International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

Prediction of Prostate Cancer using Machine Learning Algorithms

Muktevi Srivenkatesh

Abstract: ***Background/Aim:*** ***Prostate cancer is regarded as the most prevalent cancer in the word and the main cause of deaths worldwide****. The early strategies for estimating the prostate cancer sicknesses helped in settling on choices about the progressions to have happened in high-chance patients which brought about the decrease of their dangers.* ***Methods:*** *In the proposed research, we have considered informational collection from kaggle and we have done pre-processing tasks for missing values .We have* ***three missing data values in compactness attribute and two missing values in fractal dimension were replaced by mean of their column values .****The performance of the diagnosis model is obtained by using methods like classification, accuracy, sensitivity and specificity analysis. This paper proposes a prediction model to predict whether a people have a prostate cancer disease or not and to provide an awareness or diagnosis on that. This is done by comparing the accuracies of applying rules to the individual results of Support Vector Machine, Random forest, Naive Bayes classifier and logistic regression on the dataset taken in a region to present an accurate model of predicting prostate cancer disease.* ***Results:*** *The machine learning algorithms under study were able to predict prostate cancer disease in patients with accuracy between 70% and 90%.* ***Conclusions:*** *It was shown that Logistic Regression and Random Forest both has better Accuracy (90%) when compared to different Machine-learning Algorithms.*

Keywords: *Cardiovascular disease,* ***Machine Learning Algorithms, Performance Evaluators, toxins***

I. INTRODUCTION

Classification is significant component of data mining .Classification is the way toward finding a model (or capacity) that depicts and recognizes information classes or ideas. The model is inferred dependent on the investigation of a lot of preparing

Prostate data (i.e., data objects for which the class marks are known).

The model is utilized to foresee the class name of items for which the class name is having the prostate cancer malady or not having prostate cancer ailment that is obscure.

Machine Learning examines how computers can learn (or improve their exhibition) in view of cardiovascular information. The primary research zone is for computer projects to consequently figure out how to perceive complex examples and settle on clever choices dependent on Prostate Cancer data.

Manuscript published on January 30, 2020.
* Correspondence Author
Dr. M, Srivenkatesh*, Associate Professor, Department of Computer Science, GITAM Deemed to be University, Visakhapatnam, India. Email:msvenkatesh9@gmail.com

Prostate Cancer is very important health issue and needs to have very much need to take care. There has been a lot of Supervised learning is fundamentally an equivalent word for arrangement. The supervision in the taking in originates from the named models in the Prostate Cancer preparing data collection.

research on cancer diagnosis by using machine learning techniques. In [1], Decision Tree, Logistic Regression (LR) and Artificial Neural Network (ANN) are employed to evaluated prostate cancer survivability. Cancer survival forecasting may be attempted using model constructed through predictive techniques of various kinds, including statistical multivariate regression and machine learning [2].

The remaining of the research discussion is organized as follows: Section II briefs Literature , Section III describes brief description of selected machine learning algorithms Section IV describes Patient Data Set and attributes, Section V discusses Proposed Technique ,Section VI describes analysis of various algorithms, Section VII Describes Performance measure of classification, Section VIII briefs discussion and evaluated Results, and Section X determines the Conclusion of the research work and last Section describes References.

Prostate Cancer and its Symptoms

Prostate malignant growth is a typical sort of disease in guys, however it is exceptionally treatable in the beginning times. It starts in the prostate organ, which sits between the penis and the bladder.

The prostate has different capacities, including:

- producing the liquid that sustains and ship sperm.
- secreting prostate explicit antigen (PSA), a protein that assists semen withholding its fluid state.
- helping aid urine control.
- Guys who do encounter side effects may take note:
- difficulty beginning and looking after urination
- a visit desire to urine, particularly around evening time
- blood in the urine or semen
- painful urination
- in a few cases, torment on discharge
- difficulty getting or keeping up an erection
- pain or uneasiness when sitting, if the prostate is broadened

Propelled indications

Propelled prostate malignant growth can include the accompanying indications:

- bone break or bone torment, particularly in the hips, thighs, or shoulders

- edema, or growing in the legs or feet
- weight misfortune
- tiredness
- changes in gut propensities
- back torment

II. LITERATURE SURVEY

Srđan Jovi´ca, Milica Miljkovi´cb, Miljan Ivanovi´cb, Milena Šaranovi´cb, and Milena Arsi´ca [3] has explored possibility of prostate cancer prediction by machine learning Techniques and In order to improve the survival probability of the prostate cancer patients they have discussed to make suitable prediction models of the prostate cancer. If one make relevant prediction of the prostate cancer it is easy to create suitable treatment based on the prediction results. Machine learning techniques are the most common techniques for the creation of the predictive models. Therefore in their study several machine techniques were applied and compared. They obtained results were analysed and discussed. They have concluded that the machine learning techniques could be used for the relevant prediction of prostate cancer.

Huaiyu wen, Sufang li, Weili,Jianpingli, Chang yin [4] has studied deep learning method, in their study, artificial neural network and several traditional machine learning techniques are applied to SEER (the Surveillance, Epidemiology, and End Result program) database to classify mortality rate in two categories including less than 60 months and more than 60 months. Their result shows that neural network has the best accuracy (85.64%) in predicting survivability of prostate cancer patients.

Jaegeun Lee, Seung Woo Yang, Seunghee Lee, Yun Kyong Hyon, Jinbum Kim, Long Jin, Ji Yong Lee, Jong Mok Park, Taeyoung Ha, Ju Hyun Shin, Jae Sung Lim, Yong Gil Na, Ki Hak Song [5]has evaluated the applicability of machine learning methods that combine data on age and prostate-specific antigen (PSA) levels for predicting prostate cancer. They have retrospectively reviewed the patients' medical records, analyzed the prediction rate of prostate cancer, and identified 20 feature importance's that could be compared with biopsy results using 5 different algorithms, viz., logistic regression (LR), support vector machine, random forest (RF), extreme gradient boosting, and light gradient boosting machine.

Henry Barlow, Shunqi Mao and Matloob Khushi [6] have developed a pipeline to deal with imbalanced data with data set of 35,875 patients and proposed algorithms to perform preprocessing on such datasets. We evaluated the accuracy of various machine learning algorithms in predicting high-risk prostate cancer. An accuracy of 91.5% can be achieved by the their proposed pipeline, using standard scaling, SVMSMOTE sampling method, and AdaBoost for machine learning. They evaluated the contribution of rate of change of PSA, age, BMI, and filtration by race to this model's accuracy. They identified that including the rate of change of PSA and age in our model increased the area under the curve (AUC) of the model by 6.8%, whereas BMI and race had a minimal effect.

Renato Cuocolo, Maria Brunella Cipullo, Arnaldo Stanzione, Lorenzo Ugga, Valeria Romeo, Leonardo Radice, Arturo Brunetti and Massimo Imbriaco [7] have provided a synopsis of recently proposed applications of machine learning (ML) in radiology focusing on prostate magnetic resonance imaging (MRI). They explained the characteristic of deep learning (DL), a particular new type of ML, including its structure mimicking human neural networks and its 'black box' nature. Differences in the pipeline for applying ML and DL to prostate MRI are highlighted. They discussed gland segmentation; assessment of lesion aggressiveness to distinguish between clinically significant and indolent cancers, allowing for active surveillance; cancer detection/diagnosis and localisation and some more details.

III. MACHINE LEARNING ALGORITHMS

Machine Learning is modernized learning with for all intents and purposes zero human intervention. It incorporates programming PCs so they gain from the open data sources. The guideline inspiration driving AI is to research and manufacture estimations that can pick up from the past data and make desires on new information data.

The contribution to a learning calculation is preparing information, speaking to understanding, and the yield is any mastery, which typically appears as another calculation that can play out an assignment. The info information to an machine learning framework can be numerical, literary, sound, visual, or sight and sound. The relating yield information of the framework can be a gliding point number.

Concepts of Learning

Learning is the way toward changing over understanding into skill or information.

Learning can be comprehensively grouped into three classes, as referenced beneath, in view of the idea of the learning information and association between the student and the earth.

- Supervised Learning process or Supervised Learning Approach.
- Unsupervised Learning process or Unsupervised Learning Approach
- Semi-regulated Learning process or Unsupervised Learning Approach

Correspondingly, there are four classifications of Machine Learning as appeared beneath −

- Supervised learning process/Approach
- Unsupervised learning process/Approach
- Semi-directed learning process/Approach
- Reinforcement learning process/Approach

In any case, the most normally utilized ones are supervised and unsupervised learning

A. Supervised Learning

Machine Learning is normally used in genuine applications, for instance, face and talk affirmation, things or movie proposals, and arrangements assessing. Supervised learning can be moreover requested into two sorts - Regression and Classification. Regression gets ready on and predicts a reliable regarded response, for example foreseeing land costs.

Characterization endeavours to find the correct class name, for instance, looking at valuable/hostile emotions, male and female individuals, kind and undermining tumors, secure and unbound credits, etc.

Supervised learning includes building machine learning model that depends on named tests for instance on the off chance that we construct framework to discover of kind of fever dependent on different highlights of patient like temperature ,force of migraine, body agonies, hack and cool, different status parameters of blood to order quiet is having jungle fever, dingo, viral fever, sine flew and so forth .This is the incentive for class mark.

Supervised learning manages taking in a capacity from accessible preparing information. There are many supervised learning calculations, for example, Logistic Regression, Neural systems, Support Vector Machines (SVMs), and Naive Bayes classifiers.

B. Unsupervised Learning

Unaided learning is utilized to recognize inconsistencies, anomalies, for example, extortion or imperfect gear, or to aggregate clients with comparative practices for a business battle. It is something contrary to managed learning. There is no named data here.

When learning information contains just a few signs with no portrayal or names, it is up to the coder or to the calculation to discover the structure of the basic information, to find shrouded designs, or to decide how to depict the information. This sort of learning information is called unlabeled information.

Assume that we have various information focuses, and we need to characterize them into a few gatherings. We may not actually realize what the criteria of order would be. Along these lines, an unsupervised learning algorithms attempts to characterize the given dataset into a specific number of gatherings in an ideal manner.

Solo learning calculations are very amazing assets for examining information and for recognizing examples and patterns. They are most ordinarily utilized for bunching comparative contribution to consistent gatherings. Solo learning calculations incorporate K-implies, Random Forests, and Hierarchical bunching, etc.

C. Semi-supervised Learning

In the event that some learning tests are marked, yet some other are not named, at that point it is semi-supervised learning. It utilizes a lot of unlabeled data for preparing and a modest quantity of named data for testing. Semi-regulated learning is applied in situations where it is costly to get a completely named dataset while progressively pragmatic to mark a little subset.

D. Reinforcement Learning

Here learning data gives input with the goal that the framework acclimates to dynamic conditions so as to accomplish a specific goal. The framework assesses its exhibition dependent on the input reactions and responds in like manner.

A. Supervised Learning Algorithms

1. K-Nearest Neighbour Algorithm

K-closest neighbors (KNN) algorithm is a kind of supervised machine learning algorithms which can be utilized for both classification as well as regression predictive issues.

- Lazy learning calculation − KNN is a lazy learning algorithm since it doesn't have a specific training phase and uses all the data for training while classification.
- Non-parametric learning calculation − KNN is additionally a non-parametric learning algorithm calculation since it doesn't expect anything about the fundamental data.

K-closest neighbors (KNN) calculation utilizes 'highlight closeness' to anticipate the estimations of new data points which further implies that the new data point will be assigned a value based on how closely it matches the points in the training set. We can comprehend its working with the assistance of following advances –

Stage 1 − For executing any algorithm, we need dataset. So during the initial step of KNN, we should stack the preparation just as test information.

Stage 2 − Next, we have to pick the estimation of K for example the closest data points. K can be any whole number.

Stage 3 − For each point in the test information do the accompanying –

- **3.1**− Calculate the separation between test data and each row of training data with the help of any of the following methods namely:

Euclidean, Manhattan or Hamming distance. The most ordinarily utilized strategy to compute separation is Euclidean.

- **3.2**− Now, based on the distance value, sort them in ascending order.
- **3.3**− Next, it will choose the top K rows from the sorted array.
- **3.4**− Now, it will assign a class to the test point based on most frequent class of these rows.
- **3.5**− Now, it will appoint a class to the test point dependent on the most successive class of these columns.

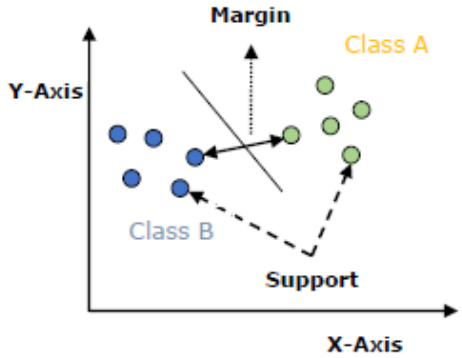
Stage 4 − End

2. Support Vector Machines

Support vector machines (SVMs) are amazing yet adaptable administered machine learning algorithms which are utilized both for classification and regression. SVMs have their one of a kind method for execution when contrasted with other machine learning algorithms. Of late, they are very famous as a result of their capacity to deal with various continuous and categorical variables.

A SVM model is essentially a portrayal of various classes in a hyper plane in multidimensional space.

The hyper plane will be created in an iterative way by SVM with the goal that the mistake can be limited. The objective of SVM is to partition the datasets into classes to locate a most extreme peripheral hyper plane hyperplane(MMH).



- Support Vectors – Data indicates that are nearest the hyperplane is called support vectors. Isolating line will be characterized with the assistance of these data points .
- Hyperplane − As we can find in the above outline, it is a choice plane or space which is isolated between a lot of articles having various classes.
- Margin − It might be characterized as the gap between two lines on the data points of different classes . It tends to be determined as the opposite good ways from the line to the help support vectors. Huge edge is considered as a decent edge and little edge is considered as a terrible edge.

The fundamental objective of SVM is to separate the datasets into classes to locate a most extreme minor hyperplane (MMH) and it very well may be done in the accompanying two stages −

- First, SVM will produce hyper planes iteratively that isolates the classes in most ideal manner.
- Then, it will pick the hyper plane that isolates the classes effectively.

3. Logistic Regression

Linear Regression isn't constantly fitting on the grounds that the data may not fit a straight line yet in addition the straight line esteems can be more prominent than 1 and under 0 .Thus ,they surely can't be utilized as the likelihood of event of the objective class .Under these circumstances logistic regression is used . Instead fitting data into straight line logistic regression uses logistic curve.

Simple Logistic Regression

Output = 0 or 1, Hypothesis => Z = WX + B hΘ(x) = sigmoid (Z)

Sigmoid Function

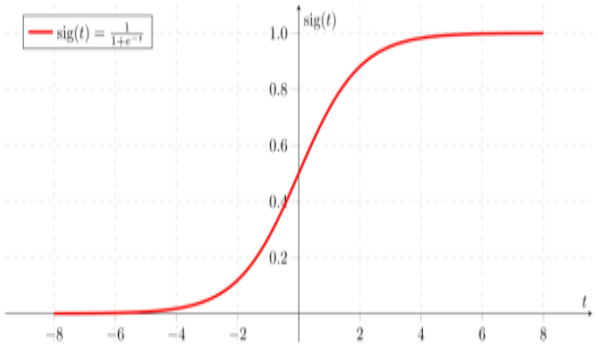


Figure 2: Sigmoid Activation Function

If 'Z' goes to infinity, Y(predicted) will become 1 and if 'Z' goes to negative infinity, Y(predicted) will become 0.

This type of regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

In basic words, the dependent variable is double in nature having information coded as either 1 (represents achievement/yes) or 0 (represents disappointment/no).

Scientifically, a calculated this model predicts P(Y=1) as an element of X. It is one of the

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems .In our example Sorts of Logistic Regression For the most part, strategic regression implies twofold calculated regression having paired objective factors, however there can be two additional classes of target factors that can be anticipated by it. In view of that number of classifications, Logistic regression can be separated into following sorts −

Parallel or Binomial

In such a sort of arrangement, a needy variable will have just two potential sorts either 1 or 0. For instance, these factors may speak to progress or disappointment, yes or no, win or misfortune and so on.

Multinomial

In such a sort of arrangement, subordinate variable can have at least 3 potential unordered sorts or the sorts having no quantitative hugeness. For instance, these factors may speak to "Type A" or "Type B" or "Type C".

Ordinal

In such a sort of characterization, subordinate variable can have at least 3 potential arranged sorts or the sorts having a quantitative centrality. For instance, these factors may speak to "poor" or "great", "generally excellent", "Superb" and every classification can have the scores like 0,1,2,3.

Numerically, a strategic relapse model predicts P(Y=1) as a component of X. It is one of the least difficult ML calculations that can be utilized for different characterization issues.

Regression Models

- Binary Logistic Regression Model − The most straightforward type of strategic regression is parallel or binomial calculated regression in which the objective or ward variable can have just 2 potential sorts either 1 or 0.
- Multinomial Logistic Regression Model − another valuable type of calculated regression is multinomial strategic regression in which the objective or ward variable can have at least 3 potential unordered sorts for example the sorts having no quantitative hugeness.

E. Naive Bayes

The Bayes Rule and Naïve Bayes Classification

The Bayes Rule is a method for going from P(X|Y), known from the preparation dataset, to discover P(Y|X).

What occurs if Y has multiple classes? we process the likelihood of each class of Y and let the most elevated success.

P(X/Y)= P(X ∩ Y)/P(Y) [P(Evidence/Outcome) (Known from Training Data)]

P(Y/X)= P(X ∩ Y)/P(X) [P(Outcome/Evidence) (To be Predicted for Test Data)]

Naïve Bayes calculations are an arrangement method dependent on applying Bayes' hypothesis with a solid supposition that every one of the indicators are autonomous to one another. In basic words, the assumption is that the nearness of a component in a class is autonomous to the nearness of some other element in a similar class

In Bayesian portrayal, the rule interest is to find the back probabilities for instance the probability of a name given some watched features, $(L \mid features)$. With the help of Bayes speculation, we can express this in quantitative structure as seeks after −

P(L|features)=P(L)P(features|L)/P(features)

Here, $(L \mid features)$ is the posterior probability of class.

(L) $L)$ is the earlier probability of class.

$(features|L)$ is the likelihood which is the probability of marker given class.

$(features)$ is the earlier probability of pointer.

F. Random Forest

Random forest is a supervised learning which is utilized for both classifications just as regression. In any case, be that as it may, it is principally utilized for classification issues. As we realize that a forest is comprised of trees and more trees implies progressively robust forest. So also, arbitrary random forest algorithm makes choice trees on data samples and afterward gets the forecast from every one of them lastly chooses the best solution by methods for casting a vote. It is an outfit strategy which is superior to anything a solitary choice tree since it decreases the over-fitting by averaging the outcome.

Random Forest Algorithm

- Step 1 − First, start with the choice of random samples from a given dataset.
- Step 2 − Next, this calculation will build a choice tree for each example. At that point, it will get the forecast outcome from each choice tree.
- Step 3 − In this progression, casting a ballot will be performed for each anticipated outcome.
- Step 4 − At last, select the most casted a ballot forecast result as the final prediction result.

IV. PATIENT DATA SET

The complete of 100 cases with ten attributes was amassed for the Prostate Cancer data set from kaggle. The attribute “diagnosis” described as the measurable and zero indicates patients are not having prostate cancer and one indicates patients are having prostate cancer .Table I suggests the attributes values of prostate cancer data set .The data set having 38 no prostate cancer cases and 62 prostate cancer yes cases .

Table 1: Prostate Cancer Data Set

Serial Number	Attribute
1	ID
2	Radius
3	Texture
4	Perimeter
5	Area
6	Smoothness
7	Compactness
8	Symmetry
9	Fractal Dimension
10	Diagnosis

V. PROPOSED TECHINQUE

The principle destinations of this examination are to propose a technique that can create best Machine Learning algorithm for prediction of Prostate Cancer disease. We have considered various machines learning algorithms and their various performance metrics have compared.

1. Selection

We have considered Prostate Cancer data set from Kaggle .We have considered 10 attributes of Prostrate data set as stated in section IV .They are 100 tuples in this data set and this set having 32 yes (having prostate cancer disease)cases and 68 no cases(not having prostate cancer disease) .

2. Pre-processing and Transformation

The prostate cancer dataset is set up in Comma Separated Document format (CSV) from Excel File. Different things required are the expulsion of right qualities for missing records, copy records evacuate pointless information field, standard information position, adjust information in a convenient way and so on. The considered prostate cancer data set have three missing data values in compactness attribute and two missing values in fractal dimension were replaced by mean of their column values .

3. Performance Evaluation

The performance evaluation of various machine learning algorithms like correctly classified instances, incorrectly classified instances, kappa statistic, Mean absolute error (MAE), Root Mean square error (RMSE),Relative Absolute Error, Root Relative Square Error are to be discussed. We are about to do calculation of True positive rate, False positive rate Precision, Recall, F-Measure and confusion matrix of various considered machine learning algorithms.

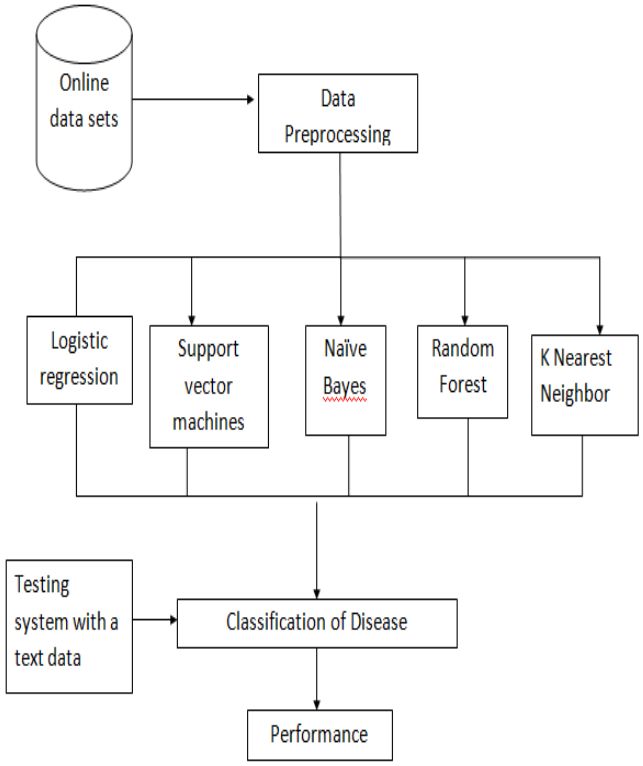


Figure 3: Architecture diagram of cardiovascular disease prediction system

VI. PERFORMANCE MEASURES FOR CLASSIFICATION

One can use following execution measures for the request and figure of imperfection slanted module as shown by his/her own need. Confusion Matrix: The confusion matrix is used to measure the introduction

of two class issue for the given instructive record. The right corner to corner parts TP (True positive) and TN (True Negative) adequately describe Instances similarly as FP (false positive) and FN (false negative) wrongly request Instances. Confusion Matrix Correctly Classify Instance TP+TN Incorrectly Classify Instances.

- True positives imply the positive prostate cancer tuples that were precisely named by the classifier,
- True negatives are the negative prostate cancer tuples that were precisely set apart by the classifier.
- False positives are the negative prostate cancer tuples that were erroneously set apart as positive tuples
- False negatives are the positive prostate cancer tuples that were incorrectly stamped negative tuples
- A confusion matrix for positive and negative tuples is as follows

Predicted Class

Table 2: Components of Confusion Matrix

		Yes	No	
Actual Class	Yes	True Positives(TP)	False Negatives(FP)	P
	No	False Positives(FN)	True Negatives(TN)	N
		P Complement	N Complement	P+N

A confusion matrix for positive and negative prostate cancer tuples for the considered data set is as follows

Table 3: Confusion Matrix of Various Algorithms

Name of the algorithm	Classes	Prostate Cancer = yes	Prostate Cancer = no
K-Nearest Neighbour	Prostate Cancer = yes	4	2
	Prostate Cancer =no	0	4
	Total	4	6
Support Vector Machines	Prostate Cancer = yes	3	3
	Prostate Cancer =no	0	4
	Total	3	7
Logistic Regression	Prostate Cancer = yes	5	1
	Prostate Cancer =no	0	4
	Total	5	5
Naive Bayes	Prostate Cancer = yes	4	2
	Prostate Cancer =no	0	4
	Total	4	6
Random Forest	Prostate Cancer = yes	5	1
	Prostate Cancer =no	0	4
	Total	5	5

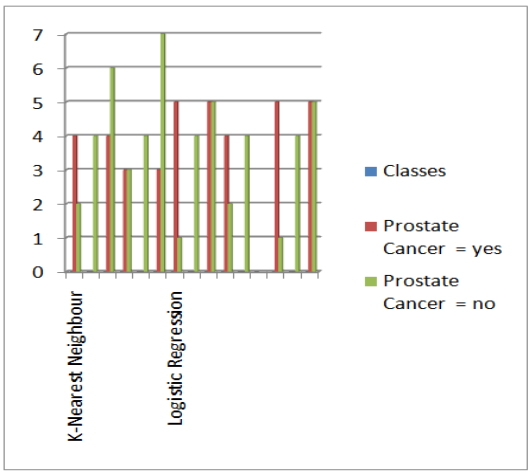


Figure 4: Graphical Presentation of various algorithms

The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. That is,

Table 4: Various Measurements Formula

Measure	Formula
Accuracy, Recognition Rate	$\frac{TP+TN}{P+N}$
Error ,Misclassification Rate	$\frac{FP+FN}{P+N}$
Sensitivity, True Positive rate, Recall	$\frac{TP}{P}$
Specificity, True Negative Rate	$\frac{TN}{N}$
Precision	$\frac{TP}{TP+FP}$
F, F1, F-score, Harmonic mean of precision and recall	$\frac{2* Precision* Recall}{Precision +Recall}$

Table 5: Results of Precision, Recall, F1-Score for various algorithms with prostate cancer data set

Name of the algorithm	Averages	Precision	Recall	F1-Score	Support
K-Nearest Neighbour		1.00	0.67	0.80	6
		0.67	1.00	0.80	4
	Micro Average	0.80	0.80	0.80	10
	Macro Average	0.83	0.83	0.80	10
	Weighted Average	0.87	0.80	0.80	10
Support Vector Machines		1.00	0.50	0.67	6
		0.57	1.00	0.73	4
	Micro Average	0.70	0.70	0.70	10
	Macro Average	0.79	0.75	0.70	10
	Weighted Average	0.83	0.70	0.69	10
Logistic Regression		1.00	0.83	0.91	6
		0.80	1.00	0.89	4
	Micro Average	0.90	0.90	0.90	10
	Macro Average	0.90	0.92	0.90	10
	Weighted Average	0.92	0.90	0.90	10
Naive Bayes		1.00	0.67	0.80	6
		0.67	1.00	0.80	4
	Micro Average	0.80	0.80	0.80	10
	Macro Average	0.83	0.83	0.80	10
	Weighted Average	0.87	0.80	0.80	10
Random Forest		1.00	0.83	0.91	6
		0.80	1.00	0.89	4
	Micro Average	0.90	0.90	0.90	10
	Macro Average	0.90	0.92	0.90	10
	Micro Average	0.92	0.90	0.90	10

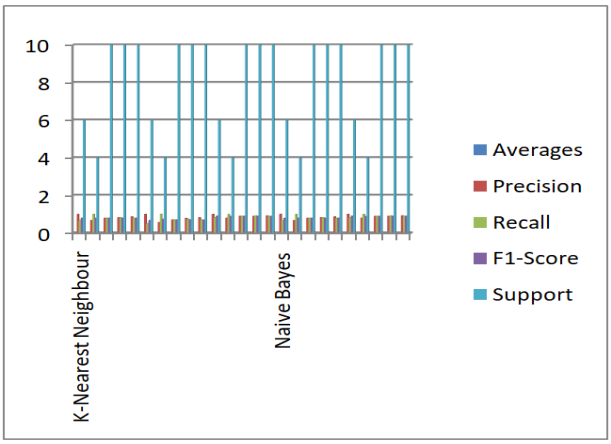


Figure 5: Comparison of Micro, Macro, and Weighted Average of various algorithms

Table 6: Accuracy Measure for Prostate Cancer Dataset

Name of the Algorithm	Correctly Classified instances (%)	Incorrectly Classified instances (%)
K-Nearest Neighbour	80.0	20.0
Support Vector Machines	70.00	30.0
Logistic Regression	90.00	10.00
Naive Bayes	80.0	20.0
Random Forest	90.00	10.00

Table 7: Accuracy Measure for Prostate Cancer Dataset

Name of the Algorithm	Kappa Statistics	Mean Absolute Error
K-Nearest Neighbour	0.61	0.2
Support Vector Machines	0.44	0.3

Logistic Regression	0.8	0.1
Naive Bayes	0.61	0.2
Random Forest	0.8	0.1

Table 8: Accuracy Measure for Prostate Cancer Dataset

Name of the Algorithm	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Square Error (%)
K-Nearest Neighbour	0.44	41.66	29.01
Support Vector Machines	0.54	62.49	43.52
Logistic Regression	0.31	20.83	14.50
Naive Bayes	0.44	41.66	29.01
Random Forest	0.31	20.83	14.50

1. Correctly and Incorrectly Classified Instances: Correctly classified instances mean the sum of True Positives and True Negatives of prostate cancer data set tuples. Similarly, incorrectly classified instances means the sum of false positive and False Negatives of prostate cancer data sets. The total number of correctly prostate cancer data instances divided by total number of prostate data instances gives the accuracy.

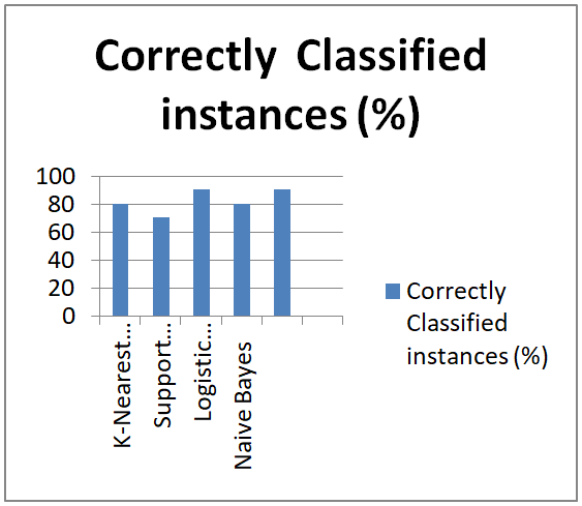


Figure 5: Comparison of incorrectly classified instances of various algorithms

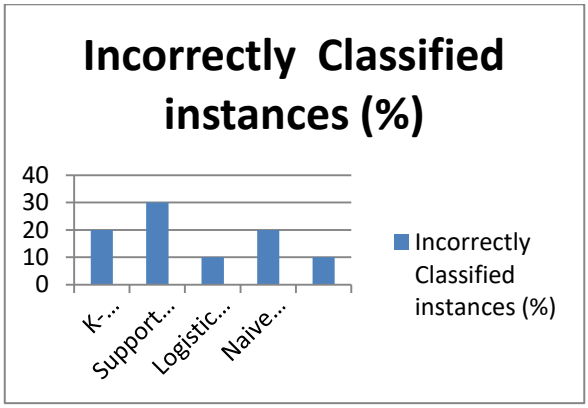


Figure 6: Comparison of incorrectly classified instances of various algorithms

2. Kappa Statistics: The kappa measurement is a proportion of how intently the prostate cancer data instances characterized by the machine learning classifier coordinated the prostate data named as ground truth, controlling for the exactness of an irregular classifier as estimated by the normal precision.

3.

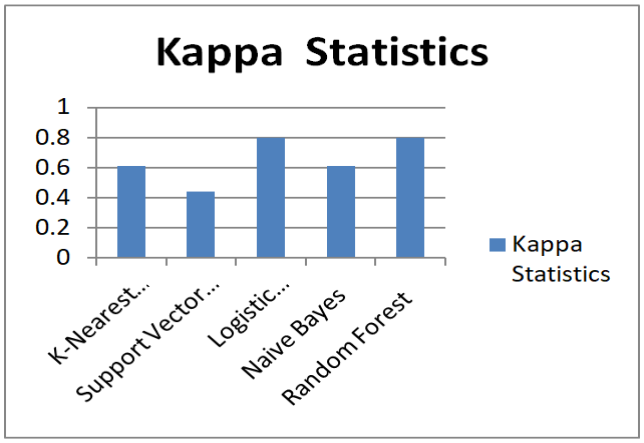


Figure 7: Comparison of kappa statistic of various algorithms

4. Mean Absolute Error: Mathematical representation of mean absolute error (MAE) is the mean prostate cancer test instances of the absolute difference between predicted and actual results.

$$MAE = \frac{1}{N}\sum_{j=1}^{n}|yi - y'i|$$

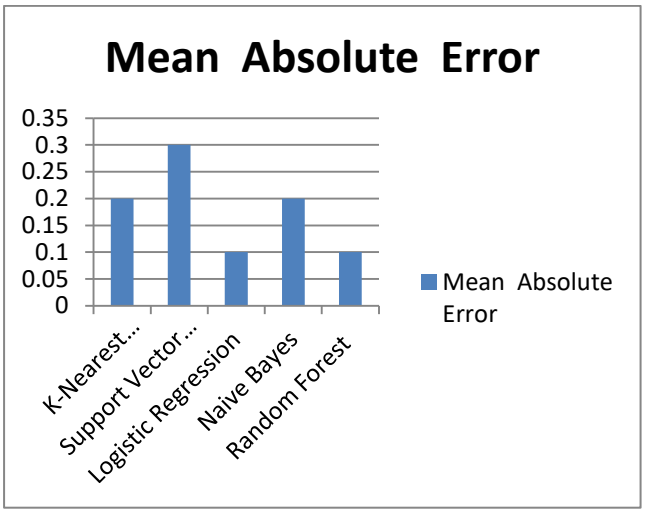


Figure 8: Comparison of Mean Absolute Error of various algorithms

5. Root Mean Squared Error: The size of root mean squared error (RMSE) is determined and It's the square base of the normal of squared contrasts among anticipated and genuine outcomes.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j-1}^{n}(yi - y'i}$$

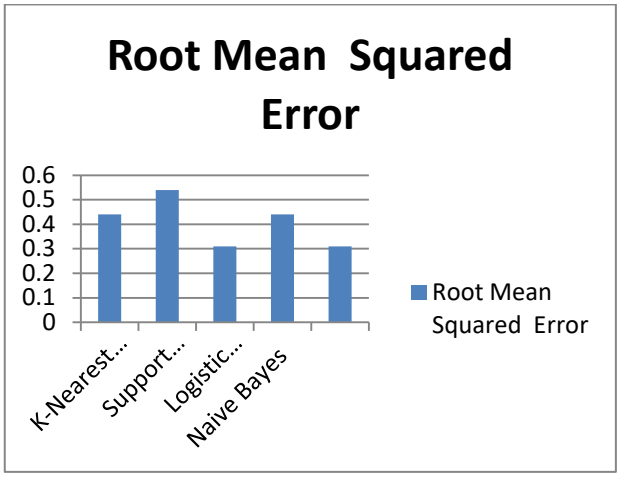


Figure 9: Comparison of Root Mean Squared Error of various Algorithms

6. Root Absolute Error: It is the root of Absolute Error. It is one of the important performances Measure for machine learning algorithms.

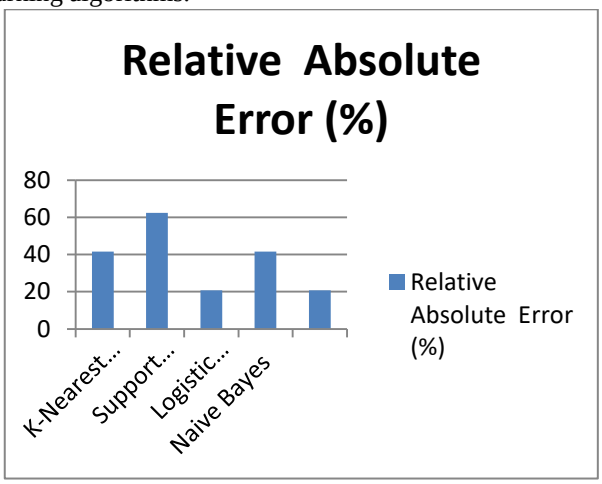


Figure 10: Comparison of Relative Absolute Error

7. Root Relative Squared Error: It is the root of relative squared Error. It is also one of the important performances Measure for machine learning algorithms.

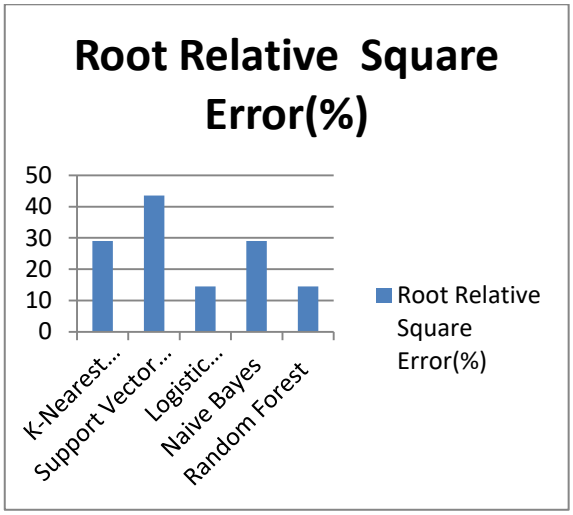


Figure 10: Comparison of Root Relative Square Error of various Algorithms

VII. DISCUSSION AND RESULTS

In this assessment, we applied Machine Learning Algorithms on prostate cancer data set to foresee patients who have interminable prostate cancer ailment, and the individuals who are not debilitated, in light of the information of each characteristic for every patient. Our objective was to think about various arrangement models and characterize the most productive one. Our examination was made based on five calculations positioned among the K-Nearest Neighbour, Support Vector Machines, Logistic Regression, Naive Bayes, Random Forest.

From the above tables 6,7,8 ,We have showed that Logistic Regression and Random Forest have highest accuracy when compared to remaining algorithms.

Table 9: Accuracy Measure for Prostate Cancer Dataset

Name of the Algorithm	Correctly Classified instances (%)	Incorrectly Classified instances (%)
Logistic Regression	90.00	10.00
Random Forest	90.00	10.00

Table 10: Accuracy Measure for Prostate Cancer Dataset

Name of the Algorithm	Kappa Statistics	Mean Absolute Error
Logistic Regression	0.8	0.1
Random Forest	0.8	0.1

Table 11: Accuracy Measure for Prostate Cancer Dataset

Name of the Algorithm	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Square Error(%)
Logistic Regression	0.31	20.83	14.50
Random Forest	0.31	20.83	14.50

Both Logistic Regression and Random Forest has highest number of correctly classified instances that is 90.00 and it has lees number of in correctly classified instances that is 10.00 and when compared to remaining three algorithms

Concerning estimation of indicators, the estimations of Mean total error(MAE), Root Mean Square Error(RMSE), Relative Absolute Error(RAE), Root Relative Square Error (RRSR) demonstrated that Logistic Regression Random Forest indicators scored the most reduced qualities (MAE = 0.1) (RMSE = 0.31, RAE =20.83%, RRSE = 14.50%) trailed by different calculations .

VIII. CONCLUSION

As end, the use of information digging systems for prescient examination is significant in the wellbeing field since it enables us to confront ailments prior and accordingly spare individuals' lives through the expectation of fix.

In this work, we utilized a few learning calculation K-Nearest Neighbour, Support Vector Machines, Logistic Regression, Naive Bayes, Random Forest to foresee patients with constant prostate cancer disappointment issue and patients who are experiencing this illness. Re-enactment results demonstrated the Logistic Regression and Random Forest classifiers demonstrated its exhibition in foreseeing with best outcomes regarding precision and least execution time.

REFERENCES

1. Chih-Jen Tseng,Chi-Jie Lu,Chi-Chang Chang, GinDen Chen, Application of machine learning to predict the recurrence-proneness for cervical cancer. Neural Computation and Application (2014) 24: PP.1311-1316.
2. Floyd S, Ruiz C, Sergio AA, Tseng J, Whalen G, Prediction of Pancreatic Cancer Survival through Automated Selection of Predictive Models. Biomedical Engineering Systems and Technologies Volume 127 of the series Communications in Computer and Information Science pp. 29–43.
3. Srđan Jovi´ca, Milica Miljkovi´cb, Miljan Ivanovi´cb, Milena Šaranovi´cb, and Milena Arsi´ca ,Prostate Cancer Probability Prediction by Machine Learning Technique, Cancer Investigation ,2017,vol 35,No .10,PP 647-651
4. Huaiyu wen, Sufang li, Weili,Jianpingli, Chang yin, Comparison of four Machine Learning Techniques for the Prediction of Prostate Cancer Survivability IEEE Truncations,2018.
5. Jaegeun Lee, Seung Woo Yang, Seunghee Lee, Yun Kyong Hyon, Jinbum Kim, Long Jin, Ji Yong Lee, Jong Mok Park, Taeyoung Ha, Ju Hyun Shin, Jae Sung Lim, Yong Gil Na, Ki Hak Song,Machine Learning Approaches for the Prediction of Prostate Cancer according to Age and the Prostate-Specific Antigen Level, Korean journal of urological oncology 2019; 17(2): PP.110-117.
6. Henry Barlow, Shunqi Mao and Matloob Khushi**,** Predicting High-Risk Prostate Cancer Using Machine Learning Methods, Data MDPI, September 2019,PP.1-15.
7. Renato Cuocolo, Maria Brunella Cipullo, Arnaldo Stanzione, Lorenzo Ugga, Valeria Romeo, Leonardo Radice,Arturo Brunetti and Massimo Imbriaco, Machine learning applications in prostate cancer magnetic resonance imaging, European Radiology Experimental (2019) 3:35

AUTHOR PROFILE

Dr. M. Srivenkatesh working as Associate Professor, Department of Computer Science, GITAM Deemed to be University, Visakhapatnam, Andhra Pradesh, India .He has Published Eleven international Journal papers. His research interest includes Data mining, Machine Learning, Software Engineering, Cloud Computing, Rough Sets. Nine Research Scholars are working for their Ph.D. in computer Science under his guidance.