PREDICTION OF DIABETES

MELLITUS

CAUSES, SYMPTOMS AND TREATMENT OPTIONS

Team members

1. Podila Harsha Vardhini
2. Changantipati Pavani
3. Sangani Sathwik Varma

|  |  |  |
| --- | --- | --- |
| S NO. | Topic | Page no |
| 1 | Abstract | 2 |
| 2 | Introduction | 2 |
| 3 | Diabetes | 4 |
| 4 | Causes | 7 |
| 5 | Symptoms and complications | 8 |
| 6 | Treatment and Prevention | 8 |
| 7 | Existing system | 9 |
| 8 | Proposed system | 9 |
| 9 | Objectives of research | 10 |
| 10 | Problem statement | 10 |
| 11 | Industry profile | 10 |
| 12 | Review of literature | 10 |
| 13 | The data | 11 |
| 14 | Related works | 13 |
| 15 | Model and algorithm | 13 |
| 16 | Data mining toolkit | 14 |
| 17 | Data Set Description | 15 |
| 18 | Data Preprocessing | 16 |
| 19 | Logistic regression algorithm | 18 |
| 20 | Experimental result | 18 |
| 21 | Conclusion and future work | 19 |
| 22 | Reference | 20 |

ABSTRACT

Data Mining now a day’s plays an important role in prediction of diseases in health care industry. Data Mining is the process of selecting, exploring, and modelling large amounts of data to discover unknown patterns or relationships useful to the data analyst. The recent report of World Health Organization (WHO) shows a remarkable hike in the number of diabetic patients and this will be in the same pattern in the coming decades also. Early identification of diabetes mellitus is an important challenge. Data Mining has played an important role in diabetes research. Data mining would be a valuable asset for diabetes researches because it can unearth hidden knowledge from a huge amount of diabetes related data. Various data mining techniques help diabetes research and ultimately improve the quality of health care for diabetes patients. However raw medical data are available widely in distributed environment which are indiscriminate in nature and voluminous for ordinary processing. This paper provides a survey of data mining methods that have been commonly applied to diabetes data analysis and prediction of the disease

INTRODUCTION

Data Mining refers to extracting knowledge from large amounts of data. It enables us to explore the large patterns and analyse the same by means of statistical and Artificial Intelligence in large datasets. The data mining technique is used to predict possible future trends or to discover hidden pattern in the behaviour of the data. Techniques such as Artificial Neural Networks, Decision Trees, Classification, Clustering, Association rule algorithms are widely utilized by experts. Data mining techniques is widely being applied by researches in Bioinformatics. Bioinformatics is the science of storing, extracting, organizing, interpreting and utilizing information from biological sequences and molecules. Biological databases. The amount of biological data is growing rapidly. Analysing these data sets requires making sense of the data by inferring structure or generalizations from the data. The interaction between data mining and bioinformatics plays a vital role in diagnosing many diseases. The World Health organization predicts that by 2030 there will be approximately 350 million people worldwide affected by diabetes. Diabetes is a state or a condition in which the body is not able to produce or utilize the insulin properly which results in developing diseases affecting kidney, eyesight, nerve system, blood vessels and also cardiac related issues. There are three types of diabetes. Type-1 diabetes used to be called juvenile-onset diabetes. It is usually caused by an auto-immune reaction where the body defence system attacks the cells that produce insulin. Type-1 is also called insulin dependent. Almost 90% of the diabetes suffers from type-II diabetes which is alternatively called as non-insulin dependent diabetes or Adult-onset diabetes. This is characterized by resistance to insulin and deficiency of relative insulin which may either or both together be present at the time of which diabetes is diagnosed. The type-3 Gestational Diabetes Mellitus (GDM) which is often seen during pregnancy resulting from high glucose levels [8]. This is often accomplished by renal complications, cardiac diseases and peripheral vascular diseases. An early identification of patients with undiagnosed type-2 diabetes or those at an increased risk of developing type-2 diabetes is an important challenge in the field of medical. The availability of huge amounts of medical data leads to the need for powerful data analysis tools to extract useful knowledge. Researchers have long been concerned with applying statistical and data mining tools to improve data analysis on large data sets. Disease diagnosis such as diabetics is one of the applications where data mining tools are proving successful results in the recent years. Data mining process.

DIABETES

Any of a group of diseases characterized by high blood sugar levels  caused by insufficient production of insulin,impaired response to insulin, or  both, especially:

**a.**Type 1 diabetes.

**b.**Type 2 diabetes.

**c.**Gestational diabetes. In all subsenses also called *diabetes mellitus*.

A disease marked by abnormal levels of sugar in the blood, caused by the body's inability to produce or use insulinproperly. If untreated, it can cause circulatory problems and nerve damage. Diabetes may be treated with medication, insulin injections, and dietary restrictions.

Ancient Greek physicians gave the name *diabētēs* to a chronic disease characterized by excessiveurinationprobably what we now know as diabetes insipidus. (Later, the name was also used for a different disease, diabetes mellitus, in which increased urination is a common symptom.) The term is ultimately derived from the verb*diabainein,* "to stride or stand with the legs apart, step across, pass over," but it is not certain how *diabētēs* came todescribe the disease. *Diabētēs* has a variety of other meanings in Greek, including "compass" (since a compass canbe likened to a person striding with the legs spread wide) and "siphon" (perhaps because a siphon straddles so to speak two containers and permits the passage of liquid from one to the other). The first known use of *diabētēs* as adesignation for a disease is found in the works of Aretaeus of Cappadocia, who probably lived in the first century AD.Aretaeus's works became standard medical texts of the ancient and medieval world. One chapter of his work on theCauses and Signs of Chronic Diseases is devoted to a condition he calls diabētēs. Aretaeus, however, was not thefirst physician to give the condition this name, for he offers his own thoughts on the etymology of the term: "Thedisease seems to me to have acquired the name *diabētēs,* as if from the Greek word for siphon (*diabētēs*), becausethe fluid does not remain in the body." Some modern scholars, on the other hand, have suggested that as a medicalterm, *diabētēs* originally made reference to the straddling stance taken during urination by those with the disease the intended meaning may have been "one standing with the legs planted firmly apart." Whatever its origin, *diabētēs*became the standard name for the disease in Greek and medieval medical Latin. *Diabetes* is first attested in Englisharound 1425 in the spelling *diabete,* found in a Middle English translation of a Latin medical text by the French physician Guy de Chauliac (ca.1300-1368):

**Auicen forsoþ in diabete graunteþ water of whey of  shepis mylke. "In thecase of diabetes, Avicenna forsooth gives water of the whey of sheep's milk**."

**diabetes** - a polygenic disease characterized by abnormally high glucose levels in the blood; any of severalmetabolic disorders marked by excessive urination and persistent thirst

[**polydipsia**](https://www.thefreedictionary.com/polydipsia) - excessive thirst (as in cases of diabetes or kidney dysfunction)[**polygenicdisease**](https://www.thefreedictionary.com/polygenic+disease), [**polygenicdisorder**](https://www.thefreedictionary.com/polygenic+disorder) - an inherited disease controlled by several genes at once

[**polyuria**](https://www.thefreedictionary.com/polyuria) - renal disorder characterized by the production of large volumes of pale dilute urine; often associatedwith diabetes

[**diabetesmellitus**](https://www.thefreedictionary.com/diabetes+mellitus), [**DM**](https://www.thefreedictionary.com/DM) - diabetes caused by a relative or absolute deficiency of insulin and characterized bypolyuria; "when doctors say `diabetes' they usually mean `diabetes mellitus'"

[**diabetesinsipidus**](https://www.thefreedictionary.com/diabetes+insipidus) - a rare form of diabetes resulting from a deficiency of vasopressin (the pituitary hormonethat regulates the kidneys); characterized by the chronic excretion of large amounts of pale dilute urine whichresults in dehydration and extreme thirst

TYPE 1 DIABETES

Occurs when the pancreas cannot make insulin. Everyone with type 1 diabetes requires insulin injections. Most people are diagnosed with type 1 diabetes during their childhood or adolescent years. Type 1 diabetes occurs most commonly in people of northern European ancestry.

TYPE 2 DIABETES

Occurs when the pancreas does not make enough insulin or the body does not use insulin properly. It usually occurs in adults, although in some cases children may be affected. People with type 2 diabetes usually have a family history of this condition and 90% are overweight or obese. People with type 2 diabetes may eventually need insulin injections. This condition occurs most commonly in people of Indigenous and African descent, Hispanics, and Asians.

Another less common form is **gestational diabetes**, a temporary condition that occurs during pregnancy. Depending on risk factors, between 3% to 13% of Canadian women will develop gestational diabetes which can be harmful for the baby if not controlled. The problem usually clears up after delivery, but women who have had gestational diabetes have a higher risk of developing type 2 diabetes later in life.

Prediabetes is a term used to describe blood sugar levels that are higher than normal, but not high enough to be classified as diabetes. Many people with prediabetes go on to develop diabetes.

Causes

**Type 1 diabetes is an autoimmune disorder.** It's believed that a combination of genetic predisposition and additional environmental (as yet unidentified) factors provoke the immune system into attacking and killing the insulin-producing cells in the pancreas. There is no way to prevent type 1 diabetes from occurring.

**Type 2 diabetes is mainly caused by insulin resistance.** This means no matter how much or how little insulin is made, the body can't use it as well as it should. As a result, glucose can't be moved from the blood into cells. Over time, the excess sugar in the blood gradually poisons the pancreas causing it to make less insulin and making it even more difficult to keep blood glucose under control.

**Obesity is a leading cause of insulin resistance – about 90% of people with type 2 diabetes are overweight or obese.** Genetic factors are also likely to be involved in the cause of type 2 diabetes. A family history of the disease has been shown to increase the chances of getting it.

Other risk factors for the development of type 2 diabetes include:

* being 40 years of age or older
* being of Indigenous,Hispanic, South Asian,Asian, orAfrican descent
* blood vessel disease (e.g., damage to blood vessels in eyes, kidneys, nerves, heart, brain, or arms and legs)
* high blood pressure
* high cholesterol
* a history of gestational diabetes
* a history of prediabetes or impaired fasting glucose
* giving birth to a large baby
* certain medical conditions (e.g., HIV infection)
* mental health disorders (e.g., bipolar disorder, depression, schizophrenia)
* acanthosis nigricans (a condition causing darkened patches of skin)
* polycystic ovary syndrome
* obstructive sleep apnea
* use of certain medications (e.g., corticosteroids such as prednisone, certain antipsychotic medications, certain antiviral medications for HIV)

## Symptoms and Complications

**People with type 1 diabetes who are not being treated urinate frequently and feel excessively thirsty.** They usually feel very tired and experience severe weight loss despite normal or excessive food intake.

**The symptoms of type 2 diabetes usually appear more gradually.**People with type 2 diabetes who do not have their blood glucose under control often have a persistent, mild thirst. They urinate frequently, and often feel mild fatigue and complain of blurred vision. Many women with the disease have recurring vaginal yeast infections.

Diabetes is a major cause of heart disease, one of the leading causes of death in Canada. It's also the biggest cause of blindness and kidney failure in Canadian adults. Older adults with diabetes are twice as likely to develop high blood pressure as people without diabetes.

People with diabetes are 20 times more likely to undergo foot and other "lower extremity" amputations due to circulatory problems. Between 34% to 45% of men who have diabetes will experience erectile dysfunction at some point.

## Treatment and Prevention

**Currently, type 1 diabetes is not preventable. However, studies have shown that type 2 diabetes can be prevented by adopting lifestyle changes** that include moderate weight loss through eating a healthy diet and regularly exercising.

In addition, studies have shown that certain oral antidiabetes medications may play a role in preventing the development of type 2 diabetes for people with prediabetes.

**Diabetes is a chronic condition, and it can last an entire lifetime.** The goal of treating diabetes is to keep blood glucose levels as close to a normal range as possible. This prevents the symptoms of diabetes and the long-term complications of the condition. If you've been diagnosed with diabetes, your doctor – working with the members of your diabetes care team – will help you find your target blood glucose levels.

More than most conditions, treating diabetes requires a significant amount of real effort on the person's part. Coping with diabetes is a lifelong challenge, so people with diabetes should not be afraid to speak with a doctor or pharmacist if they feel overwhelmed.

**People with type 1 diabetes need insulin\* continuously to survive.**The only way to cure this disease is to have a pancreas or islet cell transplant, but these operations are only recommended in a small set of circumstances.

As with many conditions, treatment of type 2 diabetes begins with lifestyle changes, particularly in your diet and exercise. If you have type 2 diabetes, speak to your doctor and diabetes educator about an appropriate diet. You may be referred to a dietitian. It is also a good idea to speak with your doctor before beginning an exercise program that is more vigorous than walking to determine how much and what kind of exercise is appropriate.

If lifestyle changes don't put blood glucose levels in the target range, medications may be required. Medications for type 2 diabetes include anti diabetes pills or injections, insulin injections, or a combination of these.

Medications are very effective at treating diabetes and reducing the symptoms and long-term effects of the condition. However, you may experience *hypoglycemia* (a blood glucose level that is too low) when taking certain medications for diabetes.

EXISTING SYSTEM

In the existing system ,we need to search for the symptoms or else visit the doctor and the doctor based on symptoms he will give precautions and medicines. This is a lengthy and time taking process.

PROPOSED SYSTEM

In the proposed system the user registered in the application can just login and select the type of symptoms he is prevailing to be taken and the necessary occulted doctor details were given to the user.

OBJECTIVES OF RESEARCH

We undertook a systematic review of individual patient data (IPD) to identify the most highly prognostic factors for foot ulceration (ie symptoms, signs, diagnostic tests) in people with diabetes.

PROBLEM STATEMENT

Large number of people were affected by the disease called diabetes.

Our project is to predict the disease of the diabetes that the number of people were affected.

INDUSTRY PROFILE

* The project is helpful to the people who is suffering from the disease of diabetes.
* The project is also helpful to the people of medical department.

REVIEW OF LITERATURE

Diabetes a non-communicable disease is leading to long-term complications and serious health problems. A report from the World Health Organisation [30] addresses diabetes and its complications that impact on individual physically, financially, economically over the families. The survey says about 1.2 million deaths due to the uncontrolled stage of health lead to death. About 2.2 million deaths occurred due to the risk factors of diabetes like a cardiovascular and other diseases. Artificial intelligence is having more effect is machine realizing, which creates calculations ready to take in examples and choice standards from information. Machine learning calculations have been implanted into information mining pipelines, which can consolidate them with established measurable techniques, to remove learning from information. Inside the EU-financed MOSAIC undertaking, an information mining pipeline has been utilized to determine an arrangement of prescient models of sort 2 diabetes mellitus (T2DM) entanglements in light of electronic wellbeing record information of almost one thousand patients.

THE DATA

D[iabetes mellitus](https://www.sciencedirect.com/topics/medicine-and-dentistry/diabetes-mellitus) (DM) is a [chronic disease](https://www.sciencedirect.com/topics/medicine-and-dentistry/chronic-disease) that is characterized by high blood glucose. Nearly half of all diabetics have [household](https://www.sciencedirect.com/topics/medicine-and-dentistry/household) [heredity](https://www.sciencedirect.com/topics/medicine-and-dentistry/heredity) factors, which is one of the most [important features](https://www.sciencedirect.com/topics/computer-science/important-feature) of DM. Failure of the pancreas to produce enough insulin and the body's inefficient use insulin are both pathologic causes of DM. There are two types of DM. The [pathogenesis](https://www.sciencedirect.com/topics/medicine-and-dentistry/pathogenesis) of [type 1 diabetes mellitus](https://www.sciencedirect.com/topics/medicine-and-dentistry/insulin-dependent-diabetes-mellitus) (T1DM) is that the pancreas secretes damaged β-cells, preventing it from lowering blood glucose level in time. Insulin resistance and [insulin secretion](https://www.sciencedirect.com/topics/medicine-and-dentistry/insulin-release) deficiency are the pathogeneses of [type 2 diabetes mellitus](https://www.sciencedirect.com/topics/medicine-and-dentistry/maturity-onset-diabetes-of-the-young) (T2DM), which is also called non-insulin dependent DM.

In the past 30 years of [development](https://www.sciencedirect.com/topics/medicine-and-dentistry/development) in China, with rising number of diabetics, people have started to realize that this chronic disease has deeply impacted every family and everyone's daily [life](https://www.sciencedirect.com/topics/medicine-and-dentistry/life). There is an ascending trend in the proportion of diabetics in the general population, and the [growth rate](https://www.sciencedirect.com/topics/medicine-and-dentistry/growth-rate) of male diabetics is higher than that of female diabetics, as shown in [Fig. 1](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "fig1). According to some official statistics, the number of diabetics in China was nearly 110 million in 2017. This means that China has the largest diabetic population in the world.

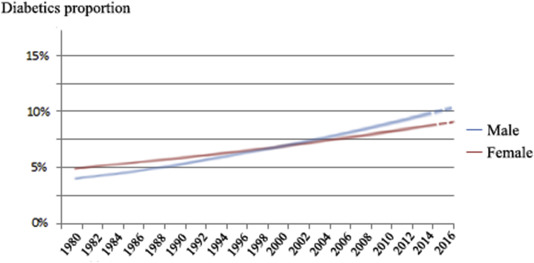


Fig. 1. Trend of [Diabetics](https://www.sciencedirect.com/topics/medicine-and-dentistry/diabetes-mellitus) proportion in China.

The International Diabetes Federation (IDF) presents the latest data on DM in the Diabetes Atlas (Seventh Edition) [[1](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "bib1)]. It shows that in 2015, the number of diabetics worldwide was close to 415 million. In terms of the population growth trend of diabetics, it predicts that the number will approach to 642 million, or one in ten adults.

In order to lower the morbidity and reduce the influence of DM, it is vital for us to focus on a high-risk group of people with DM. According to the latest World Health Organization (WHO) standard, the definitions of groups with a high risk of DM are as follows:

● [Age](https://www.sciencedirect.com/topics/medicine-and-dentistry/age) ≥ 45 and seldom exercising

● BMI ≥ 24 kg/m2

● Impaired glucose tolerance (IGT) or impaired fasting glucose (IFG)

● Family history of DM

● Lower high-density lipoprotein cholesterol or hypertriglyceridemia (HTG)

● Hypertension or cardiovascular and [cerebrovascular disease](https://www.sciencedirect.com/topics/medicine-and-dentistry/cerebrovascular-disease) ● Gestation female whose age ≥30

In order to research the high-risk group of DM, we need to utilize advanced [information technology](https://www.sciencedirect.com/topics/computer-science/information-technology). Therefore, [data mining technology](https://www.sciencedirect.com/topics/computer-science/data-mining-techniques) is an appropriate study field for us. [Data mining](https://www.sciencedirect.com/topics/computer-science/data-mining), also known as [Knowledge Discovery in Databases](https://www.sciencedirect.com/topics/computer-science/knowledge-discovery-in-database) (KDD), is defined as the [computational process](https://www.sciencedirect.com/topics/computer-science/computational-process) of discovering patterns in [large datasets](https://www.sciencedirect.com/topics/computer-science/large-datasets) involving methods at the intersection of artificial [intelligence](https://www.sciencedirect.com/topics/medicine-and-dentistry/intelligence), machine [learning](https://www.sciencedirect.com/topics/medicine-and-dentistry/learning), statistics, and [database systems](https://www.sciencedirect.com/topics/computer-science/database-systems) [[2](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "bib2)]. The main purposes of these methods are pattern recognition, prediction, association, and [clustering](https://www.sciencedirect.com/topics/medicine-and-dentistry/clustering). Data mining contains a series of steps disposed automatically or semi-automatically in order to extract and discover interesting, unknown, hidden features from large quantities of data. The high quality of data and the properly applied method are two significant aspects of data mining.

Data mining has been successfully applied to various fields in human society, such as weather prognosis, market [analysis, engineering](https://www.sciencedirect.com/topics/computer-science/engineering-analysis) [diagnosis](https://www.sciencedirect.com/topics/computer-science/diagnosis), and [customer relationship management](https://www.sciencedirect.com/topics/computer-science/customer-relationship-management). However, the application in disease prediction and medical data analysis still has room for improvement. For example, every hospital possesses a plethora of patient's basic and [medical information](https://www.sciencedirect.com/topics/computer-science/medical-information), and it is essential to revise, supplement, and extract meaningful knowledge from these data to support clinical analysis and diagnosis [[3](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "bib3),[4](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "bib4)]. It is reasonable to believe that there are various valuable patterns and waiting for [researchers](https://www.sciencedirect.com/topics/medicine-and-dentistry/researcher) to explore them.

As we all know, the number of diabetics is large, and it is continuously increasing. Additionally, most people know little about their health quality. Therefore, believe it is necessary to establish a model that can classify patients into either suspected patients or confirmed patients in 5 years from the first [examination](https://www.sciencedirect.com/topics/medicine-and-dentistry/examination) time for the high-risk DM group. In particular, we have focused on [T2DM](https://www.sciencedirect.com/topics/medicine-and-dentistry/type-2-diabetes).

Section [2](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "sec2) presents the related work of data mining in the group of diabetics and potential patients. Section [3](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "sec3) details the experimental tools, dataset, and [prediction model](https://www.sciencedirect.com/topics/computer-science/prediction-model). Section [4](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "sec4) describes the results of the experiment. Section [5](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "sec5) discusses the results and the procedures of validation. Section [6](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "sec6) concludes the paper with some directions for future work.

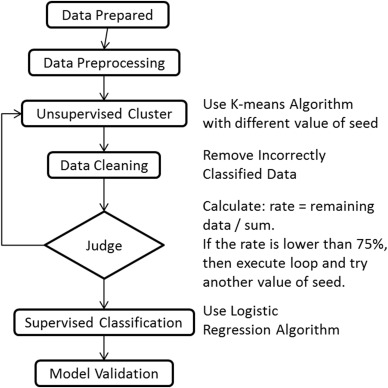
Related works

In recent years, using the [data mining technique](https://www.sciencedirect.com/topics/computer-science/data-mining-techniques) has been used with increasing frequency to predict the possibility of disease. Many [algorithms](https://www.sciencedirect.com/topics/computer-science/algorithms) and toolkits have been created and studied by [researchers](https://www.sciencedirect.com/topics/medicine-and-dentistry/researcher). These have highlighted the tremendous potential of this [research field](https://www.sciencedirect.com/topics/computer-science/field-research). In this section, a few important works that are closely related to the proposed issue are presented.

Based on several studies, we found that a commonly used dataset was the Pima Indians [Diabetes](https://www.sciencedirect.com/topics/medicine-and-dentistry/diabetes-mellitus) Dataset from the University of California, Irvine (UCI) Machine [Learning](https://www.sciencedirect.com/topics/medicine-and-dentistry/learning) [Database](https://www.sciencedirect.com/topics/computer-science/database). Patil proposed a [hybrid](https://www.sciencedirect.com/topics/medicine-and-dentistry/hybrid) [prediction model](https://www.sciencedirect.com/topics/computer-science/prediction-model) (HPM), which used a K-means [clustering](https://www.sciencedirect.com/topics/medicine-and-dentistry/clustering) algorithm aimed at validating a chosen class label of given data and used the C4.5 algorithm aimed at building the final [classifier model](https://www.sciencedirect.com/topics/computer-science/classifier-model), with 92.38% [classification accuracy](https://www.sciencedirect.com/topics/computer-science/classification-accuracy). Ahmad compared the [prediction accuracy](https://www.sciencedirect.com/topics/computer-science/prediction-accuracy) of multilayer perception (MLP) in [neural networks](https://www.sciencedirect.com/topics/computer-science/neural-networks) against the ID3 and J48 algorithms. The results showed that a pruned J48 tree performed with higher accuracy, which was 89.3% compared to 81.9%. Marcano-Cedeño proposed artificial [metaplasticity](https://www.sciencedirect.com/topics/medicine-and-dentistry/metaplasticity) on [multilayer perceptron](https://www.sciencedirect.com/topics/computer-science/multilayer-perceptron) (AMMLP) as a prediction model for diabetes, for which the best result obtained was 89.93%. All the studies presented above used the same Pima Indians Diabetes Dataset as the experimental material. The Waikato Environment for Knowledge Analysis (WEKA) toolkit was the primary tool which most researchers chose.

## MODEL AND ALGORITHM

This section is comprised of the dataset description, the [preprocessing](https://www.sciencedirect.com/topics/computer-science/preprocessing) procedure, and the [classification algorithm](https://www.sciencedirect.com/topics/computer-science/classification-algorithm). All the experimental processes have been completed using the WEKA toolkit. The proposed model is shown in.



[Algorithm](https://www.sciencedirect.com/topics/computer-science/algorithms) model.

### **Data mining toolkit**

WEKA is a free and non-commercial toolkit. It consists of standard machine [learning](https://www.sciencedirect.com/topics/medicine-and-dentistry/learning) and [data mining algorithms](https://www.sciencedirect.com/topics/computer-science/data-mining-algorithm), which are based on the JAVA environment. Using these preprocessing, classifying, [clustering](https://www.sciencedirect.com/topics/medicine-and-dentistry/clustering), associating [algorithms](https://www.sciencedirect.com/topics/computer-science/algorithms), and the visual interface, we were able to obtain useful knowledge from [databases](https://www.sciencedirect.com/topics/computer-science/database) easily and conveniently. Parts of those algorithms have been selected to establish the [prediction model](https://www.sciencedirect.com/topics/computer-science/prediction-model) for [T2DM](https://www.sciencedirect.com/topics/medicine-and-dentistry/type-2-diabetes).

In recent years, utilizing [data mining](https://www.sciencedirect.com/topics/computer-science/data-mining) algorithms in medical [predictive analysis](https://www.sciencedirect.com/topics/computer-science/predictive-analysis) has increased due to earnest research in related areas. Over the last few years, several [researchers](https://www.sciencedirect.com/topics/medicine-and-dentistry/researcher) have posited that it is possible to acquire clinically assistive supports and [predictive models](https://www.sciencedirect.com/topics/computer-science/predictive-model) from basic patient data [[[21]](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "bib21), [[22]](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "bib22), [[23]](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "bib23)]. Most papers published in the field of disease predictive analysis for DM are aimed at improving accuracy. Some researchers have obtained considerable results by using this WEKA toolkit and the Pima Indian [Diabetes](https://www.sciencedirect.com/topics/medicine-and-dentistry/diabetes-mellitus) dataset. However, the accuracy has room for improvement.

Extensive research has also been done on Pima Indian diabetes disease [diagnosis](https://www.sciencedirect.com/topics/computer-science/diagnosis), and the results obtained are presented in [Table 1](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "tbl1) [[24](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "bib24)]. We used the preprocessing method which introduced in Section [3.3](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "sec3.3) to deal with the original dataset and then simulated the same experiments as other researchers'. After that, we updated the data in [Table 1](https://www.sciencedirect.com/science/article/pii/S2352914817301405#tbl1). Most of the values of accuracy increased.

Table 1. The values of accuracy of classification made on Pima Indian [Diabetes](https://www.sciencedirect.com/topics/medicine-and-dentistry/diabetes-mellitus) Dataset.

| **Method** | **Accuracy (%)** |
| --- | --- |
| Discrim | 77.5 |
| MLP | 73.8 |
| Log disc | 78.2 |
| SMART | 76.8 |
| Bayes Net | 74.7 |
| Naïve Bay | 74.9 |
| Random Forest | 76 |
| J48 | 76.7 |
| SGD | 76.6 |
| SMO | 77 |
| Backprop | 75.2 |
| RBF | 75.7 |
| LMT | 76.6 |

### **Dataset description**

The Pima Indian Diabetes Dataset consists of information on 768 patients (268 tested positive instances and 500 tested negative instances) coming from a population near Phoenix, Arizona, USA. Tested positive and tested negative indicates whether the patient is diabetic or not, respectively. Each instance is comprised of 8 attributes, which are all numeric. These data contain personal health data as well as results from [medical examinations](https://www.sciencedirect.com/topics/medicine-and-dentistry/medical-examination). The detailed attributes in the dataset are listed as follows, and [Table 2](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "tbl2) shows some samples extracted from the dataset.

● Number of pregnancies

● Plasma glucose concentration at 2 h in an [oral glucose tolerance test](https://www.sciencedirect.com/topics/medicine-and-dentistry/oral-glucose-tolerance-test)

(PLAS)

● Diastolic blood pressure (pre)

● Triceps skin fold thickness (skin)

● 2-h serum insulin (insulin)

● Body mass index (BMI)

● Diabetes pedigree function (pedi)

● Age (age)

● Class variable (class)

Samples of dataset.

| **preg** | **plas** | **pres** | **skin** | **insu** | **bmi** | **pedi** | **age** | **class** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | Tested negative |
| 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | Tested positive |
| 3 | 78 | 50 | 32 | 88 | 31 | 0.248 | 26 | Tested positive |
| 2 | 197 | 70 | 45 | 543 | 30.5 | 0.158 | 53 | Tested positive |
| 1 | 189 | 60 | 23 | 846 | 30.1 | 0.398 | 59 | Tested positive |
| 5 | 166 | 72 | 19 | 175 | 25.8 | 0.587 | 51 | Tested positive |
| 0 | 118 | 84 | 47 | 230 | 45.8 | 0.551 | 31 | Tested positive |
| 1 | 103 | 30 | 38 | 83 | 43.3 | 0.183 | 33 | Tested negative |
| 1 | 115 | 70 | 30 | 96 | 34.6 | 0.529 | 32 | Tested positive |

### **Data Pre-processing**

The quality of the data, to a large extent, [affects](https://www.sciencedirect.com/topics/medicine-and-dentistry/affect) the result of prediction. This means that [data pre-processing](https://www.sciencedirect.com/topics/computer-science/data-preprocessing) [plays](https://www.sciencedirect.com/topics/medicine-and-dentistry/play) an important [role](https://www.sciencedirect.com/topics/medicine-and-dentistry/role-playing) in the model. The WEKA toolkit contains many kinds of filters for pre-processing purposes. In this study, we have selected some appropriate methods to optimize the original dataset.

### First, we have analysed each attribute's medical implication and its [correlation](https://www.sciencedirect.com/topics/computer-science/correlation) to DM. We determined that the number of [pregnancies](https://www.sciencedirect.com/topics/medicine-and-dentistry/pregnancy) has little connection with DM. Therefore, we transformed this [numeric attribute](https://www.sciencedirect.com/topics/computer-science/numeric-attribute) into a [nominal attribute](https://www.sciencedirect.com/topics/computer-science/nominal-attribute). The value 0 indicates non-pregnant and 1 indicates pregnant. The complexity of the dataset was reduced by this process.

Second, there are some missing and [incorrect values](https://www.sciencedirect.com/topics/computer-science/incorrect-value) in the dataset due to errors or deregulation. Most of the inaccurate experimental results were caused by these meaningless values. For example, in the original dataset, the values of [diastolic blood pressure](https://www.sciencedirect.com/topics/medicine-and-dentistry/diastolic-blood-pressure) and [body mass index](https://www.sciencedirect.com/topics/medicine-and-dentistry/body-mass-index) could not be 0, which indicates that the real value was missing. To reduce the influence of meaningless values, we used the means from the training data to replace all missing values.

After the above steps were applied, the unsupervised normalize filter for attribute was used to normalize all the data into the section [0, 1] by using [(1)](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "fd1), where x' is the mean or average value for the variable and s is the standard deviation for the variable. Value is the new normalized value. This avoids the complexity of calculation and accelerates the speed of the [operation](https://www.sciencedirect.com/topics/medicine-and-dentistry/surgery).

Value=value−x's

### **Data classification**

The model consists of double-level algorithms. In the first level, we used the improved K-means algorithm to remove incorrectly [clustered data](https://www.sciencedirect.com/topics/computer-science/clustered-data). The optimized dataset was used as input for next level. Then, we used the [logistic regression](https://www.sciencedirect.com/topics/medicine-and-dentistry/logistic-regression-analysis) algorithm to classify the remaining data.

## Logistic regression algorithm

The classification algorithm aimed to establish a model that can map data items to a given category, based on the existing data. It was used to extract significant data items from the model or to predict the tendency of data. In most cases, the dependent variable of the logistic regression algorithm is [binary-classification](https://www.sciencedirect.com/topics/computer-science/binary-classification). It means that the logistic regression algorithm is always used to solve two-category problem. The main purpose of our experiment is to predict whether one person is diabetic or not, which is a typical binary-classification problem. Besides, the logistic regression algorithm is always used in data mining, disease automatic diagnosis and economic prediction, especially predicting and classifying of medical and health problem. In conclusion, we decided to use the logistic regression as one part of our proposed model. The logistic regression algorithm is based on the [linear regression](https://www.sciencedirect.com/topics/medicine-and-dentistry/linear-regression-analysis) model expressed as.

* P=α+β1x1+β2x2+…+βmxm

The classification problem is very much like the linear regression problem. Linear regression problem can only predict a continuous value. It maintains consistent sensitivity throughout the [real number](https://www.sciencedirect.com/topics/computer-science/real-number) field. The predictive value of the classification problem can only be 0 or 1, so we may set a [critical point](https://www.sciencedirect.com/topics/computer-science/critical-point). The output is 1 if the value is greater than the threshold, otherwise the output is 0. The output variable range of logistic regression is always between 0 and 1. Logistic regression is a regression model that reduces the prediction range and limits the prediction value to [0, 1]. Based on linear regression, the logistic regression adds a layer of [sigmoid](https://www.sciencedirect.com/topics/medicine-and-dentistry/sigmoid) function (non-linearity). The features are first summed linearly and then predicted using the sigmoid function. The main formulas of the logistic regression algorithm are

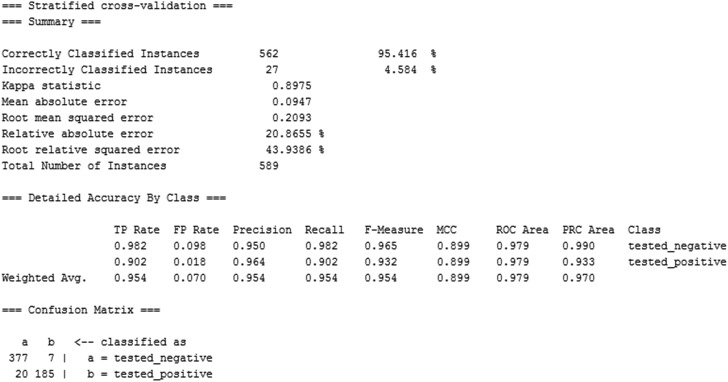
* PrY=+1|X~β·XandPrY=−1|X=1−PrY=+1|X
* ↓σx:=11+e−x∈0,1thesigmoidfunction
* PrY=+1|X~σβ·XandPrY=−1|X=1−PrY=+1|X

In this study, we have two categories, i.e., the positive group and the negative group. The Y indicates that the patient is diabetic. X independent variables represent the 8 attributes in the original dataset. Every dependent variable X is assigned a [coefficient value](https://www.sciencedirect.com/topics/computer-science/coefficient-value) called β representing the weight. After being analyzed by the logistic regression algorithm, the dataset showed every variable's value of weight. Different weights represent diverse correlation between X and Y. Once the regression model has been settled, it is efficient to input new data and predict whether the outcome is positive or negative. We set the logistic regression algorithm as the final step. The output and result are discussed in the next chapter.

## Experimental result

## Using the WEKA toolkit, it was convenient for us to study the result of the experiment through a visualized interface. We analysed and evaluated our model based on the following aspects.

The result is…



The result of the experiment.

**Conclusion and future work**

This paper aimed to establish an appropriate [prediction model](https://www.sciencedirect.com/topics/computer-science/prediction-model) for the high-risk [T2DM](https://www.sciencedirect.com/topics/medicine-and-dentistry/type-2-diabetes) group. Based on a number of researchers' experiences, we proposed a novel model, which consists of double-level algorithms, i.e., the improved K-means and [logistic regression](https://www.sciencedirect.com/topics/medicine-and-dentistry/logistic-regression-analysis) algorithms. In order to make a valid comparison with others' results, it was necessary to conduct this model using the WEKA toolkit and use the same Pima Indian [Diabetes](https://www.sciencedirect.com/topics/medicine-and-dentistry/diabetes-mellitus) Dataset. Proper filters were utilized to improve the validity and rationality of the dataset. The proposed model that consisted of both cluster and class method ensured the enhancement of [prediction accuracy](https://www.sciencedirect.com/topics/computer-science/prediction-accuracy). In Section [4](https://www.sciencedirect.com/science/article/pii/S2352914817301405#sec4), another realistic dataset provided by Dr.Schorling was used to test and verify the model. Our proposed model has proven to be appropriate for predicting T2DM. One of our proposed model's benefits is that it avoids deleting overmuch original data. It ensures the high quality of experimental data. The other benefit is that our model can apply in the Pima Indian Diabetes Dataset as well as other various datasets. While the limitation is that it consumes more time during the part of [preprocessing](https://www.sciencedirect.com/topics/computer-science/preprocessing).

We described that some papers focus on improving K-means by optimizing the initialized procedure of cluster centre in Section [2](https://www.sciencedirect.com/science/article/pii/S2352914817301405#sec2). But our improved model is based on the purpose of predicting DM2 and matches up with the logistic regression algorithm. It assures less time consuming and maximum retention of original data. Although the improved model is not so complicated, it attained well effect according to plenty of experiments.

The main problems we solved are improving accuracy of prediction model and making the model to adapt to different datasets. In this paper, we conclude that our proposed model showing higher prediction accuracy than other researchers' experimental results. And the improved K-means algorithm we proposed contributed a lot to the prediction model. Moreover, there are two more dataset applied in our proposed model and all of them obtained well effect.

For future work, it is necessary to bring in hospital's real and latest patients' data for continuous training and optimization of our proposed model. The quantity of the dataset should be large enough for training and predicting [[31](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "bib31),[32](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "bib32)]. Some advanced algorithms and models should be applied in the study of DM. Grading forecasting standards are also necessary for [potential diabetes](https://www.sciencedirect.com/topics/medicine-and-dentistry/impaired-glucose-tolerance) patients. Developing a series of rules and standards is a valid method to prevent people from developing DM. Based on that, a more effective model for predicting DM and grading potential patients is presented. This will help to lower the growth rate of diabetes and eventually decrease the risk of developing DM.

It is more convenient and efficient for people to obtain an application about [health management](https://www.sciencedirect.com/topics/medicine-and-dentistry/health-care-management) of DM on their mobile devices [[[33]](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "bib33), [[34]](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "bib34), [[35]](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "bib35), [[36]](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "bib36), [[37]](https://www.sciencedirect.com/science/article/pii/S2352914817301405" \l "bib37)]. We are currently developing an application that will provide reasonable and rational health suggestions to the high-risk group. Diabetes patients can conveniently use this application to test their blood glucose level, blood pressure, and heart rate. Furthermore, this medical data will be saved in a database for further procedures about data visualizing and model optimization. This will not only help people understand their health conditions, but will also help them create a [healthy lifestyle](https://www.sciencedirect.com/topics/medicine-and-dentistry/healthy-lifestyle).

## References

[[1]](https://www.sciencedirect.com/science/article/pii/S2352914817301405#bbib1) International diabetes federation (IDF) diabetes atlas (seventh ed.) (2015)

[Google Scholar](https://scholar.google.com/scholar?q=%22International%20diabetes%20federation%20(IDF)%20diabetes%20atlas%22)

[[2]](https://www.sciencedirect.com/science/article/pii/S2352914817301405#bbib2) <http://en.wikipedia.org/wiki/Data_mining#cite_note-acm-1>.

[Google Scholar](https://scholar.google.com/scholar?q=%22http://en.wikipedia.org/wiki/Data_mining#cite_note-acm-1.")

[[3]](https://www.sciencedirect.com/science/article/pii/S2352914817301405#bbib3) B. Riccardo, Z. Blaz **Predictive data mining in clinical medicine: current issues and guidelines**

Int J Med Inf, 77 (2008), pp. 81-97

[Google Scholar](https://scholar.google.com/scholar_lookup?title=Predictive%20data%20mining%20in%20clinical%20medicine%3A%20current%20issues%20and%20guidelines&publication_year=2008&author=B.%20Riccardo&author=Z.%20Blaz)

[[4]](https://www.sciencedirect.com/science/article/pii/S2352914817301405#bbib4) Mechelle Gittens, Reco King, Curtis Gittens and Adrian Als, Post-diagnosis Management of Diabetes through a Mobile Health Consultation Application, 2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom).

[Google Scholar](https://scholar.google.com/scholar?q=%22Mechelle%20Gittens,%20Reco%20King,%20Curtis%20Gittens%20and%20Adrian%20Als,%20Post-diagnosis%20Management%20of%20Diabetes%20through%20a%20Mobile%20Health%20Consultation%20Application,%202014%20IEEE%2016th%20International%20Conference%20on%20e-Health%20Networking,%20Applications%20and%20Services%20(Healthcom).%22)

[[5]](https://www.sciencedirect.com/science/article/pii/S2352914817301405#bbib5) <http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes>.

[Google Scholar](https://scholar.google.com/scholar?q=%22http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes.%22)

[[6]](https://www.sciencedirect.com/science/article/pii/S2352914817301405#bbib6) B.M. Patil **Hybrid prediction model for Type-2 diabetic patients**

Expert Syst Appl, 37 (2010), pp. 8102-8108

[Article](https://www.sciencedirect.com/science/article/pii/S0957417410004896) [Download PDF](https://www.sciencedirect.com/science/article/pii/S0957417410004896/pdfft?md5=814bc0afb1e5fe23738e39949d80bd20&pid=1-s2.0-S0957417410004896-main.pdf) [View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-77957831716&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Hybrid%20prediction%20model%20for%20Type-2%20diabetic%20patients&publication_year=2010&author=B.M.%20Patil)

[[7]](https://www.sciencedirect.com/science/article/pii/S2352914817301405#bbib7) Aliza Ahmad, Aida MustaphaH **Comparison between neural networks against decision tree In improving prediction accuracy for diabetes mellitus. ICDIPC 2011, Part I**

CCIS 188 (2011), pp. 537-545

[CrossRef](https://doi.org/10.1007/978-3-642-22389-1_47)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-79960297525&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Comparison%20between%20neural%20networks%20against%20decision%20tree%20in%20improving%20prediction%20accuracy%20for%20diabetes%20mellitus.%20ICDIPC%202011%2C%20Part%20I&publication_year=2011&author=Aliza%20Ahmad&author=Aida%20MustaphaH)

[[8]](https://www.sciencedirect.com/science/article/pii/S2352914817301405#bbib8) Alexis Marcano-Cedeño, Joaquín Torres, Diego Andina **A prediction model to diabetes using artificial metaplasticity. IWINAC 2011, Part II**

LNCS 6687 (2011), pp. 418-425

Cross Revi[ew Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-79956299375&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=A%C2%A0prediction%20model%20to%20diabetes%20using%20artificial%20metaplasticity.%20IWINAC%202011%2C%20Part%20II&publication_year=2011&author=Alexis%20Marcano-Cede%C3%B1o&author=Joaqu%C3%ADn%20Torres&author=Diego%20Andina)

[[9]](https://www.sciencedirect.com/science/article/pii/S2352914817301405#bbib9) Veena Vijayan V. and Anjali C., Decision support systems for predicting diabetes mellitus –a review. Proceedings of 2015 global conference on communication technologies (GCCT 2015).

[Google Scholar](https://scholar.google.com/scholar?q=%22Veena%20Vijayan%20V.%20and%20Anjali%20C.,%20Decision%20support%20systems%20for%20predicting%20diabetes%20mellitus%20–a%20review.%20Proceedings%20of%202015%20global%20conference%20on%20communication%20technologies%20(GCCT%202015).%22)

[[10]](https://www.sciencedirect.com/science/article/pii/S2352914817301405#bbib10) Zhe Wei, Guangjian Ye, Nengcai Wang **Analysis for risk factors of type 2 diabetes mellitus based on FP-growth algorithm**

China Med Equip, 13 (5) (2016), pp. 45-48

[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85043519109&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Analysis%20for%20risk%20factors%20of%20type%202%20diabetes%20mellitus%20based%20on%20FP-growth%20algorithm&publication_year=2016&author=Zhe%20Wei&author=Guangjian%20Ye&author=Nengcai%20Wang)

[[11]](https://www.sciencedirect.com/science/article/pii/S2352914817301405#bbib11) Yirui Guo **Application of artificial neural network to predict individual risk of type 2 diabetes mellitus**

J Zhengzhou Univ, 49 (3) (2014), pp. 180-183

[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85043508196&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Application%20of%20artificial%20neural%20network%20to%20predict%20individual%20risk%20of%20type%202%20diabetes%20mellitus&publication_year=2014&author=Yirui%20Guo)

[12] Shuaishuai Li, Enke Zhang, Min Li, Wei Pan **Research on the effectiveness of application of diabetes management APP**