An Efficient Data Mining Techniques - Multi-Objective KNN Algorithm to Predict Breast Cancer

T.Mohana Priya M.Punithavalli

ABSTRACT--- Breast cancer becomes most important foundation of mortality among women. The convenience of medical related dataset and data investigation support to extracting unidentified pattern in medical related or health related dataset. The objective of this research work is is to develop a health care prediction tool predicts the occurrence of the disease at near the beginning level of the criteria by analyzing the collected data set attributes to extract the disease exact level from the medical related information. The projected multi-objective KNN machine learning algorithm (classification) confirms that the highest accuracy (97.16%) is achieved compared to existing decision tree and Random Forest Techniques.

Keywords— Breast cancer, risk prediction, genetic factors, hormone receptor status

INTRODUCTION

AI is a division of man-made reasoning that is adequately connected to different characteristic issues. Only, AI centers around creating computational strategies utilizing dataset to make precise forecasts. Especially, AI methods give promising instruments in simulated intelligence connected to the early recognition of disease and improve the nature of lives of malignant growth patients in the expected years. Malignant growth is a multifaceted illness and represents an overall danger that is ascribed to the enhanced death rates related with it.

LITERATURE REVIEW

Results from genome wide affiliation considers (GWAS) are consistently adding as far as anyone is concerned of hereditary hazard factors for Breast malignancy [1– 13]. Despite the fact that impacts for single quality variations are little, in total they may in the long run clarify a sizable extent of heritable Breast disease hazard, and there is expanding enthusiasm for using data from basic hereditary polymorphisms for Breast malignancy chance forecast. Hazard expectation models can be a significant device for Breast malignant growth aversion, by distinguishing ladies at high hazard who might for the most part profit by focused preventive estimates.

Now proposals for recognizing ladies at adequately high hazard to profit by chemoprevention incorporate reference to the Breast Cancer Risk Assessment Tool (BCRAT) initially created by Gail et al. [14] with the intend to lessen costs as far as money related cost, yet additionally to advance expected health advantages against conceivable negative symptoms (for example expanded danger of endometrial malignancy) [15]. Similarly, in the light of new outcomes on the constrained benefit of mammography viewing for certain ladies [16], which should be adjusted against monetary expenses just as conceivable negative reactions, for example, radiation and over diagnosis or false positive conclusion, it seems advantageous to likewise consider the use of hazard expectation models with regards to mammography screening [17– 19].

The Cancer Cohort Consortium offers an extensive and all around described examination populace with both traditional epidemiologic hazard factor and hereditary information [20], which permit the calculation and assessment of exhaustive hazard forecast models. Here we present outcomes from this asset, assessing the group prescient nature of 32 regular quality variations that were accounted for to be related with Breast malignant growth in something like one GWAS at genome-wide noteworthiness level [1– 13]. We researched danger of Breast disease by and large just as by subtypes characterized by estrogen and progesterone receptor status. Other than investigations of the unfair capability of hereditary and non-hereditary hazard factor data, we likewise made an interpretation of our outcomes to appraisals of outright hazard.

MATERIAL AND METHODS

Breast and Prostate Cancer Cohort Consortium, proposed systems broke down 6025 constant Breast malignant growth cases and 7825 coordinated controls of European parentage, with information on traditional Breast disease risk factors and 32 normal quality variations recognized through GWAS. One-sided capacity as for Breast malignant growth of explicit hormone receptor-status was inspected with the age and accomplice balanced concordance measurement. Complete hazard scores were determined with outside reference information. Incorporated separation improvement (IDI) was utilized to figure enhancements in hazard conjecture.

PROPOSED FRAMEWORK & RESULTS

The structure is collected from the following key phases:

- Dataset Identification
- Preprocessing
- Classification using Support vector machine algorithm, Linear Regression and KNN.
- Highest accuracy
- Trained model for prediction

Revised Manuscript Received on August 19, 2019.

T.Mohana Priya, Research Scholar, Bharathiar University, Coimbatore, Assistant Professor, Dr. SNS Rajalakshmi College of Arts and Science (Autonomous), Coimbatore.

Dr.M.Punithavalli, Associate Professor, Department of Computer Applications, Bharathiar University, Coimbatore.

Data Source

UCI data repository provides the medical data set with different attributes to extract or analyze the disease based on the existing criteria. Breast Cancer Dataset contains 655 instances out of which 17 attributes of the data set suffered due to missing values. Dataset is dispersed over of 65.5% diseased samples and 34.5% of non diseased samples. The whole facts of all the eight features are shown below in table 1.

Table 1. Analysis of Dataset

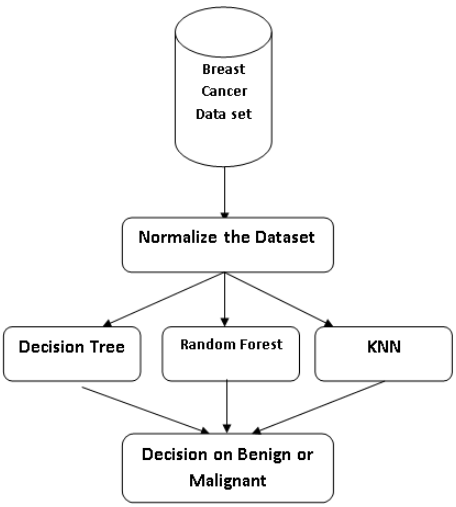
No	Attribute	Value	Mean	SD
1	Clump	1-10	4.45	2.87
2	Consistency of cell size	1-10	3.19	3.01
3	steadiness of cell shape	1-10	3.20	2.68
4	Trivial Adhesion	1-10	2.81	2.68
5	Single epithelial size	1-10	2.31	2.21
6	Bare nuclei	1-10	3.56	3.56
7	Bland Chromatin	1-10	3.45	2.54
8	Normal nuclcoli	1-10	2.98	3.15
9	mitoses	1-10	1.58	1.82

Training and Classification

Order of the informational collection is finished based on precise properties groups by the sample attribute which was extracted, that ready to arrange them, and each example variable is allotted a dangerous or favorable class. The order reason set for this learning is to accomplish improved exactness by utilizing DT, RF and KNN classifier techniques and figure out which one suits the most for diabetes characterization method. We set up the classifier with known example information in a preparation dataset and check its presentation by analytical the test dataset, which comprises of the obscure example used to foresee its class mark KNNeighbors Classifier is an administered, occasion based taking in classifier which gains from the labeled information tests.

Dataset

In this research work, dataset contains 768 instances and 8 attributes are used in this comparative analysis.



Methodology

Fig 2: Proposed Methodology

Step 1: KNeighbor Classifier(x,y,x1,n)
Step 2: For i=0 to n do
Step 3: calculate distance D(x1,x)
Step 4: while(i<=m)
Step 5: Compute set Z
Step 6: for k =1 to n
Step 7: calculate distance D(x1,x)
Step 8: return

ALGORITHM: KNN CLASSIFIER PSEUDO CODE

Performance Metrics

Accuracy

Accuracy is calculated from the extracted data attributes which is correctly classified item divided by the whole number of items are present in the dataset.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Where TP- True Positive, FP- False Positive, TN- True Negative, FN- False Negative

TP Rate

It is the capability which is used to find the high true-positive rate. The true-positive rate is also called as sensitivity.

$$TPR = \frac{TP}{TP+FN} \quad (2)$$

Precision

Precision is calculates with the correlation of number of modules properly classified to the number of entire modules classified fault-prone. It is quantity of units correctly predicted as faulty.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

F-Measure

F- Measure is the one of the mechanism to grouping of both precision and recall which is used to calculate the score from the predicted results.

$$F - Measure = 2 * recall * precision\, recall + precision \quad (4)$$

Experimental Results

Research study focused and implemented in Weka research tool. The weka tool is a common tool for all machine learning applications and has the number of feature for data pre-processing, data-classification, regression analysis, clustering, association rule mining techniques, and mechanism of visualization. The assessment of classification algorithms are based on the performance actions of different attributes in classification techniques to predict the better accuracy and execution time.

Table 2: Accuracy Measure for Classifier Algorithms

Name of the Algorithm	True instances (%)	Wrong instances (%)	TP Rate	F-Measure	IR Precision	IR Recall	Accuracy (%)	Execution Time
Decision Tree	215	71	0.85	0.81	0.78	0.85	64%	0.1
Random Forest	217	69	0.95	0.84	0.75	0.95	60%	0.2
KNN	218	63	0.78	0.76	0.81	0.79	67%	0.1

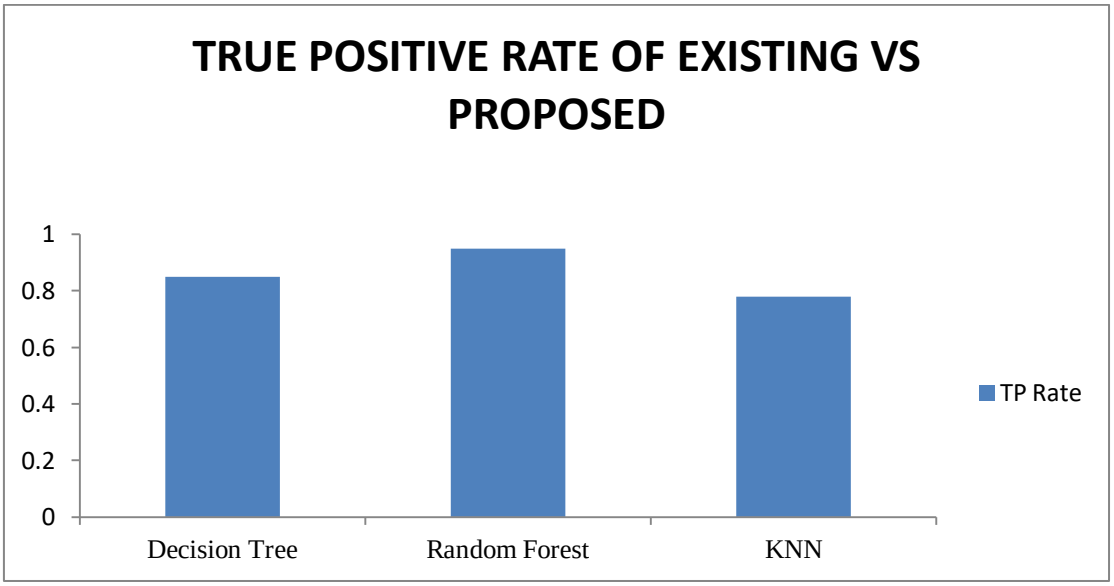


Fig 3: Comparison of True Positive Rate

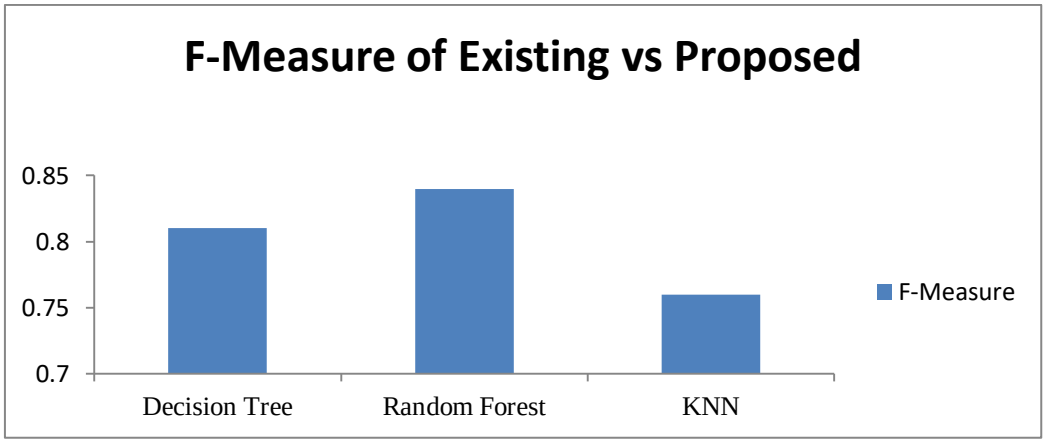


Fig 4: Comparison of F-Measure

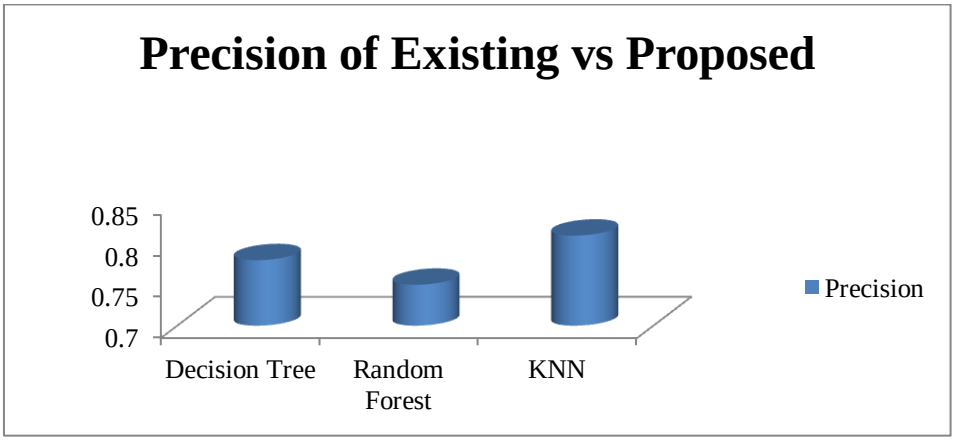


Fig 5: Comparison of Precision

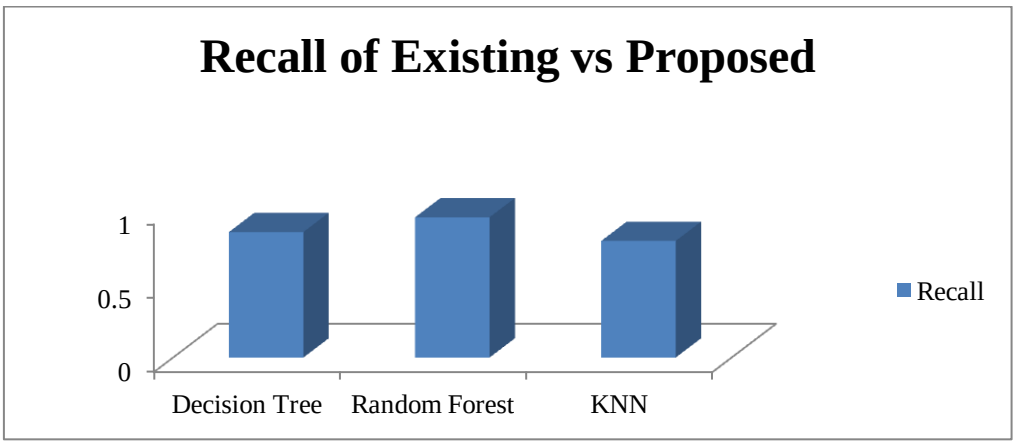


Fig 6: Comparison of Recall

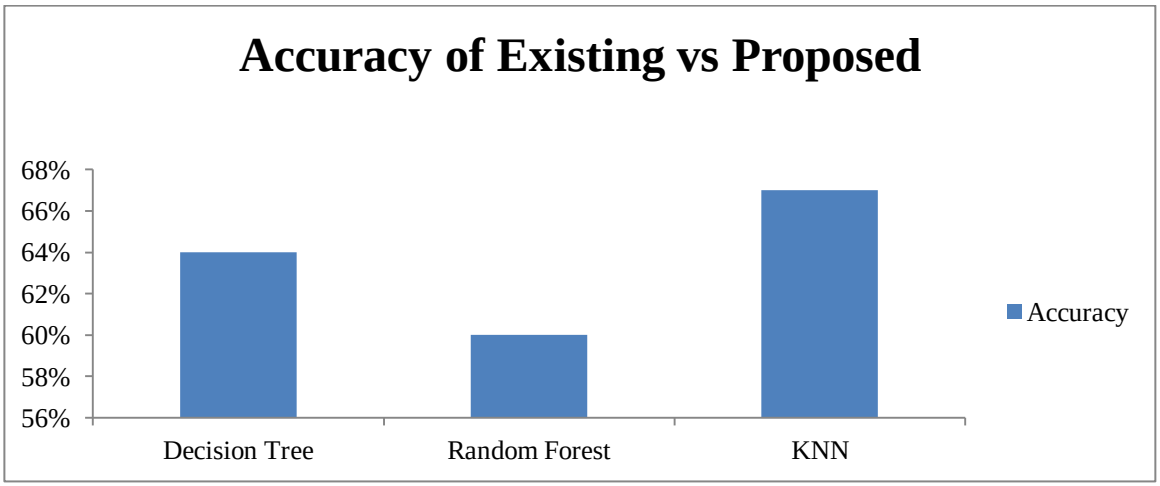


Fig 7: Comparison of Accuracy

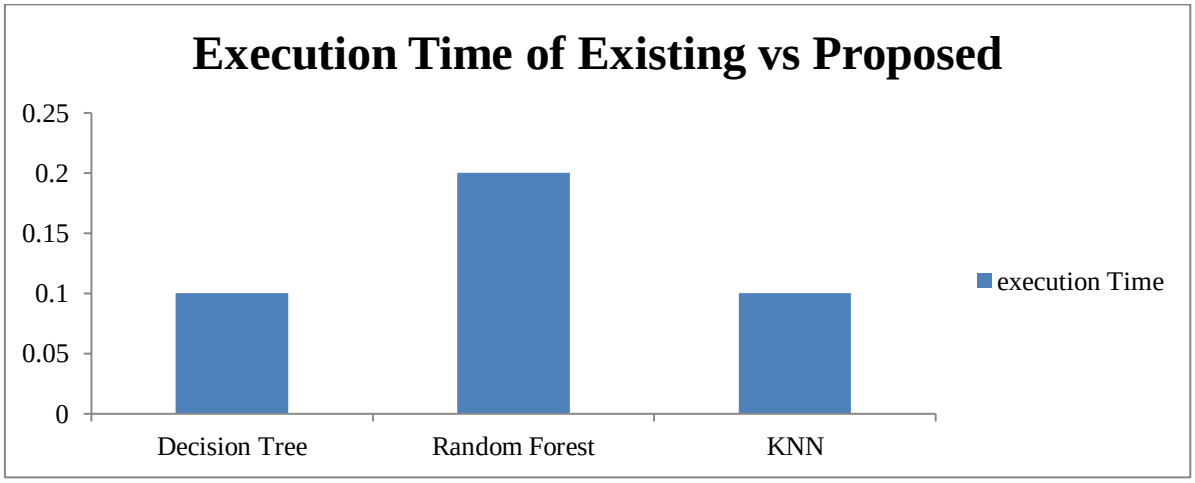


Fig 8: Comparison of Execution Time Rate

RESULT AND DISCUSSION

The executed result from the Breast Cancer Dataset shows the exact extraction of the disease prediction from the proposed methodology. The outcomes are divided into a few subordinate things for simpler investigation and assessment. The execution of the different attributes which has the higher precision with the normal execution time has picked as the best calculation. In this disease prediction execution part , KNN has the most extreme order precision, execution elements and least execution time. So it is estimated as the best order calculation.

CONCLUSION

Emotionally supportive network to test Breast cancer disease helps doctor in making ideal, exact and favorable outcome, and diminishes the general charge of action. Divergent classification mechanism has been utilized to extract the disease from the cancer dataset. It is been experiential KNN classifier yields the greatest characterization exactnesses when utilized with most prescient factors. This proposed Multi-objective KNN techniques properly predicted the cancerous cell in advance to cure the disease with better results. The opportunity to work with this dataset in future will concentrate to extract the dangerous and non dangerous cells from the large amount of the dataset attribute qualities and yielding all the more interesting results.

REFERENCES

1. Zhao Zhao, Han-Ping-Zhu," Risk factor analysis and Preventions of Breast Cancer" International Journal of Biological Sciences,2012.
2. AlirezaOsarech, Bita Shadgar,"A Computer Aided Diagnosis System for Breast Cancer", International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011
3. MandeepRana, Pooja Chandorkar, Alishiba Dsouza, "Breast cancer diagnosis and recurrence prediction using machine learning techniques", International Journal of Research in Engineering and Technology Volume 04, Issue 04, April 2015.
4. VikasChaurasia, BB Tiwari and Saurabh Pal –"Prediction of benign and malignant breast cancer using data mining methods ",Journal of Algorithms and Computational Technology
5. Haifeng Wang and Sang Won Yoon –Breast Cancer Prediction using Data Mining Method, IEEE Conference paper
6. .D.Dubey ,S.Kharya, S.Soni and –"Predictive Machine Learning techniques for Breast Cancer Detection", International Journal of Computer Science and Information Technologies,Vol.4(6),2013,1023-1028.
7. Thorsteinsdottir U, Johannsson OT, Kong A, Stefansson K. Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet. 2008 Jun;40(6):703–706.
8. Dean M, Boyd J, Offit K. Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. Proc Natl Acad Sci U S A. 2008 Mar 18;105(11):4340–4345.
9. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1) Nat Genet. 2009 May;41(5):579–584.
10. Deming SL, Haines JL, Gu K, Fair AM, Cai Q, Lu W, Shu XO. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat Genet. 2009 Mar;41(3):324–328.
11. Chanock SJ, Easton DF. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat Genet. 2009 May;41(5):585–590.
12. Antoniou AC, Altshuler DM, Offit K. Common genetic variants and modification of penetrance of BRCA2-associated breast cancer. PLoS Genet. 2010 Oct;6(10)
13. Gail MH. Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model. J Natl Cancer Inst. 2009 Jul 1;101(13):959–963.
14. 14. Jebaraj Ratnakumar, "An Implementation of Web Personalization Using Web Mining Techniques", Journal Of Theoretical And Applied Information Technology, 2014 JATIT
15. Tsuyoshi, M and Saito, K., "Extracting User"s Interest for Web Log Data", IEEE 2016, pp. 343-346, ISBN: 0-7695-2747-7