ABSTRACT

- Nowadays, e-Health service has become a booming area, which refers to computer-based health care and information delivery to improve health service locally, regionally and worldwide. An effective disease risk prediction model by analyzing electronic health data benefits not only to care a patient but also to provide services through the corresponding data-driven e-Health systems. In this paper, we particularly focus on predicting and analyzing diabetes mellitus, an increasingly prevalent chronic disease that refers to a group of metabolic disorders characterized by a high blood sugar level over a prolonged period of time. K-Nearest Neighbor (KNN) is one of the most popular and simplest machine learning techniques to build such a disease risk prediction model utilizing relevant health data. In order to achieve our goal, we present an optimal K-Nearest Neighbor (OPT-KNN) learning based prediction model based on patient's habitual attributes in various dimensions. This approach determines the optimal number of neighbors with low error rate for providing better prediction outcome in the resultant model. The effectiveness of this machine learning e-Health model is examined by conducting experiments on the real-world diabetes mellitus data collected from medical hospitals.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER – 1

INTRODUCTION

1.1 OUTLINE OF THE PROJECT

Diabetes may be a set of metabolic problems known by high blood sugar levels over a protracted period of our time. Diabetes is outlined as a bunch of metabolic disorders in the main caused by abnormal insulin secretion and/or action. Symptoms of high aldohexose incorporate excessive voiding, continually feeling thirsty and enlarged hunger. If not treated on time, diabetes will cause serious health problems in a person like diabetic acidosis, hyperosmolar hyperglycemic state, or maybe result in death. This could result in time period complications as well as vas upset, brain stroke, failure, ulcers within the foot, and eye complications etc. Diabetes is caused once the duct gland within the body is unable to come up with insulin in enough amounts or once the cells and tissues within the body fail to utilize the insulin created. Diabetes exists in 3 forms: Diabetes Mellitus Type-1 is characterized by duct gland generating insulin but what's needed by the body, a condition conjointly referred to as "insulin-subordinate diabetes mellitus"(IDDM). Folks littered with type-1 DM need external insulin indefinite quantity to form up for the less insulin created by the duct gland. Diabetes Mellitus Type-2 is marked by the body resisting insulin because the body cells react otherwise to insulin than they traditional would. This could ultimately result in no insulin within the body. This can be otherwise referred to as "non-insulin subordinate diabetes mellitus"(NIDDM) or "adult beginning diabetes”. This sort of diabetes is often found in folks with high BMI or people who lead associate degree inactive manner. Gestational diabetes is that the third principle structure that's ascertained throughout physiological state. Generally, for a traditional person, aldohexose levels vary from seventy to ninety-nine milligrams per deciliter. An individual is taken into account diabetic providing the fast aldohexose level is foundtobeover126 mg/dL. Within the practice, an individual having an aldohexose concentration of a hundred to one hundred twenty-five mg/dL is taken into account aspre-diabetic.

1.2 MACHINE LEARNING

Machine learning could be a subfield of computer science (AI). The goal of machine learning typically is to know the structure information of knowledge of information and match that data into models which will be understood and used by folks. Although machine learning could be a field inside technology, it differs from ancient process approaches.

In ancient computing, algorithms are sets of expressly programmed directions employed by computers to calculate or downside solve. Machine learning algorithms instead give computers to coach on knowledge inputs and use applied math analysis so as to output values that fall inside a particular vary. thanks to this, machine learning facilitates computers in building models from sample knowledge so as to modify decision-making processes supported knowledge inputs.

1.3 SCOPE AND OBJECTIVE

The system should be useful in many e-commercial websites for maintaining the security and reliability of customers and people online

The system should be useful in preventing online frauds leading to leakage of important and private user data

The scope of using Machine Language over other Traditional Detecting Methods

Objectives:

Diabetes mellitus (DM) is one of the most prevalent chronic non-communicable diseases (NCD) around the world; about 90% of the patients who have diabetes suffer from Type 2 DM (T2DM)

The risk of developing T2DM is strongly associated with many predispositions, behavioral, and environmental risk factors and also genetic factors

About 90% of patients who have diabetes suffer from Type 2 DM (T2DM)

Many studies suggest using the significant role of lncrnas to improve the diagnosis of T2DM.

Machine learning is the techniques are tools that can improve the analysis and interpretation or extraction of knowledge from the data

CHAPTER – 2
LITERATURE SURVEY

2.1 RELATED WORK

prediction of Diabetes using Classification Algorithms

Diabetes isn't a hereditary disorder however heterogeneous group of disorder which could ultimately result in an boom of glucose within the blood and lack of glucose inside the urine. Diabetes is typically resulting from genetics, way of life and surroundings. Eating an dangerous weight loss plan, being overweight play role in developing the diabetes. High blood sugar tiers can also result in kidney diseases, coronary heart illnesses. The excess of sugar in the blood can harm the tiny blood vessels in your frame. Signs of diabetes are blurry imaginative and prescient , extreme hunger, unusual weight reduction, common urination and thirsty. In this paper, parameters used within the facts set to locate the diabetes are Glucose, Blood pressure, pores and skin thickness, Insulin, Age. Huge volumes of statistics units are generated by health care industries. Those facts sets is a collection of patient information about the diabetes from the hospitals. Big records analytics is the processing which it examines the information units and exhibits the hidden information. Pima Indians Diabetes Database (PIDD), this dataset is taken from the national Institute of Diabetes and Digestive diseases. The objective of the dataset is to predict whether or not the patient has diabetes or not, primarily based on diagnostic measurements in the dataset. Several constraints were taken from the massive database.

A Novel Technique To Predict Diabetic Disease Using Data Mining Classification Techniques

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. The software of data mining is an analytical tool for analyzing data. Data mining has become a main strategy in many industries to improve outputs and decrease costs. Now days in healthcare management this field will become very useful. Data mining techniques has became great potential for the healthcare industry to predict health deceases by using systematic data and analytics to identify inefficiencies and best practices that improve care and reduce costs. These techniques are fast in nature and take less time for the prediction system to improve the diabetic decease with more accuracy. In this paper we are applying the various classification techniques over diabetic mellitus decease dataset for the prediction of decease and non decease person. The diabetic database is preprocessed to make the mining process more efficient. The preprocessed data is used to predict using classification algorithms like Discriminent analysis, KNN, Naïve Bayes and Support vector machine. These classifiers can be efficiently used in bioinformatics problem. We are analyzing the various classification techniques like Discriminent analysis, KNN, Naïve Bayes and Support vector machine with linear and RBF kernel function and showing their accuracy.

Review on Prediction of Diabetes using Data Mining Technique

Diabetes mellitus is one of the world's major diseases. Millions of people are affected by the disease. The risk of diabetes is increasing day by day and is found mostly in women than men. The diagnosis of diabetes is a tedious process. So with improvement in science and technology it is made easy to predict the disease. The purpose is to diagnose whether the person is affected by diabetes or not using K Nearest Neighbor classification technique. The diabetes dataset is a taken as the training data and the details of the patient are taken as testing data. The training data are classified by using the KNN classifier and secondly the target data is predicted. KNN algorithm used here would be more efficient for both classification and prediction. The results are analyzed with different values for the parameter k.

A Prediction Technique in Data Mining for Diabetes Mellitus

Diabetes mellitus is a chronic disease characterized by hyperglycemia. It may cause many complications. According to the growing morbidity in recent years, in 2040, the world's diabetic patients will reach 642 million, which means that one of the ten adults in the future is suffering from diabetes. There is no doubt that this alarming figure needs great attention. With the rapid development of machine learning, machine learning has been applied to many aspects of medical health. In this study, we used decision tree, random forest and neural network to predict diabetes mellitus. The dataset is the hospital physical examination data in Luzhou, China. It contains 14 attributes. In this study, five-fold cross validation was used to examine the models. In order to verity the universal applicability of the methods, we chose some methods that have the better performance to conduct independent test experiments. We randomly selected 68994 healthy people and diabetic patients' data, respectively as training set. Due to the data unbalance, we randomly extracted 5 times data. And the result is the average of these five experiments. In this study, we used principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) to reduce the dimensionality. The results showed that prediction with random forest could reach the highest accuracy (ACC = 0.8084) when all the attributes were used.

CHAPTER 3

METHODOLOGY

3.1 EXISTING SYSTEM

The existing system was taking in order to meets the demands of this system and solve the problems of the existing system by implementing the naïve bayes classifier.

DISADVANTAGES OF EXISTING SYSTEM

- The system is not fully automated, it needs data from user for full diagnosis

3.2 PROPOSED SYSTEM

The proposed diabetes prediction system has two main stages that work together to achieve the desired results. The first stage of the proposed system is the data preparation, and the second one is the classification. However, the input into the system is the dataset and the output will be one class which represent the healthy or the diabetic. The implementation of this new system will help to reduce the stressful process, the result of the experiment shows that the proposed system has a better prediction in terms of accuracy. We have applied K- Nearest neighbor algorithm on the training and test sample data and obtained results for different values of K which is number of nearest neighbors.

ADVANTAGES OF PROPOSED SYSTEM

- User can diagnose their diabetes and get instant result.
- K- Nearest neighbor Algorithm is a fast, highly scalable algorithm.

3.3 ALGORITHMS USED

3.3.1 DECISION TREE CLASSIFIER

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 (remove)Quite possibly the most impressive and famous calculation. The choice tree calculation is beneath the control level of the calculation. It works in two ways, similar to yield. Input:

```
STEP 1: BEGIN
STEP 2: Input: D = {(x1, c1), . . . , (xN , cN )}.
STEP 3: x = (x1. . . xn) new instance to be classified
STEP 4 FOR each labelled instance (xi, ci) calculate d (xi, x)
STEP 5: Order d (xi , x) from lowest to highest, (i = 1. . . N)
STEP 6: Select the K nearest instances to x: Dkx.
STEP 7: Assign to x the most frequent class in Dkx
STEP 8: END
```