A Classifier Model for Diabetes Prediction

Mahindapala D.P.P.
Faculty of Engineering
University of Ruhuna
Galle, Sri Lanka
pabasaramahindapala@gmail.com

Thalpavila T.W.K.M.B.K.
Faculty of Engineering
University of Ruhuna
Galle, Sri Lanka
madarathalpavila@gmail.com

***Abstract*—This classifier model was constructed through a supervised learning problem using logistic regression and support vector machines as machine learning algorithms. We used a dataset which consists several features which have medical data of a particular group. We had to make some changes the dataset in order to replace some value which seemed not practical in the real world, by more acceptable values. Accuracy of minimizing the bad impact of wrong results. We were more attentive about having low false negatives than low false positives.**

I. Introduction

Diabetes has become a major health issue all over the world. It effect not only the fitness but also mental health and efficiency of a person. We planned to construct a classifier model for diabetes prediction as our machine learning project in order to find a solution to control this health problem.

Diabetes is a non communicable disease. It is considered to be related to hereditary history. Because failure of pancreas to produce the sufficient amount of insulin and inefficient usage of the hormone in the body is what basically happens when somebody gets diabetes. Hence it is able to predict whether somebody has got a risk to be a diabetes patient by analyzing information about family history and observing the fact about physical fitness. These predictions would help people to manage their life style and get medical advises to prevent diabetes. It was what forced us to plan this project mainly.

This model can identify people who are in a higher risk to acquire diabetes. Then people would be careful to maintaining a healthy life style mainly to have a balanced diet. Junk foods and fatty foods are increase the risk of getting diabetes as well as other non communicable diseases like high blood pressure and cancers. Moving to healthy eating habits reduce the probability to get those diabetes. That is not all most of the time people are not aware that they have these diabetes in their body and don't take medicines until the situation get worse. This project will be helpful to that group of people to manage the problem before it becomes dangerous. It makes a great positive impact on the economy of the country and efficiency of the human life. This very much important as diabetes has no exact medicine.

Therefore, we hope to achieve the goal of constructing a classifier model which can successfully predict diabetes to reduce the influence of the disease to the general human life and country's economy.

In this project we used two machine learning algorithms. They are Support Vector Machines(SVM) and Logistic Regression.

We decided to use Logistic Regression because it performs well when the dataset is linearly separable. It not only gives a measure of how relevant a predictor is, but also its direction of association. This algorithm is an easy one to implement interpret and very efficient to train.

Major limitation of the algorithm is the assumption of linearity between the dependent variable and independent variables. Logistic regression can only be used to predict decrease functions. Hence, the dependent variable of logistic regression is restricted to the decrease number set.

SVM has L2 regression feature. So, it has good generalization capabilities which prevent it from overfitting. This algorithm can handle linear data efficiently using kernal trick as well as it solves both classification and regression problems. Complexity of choosing an appropriate Kernal function and algorithm complexity and memory requirements of SVM are very high when using this algorithm. So it is slightly disadvantageous.

II. Methodology

A. Data

In constructing a classifier model for diabetes prediction a data set needs to be used for training the model. Therefore we planned to use medical data of people which are useful in diagnosing diabetes. We used ”Prime Indians Diabetes Database” from Kaggle.com [5]. That dataset includes data of female persons who are at least 21 years old. This data can be defined as predictive analytics because we hope this data set to be used to predict a future situation after analyzing the fast information.

In this data set there are several features. They are mentioned bellow.

- Pregnancies: Number of times pregnant.
- Glucose: Plasma glucose concentration a two hours in and oral glucose tolerance test.
- Blood pressure: Diastolic blood pressure(mm Hg).
- Skin thickness:Triceps skin fold thickness(mm).
- Insulin: Two hour serum insulin(mm U/ml)
- BMI: Body Mass Index(Weight in Kg/ Height in m*m)
- Diabetes pedigree function: Likelihood of diabetes based on family history.

- Age: Age(In years)

The number of data points given here is 768 points.

The machine learning problem we have focused can be defined as follows.

- Task - Predicting diabetes in patients using diagnostic measures
- Performance measure - Recall of predictions
- Experience - Medical data of patients from Pima Indians Diabetes Database with labels

B. Pre-processing

In pre-processing data feature scaling is really helpful to improve the performance of algorithms which are distance based like SVM. We have standardized our data to have a mean of 0 and a standard deviation of 1. Standardization assumes that the dataset has a Gaussian (bell curve) distribution. Standardization is suitable for this project because the logistic regression algorithm we have used make assumptions about the data having a Gaussian distribution.

In the dataset we used, there are zero values for glucose, blood pressure, skin thickness, insulin and BMI. But it is impossible in the real world to have zero values for these features. Hence, we considered them as missing values.

We decided to replace the zeros of those missing values by median values. It seems fair to do that as most of the values are distributed around the center. First we have replaced the zeros by NaN (Not a Number) in Python because zero values should not be considered for calculating the median. Then, as the next step we replace the NaN data by the corresponding median value.

From these 768 data points, we have selected the first 650 records as the training dataset and the remaining 118 points are considered as the testing data.

C. Algorithms and Implementation

We have used Logistic Regression and SVM as our machine learning algorithms in creating the classifier model.

Logistic function shown in equation 1 describes the mathematical form on which the logistic model is based. Function $f(z)$ is given by 1 over 1 plus e to the minus z. When z is equal to $-\infty$, then $f(z)$ is equal to 0. And when z is $+\infty$, then logistic function $f(z)$ is 1. Always, a value between 0 and 1 is described by the model as a probability [1].

In this project, we can obtain an appealing S-shaped description of the combined effect of all features on the risk of having diabetes using logistic regression.

$$f(z) = \frac{1}{1+e^{-z}} \tag{1}$$

SVM is a machine learning algorithm used in supervised learning which is capable of both classification and regression approaches.

SVM find a hyperplane in a space different from the input data, which is a hyperplane in a feature space induced by the kernel. The hypothesis space is defined as a set of ”hyperplanes” in the feature space induced by the kernel [4]. SVM algorithm uses the kernel trick to transform a low dimensional input space to a higher dimensional space.

For implementing the two algorithms, we have used the scikit-learn module in Python [2]. We have referenced the example given in [3] when developing the code. When creating machine learning models, defining the ideal model architecture is done through hyperparameter tuning.

For finding the best hyperparameters for the two classifiers, we have used GridSearchCV in scikit-learn. GridSearchCV methodically build and evaluate models for each combination of hyperparameters specified in a grid. Hyperparameters we have selected for the two classifiers are given in Table I.

TABLE I
SELECTED HYPERPARAMETERS

Logistic Regression	SVM

C = 0.001
Solver = Liblinear
	Kernel = Sigmoid
	C = 10
	Gamma = 1

Here, C is the penalty parameter for misclassified data points. A low C means that penalty for misclassification is small and the decision surface is smooth. If C is higher, then the algorithms may try to classify all training examples correctly and lead to overfitting. Gamma is the parameter which define how much influence a single training example has on others.

III. RESULTS

In this project, our goal is having low false negatives than low false positives. Having a false positive means that a patient is falsely predicted as diabetic. The patient would have to take further tests and treatment when they didn’t need to. But a false negative means not predicting a patient as diabetic when in fact the patient is diabetic, leading to major health issues in the future.

Therefore, the main performance measure we have focused is the recall of predictions obtained by testing the two classifiers. Performance measures we have obtained are given in Table II.

TABLE II
PERFORMANCE MEASURES

Logistic Regression	SVM
Precision = 0.692	Precision = 0.786
Recall = 0.600	Recall = 0.489
Accuracy = 0.746	Accuracy = 0.754

We can identify that the logistic regression classifier has a higher recall value than the SVM classifier. This can be also seen in the precision-recall curve shown in Figure 1.

Confusion matrices from the testing of the two classifiers are shown in Figure 2 and Figure 3. According to the confusion matrices, we can see that the Logistic Regression classifier gives less false negatives than the SVM classifier.

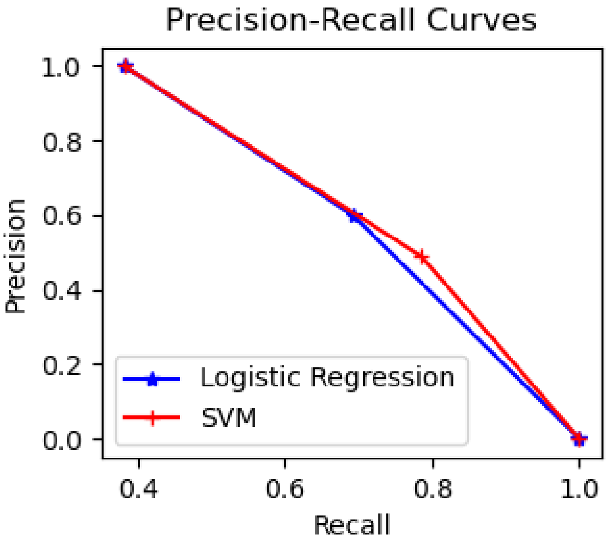

Fig. 1. Precision-Recall Curves of the two classifiers.

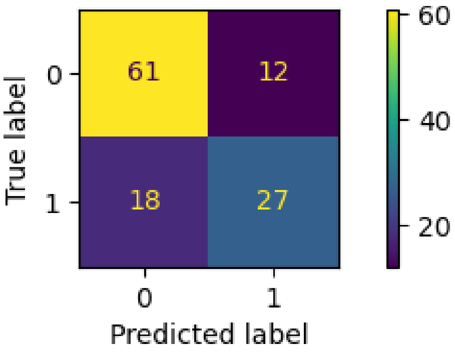

Fig. 2. Logistic Regression Confusion Matrix.

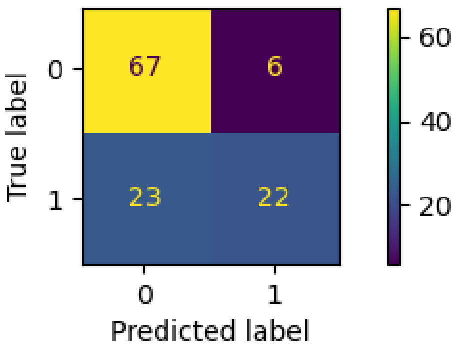

Fig. 3. SVM Confusion Matrix.

IV. Discussion and conclusion

In this project, Logistic Regression classifier is leading with a recall score of 0.6, in contrast to the recall score of 0.489 of the SVM classifier. A recall of 0.6 means that out of the patients who are diabetic, 60% will be correctly identified as diabetic using the model.

A. Limitations

We have also identified few possible limitations of this project. One possible limitation is the factor of time which affect the relevance of the dataset. This dataset was collected from the Pima Indians between 1960s and 1980s. Therefore, the results may not be entirely relevant to present conditions.

In the present, other diagnostic measures like urine tests and haemoglobin tests can be also used to identify the risk of having diabetes. But the number of features considered in this project are limited to eight. Another limitation is in the dataset, only 768 data points collected from patients in one area is available.

B. Conclusion

We can conclude that this project provides a good start on predicting the risk of having diabetes using medical data with machine learning. Using machine learning for predicting diabetes can be helpful in identifying the patients with a risk of diabetes and treat them accordingly. Machine learning can be also used to predict the risk of having other diseases in patients by considering the combined effect of all risk factors. We have identified that blood glucose level and BMI are the most prominent features used to identify a patient as diabetic. Therefore maintaining the blood glucose level at the recommended level and maintaining an average BMI is important for a healthy life.

Acknowledgment

The authors would like to thank Dr. Noeline W. Prins, module coordinator of EE7206 : Machine Learning and the Department of Electrical and Information Engineering, University of Ruhuna.

References

[1] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, Logistic regression. Springer, 2002.
[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourget al., “Scikit-learn: Machine learning in python, ”the Journal of machine Learning research, vol. 12, pp. 2825–2830, 2011.
[3] James Ng, ”Build a machine learning classifier model”, 2019. Accessed on. July 10, 2020. [online]. Available: https://github.com/JNYH/diabetes_classifier
[4] T. Evgeniou and M. Pontil, “Support vector machines: Theory and applications,” in Advanced Course on Artificial Intelligence. Springer, 1999, pp. 249–257.
[5] UCI Machine Learning, ”Pima Indians Diabetes Database”, 2016. Accessed on. July 10, 2020. [online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database/metadata