Medicine prediction based on doctor’s degree: a data mining approach

Md Shohel Arman, Kaushik Sarker, Asif Khan Shakir, Shah Fahad Hossain, Afia Hasan
Department of Software Engineering, Faculty of Science and Information Technology, Daffodil International University, Dhaka, Bangladesh

Article Info

Article history:

Received Aug 22, 2021
Revised Feb 23, 2022
Accepted Mar 17, 2022

Keywords:

Data mining
Feature selection
Machine learning
Medicine prediction
Neural network

ABSTRACT

The effective use of information mining in profoundly unmistakable fields like e-business, promoting and retail has prompted its application in different enterprises. There is an absence of powerful investigation devices to find concealed connections and patterns in information. This examination paper expects to give a review of ebb and flow systems of learning revelation in databases utilizing information mining strategies that are being used in today’s therapeutic research especially in medicine prediction. Correlation, Chi-square and Euclidean distance feature selections are used to select features and showing the comparison of the result between K-Nearest neighbors, Naïve Bayes, decision tree, artificial neural network. The result uncovers that decision tree beats and sometime Bayesian grouping is having comparative precision as of choice tree. The analysis of performance can be done in such as doctor’s degrees may vary the diseases medicine.

Corresponding Author:

Md Shohel Arman
Department of Software Engineering, Faculty of Science and Information Technology
Daffodil International University
Dhaka, Bangladesh
Email: arman.swe@diu.edu.bd

1. INTRODUCTION

The production of data innovation in different fields needs to lead the gigantic volumes of information put away in various configurations like records, archives, pictures, sound, recordings, logical information, and numerous new information groups. Learning revelation in databases (KDD), regularly known as information mining, goes for the disclosure of valuable data from huge accumulations of information [1]. Information mining procedures have been added to new fields of pattern reorganization, statistics, machine learning, databases, artificial intelligence and computation abilities and so forth [2].

The target of this examination work was partitioned into two classes initial one was the significant element choice procedure to get the key factor of changing scholastic execution and another was the way precisely the model was performed to foresee scholarly execution. In this paper, we attempt to clarify the examination work process and present the best yield by dissecting the productivity of various information mining calculation. The objective of this work is to propose a model for determining the medicine prediction depend on doctor’s degree. Also showing which classification algorithm is better for classified the student performance analysis among the decision tree Naïve Bayes and random forest [3]-[5].

Other objectives are predicting the performance using the selected feature, creating a technique for using three different types of features selection algorithm Chi-square, Euclidean distance, information gain, finally showing how to compare and combine the feature, developing a technique for classifying the dataset and predicting using a different algorithm with effective way [6], [7]. This paper is breaking down the understudy dataset and intends to locate the most important truth behind understudy execution variety and attempt to foresee understudy result utilizing those highlights [8]. According to [9] "it is information mining which is the system of discovering some connection or example among a considerable lot of fields in a major social database.

A sickness is a specific strange condition that adversely influences the structure or capacity of part or the majority of a living being, and that isn't because of any outside injury [10], [11]. A malady might be brought about by outer factors, for example, pathogens or by inward dysfunctions. For instance, inside dysfunctions of the invulnerable framework can deliver a wide range of ailments, including different types of immunodeficiency, extreme touchiness, hypersensitivities and immune system issue [12]. The deadliest ailments in people are coronary vein malady trailed by cerebrovascular ailment and lower respiratory infections [13].

Despite the fact that this encouraged in objective classification to certain gathering of illness, it opened up more inquiries on causation of a few different sicknesses [14]. Notwithstanding advancement in more speculations and ideas, there is an agreement that wellbeing and prosperity doesn't just mean the nonappearance of torment and enduring or the absence of ailment, inability, imperfection and passing, yet has a positive measurement [15]. The biomedicine utilizes the learning of pathogens as the underlying driver of the irresistible infections and in this way, utilizes anti-microbials and against viral treatments to conquer such illnesses [16]. In view of the hypothesis of Tridosha, Ayuveda investigate treatment to bring concordance of the doshas. Homeopathy considers every single incessant ailment because of miasms [17]. Information mining utilizes a blend of an express learning base, modern systematic aptitudes, and zone learning to reveal concealed patterns and examples [18]. These patterns and examples structure the premise of prescient models that empower experts to deliver new perceptions from existing information. Prescription part is moreover improved with the help of this method [19].

2. RESEARCH METHOD

Each issue solver takes after some preprocessing approach to manage deal with their issues. This examination furthermore takes after some preprocessing gadget on coherent philosophies. The examination procedure isolated into a couple of segments for getting the outcome as a great deal as exacting, for example, records gathering, data preprocessing, certainties assessment and visualizing the result. In Figure 1 displaying the thesis proposed model.

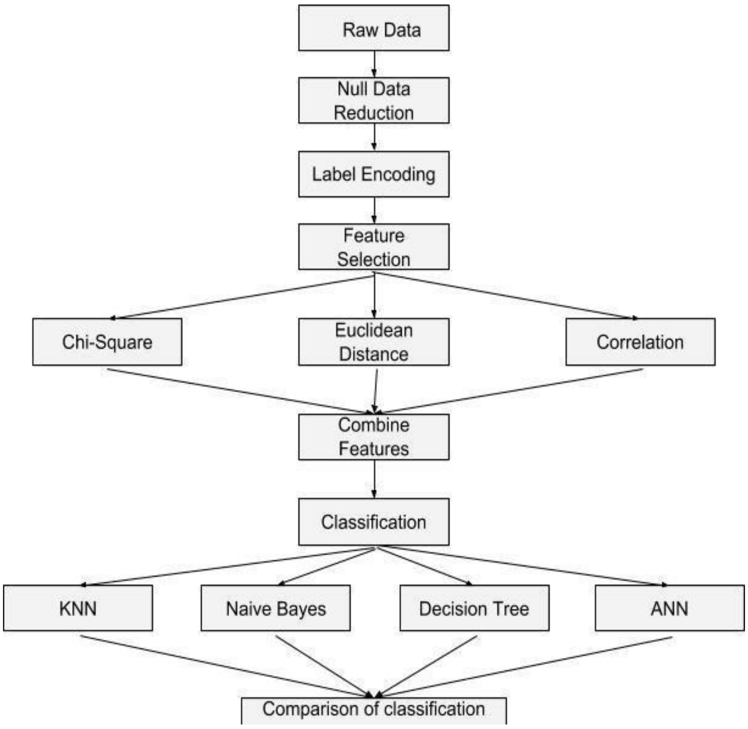

Figure 1. Proposed model for students performance analysis

2.1. Data collection

Methods of data collection can be split into two classifications: secondary data collection techniques and primary data collection methods. I used the convince a pharmacy where the whole data has been stored. For quality assurance i talk with different doctor and medicine specialist to make sure my data is proper and I can work on.

2.2. Data preprocessing

Pre-processing information is primarily focused on two problems: firstly, information should be structured into a correct form for data mining algorithms, and secondly, the data sets used should lead to the most efficient results and quality of data mining models operations The term "garbage in, garbage out" is particularly relevant to machine learning objects and data mining. Real-world information is generally incomplete, inconsistent, and lacks inbound behaviors or trends, and is likely to contain several mistakes. Preprocessing data could be a methodology tested to solve these issues.

2.2.1. Label encoding

We're dealing with a multitude of labels in general. Often these are in a multitude of numbers or phrases. The sklearn machine learning capabilities anticipate they will be numbered. Therefore, in the case that they are as of the present figures, we will use them to begin preparing specifically at that moment. This is not typically the case in any situation [20]. Name encoding refers to how the word names can be changed into the numerical frame. A data function includes a panda library to encode the data and reverse the encoded values to encode the categorical value label. The label encoding algorithm is shown in Figure 2.

```
Load_DataSet()
For each column of columns (Number of total columns/attribute) { unique_values ← Finding unique values
of the column For I = 0 to M – 1 (M number of unique values){
Encoding ← Encoded index of I unique_values
}
}
Save_DataSet ( )
```

Figure 2. Algorithm for label encoding of the dataset

2.2.2. Missing Value

We understand that real-world information tends to be incomplete, noisy, and inconsistent, and a essential job is to fill missing values, disembarrass noise, and correct inconsistencies when pre-processing the information. Choosing the right technique is a preference that depends on the domain of the issue— the domain of the information and our objective of information mining. We replace missing values of an attribute with the mean (or median if its discrete) value for that attribute in the dataset.

2.3. Feature selection

Selection of feature is also called choice variable, selection of sub-set variable or choiuce of attributes. Feature selection is a method that is often used in machine learning, selecting subsets of thecharacteristics accessible from the information to apply a learning algorithm. The goal of variable selection is threefold: to improve prediction general predictor efficiency, to provide faster and more cost-effective predictors, and to provide a better knowledge of the underlying mechanism that produced the information.

2.3.1. Chi-square and euclidean distance

Suppose we get the number O observed and the expected number E. Chi square Score measures the percentage of the anticipated counts E and the number O observed derives from each other. If O is the observed value and E is the expected value then the Chi-square is (1).

$$\sum(Oi - Ei)2/Ei \qquad (1)$$

The Chi square method is used in the contingency table to determine the null hypothesis. The null hypothesis is acceptable if the probability value p is increased by the significant degree α= 0.05. Calculate the degree to determine the p value of freedom (df) of contingency table. Degree of freedom of a contingency table is,

$$= (r - 1)(c - 1) \quad (2)$$

where c is the total column of contingency table and r is the total number of rows. P value is calculated from Chi-square distribution table [21].

Usually, distance from Euclidean is used for distance. In most instances, when discussing distance, individuals will usually refer to Euclidean distance. Euclidean distance examines within the couple of objects the nucleus of the square difference. Euclidean distance calculates from the sample to all other characteristics for each function. Euclidean distance (or), calculated characteristics between the formula and its use (3).

$$D = \{\sum(pi \ - \ qi)2\} \quad (3)$$

The Euclidean will not be calculated from standardized information, it will be calculated from raw information. Place gradually higher weight on objects, one might want to square the further portion of the normal Euclidean distance. The range is as calculated in (4).

$$\sum(v1[i] - v2[i]\,)2 \quad (4)$$

2.4. Classification

We used the classification algorithm when our expected output is a discrete label. In another word, they are helpful when a finite set of possible results drops below the answer to your company question. There are only two possible results in several use cases, such as determining whether or not an electronic mail is a spam. Multi-level classification captures everything else and is helpful for customer segmentation, client feeling text analysis, categorization of audio and picture [22].

2.4.1. Naïve Bayes

Naive Bayes is an easy approach to classifier construction: models assigning classification labels to problem cases, represented as vectors of feature values, where class labels are taken from some finite set. Naive Bayes algorithms are mostly used in sentiment analysis, spam filtering, and recommendation systems. There is no single algorithm now available to train such classifiers, but a family of algorithms supports a conventional principle. It significantly simplifies learning by assuming that, given the class variable, characteristics are autonomous. Simply put, the existence of a Naive classifier assumes. The existence of the other feature is unrelated to the particular function in a class.

2.4.2. K-nearest neighbor

KNN is a lazy learning algorithm that is non-parametric. It implies that the underlying data distribution does not make any assumptions. Its aim is to use a database to predict the classification of a fresh sample point by dividing the information points into several groups. Let the closest K neighbors have more say in influencing the query point result. This can be accomplished with a set of weights W, one for each nearest neighbor, defined by the relative closeness of each neighbor with respect to the query point. Thus:

$$W(x, p1) \ = \frac{exp(-D\ (\ x,p1))}{\sum exp(-D\ (\ x,p1))} \quad (5)$$

where D(x, pi) is the distance between the query point x and the ith case pi of the example sample. It is clear that the weights defined in this manner above will satisfy.

$$\sum W(x0\ , x1) = 1 \quad (6)$$

2.4.3. Artificial neural network

Artificial neural network (ANN) processes information in a manner that is encouraged by the biological nervous system, such as the brain, taking input and processing information that allows a computer to learn from observation data. The primary objective of the ANN strategy is to solve the issue in a manner that human brain does. Learning related to modifications to the synaptic connection between the neurons in the biological model [23]. There is one output and many inputs in an artificial neuron machine. When a cycle does not shape the input signal in an ANN is called feed forward neural network. Neural network data is addressed in only one manner in feed forward. Feed-forward network is mostly used to recognize patterns [24]. Artificial neural network has three layer: i) input layer, ii) hidden layer, and iii) output layers.

Raw data is provided to the network in the input layer. The input component and the weight relationship within the input and hidden units determines its action in the hidden layer. Output layer linked to and mapping input and hidden layer. Neurons are connected with each other in the neural network, which is said to be the link weight. In the image below we showed weight by which is connected between unit and. We depict the weight matrix W of which weights are the information. We have used back propagation algorithm technique. Output layer decide it's activity in two step.

Firstly, it compute the total weight, using the following formula:

$$Z = Bias + W1X1 + W2X2 + \dots + WnXn \tag{7}$$

while the computation of input is done, network determine the error E.

$$\frac{1}{2}\sum_{i=1}^{N}(yi - \hat{y}i)^2 \tag{8}$$

3. RESULTS AND DISCUSSION

There are many components in the outcome chapter in this document, such as the consequence of Feature Selection and the combined outcome of these three features choice algorithms. The difference between decision tree, K nearest neighbor, Naive Bayes and the processed data set's artificial neural network algorithm. We clarified the outcome with the matrix of confusion and lastly discussed the outcome.

3.1. Feature selection result

3.1.1. Chi square result

We discuss about the chi square and hot to calculate Chi-square between target attribute after applying chi square algorithm the dataset, we got the chi square value, p_value and result for each attribute. We have shown the result in Table 1. As the chart shows Chi-square value, p value, result for column attribute, we can see which attribute is essential and which, based on p value, is not essential. The important amount is set when calculating chi square 0.05. If any characteristics p value are significantly lower than it will acknowledge the null hypothesis indicating that attribute is not essential.

Table 1. Chi-square result for the dataset

Attribute	Chi Square Value	P_value	Result
PRS_ID	206.073	0.407333	Important
MONTH	716.1115	0.001314	Important
ROUND	1118.982	2.01E-12	Not Important
YEAR	382.306	0.774496	Not Important
Book_id	205.8455	0.411674	Important
Shop_id	14250.51	8.41E-12	Not Important
cdate	19857.7	2.79E-65	Not Important
pdate	236.2813	0.049435	Important
Prs_type	45204.32	0	Important
Psc_slnc	50601.24	0	Important
Phy_id	46814.51	2.60E-245	Not Important
Phy_nm	42920.57	1.03E-202	Not Important
Phy_DEGR	37236.22	2.45E-103	Not Important
Vc2	34343.47	1.30E-80	Not Important
Unit_prc	29602.47	2.83E-90	Not Important
Ing	25801.59	4.73E-41	Not Important
QT_prs	17754.5	1.41E-27	Not Important
Mpo	10381.99	1.70E-42	Not Important
Am	331.7309	2.31E-08	Not Important
RM	402.1015	0.51732	Important
Asm	970.3998	0.810027	Not Important
Sm	972.247	0.79837	Not Important
Mpo_tm	1073.378	0.081291	Not Important
Am_tm	950.2176	0.91027	Not Important
Am_nm	997.7833	0.601824	Not Important
Ch_add	1070.346	0.091496	Not Important
diagoname	1060.898	0.129502	Important
Createby	1051.621	0.176599	Important
createDate	1054.181	0.162605	Important
UpdateBy	1165.547	0.000463	Important
UpdateDate	1097.019	0.028854	Important
isRemove	1134.756	0.00364	Important

3.1.2. Euclidean distance result

In Chapter X, Section Y, we address the Euclidean Distance, the way Euclidean distance between target variable is calculated. We have prepared ED for each attribute after applying the algorithm to the dataset. We showed the Euclidean distance outcome in Table 2.

Table 2. Euclidean distance result for the dataset

Attribute	Value	Attribute	Value	Attribute	Value
PRS_ID	33.64521	Phy_nm	7.315476	Sm	26.54058
MONTH	21.44761	Phy_DEGR	7.918514	Mpo_tm	24.7367
ROUND	25	Vc2	8.387318	Am_tm	28.06852
YEAR	11.78983	Unit_prc	1.423025	Am_nm	30.31325
Book_id	36.71512	Ing	2.971011	Ch_add	24.24017
Shop_id	30.48601	Name	4.531943	diagoname	24.13225
cdate	28.05495	Dstmr	5.860485	Createby	26.19606
pdate	32.74141	Mpo	11.23278	createDate	24.43264
Prs_type	10.24674	Am	26.68333	UpdateBy	25.72597
Psc_slnc	9.010909	RM	25.88436	UpdateDate	26.40761
Phy_id	9.270138	Asm	26.0196	isRemove	26.13679

3.1.3. Correlation result

In chapter X, section Y, we discussed about the correlation algorithm and way of calculation Euclidean distance between different variable and target variable. In the following table we have showed the result that we have got by implementing the correlation algorithm. Table 3 shows the variable, its value and the result state. Some records from the result been shown in the Table 4.

Table 3. Correlation result for the dataset

Attribute	Value	Result	Attribute	Value	Result	Attribute	Value	Result
PRS_ID	0.122399	Positive	Phy_nm	0.292285	Positive	Mpo_tm	0.024914	Positive
MONTH	-0.08792	Negative	Phy_DEGR	0.219076	Positive	Am_tm	0.060245	Positive
ROUND	0.171513	Positive	Vc2	0.200257	Positive	Am_nm	0.100822	Positive
YEAR	-0.058	Negative	Unit_prc	0.08976	Positive	Ch_add	0.016186	Positive
Book_id	-0.04535	Negative	Ing	0.082376	Positive	diagoname	0.051261	Positive
Shop_id	-0.04535	Positive	Name	0.081084	Positive	Createby	0.038254	Positive
Cdate	0.25058	Positive	Dstmr	0.044712	Positive	createDate	-0.00707	Negative
Pdate	-0.06701	Negative	Mpo	nan	No Relation	UpdateBy	0.079734	Positive
Prs_type	0.854657	Positive	Am	0.091244	Positive	UpdateDate	0.033123	Positive
Psc_slnc	0.589815	Positive	RM	0.09659	Positive	isRemove	0.050858	Positive
			Asm	0.043416	Positive			
			Sm	0.053383	Positive			

We can see from the outcome that the value of the correlation is between -1 and +1. A negative linear connection will occur when the value is precisely-1, if it is -0.7 a powerful linear adverse connection between the target variable. If valuation is -0.50 a mild adverse connection will occur, if it is -3 a weak linear adverse connection will occur. This implies that there is no connection if the correlation value is 0. If the value exceeds 0, a favorable connection exists between two factors.

Table 4. Combined feature selection result

Number	Attribute	Number	Attribute	Number	Attribute
1	Book_id	7	Unit_prc	15	ASM
2	Shop_id	8	Ing	16	Mpo_tm
3	Phy_id	9	Name	17	Am_tm
4	Phy_degr	12	Mpo	18	Chdist
5	Phy_spc	13	Am	19	Chtha
6	Vc2	14	Rm	20	diagname

3.2. Classification result with confusion matrix

Classification model performance is mostly explained by a table of matrix named confusion matrix. Confusion matrix can readily describe the efficiency of our classification system [25]. It shows the right forecast as well as the wrong prediction by breaking down each and every class. Table 5 defines confusion matrix calculation techniques. Matrix of confusion in case of 2 classes.

Table 5. Classification result with confusion matrix

Actual	Negative	Positive
Negative	True Negative (TN)	False Positive (FP)
positive	False Negative (FN)	True Positive (TP)

3.3. Compariosn KNN, Naïve bayes, decision tree, and ANN

ANN, KNN, Naive Bayes, Decision tree algorithm are performed on the dataset we developed using python language. Our final version of the dataset got nine hundred records 79 entities. Half of data been collected by yourself through survey. We split our dataset in 80% as train data and 20% as test data. We used the same ration for all the model. Figure 3 represents the ratio of true positive and false positive of execution time in terms of KNN, NB, ANN and DT algorithms. In Figure 4 results are shown in the comparison of execution time between KNN, Naive Bayes, decision tree and ANN whete decision tree classifier algorithm performs better than other algorithms. Figure 5 represents the ratio of true positive and false positive of result accuracy in terms of KNN, NB, ANN and DT algorithms.

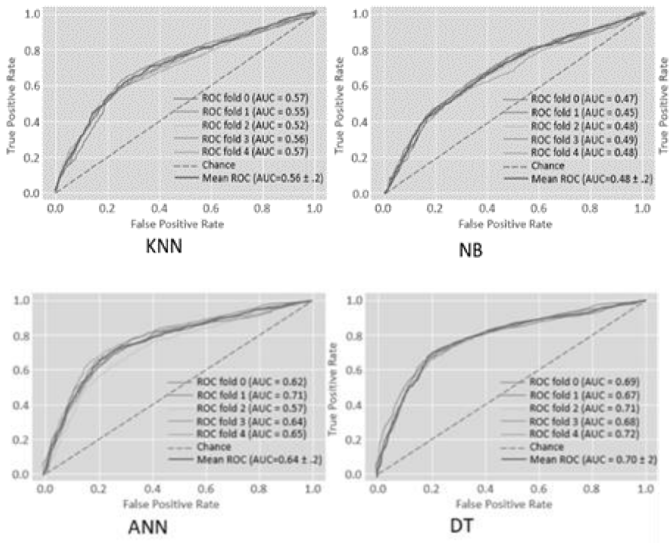

Figure 3. Execution time of different types of classification algorithm

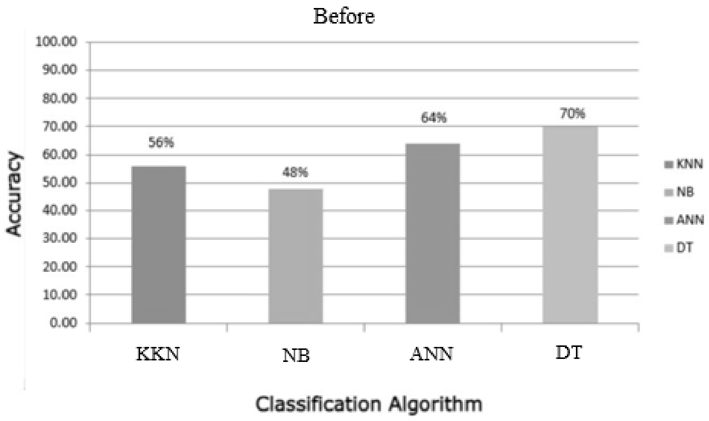

Figure 4. Execution time comparison between classification algorithms

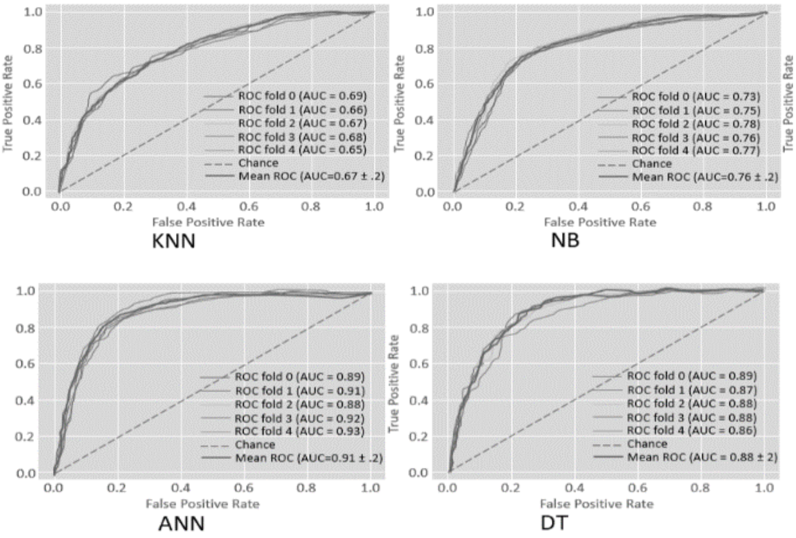

Figure 5. Result accuracy of different types of classification algorithm

In Figure 6 results are shown in the comparison of result accuracy between KNN, Naive Bayes, decision tree and ANN whete decision tree classifier algorithm performs better than other algorithms. After executing each algorithm, we calculate the execution time where ANN took 5.63 seconds, KNN took .17, Naive Bayes .14 and .21 second took by decision tree. From that result we found that decision tree predict better than another model based on the execution time and accuracy. Where accuracy of decision tree is 88.35%, ANN 91.40%, Naive Bayes 75.42% and KNN accuracy is just 63.36%.

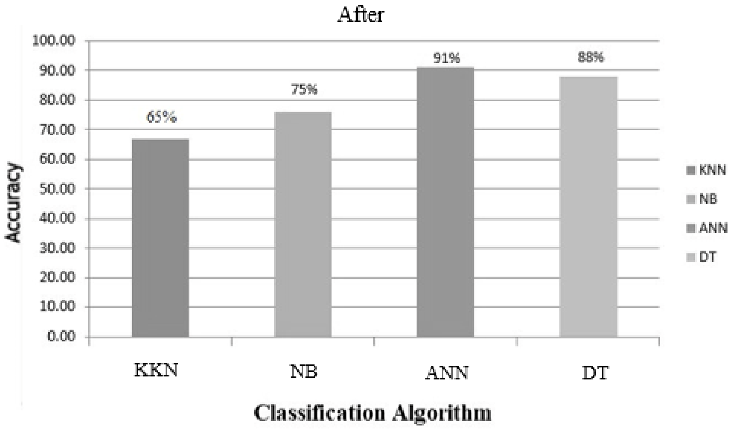

Figure 6. Result accuracy comparison between classification algorithms

4. CONCLUSION

We used medicine data mining in this study to evaluate and predict the performance of doctor. We've been using three Chi-square selection function algorithm, Eulidean distance and correlation. We discovered that in Bangladeshi Doctor performance have more effect. We have used k-nearest neighbor, Naive Bayes, decision tree and artificial neural network for 6 classes among them ANN perform the best with accuracy of 91.40%. All of that knowledge can be used to improve the to predict the medicine. The researchers will endeavor to cover and implement the information mining technique on this investigation. An approach to sum up the exploration to progressively changed exercises is to get increasingly exact results. More experimentation can be performed utilizing more data mining systems, for example, SVM, C4.5, ID3. We utilized 6 classes to anticipate the outcomes.

REFERENCES

[1] Y. Ma, B. Liu, C. K. Wong, P. S. Yu, and S. M. Lee, "Targeting the right students using data mining," in *Proceeding of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, pp. 457–464, doi: 10.1145/347090.347184.

[2] Y. Beaumont-Walters and K. Soyibo, "An analysis of high school students' performance on five integrated science process skills," *Research in Science & Technological Education*, vol. 19, no. 2, pp. 133–145, Nov. 2001, doi: 10.1080/02635140120087687.

[3] Z. N. Khan, "Scholastic achievement of higher secondary students in science stream," *Journal of Social Sciences*, vol. 1, no. 2, pp. 84–87, Feb. 2005, doi: 10.3844/jssp.2005.84.87.

[4] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting students' performance in distance learning using machine learning techniques," *Applied Artificial Intelligence*, vol. 18, no. 5, pp. 411–426, May 2004, doi: 10.1080/08839510490442058.

[5] J. A. Moriana, F. Alós, R. Alcalá, M. J. Pino, J. Herruzo, and R. Ruiz, "Extra-curricular activities and academic performance in secondary students," *Electronic Journal of Research in Educational Psychology*, vol. 4, no. 8, pp. 35–46, 2006.

[6] P. Kaur, M. Singh, and G. S. Josan, "Classification and prediction based data mining algorithms to predict slow learners in education sector," *Procedia Computer Science*, vol. 57, pp. 500–508, 2015, doi: 10.1016/j.procs.2015.07.372.

[7] N. Thai-Nghe, A. Busche, and L. Schmidt-Thieme, "Improving academic performance prediction by dealing with class imbalance," in *ISDA 2009 - 9th International Conference on Intelligent Systems Design and Applications*, 2009, pp. 878–883, doi: 10.1109/ISDA.2009.15.

[8] A. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414–422, 2015, doi: 10.1016/j.procs.2015.12.157.

[9] B. Kumar and S. Pal, "Mining educational data to analyze students performance," *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 6, 2011, doi: 10.14569/ijacsa.2011.020609.

[10] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines," *International Journal of Medical Informatics*, vol. 77, no. 2, pp. 81–97, Feb. 2008, doi: 10.1016/j.ijmedinf.2006.11.006.

[11] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43–48, Mar. 2011, doi: 10.5120/2237-2860.

[12] S. C. Satapathy, V. Bhateja, B. Janakiramaiah, and Y.-W. Chen, "Advances in intelligence systems and computing," in *Information systems design and intelligent applications*, 2016, p. 435.

[13] R. Jin, C. Xue, L. Wu, B. Zhang, Q. Lin, and S. Liu, "Many-to-one mapping: The principle of Chinese medicinal property theory learned from strong association rules," in *2013 Ninth International Conference on Natural Computation (ICNC)*, Jul. 2013, vol. 131, no. 22, pp. 951–956, doi: 10.1109/ICNC.2013.6818113.

[14] L. Sacchi, A. Dagliati, D. Segagni, P. Leporati, L. Chiovato, and R. Bellazzi, "Improving risk-stratification of Diabetes complications using temporal data mining," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Aug. 2015, vol. 2015-November, pp. 2131–2134, doi: 10.1109/EMBC.2015.7318810.

[15] M. Lindquist, M. Stahl, A. Bate, I. R. Edwards, and R. H. B. Meyboom, "A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database," *Drug Safety*, vol. 23, no. 6, pp. 533–542, 2000, doi: 10.2165/00002018-200023060-00004.

[16] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 5, pp. 241–266, Oct. 2013, doi: 10.14257/ijbsbt.2013.5.5.25.

[17] X. Wu, S. Long, and W. H. Zhu, "The use of data mining in pharmic quality analysis of traditional chinese medicines," in *Proceedings - 2012 International Conference on Computer Science and Service System, CSSS 2012*, Aug. 2012, pp. 1431–1434, doi: 10.1109/CSSS.2012.360.

[18] S. Sun *et al.*, "Exploring the associating rules of prescription and syndrome on Radix Astragali with text mining," in *2013 ICME International Conference on Complex Medical Engineering, CME 2013*, May 2013, pp. 115–118, doi: 10.1109/ICCME.2013.6548222.

[19] C. R. Meng, H. L. Zhang, L. F. Zeng, Z. P. Li, J. Huang, and Z. Liang, "Evidence-based decision support for the clinical practice of acupuncture: Data mining approaches," in *Proceedings - 2013 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2013*, Dec. 2013, pp. 180–181, doi: 10.1109/BIBM.2013.6732669.

[20] Great Learning Team, "Label encoding in python explained," *Mygreatlearning.Com*, 2020. https://www.mygreatlearning.com/blog/label-encoding-in-python/ (accessed Apr. 03, 2018).

[21] "Chi square feature selection in python." http://www.insightsbot.com/blog/2AeuRL/chi-square-feature-selection- in-python (accessed Apr. 03, 2018).

[22] S. Polamuri, "Building random forest classifier with python scikit learn," dataaspirant, 2017. https://dataaspirant.com/2017/06/26/random-forest-classifier-python-scikit-learn/ (accessed Apr. 05, 2018).

[23] "Statsoft." http://www.statsoft.com/Textbook/Neural-Networks. (accessed Aug. 08, 2018).

[24] "Developerzen." https://machinelearningmastery.com/save-load-keras-deep-learning-models (accessed Aug. 28, 2018).

[25] M. Ramaswami and R. Bhaskaran, "A CHAID based performance prediction model in educational data mining," *International Journal of Computer Science Issues*, vol. 7, no. 1, 2010, [Online]. Available: http://arxiv.org/abs/1002.1144

BIOGRAPHIES OF AUTHORS

Md Shohel Arman Md. Shohel Arman is a Lecturer and alumni of Department of Software Engineering under Faculty of Science & Information Technology in Daffodil International University, Dhaka, Bangladesh. He is an energetic and focused man since his student life. His research interests are distributed databas system, machine learning, data mining nternet of things (IoT), software security and management information system (MIS). He can be contacted at email: arman.swe@diu.edu.bd.

Kaushik Sarker Kaushik Sarker is currently a PhD fellow in Beihang University, Beijing, China. He received his B.Sc. in Electronics and Telecommunication Engineering in 2010 from Bangladesh and M.Sc. in Computer Systems and Network Engineering in 2012 from United Kingdom. Since 2013 he has been working as a faculty member in the Faculty of Science and Information Technology at Daffodil International University, Dhaka, Bangladesh. He can be contacted at email: kaushik.swe@daffodilvarsity.edu.bd.

Asif Khan Shakir is working as a Senior Lecturer in the Department of Software Engineering, Daffodil International University, Bangladesh. Currently he working as a Senior Data Analyst at Save the Children, Bangladesh. He has completed his bachelor of Information Technology in 2012 from the University of Dhaka. He can be contacted at email: shakir1232@yahoo.com.

Shah Fahad Hossain is working as a Lecturer in the Department of Software Engineering, Daffodil International University, Bangladesh. He has completed his B.Sc. in Software Engineering in 2021 from Daffodil International University, Bangladesh. His research interests are in Machine Learning, Deep Learning and Computer Vision. He has also experience in Software Development. He can be contacted at email: fahad.swe22@gmail.com.

Afia Hasan his higher studies in MS Food Engineering by coursework at the at the department of Food is working as a SQA in a private company. She has completed his B.Sc. in Software Engineering in 2018 from Daffodil International University, Bangladesh. She is ISTQB certified software tester. She can be contacted at email: afiahasan28@gmail.com.