**Predictive Data Mining in Clinical Medicine-Diabetes Detection**

**Anchal Gupta Garima Singh Tushara Duvvuri**

**SWENG 545 - Data Mining (2018)**

**Abstract**

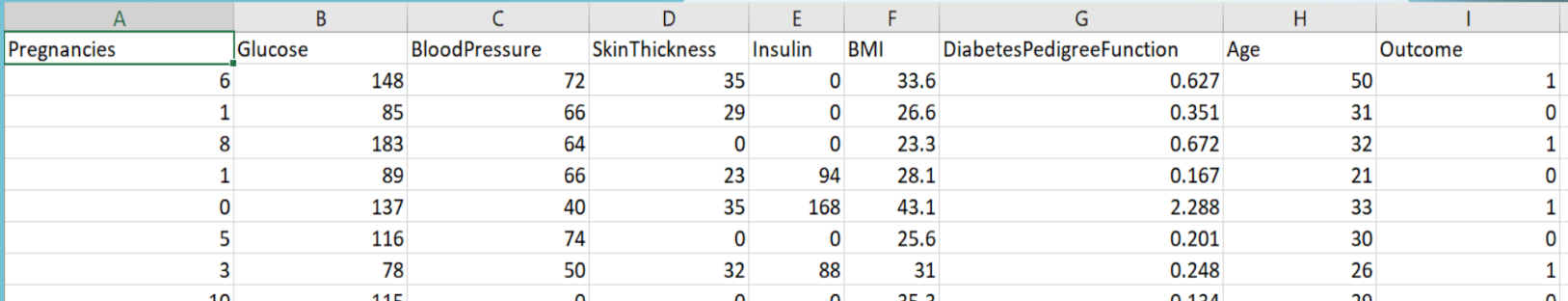
Type II Diabetes Mellitus is one of the silent killer diseases worldwide. According to the World Health Organization, 346 million people are suffering from diabetes worldwide. In this project, we have proposed a novel model based on data mining techniques for predicting type 2 diabetes mellitus. We have used the Pima Indians Diabetes Dataset, which is an open source dataset. We have compared the classification accuracies of non-processed and pre-processed data. The results clearly show that the pre-processed data gives better classification accuracy.

**Introduction**

Diabetes Mellitus which is, is a group of metabolic disorders in which there are high blood sugar levels over a prolonged period has become a common health problem nowadays, which would affect people and lead to various disablements like cardio vascular disease, visual impairments, leg amputation and renal failure if diagnosis is not done in the right time. High blood sugar is either when the pancreas does not produce insulin at all, or when the body does not respond correctly to the insulin produced by the pancreas and hence the glucose that is consumed by the person is locked inside the blood instead of entering into the cells of the body. The first case is referred to as Type-1 diabetes, while the second case is referred to as type-2 diabetes. In order to research the high-risk group of Diabetes, we need to utilize advanced information technology. Therefore, data mining technology is an appropriate study field for us.

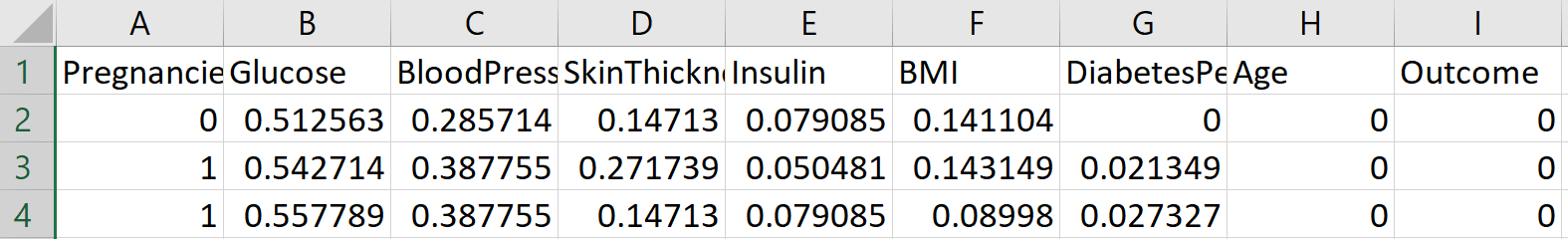
**Dataset Description**

We have used the Pima Indians diabetes dataset, which was collected from <https://www.kaggle.com/uciml/pima-indians-diabetes-database>. The data is of .csv form. The datasets consist of eight medical predictor variables and one target variable, **Outcome**. Independent variables include the number of pregnancies the patient has had, their BMI, insulin level, age, Glucose level, the Diabetes Pedigree Function, blood pressure and Skin Thickness.



**Data Preprocessing**

The data preprocessing was done using the WEKA IDE. The pregnancies column has been binarized to zero and one. Also, the zero values in Blood pressure, Skin Thickness and BMI have been replaced with the mean of the respective columns. Also, the entire dataset has been standardized by applying Z-score normalization, to bring all the data onto one scale. A snapshot of the data after preprocessing is:

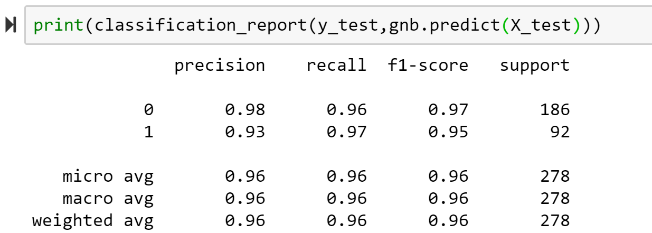


**Description of Methodology**

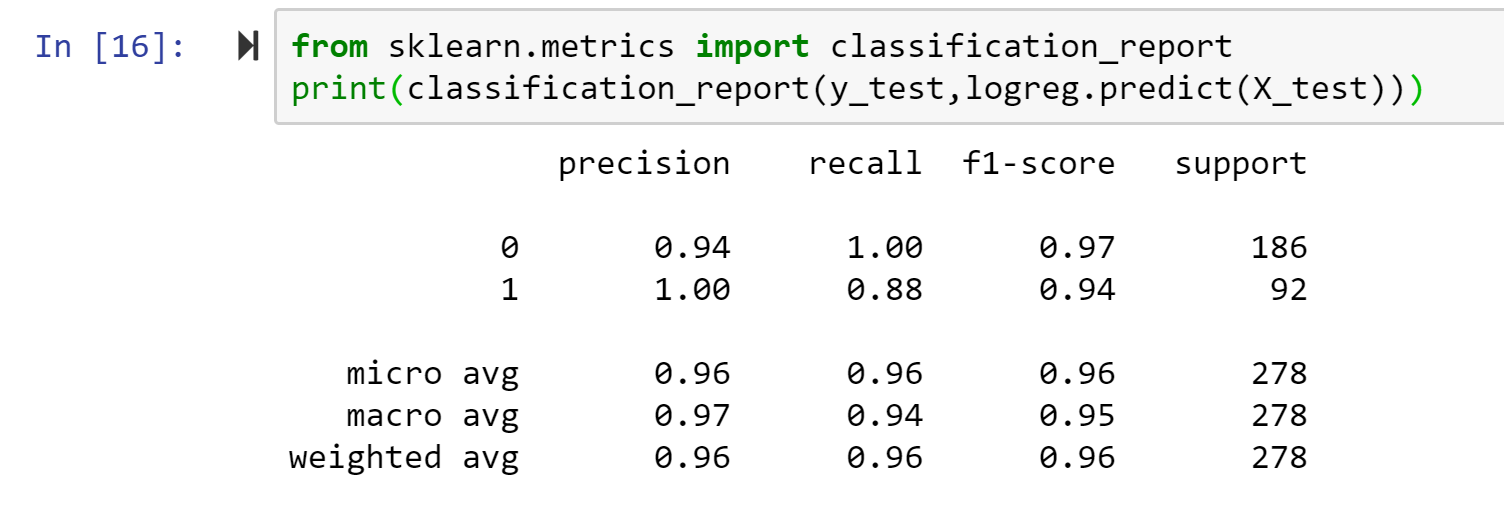
We had initially applied Naïve Baye’s prediction model and logistic regression model, which are supervised methods, but had not achieved a satisfactory accuracy score. So, what we then did is that we applied the K-means clustering algorithm, identified and removed the incorrectly clustered data, and then re-applied Naïve Baye’s Prediction model and Logistic regression, on only the correctly classified data, to finally achieve an accuracy of 96.4% and 96%, respectively. All of this was implemented using the Python programming language.

**Presentation of Results**

Classification report of Naïve Bayes



Classification report of Logistic Regression



**Limitations Faced:**

* We faced certain issues when it came to extraction of correctly clustered data with Weka.
* The size of the dataset could have been larger, to facilitate implementation of machine learning algorithms.

**Literature Review:**

[1] 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).

[2] Patil BM. Hybrid prediction model for Type-2 diabetic patients. Expert Syst Appl 2010; 37:8102–8.

[3] Ahmad Aliza, MustaphaH Aida. Comparison between neural networks against decision tree in improving prediction accuracy for diabetes mellitus. ICDIPC 2011, Part I. CCIS 188; 2011. p. 537–45.

[4] Marcano-Cede~no Alexis, Torres Joaquín, Andina Diego. A prediction model to diabetes using artificial metaplasticity. IWINAC 2011, Part II. LNCS 6687; 2011. p. 418–25.

[5] 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).

[6] Ms. K Sowjanya, MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices. 2015 IEEE International Advance Computing Conference (IACC).