Building Disease Prediction Model Using Machine Learning Algorithms on Electronic Health Records' Logs

Sabina Rakhmetulayeva[1], and Aliya Kulbayeva[1]

[1]*International Information Technology University, Manas St. 34/1, Almaty, Kazakhstan*

Abstract

The number of tasks devoted to predicting the incidence of infectious diseases is growing rapidly due to the availability of statistical data supporting the analysis. This article describes the main solutions currently available for creating short and long-term disease forecasts. Their limitations and practical applications are shown. Much attention is given to the Naïve Bayes classification, logistic regression, artificial neural network algorithm and k-means artificial neural networks as methods of model analysis based on machine learning. This article provides an overview of two popular machine learning algorithms used to predict diseases. The standard datasetis used for a wide range of diseases including fungal infection, allergy, GERD, chronic cholestasis,peptic ulcer disease, diabetes, bronchial asthma, migraine, paralysis (brain hemorrhage) and more.

Keywords

Point-to-point estimates, regression models, method of analogues, Naïve Bayes, logistic regression, support vector machine, data mining, K-Means Clustering, Artificial neural network

1. Introduction

Patients access medical services over the Internet by connecting to medical information systems. When people are sick, they frequently search the Internet for various information explaining their symptoms and develop incorrect diagnosis for themselves. As a result, the medical services system, which includes medical consultations, visits to medical facilities, drug purchases, recuperation, and treatment, is evolving [1].

When it comes to data collecting and processing, one of the most inconvenient topics is health. With the digital age, a vast amount of patient data is being generated, including hospital resource factors, diagnostic patient information records, and medical equipment. Making excellent judgments necessitates extremely complicated data processing and review. The extraction of medical data gives up a lot of possibilities for detecting duplicates of medical datathat have been saved [2]. Patients and doctors seeking information about their symptoms use automated tools that support medical diagnostic systems as they focus on several possible causes to avoid complex or premature diagnoses [3]. A lot of efforts have been made to create predictive diagnostic systems and encode relevant information for the development of forecasting methods [4][5].

Data collecting in many industries is continuously increasing as a result of recent technological advancements such as computers and satellites. Traditional data analysis approaches are obviously incapable of processing enormous volumes of data efficiently. Data mining techniques are the only option to extract knowledge from enormous volumes of data inthis scenario. In the field of data mining, the obtained machine learning algorithms are a strong instrument in prediction of diseases. The

Proceedings of the 7th International Conference on Digital Technologies in Education, Science and Industry (DTESI 2022), October 20–21, 2022, Almaty, Kazakhstan

EMAIL: ssrakhmetulayeva@gmail.com (Sabina Rakhmetulayeva); aakulbayeva@gmail.com (Aliya Kulbayeva)
ORCID: 0000-0003-4678-7964 (Sabina Rakhmetulayeva);

potential utility of these technologies has lately beendiscovered through diagnostics based on health data. Statistics on various disease data for up to ten years provide a solid opportunity to forecast data for the following 2-3 years.

The implementation of proper diagnostics for autonomous extraction of relevant information from electronic medical records is one of the ultimate aims of intelligent healthcare [6]. This is a highly important and promising duty that cannot only improve work efficiency, but also minimize doctors' diagnostic mistakes while making a diagnosis [7].

Previously, the models employed in diagnostic approaches had to be specified by hand, which took a long time and effort [8]. Although the technical details of the manual make this model very weak, it is difficult to adapt it to new diseases or clinical conditions. Automatic symptom-based disease planning can significantly accelerate the development of such diagnostic tools. The pricing is also determined by the visuals.

There are four main reasons why EMR data is difficult to interpret. Previously, the texts inmedical and medical records were shorter than in traditional textbooks, which made it difficultto determine the context of diseases and symptoms. Both textbooks and magazines often offersimplified examples that reflect only the most general features to help learning. EMR data represent real patients with all common diseases and factors that make them individual. Unlikethe third textbook, which reveals the link between disease and symptoms, EMR's link betweendisease and symptoms is statistical, making it simpler to mix up the connection with the cause.Finally, the attending physician modifies the decision-making process as part of the electronicmedical record entry procedure [9]

This article explores the use of various methods of initial center selection along with naive Bayesian methods to include the grouping of k instruments in the diagnosis of patients with diseases.

2. Naïve Bayes Algorithm

The Bayesian dynamic network is another technique to time series modeling that takes into consideration the associated data structure. Straight graphs with vertices corresponding to model variables and edges corresponding to probabilistic connections between them established by particular distribution rules are used to represent Bayesian networks. After training the Bayesian network, the likelihood of an event occurring in the observed sequence of events may be estimated. Bayesian networks are fast gaining popularity in a variety of sectors of knowledge and are being utilized to tackle the challenge of forecasting morbidity, particularly in their most basic version - the Hidden Markov mode.

The basic idea of HMM is to compare any random variable Yt with an unobservable randomvariable St t which determines the conditional distribution of Yt [10].

This parameter can be estimated from the observations of yt by explaining the distribution laws of Yt and St. Thus, Yt can be the number of citizens seeking medical care and St is an important characteristic of an epidemic situation, for example, the total number of infected citizens. It is assumed that the Yt values are based only on the values of the latent variables Stat time t and that the sequences St are Markov features. That is, the value of St is based only on St-1 (Figure 1).

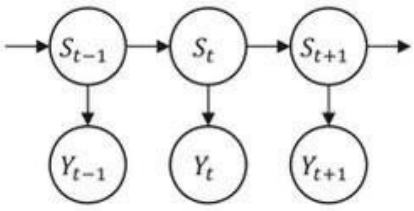

Figure 1: Dependency diagram in a hidden Markov model

Bayes' theorem is the foundation of the naïve Bayesian classifier. He has a strong sense of self-sufficiency. An autonomous functional model is another name for it [11]. The presenceor absence of an element of a certain class is considered independent of the presence or absenceof any other element of this class. Under controlled learning settings, naive Bayesian classifiersmay be taught. It employs the

method of maximal equality. This is done in challenging real- world scenarios. A limited quantity of training data is required for this [12]. Parameters for rating just the variance of the variable, not the complete array, must be determined for each class [13]. When the input data is large, naive Bayesian analysis is utilized. This makes the outcome more difficult. The likelihood of each input characteristic leaving the expected state. naïve Bayesian classification-based machine learning and data mining methods.

Bayes theorem:

$$P(C|X) = P(X|C)P(C)P(X), \quad (1)$$

where:

P(C|X) - posterior probability;P(X|C) - likelihood;

P(X) - predictor prior probability;P(C) - class prior probability.

The Naive Bayesian classification algorithm is based on Bayesian theory with the concept of attribute independence. In other words, the NBA recognizes that the existence of one attribute does not depend on the existence of another attribute [14].

Naive Bayes primarily predicts whether a patient is at risk of a certain type of disease. After applying the K-means algorithm, we get a model dataset that compares the values of the datasetwith the trained dataset. There will apply the Bayesian principle and determine if the patient has a disease.

How does the naive Bayes algorithm work in our research?

We have the dataset from kaggle.com with symptoms which are mapped to 42 diseases [15]. Using this dataset, we created a likelihood table with 10 diseases (Table 1) which contains theprobability of certain disease under symptoms. Overall, dataset contains 149 symptoms, and likelihood table describes in how many symptoms the disease will be indicated as probable.

Table 1

Likelihood table

	Disease	No
1	Fungal infection	125
2	Allergy	125
3	GERD	125
4	Chronic cholestasis	125
5	Peptic ulcer disease	125
6	Diabetes	125
7	Bronchial asthma	125
8	Migraine	125
9	Paralysis (brain hemorrhage)	125

Based on the data set of disease markers, we need to determine which 129 markers identify the types of diseases. We need to determine the accuracy and probability of the disease using the Bayes algorithm of the dataset. We can solve this problem using the approach described above.

P(Yes|Disease (n))=P(Disease(n) | Yes)*P(Yes) /P(Disease(n))P(Disease(1) | Yes))=4/55=0.07

P(Disease (1))=4/129=0.03P(Yes)=40/129=0.43

P(Yes|Disease (1))=0.07*0.43/0.03=1.003

As we can see from the calculation of disease prediction using naive Bayes algorithm (Figure 2).

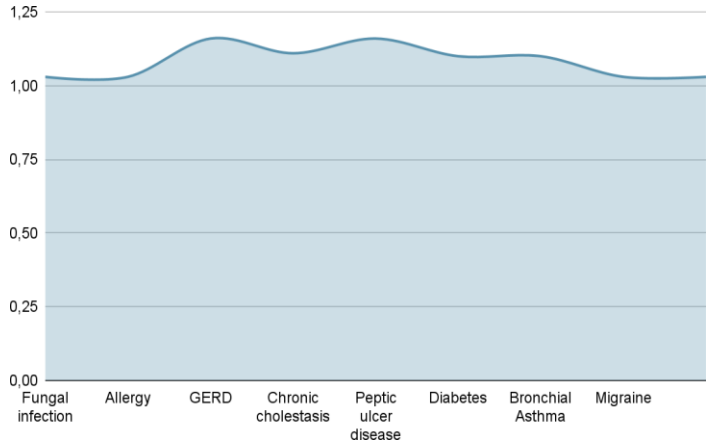

Figure 2: Naive Bayes classification

As we can see the accuracy of Naïve Bayes is 97.6% which represent the good result for ourstudy. It means that in future when we do comparative work on other machine learning algorithms we will choose the best algorithm (Figure 3).

```
In [68]:
from sklearn.naive_bayes import MultinomialNB
nb = MultinomialNB()
nb.fit(X_train_final, y_train)

Out[68]:
MultinomialNB()

In [69]:
nb_train = accuracy_score(y_train, nb.predict(X_train_final))
nb_test = accuracy_score(y_test, nb.predict(X_test_final))
print("Train accuracy :{:.2f}".format(nb_train))
print("Test accuracy :{:.2f}".format(nb_test))

Train accuracy :0.98
Test accuracy :0.95
```

Figure 3: Accuracy and model visualization NB

3. K-means classification

K-Means is a method of unsupervised learning that is often used in the process of collecting information about the nearest neighbors. The data can be grouped into k groups based on similarity. K is the number that you need to know for the algorithm to work [16][17]. K-mean is the most commonly used cluster algorithm capable of detecting new data collected with accuracy at most distances [18,19,20]. The first selection of k cluster centers is done at random;thereafter, all points are assigned to their nearest cluster centers and recomputed cluster centersfor the newly constructed group, because some cluster centers impact K means, they are particularly susceptible to noise and outliers [21]. One of the advantages of the K method is that it is easy to implement and interpret deductively. The disadvantage of this approach is the complexity of estimating K. It works with clusters of spherical meshes [22]. The K-means method is depicted graphically in Figure 4. There are two sets of themes in the first level. Then,on both sides, define the center. Groups that form additional clusters in the dataset are regenerated based on their center of gravity. Repeat this method until you get the ideal pair [23].

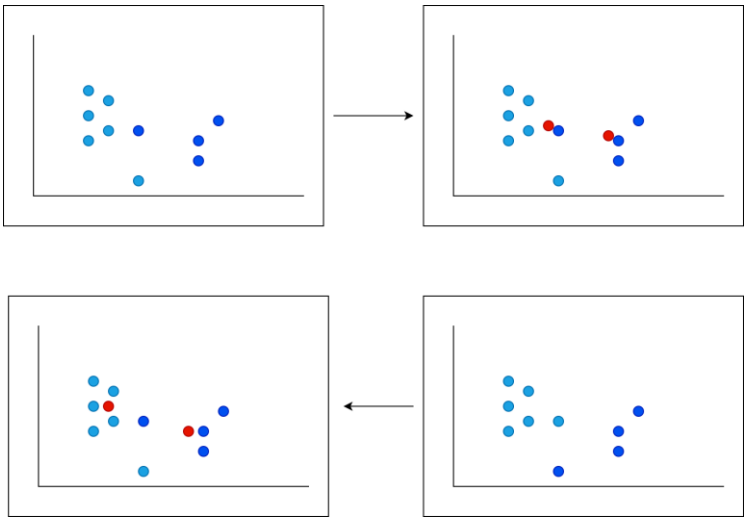

Figure 4: Accuracy K-means clustering process

Based on the data, you need to build an algorithm for solving the problem. An example of numbered list is as following.

1. Select a value for K2.
2. Randomly select a data point K representing the center of gravity of cluster.
3. Assign all other data points to the nearest center of gravity of the cluster.
4. Repeat steps 3 and 4 until there are no changes left in each cluster.

The cluster technique does not determine the number of clusters in the k-means approach since it must be specified before the start. As a result, I'll employ the elbow strategy. The elbowapproach is a popular strategy to figure out how many clusters are needed.

Based on this, we got such a graph for calculating the dataset in Python and built a graph (Figure 5).

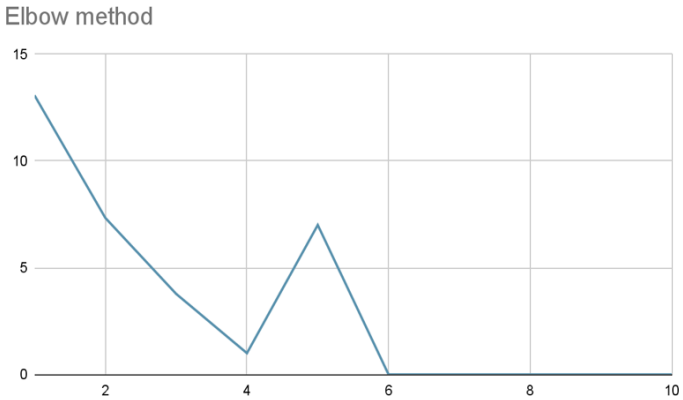

Figure 5: K-means clustering process

When calculating cluster analysis, the big question often arises how many clusters to take and the elbow method helps in this matter! With each new cluster, the total difference in eachcluster becomes smaller. In extreme cases, when there are many clusters compared to the result, the score is zero. However, in most cases, the decrease in general fluctuations after a certain moment is very small. This point is used as the best cluster number [23].

4. Logistic regression

Logistic regression (LR) is a strong and well—established technique of classifying data with a teacher.

This is an extension of classic regression, and only binary variables representing the presence or absence of events are often modeled.

LR aids in determining the likelihood that a new instance will belong to a specific class. Given that this is a probability, the outcome will be between 0 and 1. As a result, in order to employ LR as a binary classifier, a threshold must be set to discriminate between the two classes. For example, if the input instance's probability value is larger than 0.50, it is classed as "Class A." Otherwise, it's "Class B."

The LR model may be extended to represent categorical variables with three or more values. Polynomial logistic regression is the name given to this expanded variant of LR.

So, we made some research with dataset and determined that RF algorithm has that accuracywhich is presented in Figure 6 and as we can see the accuracy of RF is 100% which representbetter result for our study.

```
12]:
# Logistic Regression

logreg = LogisticRegression()
logreg.fit(X_train, y_train)
Y_predLR = logreg.predict(X_test)
print("Train Accuracy: ",round(accuracy_score(y_train, logreg.predict(X_train))*100,2))
print("Test Accuracy: ",round(accuracy_score(y_test, Y_predLR) * 100,2))

Train Accuracy:  100.0
Test Accuracy:  100.0
```

Figure 6: Accuracy and model visualization LR

5. Support vector machine

Support Vector Machines (SVMs) can classify both linear and nonlinear data. First, map each data element to an n-dimensional feature space. n is the number of objects. Detects hyperplanes that divide data into two classes, maximizing the boundary distance between both classes and minimizing classification errors.

The distance constraint for a class is the distance between the solution's hyperplane and theclass's nearest instance. Each data point is initially mapped as a point in n-dimensional space (where n is the number of items), with the value of each object being the value of the suppliedcoordinate. Figure 7 is a simple illustration of the SVM classifier.

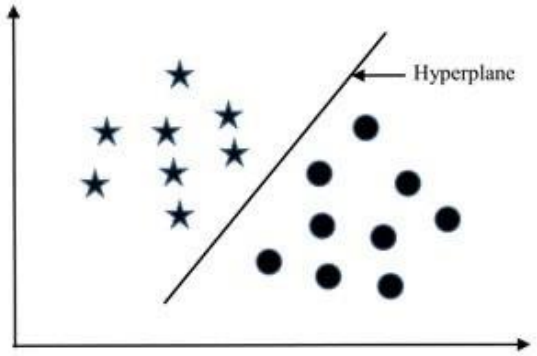

Figure 7: A basic explanation of how the support vector machine works. SVM discovered a hyperplane (a straight line) that optimizes the separation between the "star" and "circle" layers

So, we made some research with dataset and determined that SVM algorithm has that accuracy which is presented in Figure 8 and as we can see the accuracy of SVM is 100% whichrepresent better result for our study.

```
In [11]:
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score
from sklearn.model_selection import cross_val_score

svc_model = SVC()

svc_model.fit(X_train,y_train)

pred = svc_model.predict(X_test)

score = accuracy_score(y_test, pred)
print("Accuracy score for SVC is {}%".format(score*100))

Accuracy score for SVC is 100.0%
```

Figure 8: Accuracy and model visualization SVM

6. Artificial neural networks

Artificial neural networks (ANNs) are a class of machine learning algorithms that are modeled on how neural networks function in the human brain. McCulloch and Pitts [68] presented them initially, followed by Rumelhartetal. According to the research. As in architecture these associations can be reprogrammed (for example, through neuroplasticity) toaid in information adaption, processing, and storage.

ANN algorithms, similarly, may be depicted as a network of interconnected nodes. Depending on the link, the node output is used as input to another annotation for further processing. Depending on the alterations they execute, nodes are typically organized into a matrix known as a layer. The ANN structure may have one or more hidden layers in addition to the input and output layers.

Nodes and edges are hefty weights that enable you to control the strength of the communication signal. Repeated training can enhance or decrease communication signals. Predictions of test data can be made using training and subsequent selection matrices, node weights, and edges. Figure 9 depicts an ANN (with two hidden layers) and its corresponding set of nodes.

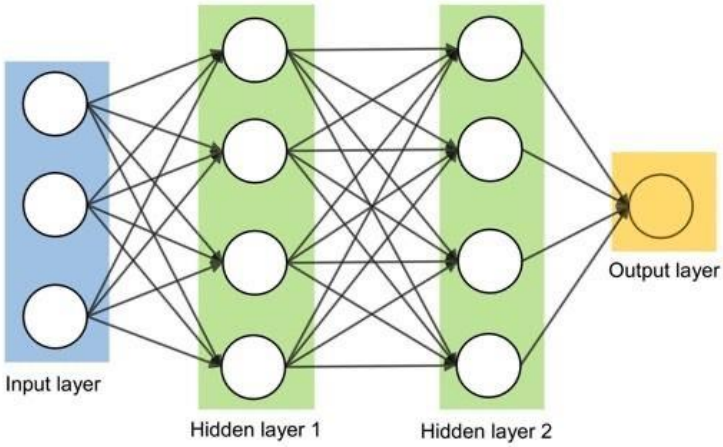

Figure 9: The structure of an artificial neural network with two hidden layers is represented. The arrow connects one node level's output to the input of another node level

So, we made some research with dataset and determined that ANN algorithm has that accuracy which is presented in Figure 10 and as we can see the accuracy of ANN is 64.1% which represent worse result for our study.

```
In [24]: history.history["accuracy"]

Out[24]: [0.6414633989334106, 1.0, 1.0, 1.0, 1.0]

In [25]: import matplotlib.pyplot as plt
         plt.plot(history.history["accuracy"])
         plt.title("Model accuracy")
         plt.xlabel("Epoch")
         plt.ylabel("Accuracy score")
         plt.show()
```

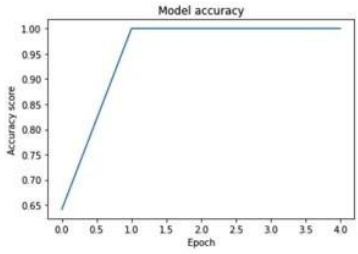

Figure 10: Accuracy and model visualization SVM

7. Results

Table 2
The benefits and drawbacks of various supervised machine learning methods

	Advantages	Limitations
Artificial neural network (ANN)	• Understanding of intricate nonlinearinteractions between dependent and independentvariables. • Less formal statistical training is required. • There are numerous learning algorithms present. • It is also applicable toclassification problems andthe regression problem.	• Because they resemble a "black box," and consumers do not have access to the exact decision-making process. • Training neural networks to Handle complicated classification tasks necessitates a substantial amount of computational power. • Preprocessing is required for variables that are predictive or explanatory.
Logistic regression (LR)	• Ease of implementation and comprehension. • Models based on LR may be simply updated. • There are no assumptions concerning the distribution of independent variables. • This Is an excellent probabilistic explanation of the model parameters.	• The accuracy is low when the connection between the input variables is complicated. • Linear relationships between variables are not considered. • The logical LR model's primary components are excessively reliant. • Overestimation of forecast accuracy can result from sampling mistake. • A standard LR can only categorize variables with two states if it is not a polynomial (i.e. halves).
Naïve Bayes (NB)	• Extremely handy for both basic	• Classes should be mutually

Support vector machine (SVM)	and huge datasets. • It is applicable to both binary and multiclass classification issues. • This necessitates less training data. • It is capable of making probabilistic predictions and processing both continuous and discrete input. • More dependable tan LR • Capable of handling numerous spatial objects. • Reduced danger of retraining. • It is effective for categorizing semi-structured or unstructured data such as words, photos, and so on.	exclusive of one another. • Dependence between characteristics has a detrimental impact on classification performance. • Assuming a normal distribution of numerical characteristics. • The computing cost for huge and complicated datasets. • It will not function if the data is noisy. • The consequences of the ensuing patterns, weights, and variables are frequently difficult to comprehend. • If there is no extension, common SVMs cannot categorize more than two classes.

In the end of the research in Figure 11 we can see the comparison graph of the methods.

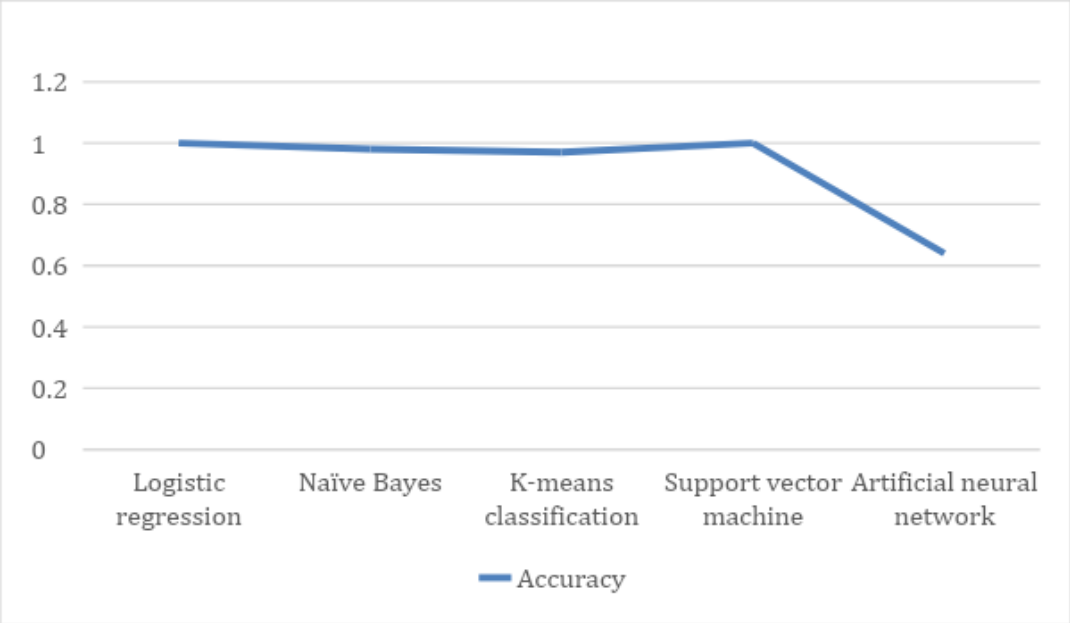

Figure 11: Accuracy and model visualization of all machine learning algorithms

Therefore, it should be noted that the accuracy of the algorithm depends on the size of the dataset, the number of objects and the results of the model as a whole, it is better to use only one model.

8. Conclusion

Machine learning works with health departments to provide disease-related tools and dataanalysis. Therefore, machine learning algorithms play an important role in the early detectionof diseases. This article provides an overview of two popular machine learning algorithms usedto predict diseases. The standard dataset is used for a wide range of diseases including fungal infection, allergy, GERD, chronic cholestasis, peptic ulcer disease, diabetes, bronchial asthma, migraine, paralysis (brain hemorrhage) and more. Furthermore, the accuracy of the same method might differ from data set to

data set since many critical aspects influence the model'saccuracy and performance. Data set function selection and function computation Another significant finding in this analysis is that the model's accuracy and performance may be enhanced by applying specific algorithms that create single pairings.

The list of results found by the researchers is divided into tables for the diagnosis of diseases using machine learning algorithms Naive Bayes and K-Means After comparing data sets of 129 columns of two models predicting Nave-Bayes disease, it was shown that they haveexcellent prediction accuracy.

Moreover, as guidance for future study, some of the limitations of this work are outlined.

9. References

[1] S. Mertens, F. Gailly, and G. Poels, Supporting and assisting the execution of flexible healthcare processes, in: Proc. Int. Conf. Pervas. Comput. Technol. Healthcare, 2015, pp.375–388.

[2] R. Manne, S.C. Kantheti, Application of artificial intelligence in healthcare: chances and challenges, Curr. J. Appl. Sci. Technol. 40.6 (2021) 78–89. URL: https://doi.org/10.9734/cjast/2021/v40i631320.

[3] J. Paparrizos, R. W. White, and E. Horvitz, Screening for Pancreatic Adenocarcinoma Using Signals From Web Search Logs: Feasibility Study and Results, Journal of Oncology Practice (2016) JOPR010504.

[4] L. J. Bisson, et al., Accuracy of a computer-based diagnostic program for ambulatory patients with knee pain, The American journal of sports medicine (2014) 0363546514541654.

[5] A. Lally, et al., WatsonPaths: scenario-based question answering and inference overunstructured information, Yorktown Heights: IBM Research, 2014.

[6] P. B. Jensen, L. J. Jensen, and S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nat. Rev. Genet. 13 (2012) 395.

[7] W. F. Stewart, N. R. Shah, M. J. Selna, R. A. Paulus, and J. M. Walker, Bridging the inferential gap: The electronic health record and clinical evidence, Heal. Aff. 26 (2007) 181–191.

[8] M. Rotmensch, Y. Halpern, A. Tlimat, et al., Learning a Health Knowledge Graph from Electronic Medical Records, Sci Rep 7 (2017) 5994.

[9] N. G. Weiskopf, A. Rusanov, and C. Weng, Sick patients have more data: the non-random completeness of electronic health records, in: AMIA Annu Symp Proc, 2013.

[10] Le Strat, Carrat, 1999; Siettos, Russo, 2013.

[11] R. Shinde, S. Arjun, P. Patil, J. Waghmare, An IntelligentHeart Disease Prediction System Using K-Means Clustering and Naïve Bayes Algorithm, International Journal of Computer Science and Information Technologies 6.1 (2015) 637-63.

[12] G. Subbalakshmi, K. Ramesh, M. Chinna Rao, Decision Support in Heart Disease Prediction System using Naive Bayes, Indian Journal of Computer Science and Engineering (IJCSE) 2.2 (2011) 170-176.

[13] Sh. A. Pattekari and A. Parveen, Prediction System For Heart Disease Using Naïve Bayes, International Journal of Advanced Computer and MathematicalSciences 3.2 (2012) 290-294.

[14] URL: http://datareview.info/article/6-prostyih-shagov-dlya-osvoeniya-naivnogo-bayesovskogo-algoritma-s-primerom-koda-na-python.

[15] URL https://www.kaggle.com/kaushil268/disease-prediction-using-machine-learning.

[16] S. B. Rakhmetulayeva, K. S. Duisebekova, A. M. Mamyrbekov, G. N. Astaubayeva, K. Stamkulova, Application of Classification Algorithm Based on SVM for Determining the Effectiveness of Treatment of Tuberculosis, Procedia Computer Science 130 (2018) 231–238.

[17] S. NagaMallik Raj, N. Thirupathi Rao, V. N. Mandhala and D. Bhattacharyya, Machine Learning Algorithms To Enhance Security In Wireless Network, Journal of Critical Reviews, 7.14 (2020) 425-432.

[18] D. Q. Zeebaree, H. Haron, A. M. Abdulazeez and S. R. M. Zeebaree, Combination of K-means clustering with Genetic Algorithm: A review, International Journal of Applied Engineering Research 12.24 (2017) 14238-14245.

[19] S. B. Rakhmetulayeva, K. S. Duisebekova, D. K. Kozhamzharova, M. Zh. Aitimov, Pollutant transport modeling using Gaussian approximation for the solution of the semi- empirical

equation, Journal of Theoretical and Applied Information 99.8 (2021) 1730–1739.
[20] A. F. Jahwar, A. M. Abdulazeez, Meta-Heuristic Algorithms for K-Means Clustering: A Review, Palarch's Journal of Archaeology of Egypt/Egyptology, 2021.
[21] N. Valarmathy and S. Krishnaveni, Performance Evaluation and Comparison of Clustering Algorithms used in Educational Data Mining, International Journal of Recent Technology and Engineering 76S5 (2019).
[22] S. Ray, A Quick Review of Machine Learning Algorithms, in: Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, Com-IT-Con, India, 14th -16th Feb, 2019.
[23] Sh. Shukla and S. Naganna, A Review on K-means Data Clustering Approach, International Journal of Information & Computation Technology 4.17 (2014) 1847-1860.