

A New Approach For Smart Health Prediction Using Data mining Techniques

Dr. K. Naveen durai¹, Deepika B², Divya S³, Divyabharathy M⁴, Indhumathi K⁵

¹Assistant Professor, Dept of Computer Science and Engineering

^{2,3,4,5}Dept of Computer Science and Engineering

^{1,2,3,4,5} Sri Eshwar College of Engineering, Tamilnadu.

Abstract- Data Mining is one of the most propelling territories of research that is gotten progressively well known in health association. Health care industry creates a lot of complex data about patients, emergency clinics assets, ailments, finding strategies, electronic patient's records, and so on. Utilizing this health care data, data mining advancements are to give advantages to gathering the patients having comparative kind of ailments or health issues so healthcare association gives them powerful medicines. In this paper, the health prediction has exactness when utilizing Support vector machine and K-Means Clustering. The data mining systems are valuable to make restorative choices in relieving maladies. The healthcare associations increment their capacity for taking choice about patient health utilizing different data mining systems, for example, characterization, clustering, affiliation, and relapse in health space. This overview additionally features applications, difficulties and future issues of Data Mining in healthcare

Keywords- SVM Classification, K means Clustering, Health Prediction, and Data mining

I. INTRODUCTION

Data Mining is one of the most indispensable and persuading zone of research with the target of finding significant data from immense data sets [4]. Here, Mining is getting main stream in healthcare field on the grounds that there is a need of proficient logical strategy for distinguishing unknown and important data in health data. In health industry, Data Mining gives a few advantages, for example, recognition of the misrepresentation in health protection, accessibility of clinical answer for the patients at lower cost, discovery of reasons for maladies and ID of clinical treatment strategies [2]. It likewise helps the healthcare specialists for making proficient healthcare strategies, building drug proposal frameworks [3], creating health profiles of people and so forth. The data produced by the health associations is exceptionally immense and complex because of which it is hard to examine the data so as to make significant choice with respect to quiet health. This data contains insights about emergency clinics, patients, clinical cases, treatment cost and so on [9]. Along

these lines, there is a need to produce an incredible asset for breaking down and removing significant data from this perplexing data [7]. The investigation of health data improves the healthcare by upgrading the presentation of patient administration tasks [6]. The result of Data Mining advances are to give advantages to healthcare association to gathering the patients having comparative kind of maladies or health issues with the goal that healthcare association gives them successful medications [8]. It can likewise helpful for foreseeing the length of remain of patients in emergency clinic, for clinical determination and making plan for compelling data framework the board. Late advances are utilized in clinical field to upgrade the clinical administrations in savvy way [1] [5]. Data Mining methods are additionally used to examine the different variables that are answerable for maladies for instance sort of nourishment, diverse working condition, training level, living conditions, accessibility of unadulterated water, health care administrations, social ,ecological and rural elements.

II. RELATED WORK

2.1 Data Mining on Health Examination Records

Albeit Electronic Health Records have pulled in expanding research consideration in the data mining and machine learning networks lately mining general health assessment data is a region that has not yet been all around investigated, aside from a couple of studies on risk prediction, for example, the interminable infection early admonition framework proposed in and our past work on health score arrangement framework. Be that as it may, none of them thought about unlabeled data. Moreover, the methodology introduced in is constrained to a parallel grouping issue (utilizing alive/perished names) and therefore it isn't useful about the particular illness zone in which an individual is at risk. The current examinations on healthcare data that took care of unlabeled data are talked about in the following area.

2.2 Classification with Unlabeled Healthcare

Data Unlabeled data grouping are generally taken care of by means of Semi-Supervised Learning that gains from both named and unlabeled data and Positive and Unlabeled learning, an uncommon instance of SSL that gains from positive and unlabeled data alone . PU learning is regularly adjusted for sickness quality order when just the names for infection qualities are accessible. As of late, Nguyen et al. proposed mPUL a multi-class PU learning model for action acknowledgment. The technique trains m 1-versus others twofold probabilistic base classifiers, each prepared with a positive set and a combined arrangement of negative and unlabeled examples. The class choice depends on the most extreme class likelihood more prominent than 0.5; in any case the unknown class is anticipated. Be that as it may, it isn't straightforwardly material to our concern, since we don't have negative occasions accessible for preparing.

III. OUR SYSTEM MODEL

ARCHITECTURE DIAGRAM

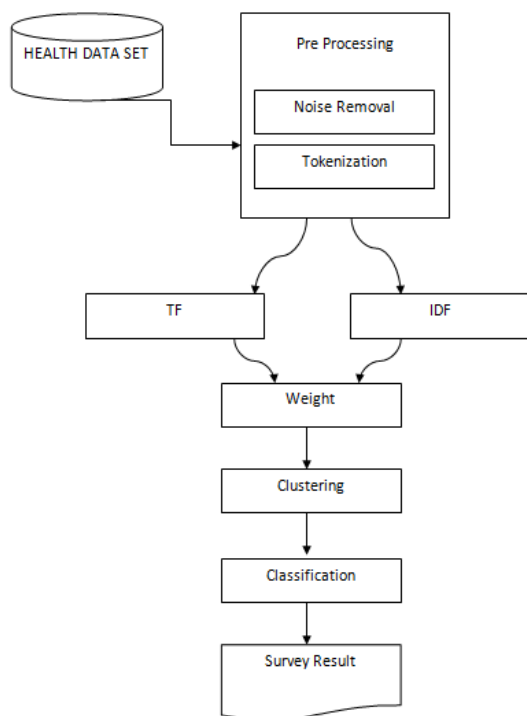


Figure 1: Proposed Health Prediction Model diagram.

It is a non-benefit association of healthcare that offers patients to online access their clinical data, online fill the solution frame and permit safe trading of email with the healthcare supplier. Data mining, it is conceivable to recognize the incessant infection and dependent on the inconvenience of the patient sickness organize the patients with the goal that they will get powerful treatment in convenient and exact way. Data mining procedures are useful

to give the data to quiet in regards to different ailments and their counteraction K-Means with SVM classifier is one of the least difficult classifier that finds the unidentified data point utilizing the recently known data focuses (mean) KMeans with SVM groups the data focuses utilizing more than one closest middle. K-Means has various applications in various territories, for example, health datasets, picture field, group examination, design acknowledgment, web based marketing utilized SVM and it is utilized to break down the connection between cardiovascular illness and hypertension and the risk components of different interminable ailments and dissecting the patients experiencing coronary illness and for diagnosing thyroid sickness.

1. DATA PREPROCESSING:

The dataset contains the tremendous measure of data. The data might be organized or unstructured in Dataset. On the off chance that dataset will be unstructured means the preprocessing takes place. In preprocessing stage every single exchange's are broke down and decide the parameters are utilized in the exchanges. In this way, the unstructured dataset is changed over into structure dataset..

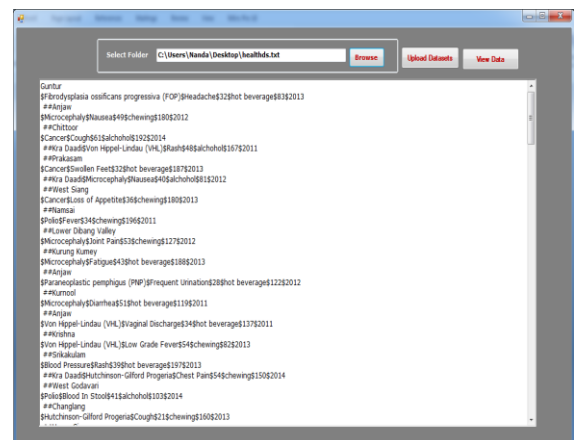


Figure 2: Dataset has been uploaded into database

2. HER record

Gigantic measures of Electronic Health Records (EHRs) gathered throughout the years have given a rich base to risk investigation and prediction. An EHR contains carefully put away healthcare data about an individual, for example, perceptions, research facility tests, demonstrative reports, meds, methodology, quiet recognizing data, and sensitivities. An uncommon kind of HER is the Health Examination Records (HER) from yearly broad health check-ups. Distinguishing members at risk dependent on their present and past HER's is significant for early notice and

preventive mediation. By "risk", we mean undesirable results, for example, mortality and dreariness.

3. WEIGHT CALCULATION:

In this module, the weight has been calculated as area and disease wise, so we find the TF for area and disease and IDF for area and disease.

Term frequency:

In the case of the **term frequency** $tf(t,d)$, the simplest choice is to use the *raw count* of a term in a document, i.e. the number of times that term t occurs in document d .

Inverse document frequency:

The **inverse document frequency** is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

Weight calculation: TF*IDF.

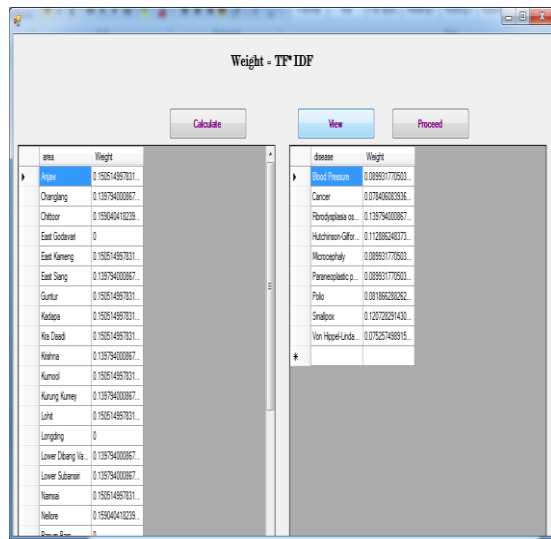


Figure 3: Word Weight Calculation using TF* IDF.

4. CLUSTERING:

We are utilizing K means Clustering for Cluster into three gatherings data, here K-means clustering is a sort of solo realizing, which is utilized when you have unlabeled data (i.e., data without characterized classes or gatherings). The objective of this calculation is to discover bunches in the data, with the quantity of gatherings spoke to by the variable K. The

calculation works iteratively to relegate every datum point to one of K bunches dependent on the highlights that are given.

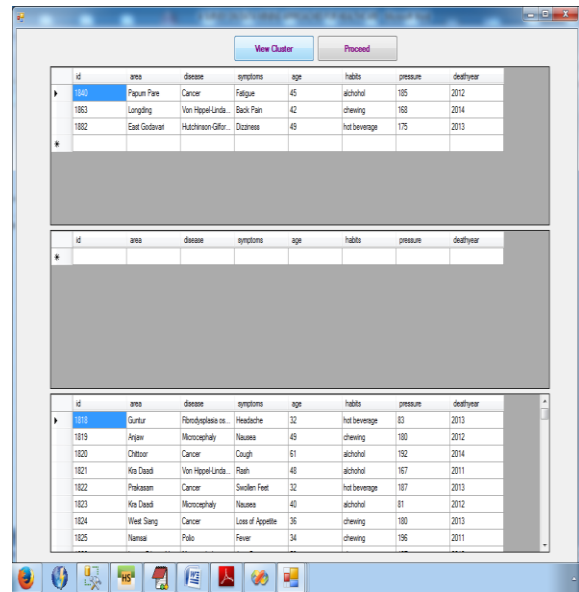


Figure 4: By applying K means clustering Results

5. SHG Based SVM Classification:-

Given a lot of preparing models, each marked as having a place with either of two classes, a SVM preparing calculation manufactures a model that allocates new guides to one classification or the other, making it a non-probabilistic paired straight classifier (in spite of the fact that strategies, for example, Platt scaling exist to utilize SVM in a probabilistic characterization setting). A SVM model is a portrayal of the models as focuses in space, mapped so the instances of the different classifications are separated by a reasonable hole that is as wide as could be allowed.

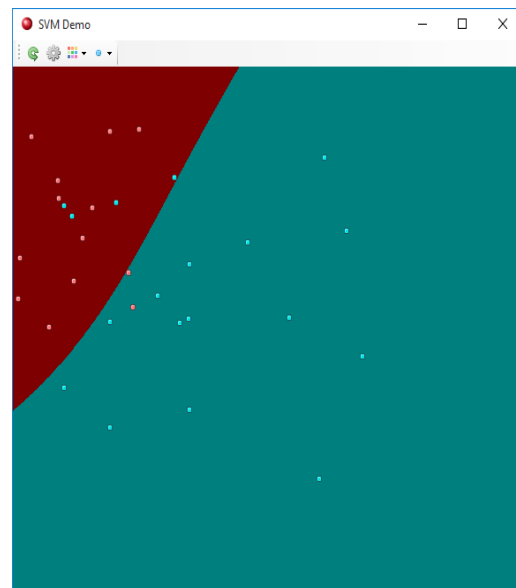


Figure 5: SVM Classification Result

Our proposed (semi supervised heterogeneous graph) SHG-Health calculation can be viewed as joining the benefits for taking care of a functional clinical issue of risk prediction from longitudinal health assessment data with heterogeneity and huge unlabeled data issues. To tackle the issue of health risk prediction dependent on health assessment records with heterogeneity and huge unlabeled data issues.

Figure 6: User input Form design

6. RESULT:

To begin with, health assessment records are spoken to as a chart that relates every single significant case together. This is particularly helpful for displaying unusual outcomes that are frequently inadequate. Second, multi-composed connections of data things can be caught and normally mapped into a heterogeneous chart. Especially, the health assessment things are spoken to as various kinds of hubs on a chart, which empowers our technique to abuse the fundamental heterogeneous sub diagram structures of Individual classes to accomplish better. Third, highlights can be weighted in their own sort through a name spread procedure on a heterogeneous diagram. These in-class weighted highlights at that point add to the successful order in an iterative union procedure. Another method for anticipating risks for members dependent on their yearly health assessments.

Figure 7: Health Prediction Report

IV. DISCUSSION

The outcomes are isolated into three sections. The first is the successive and noteworthy example revelation. The second is mapping the illness to its group and the third is prediction by giving risk score as yield. Toward the starting all the info data is put away in the non sickness bunch further it gets characterized and grouped by the model. A solitary client input data is taken care of into the framework and gets arranged by the huge example, to which it coordinates through choice tree, gets investigated for its risk score converged with both of the Non ailment and malady clusters.

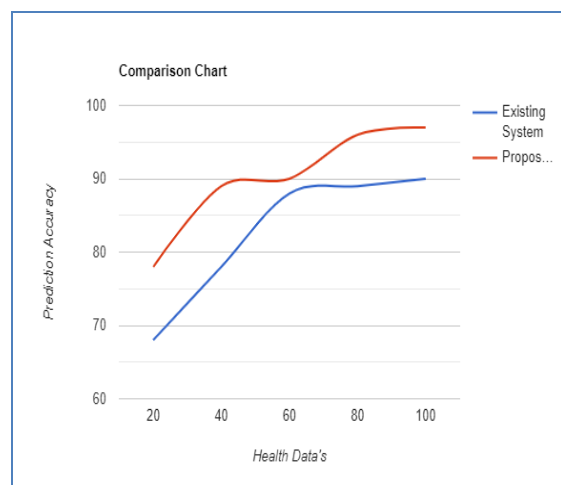


Figure 8: Accuracy comparison for existing and proposed system

In Figure 8 demonstrates the comparison of health prediction accuracy. In x axis denotes the data volumes and Y axis denotes the prediction accuracy.

V. CONCLUSION

Mining health assessment data is testing particularly because of its heterogeneity, natural clamor, and especially the enormous volume of unlabeled data. Right now, presented a viable and proficient chart based semi-managed calculation to be specific SHG-Health to address these difficulties. Our proposed diagram put together order approach with respect to mining health assessment records has a couple of noteworthy favorable circumstances. To start with, health assessment records are spoken to as a diagram that relates every single applicable case together. This is particularly helpful for displaying unusual outcomes that are regularly meager. Second, multi-composed connections of data things can be caught and normally mapped into a heterogeneous diagram. Especially, the health assessment things are spoken to as various sorts of hubs on a diagram, which empowers our technique to abuse the basic heterogeneous subgraph structures of individual classes to accomplish better. Third, highlights can be weighted in their own kind through a name proliferation process on a heterogeneous diagram. These in-class weighted highlights at that point add to the compelling characterization in an iterative combination process. Our work shows another method for anticipating risks for members dependent on their yearly health assessments. Our future work will concentrate on the data combination for the health assessment records to be coordinated with different sorts of datasets, for example, the clinic based electronic health records and the members' living conditions (e.g., abstains from food and general activities). By coordinating data from different accessible data sources, increasingly successful prediction might be accomplished.

REFERENCES

- [1] H. C. Koh and G. Tan, "Data Mining Application in Healthcare", *Journal of Healthcare Information Management*, vol. 19, no. 2, (2005).
- [2] D. Hand, H. Mannila and P. Smyth, "Principles of data mining", MIT, (2001).
- [3] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "The KDD process of extracting useful knowledge from volumes of data.commun.", *ACM*, vol. 39, no. 11, (1996), pp. 27-34.
- [4] J. Han and M. Kamber, "Data mining: concepts and techniques", 2nd ed. The Morgan Kaufmann Series, (2006).
- [5] M. Silver, T. Sakara, H. C. Su, C. Herman, S. B. Dolins and M. J. O'shea, "Case study: how to apply data mining techniques in a healthcare data warehouse", *Healthc. Inf. Manage*, vol. 15, no. 2, (2001), pp. 155-164.

- [6] P. R. Harper, "A review and comparison of classification algorithms for medical decision making", *Health Policy*, vol. 71, (2005), pp. 315-331.
- [7] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: current issues and guidelines", *Int. J. Med. Inform.*, vol. 77, (2008), pp. 81-97.
- [8] R. D. Canlas Jr., "Data Mining in Healthcare:Current Applications and Issues", (2009).
- [9] J. Soni, U. Ansari, D. Sharma and S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", (2011).