SOFT OPTIMIZATION OF TREATMENT METHODS FOR A GROUP OF NEUROLOGICAL DISEASES USING DATA MINING

OTIMIZAÇÃO SUAVE DE MÉTODOS DE TRATAMENTO PARA UM GRUPO DE DOENÇAS NEUROLÓGICAS UTILIZANDO MINERAÇÃO DE DADOS

Shabnam Zarghami*
ORCID 0000-0001-9409-9512
Ph.D. Candidate, Department of Mathematics,
University of Qom
Qom, Iran
sh.zarghami1991@gmail.com
*Corresponding Author

Gholam Hassan Shirdel
ORCID 0000-0003-2759-4606
Associate Professor, Department of Mathematics,
University of Qom
Qom, Iran
shirdel81math@gmail.com

Mojtaba Ghanbari
ORCID 0000-0001-5874-4182
Department of Mathematics, Islamic Azad
University, Farahan Branch
Farahan, Iran
Ghanbari.moji889@gmail.com

Abstract. The growing healthcare industry is generating large amounts of useful data on patient demographics, treatment plans, payment, and insurance coverage, attracting the attention of clinicians and scientists alike. The purpose of this research is to predict diseases of the brain and nerves using data mining techniques. In this research, after data preparation, disease prediction has been attempted using large matrix methods and data mining techniques. By examining the new vector, we can find out which of the diseases in the matrix will be closer to this new disease with new symptoms using the rows of the matrix. The conducted research is one of descriptive-analytical and applied studies. In this research, we used different meters such as Manhattan, cosine similarity, Pearson, Minkowski and K nearest neighbor and implemented a program to predict neurological diseases using Python software. In the algorithm implemented by Python software, the doctor enters the symptoms of the patient and the program output of each meter shows three diseases close to the input symptoms and finally all the meters are compared and each time the meter is executed, which has a weaker result is determined. The advantages of each of these meters are explained below.

Keywords: OR in medicine, soft optimization, treatment methods, data mining, prediction of neurological diseases.

Resumo. A crescente indústria dos cuidados de saúde está a gerar grandes quantidades de dados úteis sobre dados demográficos dos pacientes, planos de tratamento, pagamentos e cobertura de seguros, atraindo a atenção tanto de médicos como de cientistas. O objetivo desta pesquisa é prever doenças do cérebro e dos nervos usando técnicas de mineração de dados. Nesta pesquisa, após a preparação dos dados, foi tentada a previsão de doenças usando métodos de matrizes grandes e técnicas de mineração de dados. Ao examinar o novo vetor, podemos descobrir quais das doenças da matriz estarão mais próximas desta nova doença com novos sintomas usando as linhas da matriz. A pesquisa realizada é do tipo descritivo-analítico e aplicado. Nesta pesquisa, utilizamos diferentes medidores como Manhattan, similaridade de cosseno, Pearson, Minkowski e K vizinho mais próximo e implementamos um programa para prever doenças neurológicas utilizando software Python. No algoritmo implementado pelo software Python, o médico insere os sintomas do paciente e a saída do programa de cada medidor mostra três doenças próximas aos sintomas de entrada e finalmente todos os medidores são comparados e cada vez que o medidor é executado, que tem um sinal mais fraco resultado é determinado. As vantagens de cada um desses medidores são explicadas a seguir.

DOI http://dx.doi.org/10.14571/brajets.v17.n1.302-316
ISSN 2317-9907

Palavras-chave: OR em medicina, otimização suave, métodos de tratamento, mineração de dados, previsão de doenças neurológicas.

1. INTRODUCTION

Vast amounts of data and information are collected and analyzed daily in recent decades, all of which play an important role in business management (Ghorbani, & Ghousi, 2019). Traditional methods were used to analyze data in the past, and manual operations were trusted by researchers. Analyzing data using traditional methods is a long and time-consuming process. Furthermore, non-scientific results have been reported by many researches. The purpose of acquiring new knowledge and information is to discover useful and effective knowledge and data mining of different stages to obtain useful information. In fact, data mining can be defined as the process of discovering and extracting hidden information, patterns and specific data connections according to the predicted ideas.

Data mining is considered as a novel field with different applications, which is known as one of the top ten influential sciences in the field of technology and technology. Data mining can be found wherever data and information are available, for example, the field of data mining can include: market portfolio analysis, training, manufacturing engineering, customer relationship management, penetration detection, financial sector, banking, corporate monitoring, investigative and criminal analysis, telecommunications and healthcare and many more.

Nowadays, the healthcare industry extracts complex data about patients, hospital resources, diagnosis of diseases, electronic patient records, and medical devices. These data in large amount are considered as an essential source for data mining. There is a lot of potential in healthcare data mining programs and some of the most important ones in healthcare include disease prediction and diagnosis, treatment effectiveness, healthcare management, fraud and abuse, customer relationship management (CRM) and the medical device industry (Islam et al., 2018).

Choosing the wrong methods leads to waste of time and money, furthermore, this situation can lead to side effects such as the death of patients. Therefore, it is necessary to diagnose and choose the appropriate treatment method for patients. Data mining can help medical science in predicting and determining various diseases. Health care is considered as a booming part of the economy of many countries. Challenges facing the growing health system include increased costs, inefficiencies, poor quality, and increased complexity (Yang et al., 2015). US health care costs increased by 123% between 2010 and 2015, from $2.6 trillion to $3.2 trillion (Cortada et al., 2012). Inefficiency and non-value-added tasks (eg, readmissions, inappropriate use of antibiotics, and fraud) account for 21–47% of these enormous costs (CMS, 2017).

Some of these costs were related to poor quality care – in many studies, researchers found that approximately 251,454 patients in the United States die each year due to medical errors (Berwick & Hackbarth, 2012). Making better decisions based on available information can reduce these challenges and facilitate the transition to a value-based healthcare industry. Health care institutions are accepting information technology in their management system (Makary, & Daniel, 2016). Big data is collected through this system on a regular basis. It provides tools and techniques to extract information from this complex and voluminous data and transform it into information to aid decision making in healthcare.

Analytics aims to develop insights through the effective use of data and the application of quantitative and qualitative analysis (Prokosch & Ganslandt, 2009). This method can make fact-based decisions for the purposes of "planning, management, measurement and learning", for example, medicare service centers and Medicaid for the Elderly & People with Disabilities or Centers for Medicare & Medicaid Services (CMS) and analyticsis used to reduce hospital readmissions and prevent $115 million in fraudulent payments.

Analytics – including data mining, text mining, and big data mining - helps healthcare professionals predict, diagnose, and treat disease, leading to improvements in service quality and cost reductions (Simpao et al., 2014). Some estimates show that the application of data mining can save 450 billion dollars annually from the United States health care system (Ghassemi et al., 2015). In the last ten years, data mining and exploration researchers have studied big data from both practical (such as application in pharmaceutical care (drug side effects) or mental health) and theoretical (such as reflection on methodological or philosophical challenges of data mining). Researchers and statisticians predict that a large and increasing number of people will have various neurological disorders, which is very important to develop advanced methods and techniques for early diagnosis of neurological diseases in the early stages in order to reduce this number of deaths. Because the high number of deaths of patients in this field is due to the late diagnosis of this disease. Therefore, there is a need to use advanced information technology solutions such as data mining as a suitable tool in the field of information technology in this situation.

The commercialization of data mining techniques plays a prominent role in the medical and healthcare sector, and because data mining techniques offer a wide range of methods, solutions and applications to obtain information, collect, preserve, analyze and provide easy access to data to help users make better decisions through electronic patient health records. Various activities and functions of decision support system (DSS) such as querying and reporting include online analytical processing "OLAP", statistical analysis, forecasting and data and text mining. Also, this situation is shown as a system to emphasize solid and regular methods, successful management, health care with the aim of ensuring the effect of this management in improving quality and controlling costs. Then, data mining plays its role in discovering hidden patterns and relationships in a wide range of diseases.

On the other hand, data mining supports the ability to extract and discover hidden unknowns, in other words, it discovers interesting patterns from a large databases repository. These patterns can help in medical diagnosis and decision-making regarding diseases. Different data mining techniques and interesting algorithms have been designed and built for many applications in different sectors to extract serious knowledge and information in the diagnosis and treatment of diseases based on large medical data sets.

Data mining is considered a tool in business intelligence to discover successful knowledge. The predictive power of data mining comes from the principles of pattern recognition, machine learning and statistics, all of which can be extracted automatically in knowledge, and it is also possible to determine relationships and patterns in large databases. Data mining includes a number of complex data and provides advanced data analysis tools to discover patterns, hidden and unknown valid relationships in previously available large data sets. These analysis tools are classified into mathematical algorithms, statistical methods and learning algorithms and are used for data mining for data from extensive repositories that provide data in the EHR structure with the support of many DM tools such as Waikato environment for WEKA knowledge analysis, Constance mining information, KNIME, Rapid Miner, Orange, etc.

An interesting analysis model using DM techniques such as "classification, clustering and communication law" can be used to present the analysis results of the considered cases. Data mining techniques are responsible for data mining tasks and activities that can be presented as prediction or description of models. According to predictive models; predictions related to data values using known and specific results from existing data sets can be used to discover descriptive models of patterns or relationships in the data from them. Therefore, predictive data mining models, unlike the predictive model, include classification, prediction, regression and analysis in a specific time series. Classification is probably considered as the best classical prediction method compared to all data mining techniques based on machine learning.

The data is supported in discovering hidden patterns and finding interesting sets of subgroups in the bulk of the data. It is hoped that you will get useful and effective information for future research by reading this article. This study is organized as follows: Section 2 explains the methods and techniques of knowledge discovery in databases and the concepts of data mining and describes the research strategy used in this research. Section 3 presents part of the data related to symptoms and diseases extracted from the Aminoff's book and expert consultations during many meetings with a neurologist and it was prepared using the consultation of a doctor and using clinical data in the archive.

A matrix with about 150 rows and 500 columns and the entries of this matrix represent the jth symptoms of the ith disease. Section 4 deals with the diagnosis and prediction algorithm of the disease. Section 5 describes the code execution implemented by Python software to predict diseases. Section 6 presents the conclusion. After building this matrix using different methods in data mining, it was discussed that if the disease has its own characteristic symptoms, it will be entered into the software as an input and an algorithm that will be implemented using Python software, and The Python program implemented by us has three outputs, which are the closest diseases to the input symptoms, or to be more precise, using large matrix methods and data mining techniques after preparing the matrix, if any.

A disease with symptoms can be found by examining the new vector, which of the matrix diseases will be closer to this new disease with new symptoms, respectively. In the next step, different data mining methods that will be used for this matrix have been compared and it can be seen which one gives a more optimal answer or its error is less. The important result of this research is that the best method should be chosen that has the least time complexity to achieve the result.

The remarkable thing is that, this method can be generalized to many other situations. Section 4 deals with the results and execution method and explains the execution method. The ever-increasing advances in health science have led to an increase in the lifespan of human societies, a decrease in deaths, and an increase in the elderly population. With the increasing advancements in health science, the life expectancy of human societies has increased, and mortality has decreased, as well as the elderly population has increased.

2. MATERIAL AND METHODS

This study is considered as a descriptive-analytical and applied study. One of the best study methods is data mining that has been used. This study has used different data mining methods and techniques for early disease prediction and diagnosis.

2.1. System recognition

It is very vital and necessary to know the field in which data mining should be done, as well as having the relevant knowledge to carry out this study. Therefore, in the first step, an attempt has been made to properly understand the field of study by consulting with a neurologist and carefully studying the Aminoff's book, as well as studying patients in the field of neurology and determining the factors affecting the disease, therapeutic and diagnostic methods as well as methods of preventing this disease.

2.2. Data preparation

In this study, the data used is taken from the Aminoff's clinical neurology book and consultation with neurologist and clinical data. A matrix has been formed with about 150 rows and 500 columns after consulting with the relevant doctor and using the clinical data in the archive, and the entries of this matrix represent the jth symptom for the ith disease. It was designed by reading the literature and consulting with a specialty doctor.

2.3. Modeling

Thus, there are various data mining methods for modeling. Therefore, in this study, modeling has been done in Python and predictive models have been presented using data mining techniques.

2.4. Manhattan Distance

Manhattan distance (also known as the taxicab or city block distance) is the simplest measure of distance calculation, the advantage of this method is its high speed. Also, the use of Euclidean distance has contributed to the high speed of calculations.

Manhattan distance calculation: It is defined as the sum of absolute distance between coordinates in corresponding dimensions. For example, In a 2-dimensional space having two points Point1 (x1,y1) and Point2 (x2,y2), the Manhattan distance is given by:

$$|x1 - x2| + |y1 - y2|. \qquad (1)$$

Manhattan and Euclidean distance can be generalized. This generalization is called Minkosovki distance criterion (equation (2)).

$$d(x.y) = \left(\sum_{k=1}^{n} |x_k - y_k|^r\right)^{\frac{1}{r}} (2)$$

Pearson's correlation coefficient is a measure to check the correlation between two variables. Therefore, this criterion can be used to identify people who are more similar to the person in question. Using Pearson's correlation coefficient is one of the ways to overcome the problem of diversity in scoring.

Pearson's correlation coefficient is calculated as equation (3):

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (3)$$

2.5. Cosine similarity

Cosine similarity is very famous in the field of text processing, which is tried to be introduced. This formula is used in collaborative filtering as well. Cosine similarity ignores the 1-1 criteria. This criterion is introduced as equation (4):

$$\cos(x.y) = \frac{x \cdot y}{||x|| \times ||y||} (4)$$

2.6. K-Nearest Neighbor

The k-nearest neighbor algorithm is used to identify suitable suggestions for collaborative filtering from K people, which are most similar to the target person. The best value for k depends on the problem, different tests should be done to get the best k. Each of the mentioned methods of recommender systems is implemented in Python. The cosine similarity has been used to calculate the similarity between items in this study.

2.7. Grade inflation

Grade inflation means the awarding of higher grades higher than expected. To compensate for this inflation, the average grades of the user should be subtracted from each of the grades, the adjusted cosine similarity value is obtained from the result.

3. RESULTS

3.1. Data and symptoms

The set of diseases and data in the matrix is defined as a table in Excel, Table (1), (2) whose rows represent diseases and its columns represent symptoms, and the way to collect data is as described. Table (3) shows the symptom codes. Table (4) shows the disease codes

Disease diagnosis code was modeled using Python and data mining techniques including Euclidean distance, k nearest neighbor, Pearson distance and cosine similarity. The data (matrix row) from 1 to 150 have been obtained for the types of symptoms (matrix column) that indicate diseases.

Figures (figer 1,2,3) help us find the correct understanding of the data. By drawing graphs, we try to make numerical information in a state that is possible for humans to understand, because mere numerical information does not help us, and it is by modeling and analyzing the structure of these data that we can have a correct understanding of the reality behind these numbers. One of the most important diagrams is the heatmap diagram. In fact, the purpose of this chart is to create an initial clustering and display numerical information in the form of color. In the figure below, you can see a heat map, which is displayed in the column and row of numbers in the form of colors. Each cell of this spectrum chart represents a numerical value. This spectrum is shown in the figure with different colors. Numbers with values below zero are displayed in red and above in blue, and zero values are in black. By looking at this graph, we can see that numbers with values below zero are displayed in yellow and above in blue. By looking at this chart, we can see which part has what value.

There are several varieties for the mean in mathematics and especially in statistics. In the study of the distribution of a statistical population, the representative value around which the measurements are distributed is called the central value, and any numerical measure that represents the center of the data set is called the measure of central tendency. Mean and median are the most common measures of center tendency.

Median in statistics and probability theory is one of the measurements of the tendency to the center. The median is a number that divides a statistical population or a probability distribution into two equal parts. One of the important advantages of the median over the mean is that the mean is not affected by very large and very small numbers in the set of measurements. One of the most important properties of the median is that the sum of the absolute value of the differences of different values of the random variable is the minimum.

Standard deviation (figure 1) (symbol σ) is one of the dispersion indices that shows how far the data is from the average value. If the standard deviation of a set of data is close to zero, it is a sign that the data are close to the mean and have little dispersion; While the large standard deviation indicates the significant dispersion of the data. The standard deviation is equal to the square root of the variance. Its advantage over variance is that it is also dimensional with data. Standard deviation is also used to determine the reliability coefficient in statistical analysis. In scientific studies, usually data with a difference of more than two standard deviations from the mean value are considered as outliers and are excluded from the analysis.

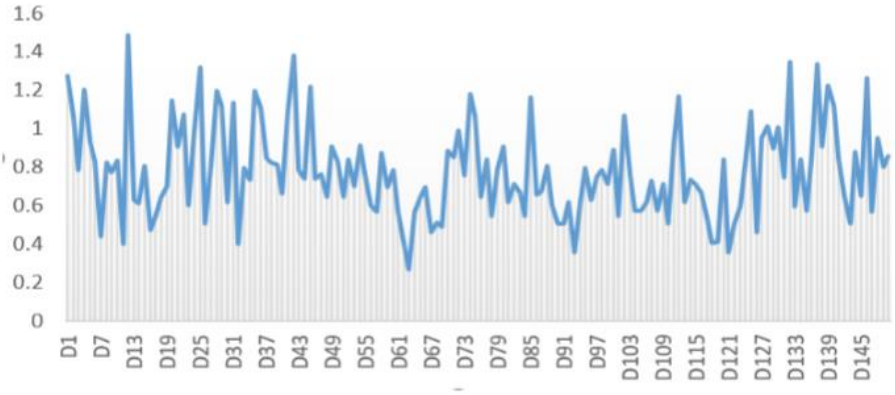


Figure 1. Standard deviation data

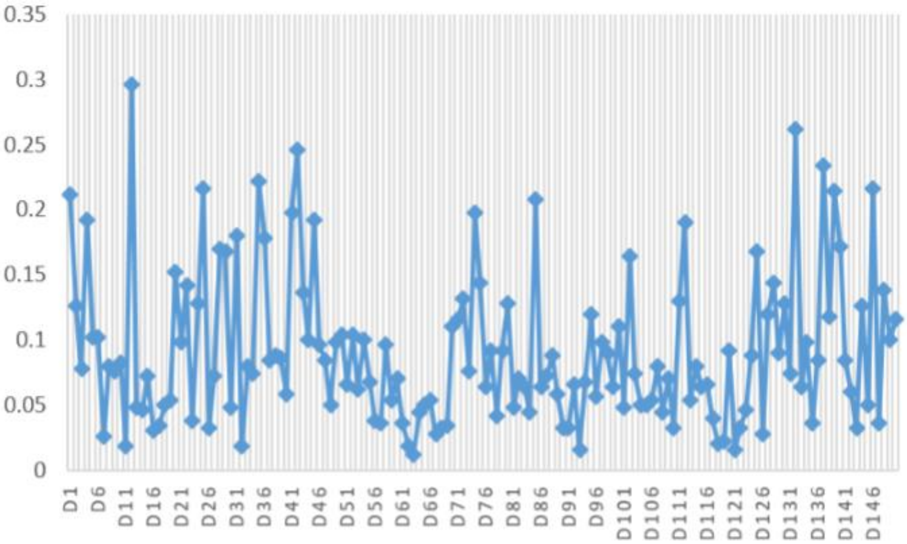


Figure 2. Mean data

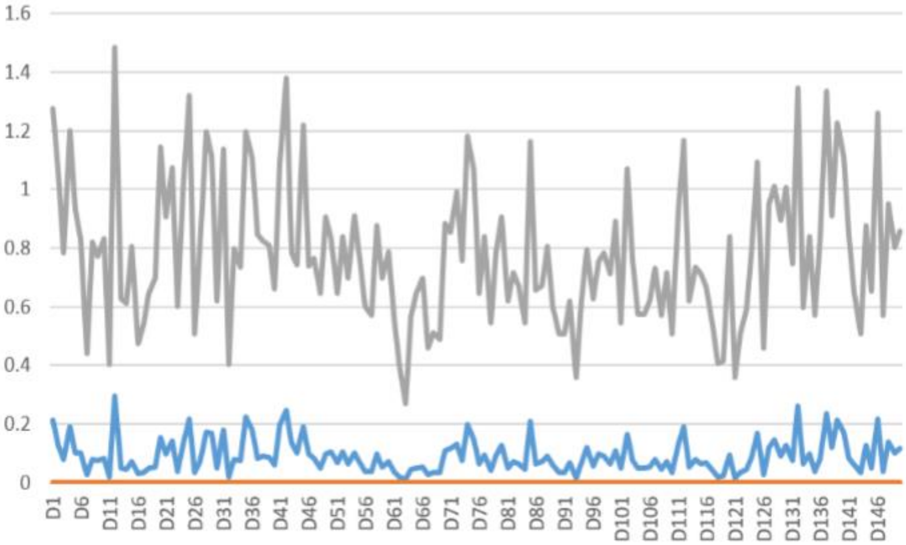


Figure 3. Check the median data and the standard deviation data and the mean data

Table 1. Symptoms A

Headache	Diseases
10	Subarachnoid hemorrhage
9	Meningitis or encephalitis
9	Hypertensive encephalopathy
9	Giant cell arteritis
8	Brain tumors
9	Pseudotumor of the brain(pseudotum or of the brain)
0	Trigeminal neuralgia (TN)
0	Trigeminal neuralgia
0	Neuralgia after herpes
9	Hypertension

Table 2. Symptoms B (Confusion's reduced level of consciousness

Confusion's reduced level of consciousness	Diseases
9	Subarachnoid hemorrhage
9	Meningitis or encephalitis
0	Hypertensive encephalopathy
0	Giant cell arteritis
7	Brain tumors
0	Pseudotumor of the brain(pseudotum or of the brain)
0	Trigeminal neuralgia (TN)
0	Trigeminal neuralgia
0	Neuralgia after herpes
0	Hypertension

Table 3. Some symptoms

Symptoms	Code	Symptoms	Code
Plantar reaction (bilateral extensor or plantar reflex)	G	Headache	A
Hemiparesis (paralysis of one limb or one half of the body)	H	Confusion's reduced level of consciousness	B
Aphasia (language disorder)	I	Vomit	C
Visual field defect or visual changes	J	neck stiffness	D
Tentorial herniation	K	high blood pressure	E
Progressive drowsiness	L	Fever	F

Table 4. Some Diseases

1) Subarachnoid hemorrhage	9) Neuralgia after herpes
2) Meningitis or encephalitis	10) Hypertension
A3) Hypertensive encephalopathy	11) Atypical facial pain
4) Giant cell arteritis	12) Migraine
5) Brain tumors	13) Cluster headache
6) pseudotumor of the brain (pseudotumor of the brain)	14) Tension-type headache
7) Trigeminal neuralgia (TN)	15) Ice pick headaches
8) Trigeminal neuralgia	

3.2. Python code implemented on the data

We implemented the algorithm in the Table (A1) after collecting data using Python. For additional explanations, we implemented this algorithm with different Manhattan, cosine similarity, Minkowski, Pearson and k-nearest neighbor metrics with Python software that is explained in the above sections completely. These meters show the closest similarities to our input signals.

3.3. Code execution and disease prediction

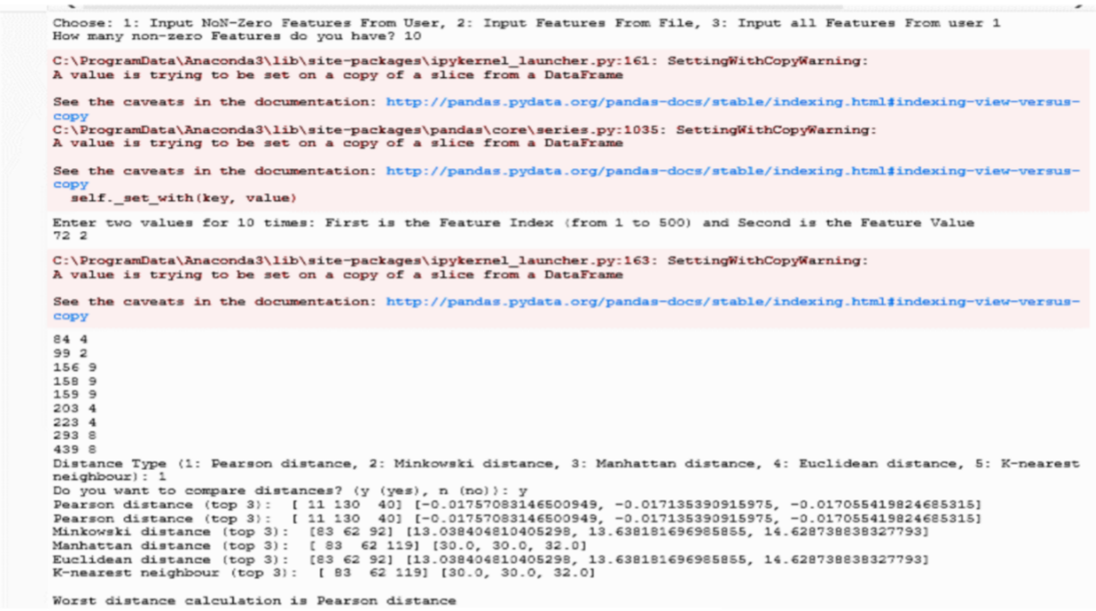
The program execution is included in this section for additional explanations of the previous sections. The signs are entered as input and then each of five different meters including Manhattan and Minkowski and cosine similarity and Pearson and the nearest neighbor explained in the previous sections are implemented, and three output diseases are presented close to the input symptoms by our Python program. And finally, all the meters are compared and the worst meter is presented, it has poorer results than the rest of the meters.

It is possible to discover and extract new knowledge from retrospective data by data mining. The way of data preprocessing, as well as selected variables, has a significant effect on knowledge discovery. There are various data techniques that are used to predict diseases. This study has used five data-mining algorithms that will be explained below. The experimental results show the efficiency and effectiveness of all three methods that have been compared based on sensitivity, specificity and accuracy. Pruning and strengthening methods were used to find the desired structure and increase the accuracy of the results. The existing database has been reviewed. The present study has reviewed the predictive approaches of data mining in neurological diseases and their diagnosis. Therefore, new research can be done in this field according to the notes mentioned about research gaps and the use of predictive data mining approaches in the early diagnosis of various (medical) diseases.

Example-1-5- As an example, a sample has been examined in Figure 1, and some symptoms of the patient have been entered as input into the software implemented with Python, which include, respectively, ophthalmoplegia (CF=4), sensory loss (CU=2)), hypotonia (EZ=2), Romberg's sign in the legs (FB=9), obvious disorder of the vibration sense in the legs (FC=9), obvious disorder of lower vibration sense (9FC=), ptosis (GU=4), optic atrophy (HO=4) , abdominal pain (KG=8) and urinary incontinence (PW=8) respectively, to show the diseases close to this patient and the output is shown with different meters respectively.

The output of Pearson shows three diseases, namely optic facial pain, entrapment syndromes in the lower limb and autosomal dominant spinocerebellar ataxia (also known as the SCAs) (Zacharski, 2021). The output of Minkowski is spinal muscular atrophy (SMA), oculomotor fascicular lesions and neuralgic amyotrophy. The output of the Manhattan is Spinal Muscular Atrophy (SMA), oculomotor nerve lesions and myoglobinuria. The output of the Cosine Similarity is spinal muscular atrophy, oculomotor nerve lesions and neuralgic amyotrophy. The output of the k-nearest neighbor (k-NN) is Spinal muscular atrophy (SMA), oculomotor fascicular lesions and myoglobinuria (Falk, 2019).

All the output diseases are close to the input symptoms in the field of movement disorders, physical sense disorders, movement defects and visual field defects. Our software has determined that the weakest meter that shows weaker results than other meters is the Pearson meter. The results were reviewed by a neurologist and the result of the review showed that the results are completely acceptable and accurate.



```
Choose: 1: Input NoN-Zero Features From User, 2: Input Features From File, 3: Input all Features From user 1
How many non-zero Features do you have? 10

C:\ProgramData\Anaconda3\lib\site-packages\ipykernel_launcher.py:161: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-
copy
C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\series.py:1035: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-
copy
  self._set_with(key, value)

Enter two values for 10 times: First is the Feature Index (from 1 to 500) and Second is the Feature Value
72 2

C:\ProgramData\Anaconda3\lib\site-packages\ipykernel_launcher.py:163: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-
copy

84 4
99 2
156 9
158 9
159 9
203 4
223 4
293 8
439 8
Distance Type (1: Pearson distance, 2: Minkowski distance, 3: Manhattan distance, 4: Euclidean distance, 5: K-nearest
neighbour): 1
Do you want to compare distances? (y (yes), n (no)): y
Pearson distance (top 3):  [ 11 130  40] [-0.01757083146500949, -0.017135390915975, -0.017055419824685315]
Pearson distance (top 3):  [ 11 130  40] [-0.01757083146500949, -0.017135390915975, -0.017055419824685315]
Minkowski distance (top 3):  [83 62 92] [13.038404810405298, 13.638181696985855, 14.628738838327793]
Manhattan distance (top 3):  [ 83  62 119] [30.0, 30.0, 32.0]
Euclidean distance (top 3):  [83 62 92] [13.038404810405298, 13.638181696985855, 14.628738838327793]
K-nearest neighbour (top 3):  [ 83  62 119] [30.0, 30.0, 32.0]

Worst distance calculation is Pearson distance
```

Figure 4. Python code execution and disease prediction

4. CONCLUSION

This study has reviewed the predictive methods in data mining in neurodegenerative diseases and their diagnosis. Therefore, new research can be carried out in this field according to the notes mentioned about the research gaps and the use of predictive data mining approaches in the early diagnosis of various (medical) diseases. The present study has been carried out with the aim of designing an efficient model to discover the knowledge of predicting diseases in the brain and nervous system based on the latest data set of indicators in this field that are related to the community health, and the analysis of detailed data mining techniques is presented in this study to predict brain and nerve diseases. In other words, research efforts have been presented to apply data mining techniques based on the use of data sets of diseases and symptoms through the business intelligence program to provide important results in making the right decisions at the right time. This study has used different meters, including Manhattan, cosine similarity, Pearson, Minkowski and K nearest neighbor for complete explanations and general conclusions, and has implemented a program using Python, so that; the doctor enters the patient's symptoms and the output of the program shows three diseases of each meter close to the input symptoms, and finally, all the meters are compared, and the meter with the weaker result is determined each time it is run. The advantages of each of these meters are explained in the following.

Distance criteria such as Euclidean or Manhattan are used for dense data (that is, most features have non-zero numbers) and the coefficient of feature values is important. Pearson's criterion is used for data that has grade inflation. The cosine similarity measure is used if the data is sparse.

Hence, it is possible to discover very large medical data, analyze it and obtain very useful knowledge and decision-making using data mining methods.

Data mining can be used as a guide for doctors in predicting neurological diseases and the accuracy of the results obtained from the models is also quite close to reality. These results are shown in figure 5.

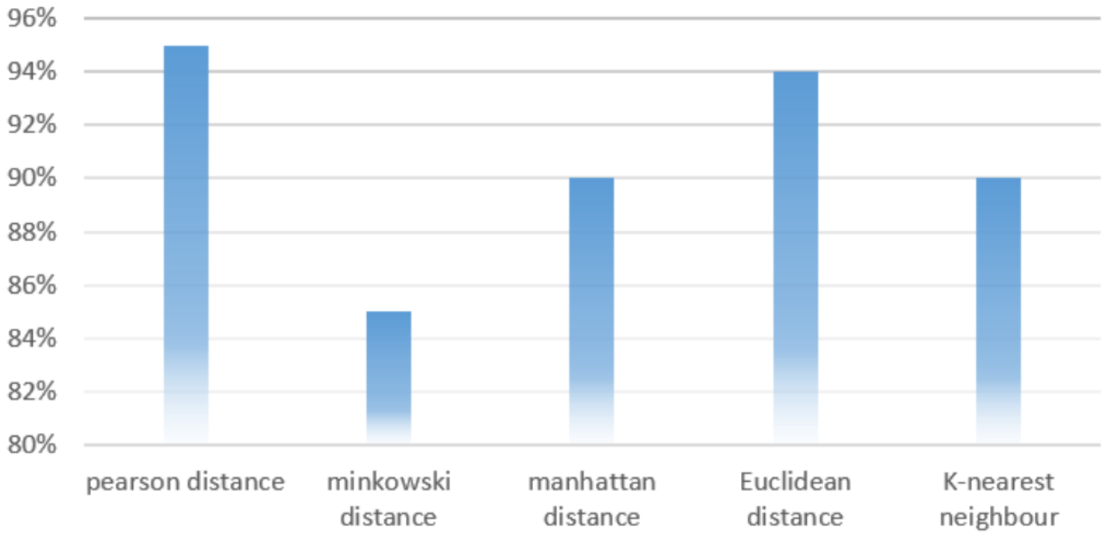


Figure 5. The result of comparison of algorithm and meters in predicting diseases using data mining

ACKNOWLEDGMENTS

I would like to express my special appreciation and thanks to my professor, Dr. Mohammad Reza Eskandari for his help in analyzing the data and medical parts.

DECLARATIONS

Regarding the statements given below, it is not applicable to the content of our submitted article, financial resources, relational benefits, ethical approval because the financial aid in our article was not taken from anywhere, and also the data collection was collected from the main reference and the main book of clinical neurology of Aminoff with the help of a neurologist so animal and human cases have not been used. The complete data table has been uploaded in related files (optional). Regarding the participation of the authors, the consent of the authors has been taken to include their names in the relevant article. The names of the authors along with their information are listed above and the authors have consented to use their names in this research.

ETHICAL APPROVAL

According to the above explanation, the data collection was collected from the main reference and the main book of clinical neurology of Aminoff with the help of a neurologist so animal and human cases have not been used therefore, this instruction is not applicable.

AUTHORS’ CONTRIBUTIONS

Study concept and design: Z.SH., SH.GH.,GH.M.,; Acquisition of data Z.SH.; Analysis and interpretation of data: GH.M.,Z.SH.,; Drafting of the manuscript: Z.SH.,; Critical revision of the manuscript for important intellectual content: SH.GH.; Statistical analysis: Z.SH.,; Administrative, technical, and material support: SH.GH., GH.M.; Study supervision: SH.GH

FUNDING

The financial aid in our article was not taken from anywhere, so this instruction is not applicable.

AVAILABILITY OF DATA AND MATERIALS

Data sets generated or analyzed during the current study are available from the corresponding author upon request.

COMPETING INTERESTS

This research has no conflict of interest.

REFERENCES

Berwick, D. M., & Hackbarth, A. D. (2012). Eliminating waste in US health care. Jama, 307(14), 1513-516.2012.362.http://doi.org/10.1001/jama.

CMS. (2017). Center for Medicare and Medicaid Services. Retrieved from https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-andReports/NationalHealthExpendData/NationalHealthAccountsHistorical.html/

Cortada, J., Gordon, D., & Lenihan, B. (2012). The Value of Analytics in Healthcare. USA: IBM Institute for Business Value Armonk. NY.

Falk K. (2019), Practical Recommender Systems.1st ed.USA, 432.

Ghassemi, M., Celi, L. A., & Stone, D. J. (2015). State of the art review: the data revolution in critical care. Critical Care,19(1),9-1. http://doi.org/10.1186/s13054-015-0801-4.

Ghorbani, R., & Ghousi, R. (2019). Predictive data mining approaches in medical diagnosis: A review of some diseases prediction. 3(2), 70-47. http://doi.org/10.5267/j.ijdns.2019.1.003.

Islam, M. S., Hasan, M. M., Wang, X., Germack, H. D., Noor, E., & Alam, M. (2018). systematic review on healthcare analytics: application and theoretical perspective of data mining.Healthcare,6(2), 43-41http://doi.org/ 10.3390/healthcare6020054.

Michael, J., Aminoff Md DSc, F., & Andrew Josephson, S. (2021). Aminoff's Neurology and General Medicine (6 ed.): Academic Press.

Makary, M. A., & Daniel, M. (2016). Medical error—the third leading cause of death in the US. 33, i2139. http://doi.org/10.1136/bmj.i2139.

Prokosch, H. U., & Ganslandt, T. (2009). Perspectives for medical informatics. Methods of information in medicine, 48(01), 44-38.

Simpao, A. F., Ahumada, L. M., Gálvez, J. A., & Rehman, M. A. (2014). A review of analytics and clinical informatics in health care. Journal of medical systems, 38(4), 7-1.

Yang, J., Li, J., Mulder, J., Wang, Y., Chen, S., Wu, H., & et al. (2015). Emerging information technologies for enhanced healthcare. Computers in industry, 69, 11-13.

Zacharski, R. (2021). A prorammers Gui to Data Mining: the ancient Art of the Numeriati. 395.

APPENDIX A: TABLE A1: PYTHON CODE EXECUTED

```
from math import sqrt
import pandas as pd
import num
from collections import Counter
from itertools import chain
def pearson_correlation_coefficient(arr1, arr2):
    sum_xy = 0
    sum_x = 0
    sum_y = 0
    sum_x2 = 0
    sum_y2 = 0
    n = 0
    if len(arr1) == len(arr2):
        for i in range(len(arr1)):
            n += 1
            x = arr1[i]
            y = arr2[i]
            sum_xy += x * y
            sum_x += x
            sum_y += y
            sum_x2 += pow(x, 2)
            sum_y2+= pow(y, 2)
        denominator = sqrt(sum_x2- pow(sum_x, 2) / n) * sqrt(sum_y2 - pow(sum_y, 2) / n)
        if denominator == 0
            return 0
        else:
            return (sum_xy - (sum_x * sum_y) / n) / denominator
    else:
        return -1
def minkowski_distance(arr1, arr2, p):
    distance = 0
    if len(arr1) == len(arr2):
        for i in range(len(arr1)):
            distance += pow(abs(arr1[i] – arr2[i]), p)
        return pow(distance, 1 / p)
    else:
        return -1

def manhattan_distance(arr1, arr2):
    distance = 0
    if len(arr1) == len(arr2):
        for i in range(len(arr1)):
            distance += abs(arr1[i] – arr2[i])
        return distance
    else:
        return -1

def euclidean_distance(arr1, arr2):
    distance = 0
    if len(arr1) == len(arr2):
        for i in range(len(arr1)):
            distance += pow((arr1[i] – arr2[i]), 2)
        return sqrt(distance)
    else:
        return -1
def k_nearest_neighbor(ref_arr_index, total_arr):
    distances = []
    for i in range(len(total_arr)):
        if i != ref_arr_index:
            distance = manhattan_distance(total_arr.iloc[ref_arr_index].values, total_arr.iloc[i].values)
            distances.append(distance)
    return distances
def recommend(df, disease_name):
   distance_type = input(
'        Distance Type (1: Pearson distance, 2: Minkowski distance, 3: Manhattan distance, 4: Euclidean 
distance, '
'        5 :K-nearest neighbour): ')
    do_compare = input('Do you want to compare distances? (y (yes), n (no)): ')
    distance = []
    distances = []
    for i in range(df.shape[0]):
        if i != disease_name:
            distance.append(pearson_correlation_coefficient(df.iloc[disease_name].values, 
df.iloc[i].values))
    distances.append(distance)
    distance = []
    for i in range(df.shape[0]):
        if i != disease_name:
            distance.append(minkowski_distance(df.iloc[disease_name].values, df.iloc[i].values, 2))
    distances.append(distance)
    distance = []
    for i in range(df.shape[0]):
        if i != disease_name:
            distance.append(manhattan_distance(df.iloc[disease_name].values, df.iloc[i].values))
    distances.append(distance)
    distance = []
    for i in range(df.shape[0]):
        if i != disease_name:
            distance.append(euclidean_distance(df.iloc[disease_name].values, df.iloc[i].values))
    distances.append(distance)
    distance = k_nearest_neighbor(disease_name, df)
    distances.append(distance)
    if distance_type == '1:'
        print('Pearson distance (top 3): ', numpy.argsort(distances[0])[0:3], sorted(distances[0])[0:3])
    if distance_type == '2:'
        print('Minkowski distance (top 3): ', numpy.argsort(distances[1])[0:3], sorted(distances[1])[0:3])
    if distance_type == '3:'
        print('Manhattan distance (top 3): ', numpy.argsort(distances[2])[0:3], sorted(distances[2])[0:3])
    if distance_type == '4:'
        print('Euclidean distance (top 3): ', numpy.argsort(distances[3])[0:3], sorted(distances[3])[0:3])
    if distance_type == '5:'
        print('K-nearest neighbour (top 3): ', numpy.argsort(distances[4])[0:3], sorted(distances[4])[0:3])
   if do_compare == 'y':
        print('Pearson distance (top 3): ', numpy.argsort(distances[0])[0:3], sorted(distances[0])[0:3])
        print('Minkowski distance (top 3): ', numpy.argsort(distances[1])[0:3], sorted(distances[1])[0:3])
        print('Manhattan distance (top 3): ', numpy.argsort(distances[2])[0:3], sorted(distances[2])[0:3])
        print('Euclidean distance (top 3): ', numpy.argsort(distances[3])[0:3], sorted(distances[3])[0:3])
        print('K-nearest neighbour (top 3): ', numpy.argsort(distances[4])[0:3], sorted(distances[4])[0:3])
        counts = Counter(chain(*map(set,
]                                    sorted(distances[0])[0:3], sorted(distances[1])[0:3], sorted(distances[2])[0:3],
                                     sorted(distances[3])[0:3], sorted(distances[4])[0:3])
        common_remove = [[i for i in sublist if counts[i] == 1] for sublist in
]                         sorted(distances[0])[0:3], sorted(distances[1])[0:3], sorted(distances[2])[0:3],
                          sorted(distances[3])[0:3], sorted(distances[4])[0:3]
        list_size = []
        for i in range(len(common_remove)):
            list_size.append(len(common_remove[i]))
        max_size_list = list_size.index(max(list_size))
        if max_size_list == 0:
            print('\n'+'Worst distance calculation is Pearson distance')
        if max_size_list == 1:
            print('\n'+'Worst distance calculation is Minkowski distance')
        if max_size_list == 2:
            print('\n'+'Worst distance calculation is Manhattan distance')
        if max_size_list == 3:
            print('\n'+'Worst distance calculation is Euclidean distance')
        if max_size_list == 4:
            print('\n'+'Worst distance calculation is K-nearest neighbour')
disease_name = 100
dfs = pd.read_excel('E:/final_projectFuture_Signs_final(1).xlsx', header=None)
first_row = dfs.iloc[1,0]:
first_column = dfs.iloc[0,1]:
first_row.to_excel("E:/final_project/features.xlsx", sheet_name='Sheet1')
first_column.to_excel("E:/final_project/disease_names.xlsx", sheet_name='Sheet1')
df = dfs.iloc[1 , :1]
show_type = input('Choose: 1: Input NoN-Zero Features From User, 2: Input Features From File, 3:
Input all Features From user ')
if show_type == '1:'
    how_many_number = int(input("How many non-zero Features do you have? "))
    print("Enter two values for %d times: First is the Feature Index (from 1 to %d) and Second is the
Feature Value" %(how_many_number,(len(df.columns))))
    df.iloc[disease_name][1:len(df)] = 0
    for i in range(how_many_number):
        index, df.iloc[disease_name][index] = [int(x) for x in input().split()]
    recommend(df, int(disease_name))
    if show_type == '2:'
    disease_name = input('Enter the Number of the disease: ')
    recommend(df, int(disease_name))
if show_type == '3:'
    print("Enter %d Feature Values" %(len(df.columns)))
    for i in range(len(df.columns)):
        df.iloc[disease_name][i] = input()
    recommend(df, int(disease_name)
 In[ ]:
```