Prediction of Diabetes Mellitus Using Machine Learning



**Chapter 1**

**INTRODUCTION**

* 1. **Type-1 Diabetes**

Type 1 diabetes occurs when the body’s immune system attacks and destroys certain cells in the pancreas, an organ about the size of a hand that is located behind the lower part of the stomach. These cells - called beta cells - are contained, along with other types of cells, within small islands of endocrine cells called the pancreatic islets. Beta cells normally produce insulin, a hormone that helps the body move the glucose contained in food into cells throughout the body, which use it for energy. But when the beta cells are destroyed, no insulin can be produced, and the glucose stays in the blood instead, where it can cause serious damage to all the organ systems of the body. Type 1 diabetes is usually diagnosed in children and young adults, and was previously known as juvenile diabetes. Scientists do not yet know exactly what causes Type 1 diabetes, but they believe that autoimmune, genetic, and environmental factors are involved. Incidence of Type-1 diabetes in the US is estimated at about 30,000 cases annually and about 40 per 10,000 children. Type 1 diabetes accounts for 5% to 10% of all diagnosed cases of diabetes. According to the National Institute of Allergy and Infectious Disease (NIAID) the prevalence rate of Type 1 diabetes is approximately 1 in 800 or around 340,000 people in the US. The risk of developing type 1 diabetes is higher than virtually all other severe chronic diseases of childhood. Peak incidence occurs during puberty, around 10 to 12 years of age in girls, and 12 to 14 years of age in boys. According to the Juvenile Diabetes Research Foundation as many as 3 million Americans may have type-1 diabetes. Each year over 13,000 children are diagnosed with diabetes in the U.S. That’s 35 children each and every day. 1 As is evident from the above discussion, Type 1 diabetes is a serious medical concern not only in the United States but across the world. There has been quite a lot of research in the medical community on how to treat this disease. Our objective has been to analyze a Type 1 diabetes dataset from the Diabetes Prevention Trial -1 and study whether it is possible to predict the onset of Type-1 diabetes from medical test results using learned classifiers.

* 1. **Project Overview**

There has been a lot of study in the area of machine learning on data from the domain of diabetes, especially on the Pima Indian Diabetes dataset [25], [2], [11] from the University of California, at Irvine (UCI) repository. There has also been tremendous interest in using machine learning algorithms for post diagnosis care, like prediction of blood glucose levels to control the dosage of insulin [21] and the use of association rules to predict the occurrence of certain diseases in diabetic patients [28], [16]. However our study differs from both these approaches in the sense that we use a dataset that is not restricted to a particular ethnicity but to a specific type of diabetes, namely Type 1 diabetes in the juvenile population and our objective is not to monitor diabetic patients but learn a model to predict the occurrence of this type of diabetes by taking the patient’s past medical records into consideration. We primarily used two types of classifiers: C4.5 based decision trees [23] and Cascade Correlation [24] based neural networks, to predict diabetic cases from non diabetic ones by using subject test results. This is not known to be a difficult problem for Physicians, but the reader will see building a good predictive model was not trivial. Next, we used the same base classifiers to predict diabetes, but this time the attributes were the differences in test results, between consecutive tests of the same type for a subject. This approach has the promise of allowing a prediction that someone is susceptible to diabetes before any test results indicate they may have it. Both random forests of decision trees [2] and bagged classifiers [1] for both neural networks and decision trees were used. Surprisingly, it was necessary to explicitly encode missing attributes to achieve over 95% accuracy in diabetes prediction for both decision trees and neural networks. 2 The ensemble classifiers provided the best accuracy. Decision tree classifiers were comparable to cascade correlation neural network classifiers. Of interest was the fact that approximately 80% accuracy can be obtained in predicting diabetes from differences in test results over time without using data from the last time period before diagnosis as diabetic. Another aspect that was tested was whether oversampling of the minority class examples (i.e. the diabetic examples) would improve the prediction accuracy of the classifiers. The oversampling technique that was used was the Synthetic Minority Oversampling Technique (SMOTE) [5], which uses a nearest neighbor approach to generate synthetic examples. Although the oversampling technique did not improve the overall accuracy of prediction (using Bagging and Random Forests), the number of True Positives did show a significant increase, with comparably higher F-measure values..



Dept. of CSE, KIT, Tiptur Page 6

Prediction of Diabetes Mellitus Using Machine Learning



**Chapter 2**

**LITERATURE SURVEY**

Defusal Faruque and Asaduzzaman, Iqbal H.Sarker has discussed that diabetes is one of the most common disorder of the human body it is caused due the metabolic disorder .Hence that they used various and important ML algorithms that are Support Vector machine, NB,KNN and DT to predict the diabetes[1].

Sidong Wei,Xuejiao Zhao and Chunyan Miao presented that diabetes is commonly called as disorder in which glucose level in body is high. In this paper they use popular methods such as SVM and deep neural network for identify the disease and data processing. [2].

Lakshmi K.S and G.Santhosh Kumar according to them Hospital databases serve as wealthy information source for the fruitful medication diagnosis. IN this they used NLP tools along with combined with data mining algorithms for the extraction of rules [3].

Jian-xunChen , Shih-LiSu and Che-Ha Chang discussed about Ontology that generate a primary care planning to the medical professional’s for the accustoming. The result of the research paper shows the model can be provided personalize diabetes mellitus care planning efficiently [4].

MM Alotaib, RSH.Istepanian, and A.Sungoor they are present a clever based mobile polygenic disease control system & tutoring model for the patients with diabetes. In this, system is able to store the clinical information about the diabetes system, such an often blood sugar level and BP measured and hypo glycaemia event [5].

Berina Alic and Lejila Gurbea,Almir Badnjevic they presented the overview of techniques in machine learning in the diabetes classification and cardiovascular diseases using BNs and ANN [6]. M.Durgadevi and Dr.R.Kalpana In this paper they estimate that risks, So gigantic cat goring and detection algorithms have been develop in the domain of DM. So, that this paper aim is to compare the fruition and 5 classification way are anti-miner, Ad boost, RBF network, CN2 and Bagging for the diabetes prediction [7].

ElliotB.Sloane, Nilmini Wickramasingle and Steve Goldberg they presented Wireless diabetes monitoring which is a cloud-based diabetes, it’s a coaching platform for diabetes management and its a low cost, innovative, cloud-based diabetes support system [8].

Minyechil Alehegn and Rahul Joshi had present about the ML technology that help to identify a dataset at the elementary so that rescue the life.By implementing NB and K-nn algorithms.[9]. Umatejaswi and P.Suresh Kumar had discussed about algorithms such as SVM, NB, DT for identify the mellitus make use of technique like data mining [10].

Dept. of CSE, KIT, Tiptur Page 14

Prediction of Diabetes Mellitus Using Machine Learning



**Chapter 3**

**SYSTEM ARCHITECTURE**

Design is a meaningful engineering representation of something that is to be built. It is the most crucial phase in the developments of a system. Software design is a process through which the requirements are translated into a representation of software. Design is a place where design is fostered in software Engineering. Based on the user requirements and the detailed analysis of the existing system, the new system must be designed. This is the phase of system designing.

**3.1 Introduction to System Architecture**

System Architecture design-identifies the overall hypermedia structure for the Application. Architecture design is tied to the goals establish for a Application, the content to be presented, the users who will visit, and the navigation philosophy that has been established. Content architecture, focuses on the manner in which content objects and structured for presentation and navigation. Application architecture, addresses the manner in which the application is structure to manage user interaction, handle internal processing tasks, effect navigation, and present content. Application architecture is defined within the context of the development environment in which the application is to be implemented.

**3.2 System Requirements**

Software Requirement Specification (SRS) is a fundamental document, which forms the foundation of the software development process. SRS not only lists the requirements of a system but also has a description of its major features. These recommendations extend the IEEE standards. The recommendations would form the basis for providing clear visibility of the product to be developed serving as baseline for execution of a contract between client and the developer. SRS constitutes the agreement between clients and developers regarding the contents of the software product that is going to be developed. SRS should accurately and completely represent the system requirements as it makes a huge contribution to the overall project plan. The software being developed may be a part of the overall larger system or may be a complete standalone system in its own right.



Dept. of CSE, KIT, Tiptur Page 15

Prediction of Diabetes Mellitus Using Machine Learning



**3.2.1 Hardware Requirements**

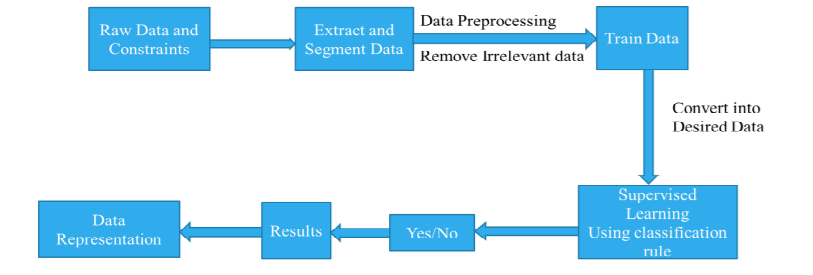
|  |  |  |
| --- | --- | --- |
| • | Processor | : Intel Core 2 Duo |
| • | RAM | :2GB |
| • | Hard Disk | :80GB |

**3.2.2 Software Requirements**

|  |  |  |
| --- | --- | --- |
| • | Operating System | : Windows (Any Version) |
| • | Programming Language | : Python |
| • | IDE | : Pycharm IDE |

**3.3 System Architecture**

System Architecture design-identifies the overall structure for the Liver disease as shown in below diagram.

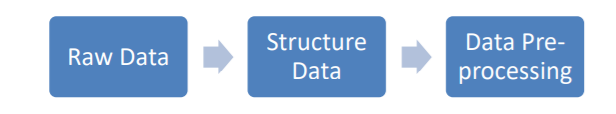


**Figure 3.1: System Architecture**

Figure 3.1 System Architecture for Diabetes disease ,Compares Naïve Bayes and SVM algorithms are done and it is based on the performance factors classification accuracy and execution time.

Data Constraints Data is a collection global dataset. IN this system use Pima Indian data set is used for training a model. Data set contain 21 parameters and around 1000 dataset. The dataset feature/parameters are: • Age • Gender • Relation • DOB • Sugar tested value • Symptoms • Family history etc. This are data is trained to the model for the prediction of diabetes. Train Dataset and Test Dataset The training data is a initial set of data which is used to understand the program. This is the one in which we have to train the model first because to set the feature and this data is available on system. This data is used to teach the machine for do different actions. It is the data in which model can learn with algorithm to teach the model and doing work automatic.

Testing data is the input given to a software. It shows the data affects when the execution of the module that specifying and this is basically used for testing. Pre-processing of data Data preprocessing is a process in which that is actual use for converting the basic data into the clean data set. It is the step in which the data transform or an encode to the state that the machine can be easily parse. The major task of data preprocessing in learning process is to remove the unwanted data and filling the missed value. So that it help to machine can be trained easily.



**Figure 3.2: Data Pre-Processing**

Feature Extraction

Feature Extraction is the method in which it used for alter the key data for features of outcomes. This, trait square is used to compute the characteristics of designs given that facilitate in different amid the class of key pattern details. This method involving to decrease the counts of resource required to describe the huge set of data. Feature extraction is an attribute reduction process. This is also used to increasing the speed and effectiveness of supervised learning.

ML Algorithm: KNN

The k-nearest neighbor’s is a ML algorithm is the non-parametric method proposed by Thomas Cover used for Regression and Classification. This algorithm is mainly used for the classification of problems in the industry. KNN algorithm is a type of instance-based learning method. This algorithm relies on the distance for objects classification, training data normalizing to the improve its accuracy dramatically. The neighbors are derived from the set of things for which classes or object property values are known. It can be thought of as a training set for the algorithm, although no explicit training steps are required.

Result

After taking that input data from the system will able to divine the statistics by appeal the ML algorithm & also provided the foremost output in the devise of different in between to detection the most accurate to treatment to diabetes millets.



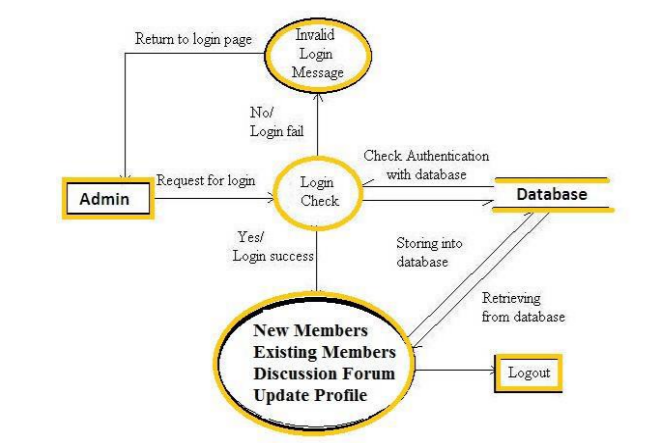
Dept. of CSE, KIT, Tiptur Page 16

Prediction of Diabetes Mellitus Using Machine Learning



**3.4 Data Flow Diagram**

Data flow diagram is a graphical portrayal of the flow of information through a data framework, displaying its procedure angles. A DFD is regularly utilized as a preparatory advance to make an outline of the framework without broadly expounding, which can later be explained. A DFD demonstrates what sort of data will be contribution to and yield from the framework, how the information will progress through the framework, and where the information will be put away. It doesn't indicate data about process timing or whether procedures will work in succession or in parallel, dissimilar to a customary organized flowchart which centers around control stream, or an UML movement work process outline, which presents both control and information streams as a brought together model.



**Figure 3.3: Data flow diagram**

Figure 3.3 is the data flow diagram the output of the results is feedback to the dataset the flow of data presented above.A DFD demonstrates what sort of data will be contribution to and yield from the framework

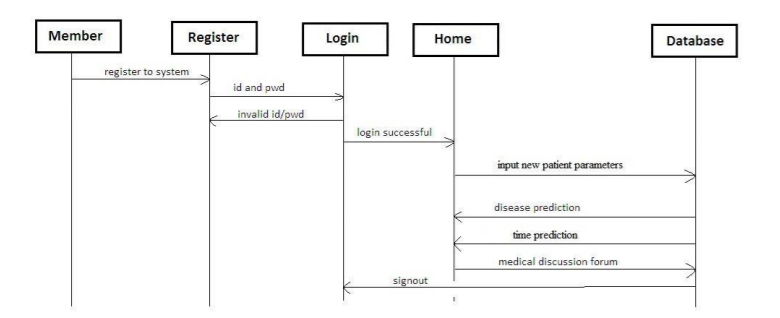


Dept. of CSE, KIT, Tiptur Page 17

Accuracy Prediction using Machine Learning Techniques for Indian Patient Liver Disease



**3.5 Sequence diagram**



**Figure 3.4: Sequence diagram**

In Figure 3.4 a sequence diagram demonstrates protest collaborations organized in time arrangement.

It delineates the articles and classes engaged with the situation and the succession of messages traded between the items expected to complete the usefulness of the situation. Arrangement charts are regularly connected with utilize case acknowledge in the Logical View of the framework a work in progress.

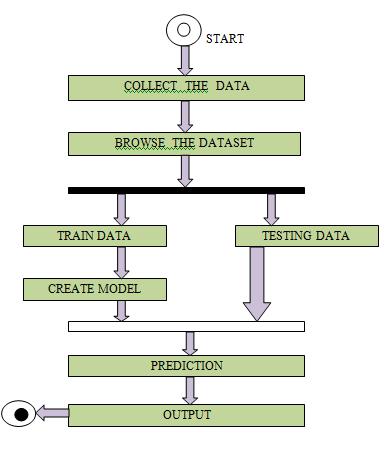


Dept. of CSE, KIT, Tiptur Page 19

Prediction of Diabetes Mellitus Using Machine Learning



**3.6 Activity diagram**



**Figure 3.5: Activity diagram**

In Figure 3.5 Activity diagram is essentially a flowchart to speak to the spill out of one action to another movement.

The action can be portrayed as an activity of the framework. The control stream is attracted starting with one task then onto the next. This stream can be consecutive, fanned, or simultaneous. Action graphs manage all sort of stream control by utilizing diverse components, for example, fork, join, and so on .The fundamental reasons for action charts is like other four graphs. It catches the dynamic conduct of the framework. Other four charts are utilized to demonstrate the message spill out of one protest another however action graph is utilized to indicate message spill out of one movement to another. Movement is a specific task of the framework. Movement graphs are not just utilized for envisioning the dynamic idea of a framework, yet they are additionally used to build the executable framework by utilizing forward and figuring out procedures. The main missing thing in the movement chart is the message part. It doesn't demonstrate any message spill out of one movement to another. Movement graph is some of the time considered as the flowchart. In spite of the fact that the graphs



Dept. of CSE, KIT, Tiptur Page 20

Prediction of Diabetes Mellitus Using Machine Learning



resemble a flowchart, they are definitely not. It indicates diverse streams, for example, parallel, fanned, simultaneous, and single. The reason for a activity diagram can be depicted as ,

* Draw the movement stream of a framework.
* Depict the succession starting with one action then onto the next.
* Depict the parallel, extended and simultaneous stream of the framework.

**Summary**

This chapter gives the description about the system architecture and system design. It gives description about the flow diagrams and use case diagrams,sequence diagrams and activity diagram.



Dept. of CSE, KIT, Tiptur Page 21

CONCLUSION

The prediction of diabetes is one the of great importance in today scenario, and concerning with its severe complications. Due to the biggest reason for the death in worldwide is diabetes. The System model is mainly focus to identification of diabetes using some of the parameters. System is useful to physicians to predict the diabetes in initial dais. So, that conventional treatments and solutions may be given to the patients. System used some of the techniques like ML for the prediction, so that to get the more precise results. There have been fortune of investigation on the diabetes imprint. Building diabetes disease prediction system is useful for hospitals and doctors. System predicts disease at early stages, so doctors can treat patients in a better way. Proposed model is the real time application in which is meant for multiple hospitals and predicts disease in less time. As we use machine learning algorithms for disease prediction, we will get more accurate and efficient results.



 Prediction of Diabetes Mellitus Using Machine Learning



**REFERENCES**

[1] “Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus” Md Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, IEEE 2019.

[2] “A Comprehensive Exploration to the Machine Learning Techniques for Diabetes Identification” Sidong Wei1, Xuejiao Zhao, Chunyan Miao Shanghai Jiao Tong University, China.

[3] “Association Rule Extraction from Medical Transcripts of Diabetic Patients” Lakshmi K S, G Santhosh Kumar, 2014.

[4] “Diabetes Care Decision Support System” 2nd International Conference on Industrial and Information Systems IEEE 2010.

[5] “An Intelligent Mobile Diabetes Management and Educational System for Saudi Arabia: System Architecture” M.M. Alotaibi, R.S.H. Istepanian, A.Sungoor and N. Philip, IEEE 2014.

[6] “Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases” by BerinaAlic, Lejla Gurbeta, IEEE 2017.

[7] “Performance Analysis of Classification Approaches for the Prediction of Type II Diabetes” by M. Durgadevi, M. Durgadevi, IEEE 2017.

[8] “Cloud-Based Diabetes Coaching Platform for Diabetes Management” Elliot B. Sloane Senior Member IEEE, Nilmini Wickramasinghe, Steve Goldberg 2016.

[9] Minyechil Alehegn and Rahul Joshi, “Analysis andprediction of diabetes diseases using machine learning algorithm”:International Research Journal of Engineering and Technology Volume: 04 Issue: 10 | Oct -2017

[10] P. Suresh Kumar and V. Umatejaswi, “Diagnosing Diabetes using Data Mining Techniques”,International Journal of Scientific and ResearchPublications, Volume 7, Issue 6, June 2017 705 ISSN 2250-3153.

[11] “Clustering Medical Data to Predict the Likelihood of Diseases” by Razan Paul, Abu Sayed Md. Latiful Hoque, IEEE 2010.

[12] “Robust Parameter Estimation in a Model for Glucose Kinetics in Type 1 Diabetes Subjects” Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA, Aug 30-Sept 3, 2006.

[13] Anjali C And Veena Vijayan V, Prediction and Diagnosis of Diabetes Mellitus, “A Machine Learning Approach” ,2015 IEEE in Intelligent Computational Systems (RAICS) | Trivandrum.

[14] Ridam Pal ,Dr. Jayanta Poray, and Mainak Sen, ,“Application of Machine Learning Algorithms on Diabetic Retinopathy”, 2017 2nd IEEE International Conference On Recent Trends In Electronics Information & Communication Technology, May 19-20, 2017, India.