IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MEDICAL DATA ANALYSIS AND PREDICTION

1.RAYAPUREDDY HEMA NAGA DURGA.

PG Scholar, Department of Computer Science,

SVKP & Dr K S Raju Arts & Science College,

Penugonda, W.G.Dt., A.P, India

2.K. LAKSHMANA REDDY.

Associate Professor in Computer Science,

SVKP & Dr K S Raju Arts & Science College,

Penugonda, W.G.Dt., A.P, India

ABSTRACT

Idea of visualizing the medical data by applying machine learning and pandas in python. Taking dataset from medical background of different people (prime Indians dataset from UCI repository). This data set consists of information of user who age, gender type of symptoms related to diabetes. Design a testing and training set and predict what are chances of patients having diabetes in coming five years. Data is classified and shown in the form of different graphs. The Pima Indians of Arizona and Mexico have the highest reported prevalence of diabetes of any population in the world. A small study has been conducted to analyse their medical records to assess if it is possible to predict the onset of diabetes based on diagnostic measures.

INTRODUCTION

1.1 Introduction

Data analysis is playing important part in analysing dataset and predicting what are situations in coming years. This analysis can give option for departments and organizations to take steps in dealing with these problems. In this project prediction of diabetes in coming years is considered as main problem. Diabetes is one of deadliest diseases in the world. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc. The normal identifying process is that patients need to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports. Moreover, every time they want to get their diagnosis report, they have to waste their money in vain.Diabetes Mellitus (DM)

is defined as a group of metabolic disorders mainly caused by abnormal insulin secretion and/or action. Insulin deficiency results in elevated blood glucose levels (hyperglycemia) and impaired metabolism of carbohydrates, fat and proteins. DM is one of the most common endocrine disorders, affecting more than 200 million people worldwide. The onset of diabetes is estimated to rise dramatically in the upcoming years. DM can be divided into several distinct types. However, there are two major clinical types, type 1 diabetes (T1D) and type 2 diabetes (T2D), according to the etiopathology of the disorder. T2D appears to be the most common form of diabetes (90% of all diabetic patients), mainly characterized by insulin resistance. The main causes of T2D include lifestyle, physical activity, dietary habits and heredity, whereas T1D is thought to be due to autoimmunological destruction of the Langerhans islets hosting pancreatic-β cells. T1D affects almost 10% of all diabetic patients worldwide, with 10% of them ultimately developing idiopathic diabetes. Other forms of DM, classified on the basis of insulin secretion profile and/or onset, include Gestational Diabetes, endocrinopathies, MODY (Maturity Onset Diabetes of the Young), neonatal, mitochondrial, and pregnancy diabetes. The symptoms of DM include polyuria, polydipsia, and significant weight loss among others. Diagnosis depends on blood glucose levels (fasting plasma glucose = 7.0 mmol/L.

2. LITERATURE SURVEY

[1]. Komi, Zhai. 2017. Application of Data Mining Methods in Diabetes Prediction

Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is help to make predictions on medical data. Diabetes mellitus or simply diabetes is a disease caused due to the increase level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. The methods strongly based on the data mining techniques can be effectively applied for high blood pressure risk prediction. In this paper, we explore the early prediction of diabetes via five different data mining methods including: GMM, SVM, Logistic regression, ELM, ANN. The experiment result proves that ANN (Artificial Neural Network) provides the highest accuracy than other techniques.

[2]. Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus, Omar Kassem Khalil Aissa Boudjella, 2016 Sixth International Conference on Developments in eSystems Engineering

Data mining approach helps to diagnose patient's diseases. Diabetes Mellitus is a chronic disease to affect various organs of the human body. Early prediction can save human life and can take control over the diseases. This paper explores the early prediction of diabetes using various data mining techniques. The dataset has taken 768 instances from PIMA Indian Dataset to determine

the accuracy of the data mining techniques in prediction. The analysis proves that Modified J48 Classifier provide the highest accuracy than other techniques.

[3]. Alan Siper, Roger Farley and Craig Lombardo, "Machine Learning and Data Mining Methods in Diabetes Research", Proceedings of Student/Faculty Research Day, CSIS, Pace University, May 6th, 2005.

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes mellitus or simply diabetes is a disease caused due to the increase level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is to help make predictions on medical data. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. This project aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project also aims to propose an effective technique for earlier detection of the diabetes disease.

[4]. Devi, M. Renuka, and J. Maria Shyla. "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus." International Journal of Applied Engineering Research 11.1 (2016): 727-730.

Diabetes mellitus or simply diabetes is a disease caused due to the increase level of blood glucose. Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is to help make predictions on medical data. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. This project aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression. This project also aims to

propose an effective technique for earlier detection of the diabetes disease.

3. OVERVIEW OF THE SYSTEM

3.1 Existing System

There were no chances of prediction in existing studies it was just by manual analysis based on existing data but analysing large amount of dataset is not considered. Data mining plays a huge role in predicting diabetes in the healthcare industry. There are many algorithms developed for prediction of diabetes. But most of the algorithms failed in case of the accuracy estimation. Also, there is a need to automate the overall process of diabetes prediction.

Disadvantages:

- But in these existing methods have to wait for long time in making a queue at time they enter the office. Many biometric systems are available but the key authentications are same is all the techniques.
- Every biometric system consists of enrolment process in which unique features of a person is stored in the database and then there are processes of identification and verification..

3.2 Proposed System:

Data analysis and machine learning libraries and algorithm are used for prediction on diabetes and information is shown in detail in the form of different types of graphs (histogram, density plots, box and whisker plots and correlation matrix plots).

Classification is one of the most important decision making techniques in many real world problem. In this work, the main objective is to classify the data as diabetic or non-diabetic and improve the classification accuracy. For many classification problem, the higher number of samples chosen but it doesn't leads to higher classification accuracy. In many cases, the performance of algorithm is high in the context of speed but the accuracy of data classification is low. The main objective of our model is to achieve high accuracy. Classification accuracy can be increase if we use much of the data set for training and few data sets for testing. This survey has analyzed various classification techniques for classification of diabetic and non-diabetic data..

Advantages:

- • In this way a lot of time is saved and this is highly secure process and effective way of user authentication is performed.
- Attendance is maintained on the excel sheet so anyone can access it for purposes like administration, employees themselves.

.

3.3 System Modules

User module:

User module can be any user who wants to analyses data by collecting data from repositories and then loading data set and view analyses data in graphical representation.

Pandas Module:

Pandas module is a calculation module which performs data analysis like view required dimensions of data , view data type of each attribute, calculate and display statistical summary and then perform class distribution , pair wise person correlation , skew for each attribute.

Graphical representation:

Analysis data is displayed in density plots/ whisker plots, matrix plots , histogram plots , correlation matrix by using matplotlib library.

. Histograms

A fast way to get an idea of the distribution of each attribute is to look at histograms.

Histograms group data into bins and provide you a count of the number of observations in each bin. From the shape of the bins you can quickly get a feeling for whether an attribute is Gaussian', skewed or even has an exponential distribution. It can also help you see possible outliers.

Density Plots

Density plots are another way of getting a quick idea of the distribution of each attribute. The plots look like an abstracted histogram with a smooth curve drawn through the top of each bin, much like your eye tried to do with the histograms.

Box and Whisker Plots

Another useful way to review the distribution of each attribute is to use Box and Whisker Plots or boxplots for short.

Boxplots summarize the distribution of each attribute, drawing a line for the median (middle value) and a box around the 25th and 75th percentiles (the middle 50% of the data). The whiskers give an idea of the spread of the data and dots outside of the whiskers show candidate outlier values (values that are 1.5 times greater than the size of spread of the middle 50% of the data).

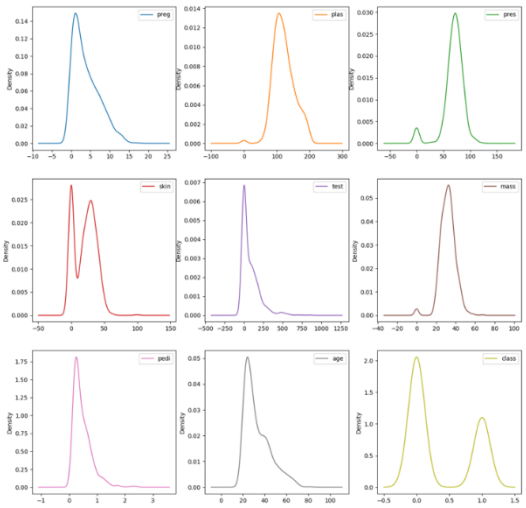
Correlation Matrix Plot

Correlation gives an indication of how related the changes are between two variables. If two variables change in the same direction they are positively correlated. If the change in opposite directions together (one goes up, one goes down), then they are negatively correlated.

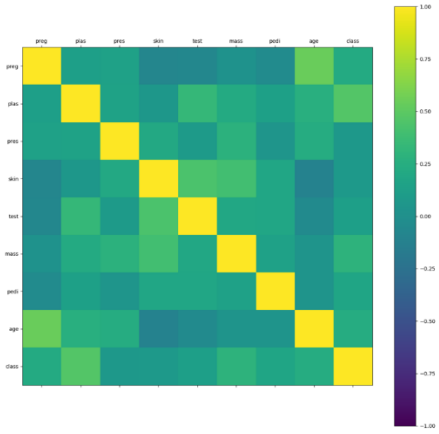
You can calculate the correlation between each pair of attributes. This is called a correlation matrix. You can then plot the correlation matrix and get an idea of which variables have a high correlation with each other.

This is useful to know, because some machine learning algorithms like linear and logistic regression can have poor performance if there are highly correlated input variables in your data.

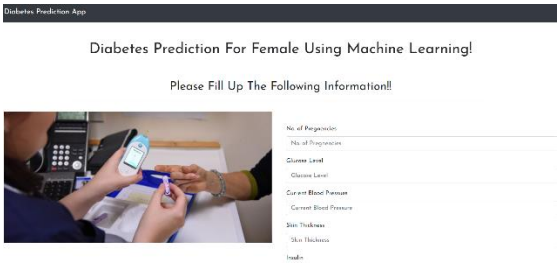
4. RESULTS



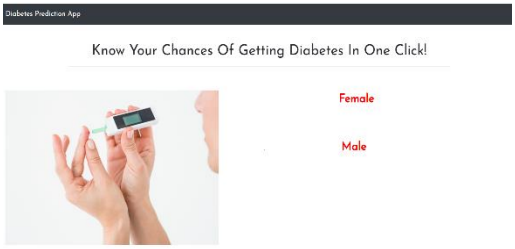
Factor analysis



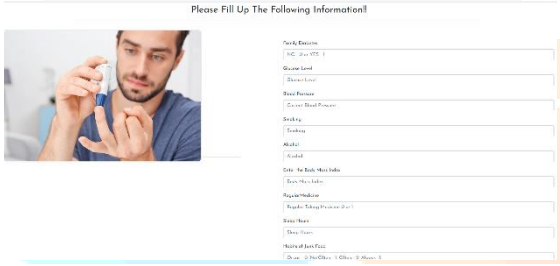
Correlation graph



Enter test values



Main Page



Male input

5. CONCLUSION

We presented a comprehensive, timely survey on visualization and visual analytics in deep learning research, using a human-cantered, interrogative framework. Our method helps researchers and practitioners in visual analytics and deep learning to quickly learn key aspects of this young and rapidly growing medical data of research, whose impact spans diabetics. Our survey goes beyond visualization-focused venues to extend a wide scope that also encompasses relevant works from top venues in AI, ML .We highlighted visual analytics as an integral component in addressing pressing issues in modern AI, helping to discover and communicate insight, from discerning model bias, understanding models, to promoting AI safety. We concluded by highlighting impactful research directions and open problems.

Future Enhancements:

Proposed system uses “KNN algorithm” to find the diabetes disease, in data science we have many algorithms for classification such as Naive Bayes, SVM, Decision Tree, ID3 etc… in future we can add more algorithms to find outputs and algorithms can be compared to find the efficient algorithm. We can add visitor query module, where visitors can post queries to administrator and admin can send reply to those queries. We can add treatment module, where doctors upload treatment details for patients and patient can view those treatment details..

REFERENCES

1. W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” The bulletin of mathematical biophysics, vol. 5, no. 4, 1943.
2. W. Rawat and Z. Wang, “Deep convolutional neural networks for image classification: A comprehensive review,” Neural computation, vol. 29, no. 9, 2017.
3. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in NIPS, 2012.
4. K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” arXiv:1312.6034, 2013.
5. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going

deeper with convolutions," in CVPR, 2015.

ABOUT AUTHORS:

RAYAPUREDDY HEMA NAGA DURGA is currently pursuing MCA in SVKP & Dr K S Raju Arts & Science College, affiliated, to Adikavi Nannaya University, Rajamahendravaram. His research interests include Data Structures Web Technologies, Operating Systems and Artificial Intelligent.

K. Lakshmana Reddy is working as an Associate Professor in the Department of Computer Science in SVKP & Dr K S Raju Arts &Science College, Penugonda, A.P. He received MCA from Andhra University, 'C' level from DOEACC, New Delhi and M.Tech from Acharya Nagarjuna University, A.P. He attended and presented papers in conferences and seminars. He has done online certifications in several courses from NPTEL. His areas of interests includes Computer Networks ,Network Security and Cryptography, Formal Languages and Automata Theory and Object Oriented programming languages.