ISSN: 2508-7894 © 2020 KODISA & KAIA.
KJAI website: http://www.kjai.or.kr
doi: http://dx.doi.org/10.24225/kjai.2020.vol8.no2.7

# A Study on Methods to Prevent Pima Indians Diabetes using SVM

**Sanghyuck YOU[1], Minsoo KANG[2]**

**Received: January 15, 2020. Revised: November 29, 2020. Accepted: December 05, 2020**.

## Abstract

In this paper, a study was conducted to find main factors to Pima Indians Diabetes based on machine learning. Diabetes is a type of metabolic disease such as insufficient secretion of insulin or inability to function normally and is characterized by a high blood glucose concentration. According to a situation report from WHO(World Health Organization), Diabetes is a chronic, metabolic disease characterized by elevated levels of blood glucose (or blood sugar), which leads over time to serious damage to the heart, blood vessels, eyes, kidneys and nerves. And also about 422 million people worldwide have diabetes, the majority living in low-and middle-income countries, and 1.6 million deaths are directly attributed to diabetes each year. Both the number of cases and the prevalence of diabetes have been steadily increasing over the past few decades. Therefore, in this study, we used Support Vector Machine (SVM), Decision Tree, and correlation analysis to discover three important factors that predict Pima Indians diabetes with 70% accuracy. Applying the results suggested in this paper, doctors can quickly diagnose potential Pima Indians diabetics and prevent Pima Indians diabetes.

**Keywords:** Pima Indians Diabetes, SVM, Correlation Analysis

**Major Classifications:** Artificial Intelligence, Supervised Learning, Support Vector Machine

## 1. Introduction

Diabetes is a type of metabolic disease such as insufficient secretion of insulin or inability to function normally and is characterized by a high blood glucose concentration. Diabetes is classified into type 1 and type 2, and type 1 diabetes was previously called 'pediatric diabetes'. It is a disease caused by the inability to produce insulin at all Type 2 diabetes, which is relatively insufficient in insulin, is characterized by insulin resistance (the ability of cells to burn glucose effectively due to poor insulin-lowering function). Arizona's Pima Indians have the highest incidence of type 2 diabetes in the world. Diabetes mellitus increases the risk of developing kidney disease, blindness, nerve damage, and blood vessel damage. It contributes to heart disease, so the diagnosis of diabetes is a very important classification problem. Diabetes of all types can lead to complications in many parts of the body and can increase the overall risk of dying prematurely. Possible complications include kidney failure, leg amputation, vision loss and nerve damage. Adults with diabetes also have two- to three-fold increased risk of heart attacks and strokes. In pregnancy, poorly controlled diabetes increases the risk of fetal death and other complications.

These symptoms are seen in millions of cases around the world. Nearly 3% of global blindness can be attributed to diabetic retinopathy, which occurs as a result of long-term accumulated damage to the blood vessels in the retina. Diabetes is also among the leading causes of kidney failure. Reduced blood flow and nerve damage in the feet caused by diabetes can lead to foot ulcers, and the associated

[1] First Author, Student, Eulji University, Korea. Email: gnb333@naver.com
[2] Corresponding Author, Professor, Eulji University, Korea. Email: mskang@eulji.ac.kr

infections and complications can lead to the need for limb amputation, as well as severe and life-long health problems. Type 1 diabetes cannot currently be prevented. Effective approaches are available to prevent type 2 diabetes and to prevent the complications and premature death that can result from all types of diabetes. These include policies and practices across whole populations and within specific settings (school, home, workplace) that contribute to good health for everyone, regardless of whether they have diabetes, such as exercising regularly, eating healthily, avoiding smoking, and controlling blood pressure and lipids.

The starting point for living well with diabetes is an early diagnosis – the longer a person lives with undiagnosed and untreated diabetes, the worse their health outcomes are likely to be. Easy access to basic diagnostics, such as blood glucose testing, should therefore be available in primary health care settings. Patients will need periodic specialist assessment or treatment for complications. A series of cost-effective interventions can improve patient outcomes, regardless of what type of diabetes they may have. These interventions include blood glucose control, through a combination of diet, physical activity and, if necessary, medication; control of blood pressure and lipids to reduce cardiovascular risk and other complications; and regular screening for damage to the eyes, kidneys and feet, to facilitate early treatment. So this paper going to study the factors (Pregnancies, Glucose, Blood pressure, Skin thickness, Insulin, Bmi, Diabetes pedigree function, age) most affecting the occurrence of diabetes and prevention.

For these reasons, it is necessary to study for finding major factors to prevent Pima Indians Diabetes. There are *9* attributes that contain a demographic feature, pregnancies, glucose, blood pressure, skin thickness, insulin, bmi, diabetes pedigree function and outcom. First, correlation analysis was conducted to determine which attributes profoundly affect Pima Indian Diabetes, and then, outcome feature was predicted using SVM(Support Vector Machine) through variables that have a potent effect.

## 2. Supervised Learning

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples. In supervised learning, each example is a pair consisting of an input object(typically a vector) and a desired output value(also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances.

### 2.1. SVM (Support Vector Machine)

In machine learning, Support-Vector Machine is supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, and an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the separate categories that are divided by a clear fap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall. The reason for using a two-class support vector machine is that technically it can be used in both classification and forecasting problems, and the second is less likely to be overfitted than neural network techniques, and thirdly, it is more accurate to predict, and lastly for its simplicity. For its experimental usage, the performance of two-class logistic regression and two-class neural networks were lower than the two-class support vector machine.

## 3. Related research

Diabetes is an endocrine disorder disease in which glucose absorbed in the body cannot be used because the secretion of insulin is abnormal or decreased, and it is accumulated in the blood and discharged into the urine(Ozougwu, Obimba, Belonwu, & Unakalamba, 2013). Unlike type 1 diabetes, which is caused by absolute deficiency of insulin due to autoimmune destruction of beta cells, type 2 diabetes increases insulin demand due to insulin resistance of peripheral tissues and it occurs due to a defect in the insulin secretion function due to the increase in insulin secretion due to the compensation action(Scheen, 2003).

As the incidence rate of type 2 diabetes was confirmed to be high in the same family, efforts to find the genetic cause began by comparing phenotypic variations through family tree analysis from the 1980s(Bodansky & Kelly, 1987). One large-scale leading study found that if one of the parents had type 2 diabetes, the risk increased by 40%, and the risk increased by 70% when both parents had diabetes(Valdez, Yoon, Liu, & Khoury, 2007). In a study that analyzed the family line of twins, the concordance rate of type 2 diabetes incidence was 70% for identical twins, and 25% for fraternal twins, suggesting the importance of genetic factors in the occurrence of diabetes(Kaprio, Koskenvuo, & Rose, 1990).

## 4. Experiment

### 4.1. Data Preprocessing

Pima Indians Diabetes data utilized in this study were collected from the open source site, Kaggle specifically 769 rows and 9 columns, including demographic data, various symptoms, and outcomes. Table 1 illustrates the data collected. And the sample shows 769 women under the age of 21 from the Pima Indian tribe.

**Table 1.** Pima Indians Diabetes DATA

| Data Group | Example |
|---|---|
| Demographic data | Age |
| Various Symptoms | Pregnancies, Glucose, Blood Pressure, Skin Thickness,<br>Insulin, BMI, Diabetes Pedigree Function |
| Result | Outcome |

Pre-processing of data collected was executed using correlation analysis. Figure 2 illustrates the pre-processing results of the data.

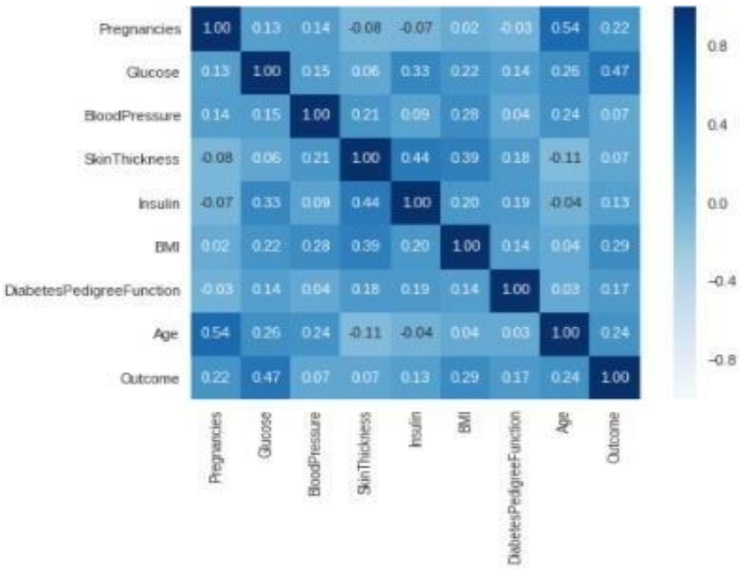
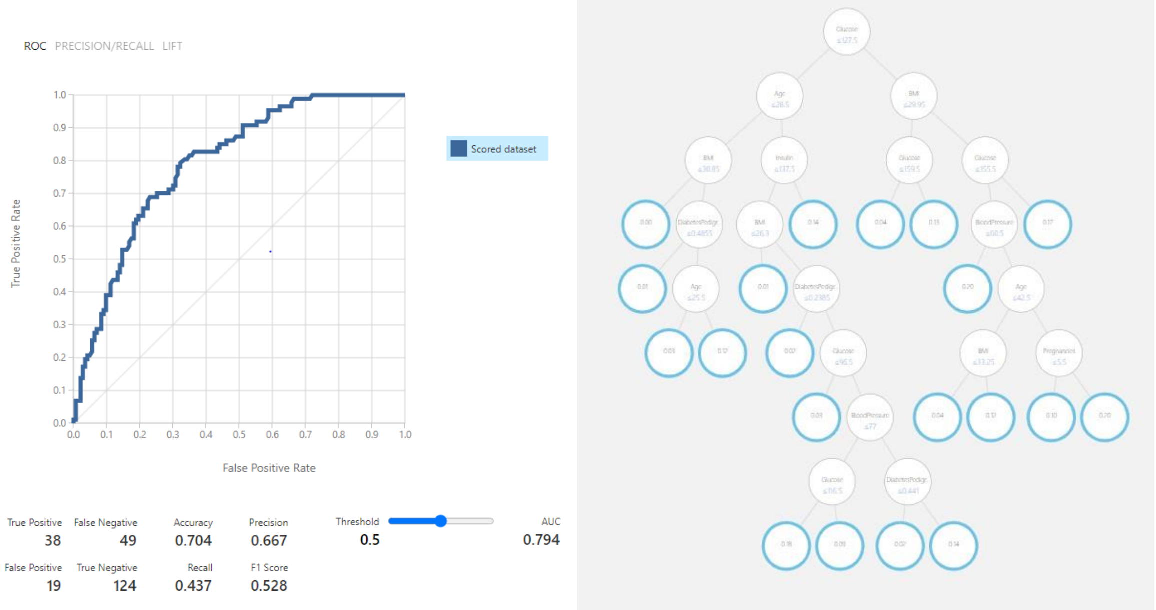


**Figure 2.** Correlation Analysis Results Using Pima Indians Diabetes Data

Through the correlation result analysis result Glucose, BMI, Age attributes have 0.47, 0.29, 0.24 relation to the outcome. Analyzing the values from correlation analysis, Glucose, bMI, and age have the greatest influence on diabetes in the correlation between variables. So, 1 experiment was executed using these three attributes applying SVM and Decision tree to predict Pima Indians Diabetes.

## 5. Results

The experiment with Glucose, BMI, Age feature using SVM resulting in 0.704 accuracy, 0.667 precision, 0.437 recall, and finally 0.528 F1 score. Figure 3 visualizes how evaluate model came out. And with Decision Tree the first step to diagnose Pima Indian Patients was with Glucose then BMI and lastly Age.



True Positive 38
False Negative 49
Accuracy 0.704
Precision 0.667
Threshold 0.5
AUC 0.794
False Positive 19
True Negative 124
Recall 0.437
F1 Score 0.528

**Figure 3.** Evaluate Model

## 5. Conclusion

In this paper, 1 experiment was made to predict Pima Indian Diabetes patients using two – class support vector machine and two-class boosted decision tree. The result showed means that when classifying Pima Indian Diabetes patients, only checking the potential patient's glucose, BMI, age is more efficient than doing all medical check outs, which takes time. In particular, active screening and early detection in high-risk areas of diabetes are the top priorities in preventing diabetes, as diabetes, especially in young adults, can prolong the transfer period of chronic diabetes complications, thereby increasing the mortality rate, disease rate, and health care costs and burdens. So, applying the result proposed in this paper, doctors could be able to reduce time when diagnosing Pima Indian Diabetes patients and diagnose other potential Pima Indian Diabetes patients more. The conclusion is that this study will help accelerate the diagnosis of diabetes in women in the Pima Indian tribe and further prevent diabetes in the early stages.

## References

Bodansky, H., & Kelly, W. (1987) Familial diabetes mellitus with variable B cell reserve: Analysis of a pedigree. *Diabetologia, 30*(8), 638-640.

Cortes, Corinna, V. & Vladimir N. (1995). Support-vector networks. *Machine Learning. 20*(3), 273–297. CiteSeerX 10.1.1.15.9362. doi*:*10.1007/BF00994018.

Kaggle. (2020). Parkinson's Disease Dataset. Retrieved July 27, 2020 from https://www.kaggle.com/salihacur/diabetes

Lee, W. S. (2014). A Comparative Study on Multiple Logistic Regression Model Using Pima Indian Diabetes Data

Naver Terms. (2020). Definition of Diabetes. Retrieved December 01, 2020 from https://terms.naver.com/entry.nhn?docId=926835&cid=51007&categoryId=51007

Ozougwu, J. C., Obimba, K. C., Belonwu, C. D., & Unakalamba, C. B. (2013). The pathogenesis and pathophysiology of type 1 and type 2 diabetes mellitus. *Journal of Physiology and Pathophysiology, 4*(4), 46-57. doi:10.5897/jpap2013.0001

Scheen, A. J. (2003). Pathophysiology of type 2 diabetes. *Acta Clinica Belgica, 58*(6), 335-341.

Valdez, R., Yoon, P. W., Liu, T., & Khoury, M. J. (2007). Family history and prevalence of diabetes in the US population: The 6-year results from the National Health and Nutrition Examination Survey (1999–2004). *Diabetes Care, 30*(10), 2517-2522. doi:10.2337/dc07-0720

Kaprio, J., Koskenvuo, M., & Rose, R. (1990). Population-based twin registries: Illustrative applications in genetic epidemiology and behavioral genetics from the Finnish Twin Cohort Study. *Acta Geneticae Medicae et Gemellologiae: Twin Research, 39*(4), 427-439. doi:10.1017/s0001566000003652

Wikipedia. (2020). Retrieved December 01, 2020 from https://en.wikipedia.org/wiki/Support-vector_machine