., Jansen, J. M., Faria, A. C., & Melo, P. L. (2012). Machine learning algorithms and forced oscillation measurements applied to the automatic identification of chronic obstructive pulmonary disease. Computer methods and programs in biomedicine, 105(3), 183-193.

[16] Amaral, J. L., Lopes, A. J., Faria, A. C., & Melo, P. L. (2015). Machine learning algorithms and forced oscillation measurements to categorise the airway obstruction severity in chronic obstructive pulmonary disease. Computer methods and programs in biomedicine, 118(2), 186-197.

[17] Kim, S. Y., Diggans, J., Pankratz, D., Huang, J., Pagan, M., Sindy, N., ... & Steele, M. P. (2015). Classification of usual interstitial pneumonia in patients with interstitial lung disease: assessment of a machine learning approach using high-dimensional transcriptional data. The Lancet Respiratory Medicine, 3(6), 473-482.

[18] Ramana, B. V., Babu, M. S. P., & Venkateswarlu, N. B. (2011). A critical study of selected classification algorithms for liver disease diagnosis. International Journal of Database Management Systems, 3(2), 101-114.

[19] Abdar, Moloud, Mariam Zomorodi-Moghadam, Resul Das, and I-Hsien Ting. "Performance analysis of classification algorithms on early detection of liver disease." Expert Systems with Applications 67 (2017): 239-251.

[20] Vijayarani, S., & Dhayanand, S. (2015). Liver disease prediction using SVM and Naïve Bayes algorithms. International Journal of Science, Engineering and Technology Research (IJSETR), 4(4), 816-820.

[21] Kukar, M., Kononenko, I., Grošelj, C., Kralj, K., & Fettich, J. (1999). Analysing and improving the diagnosis of ischaemic heart disease with machine learning. Artificial intelligence in medicine, 16(1), 25-50.