Genome-based Prediction of Diabetes Using Machine Learning

Submitted By

Kaveri Kosambi

17MCEC09

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2019

Genome-based Prediction of Diabetes Using Machine Learning

Final Project

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology in Computer Science and Engineering (CSE)

Submitted By

Kaveri Kosambi

(17MCEC09)

Guided By

Prof. Tejal Upadhyay

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

Certificate

This is to certify that the final project entitled **"GENOME BASED PREDICTION OF DIABETES USING MACHINE LEARNING"** submitted by **KAVERI KOSAMBI (17MCEC09)**, towards the partial fulfillment of the requirements for the degree of Master of Technology in Computer Science and Engineering of Nirma University is the record of work carried out by her under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination.

Prof. Tejal Upadhyay
Assistant Professor,
Department of CSE,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. Madhuri Bhavsar
Professor and Head,
Department of CSE,
Institute of Technology,
Nirma University, Ahmedabad

Statement of Originality

I, **Kaveri Kosambi**, Roll. No. **17MCEC09**, give undertaking that the Final Project entitled "**Genome based Prediction of Diabetes Using Machine Learning**" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering (CSE)** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made. It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

_____

Signature of Student

Date:

Place:

Endorsed by

Prof. Tejal Upadhyay

(Signature of Guide)

Acknowledgements

It gives me immense pleasure in expressing thanks to **Prof. Tejal Upadhyay**, Assistant Professor, Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad for her valuable guidance and continual encouragement throughout this work. The appreciation and continual support she has imparted has been a great motivation to me in reaching a higher goal. Her guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

It gives me an immense pleasure to thank and profound gratitude to **Dr.Madhuri Bhavsar**, Head of the Department, Institute of Technology, Nirma University, Ahmedabad for her kind support and providing basic infrastructure and healthy research environment.

I would also thank the Institution, all faculty members of Computer Science and Engineering Department, Nirma University, Ahmedabad for their special attention and suggestions towards the project work.

- Kaveri Kosambi

17MCEC09

Abstract

Genome based prediction is widely popular in public health care sector due to its ability to predict the onset of common diseases, thus helping patients in early diagnosis and recovery. In this paper, we have created a simple Keras Neural Network with one hidden layer having 5 hidden nodes. A sequential model is used and after it trains on the data it predicts the likelihood of diabetes on the testing data. In the end, a graph is plotted to prove that more the diabetes pedigree function, more likely it is for a patient to get diabetes. This proves that ancestor history plays an important role when it comes to prediction of diabetes.

Abbreviations

DPF	Diabetes Pedigree Function.
BMI	Body Mass Index.

Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction

The dataset that we have used for diabetes prediction is an updated version of the Pima Indians Diabetes dataset. In this dataset the independent variables are Pregnancies, BMI, Insulin, Age, Blood Pressure, Glucose, Skin Thickness and Diabetes Pedigree Function. The dependent or target variable is Outcome which shows us whether the patient has diabetes or not. This is classified with the help of class label indicating 0 or 1, corresponding to no diabetes and having diabetes respectively. The dataset has dimensions of 15000 x 9. There are 10000 patients that do not have diabetes and remaining 5000 have diabetes.

In this dataset, there are no records that have null values or missing values. Also, there are no unnecessary zero values, which were there in the original Pima Indians dataset with many attributes. They have all been replaced with median values, the dataset is completely clean.

1.2 Diabetes Pedigree Function

The main genetic attribute used in this diabetes dataset is known as Diabetes Pedigree Function. Diabetes Pedigree Function gives data based on the family history of the patient. It checks whether diabetes is present in relatives or in other genetic relationships of the given patient. This attribute gives us an idea of the genetic risk that is hereditary

to the patient to check whether the patient is also likely to get diabetes or not based on their ancestor history. Our main goal with this project, is to show that it plays a very important role in predicting whether a patient is likely to get diabetes in the future, hence proving that genetic profile and family history directly affect the chances.

Chapter 2

Implementation

2.1 Using Neural Network

Artificial neural network is one the most popular machine learning algorithm. It has various application areas and is widely used in Robotics, Artificial Intelligence, Health care, etc. with the help of deep learning.

Here, we have implemented a three layer simple Keras neural network. The sequential model consists of a linear stack of layers. We initialize the number of units of all layers except the last one with the (number of features + number of output nodes)/2 which equals to 4 in our case. However, my results were improved by setting units = 16 for the first layer and decreasing the units in the hidden layer. The first layer requires input dimension and 'relu' stands for rectified linear unit and it is the most used activation function in the world with range from 0 to infinity. Final layer has sigmoid activation function and it is used as it is a standard for binary classification as it keeps the answer between the range of 0 and 1, making it easier for us.

The model is then compiled by specifying the training configuration: optimizer, loss, metrics. Here, 'adam' is the optimizer we have used as it outperforms other optimizers and performs well in practice. Also, 'binary_crossentropy' is the loss function that we try to minimize and in the list of metrics to monitor we have chosen 'accuracy' since we have classification problem.

Now, we train the model for 10 epochs(no. of times we iterate on a dataset). The data is divided into batches of "batch_size", which we have taken as one for our model. The validation data is passed at the end of each iteration to monitor and evaluate validation loss and metrics, as shown in figure 2.1.

```
model = Sequential()
model.add(Dense(16, input_dim=8, activation='relu'))
model.add(Dense(8, activation='relu'))
model.add(Dense(1, activation='sigmoid'))


model.compile(optimizer = 'adam', loss = 'binary_crossentropy', metrics=['accuracy'])

model.fit(X_train, y_train, epochs = 10, batch_size=1, validation_data=(X_test, y_test))
```

```
Using TensorFlow backend.
```

Figure 2.1: Model

2.2 Predictions and Accuracy

Now y_pred contains the probability of Outcome being 0 or 1 class label. We assign the results as 0 or 1 based on their probabilities. If the probability is ¿= 0.5 then it will be 1 otherwise it will be 0. The accuracy of our model is 84.23 or roughly 85% as shown in figure 2.2.

```
In [9]: from sklearn.metrics import accuracy_score

        accuracy_score(y_test,y_pred)

Out[9]: 0.8423333333333334
```

Figure 2.2: Accuracy

The y_pred or "Prediction" column is appended at the end of diabetes testing samples dataset. Then we compare the plot for Prediction=0 and Prediction=1 against Diabetes Pedigree Function. The results can be seen in figure 2.3 and 2.4.

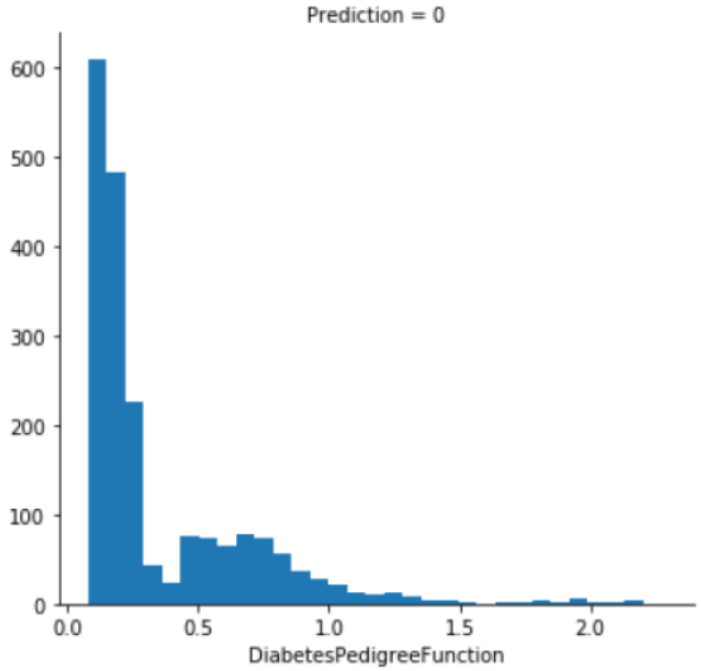

Figure 2.3: Diabetes Pedigree Function and Prediction=0

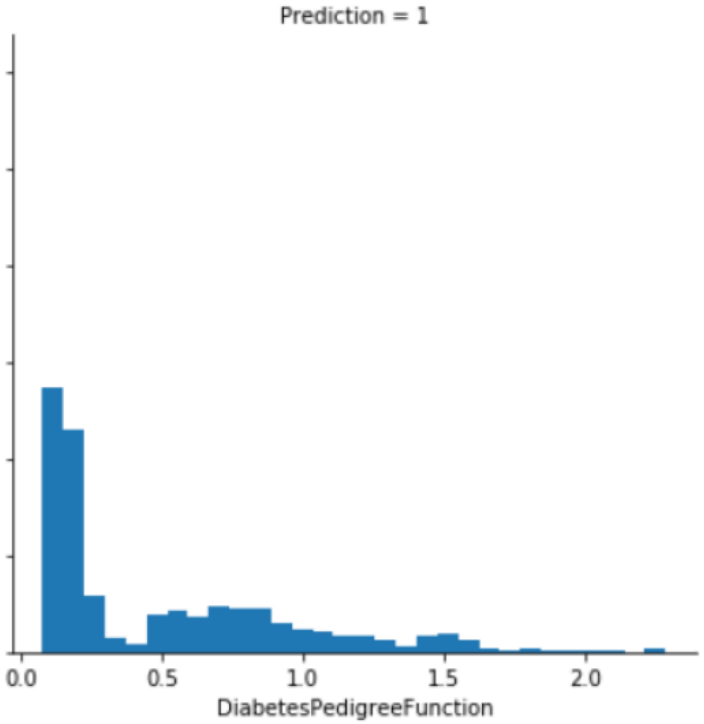

Figure 2.4: Diabetes Pedigree Function and Prediction=1

Chapter 3

Results

3.1 Results

As we can see from the graph plot, when the value of Diabetes Pedigree Function is low, then there are higher chances of the patient to not get diabetes (Prediction = 0). On the other hand we can see that the probability of diabetes increases as the value of Diabetes Pedigree Function also increases.

To further solidify our claims, we predict on two new diabetes patients. Both the patients have same values for all the other attributes except Diabetes Pedigree Function. However, the value of Diabetes Pedigree Function differs in both with one being low and one being high. After our model predicts on these two patients, the results clearly prove that patient having high diabetes pedigree function is more likely to get diabetes, as they have class label as 1 whereas the patient with lesser value was not likely to get diabetes with class label shown as 0. This can be shown in figure 3.1 Hence, we can conclude that family history of diabetes and having similar genetic mutations to them can lead to the patient at higher risk of getting diabetes.

```
new_person = pd.DataFrame([[1, 80, 100, 120, 70, 90, 4.27, 35]])
result = model.predict_classes(new_person)
print(result)

if result==1:
    print("Diabetes")
else:
    print("No Diabetes")
```

```
[[1]]
Diabetes
```

```
new_p= pd.DataFrame([[1, 80, 100, 120, 70, 90, 0.6, 35]])
r2 =model.predict_classes(new_p)
print(r2)



if r2==1:
    print("Diabetes")
else:
    print("No Diabetes")
```

```
[[0]]
No Diabetes
```

Figure 3.1: New Patient Predictions

Chapter 4

Future

4.1 Challenges

A limitation of this dataset is that it is limited to Pima Indians and hence the results are also very limited. If the dataset can be expanded to include people belonging to other ethnic classes and populations as well, then the models can be even better.

Another challenge with this dataset is that the data was collected between the time period of 1960s and 1980s. Hence we can say that the data is pretty outdated since it is quite a few years old. And hence even the features that are a part of this dataset are not enough. In the current medical scenario, not just these features but others are used as well. For example, these days a haemoglobin test is taken to monitor average sugar levels. Other factors are also taken into consideration, such as a patient's urine test is also taken to be tested for the likelihood of diabetes.

Finally, the Pima Indians diabetes dataset is also rather small, and hence some of the models and the algorithms may have only limited performance due to it. It also only contains data of only women who belong to Pima Indian heritage.

4.2 Future

In the future, if the dataset is improved to include more of the current medical attributes along with more genetic attributes as well then it may be beneficial. If the dataset can be

expanded for the whole population of different ethnicities then it may provide valuable insight on whether the people of that geographical location are more susceptible to have diabetes or not. This may help in using precision medicine and targeting the disease as early as possible so as to avoid its onset in the first place. For example, people that tend to show more likelihood to develop diabetes may start prompting lifestyle changes early on: like reduce sugar intake, exercise as much as possible, etc.

4.3 Conclusion

With this implementation, we can conclude that we can predict the likelihood of a patient to get diabetes or not based on their features including their genetic attribute- called as Diabetes Pedigree Function. The more the value of it, the more likely the patient is expected to have diabetes in the future. We can see this through multiple new data entries where we raise the value of DPF. Hence, we can conclude that people with diabetes in their heredity are at a greater risk. This proves that genetic attributes and our genes play an important role in our overall health in the future.

Bibliography

[1] Pima Indians Diabetes Dataset- https://www.kaggle.com/uciml/pima-indians-diabetes-database

[2] Associations between Variables Regarding Diabetes for Pima Indian Women- https://rpubs.com/ikodesh/53189

[3] Rescaling Data for Machine Learning in Python with Scikit-Learn- https://machinelearningmastery.com/rescaling-data-for-machine-learning-in-python-with-scikit-learn/

[4] Getting started with the Keras Sequential model- https://keras.io/getting-started/sequential-model-guide/

[5] Activation Functions in Keras- https://keras.io/activations/

[6] Usage of optimizers- https://keras.io/optimizers/

[7] Associations between Variables Regarding Diabetes for Pima Indian Women- https://rpubs.com/ikodesh/53189

[8] Metrics Classification Report-
http://joshlawman.com/metrics-classification-report-breakdown-precision-recall-f1/

[9] Visualize Machine Learning Data in Python With Pandas-
https://machinelearningmastery.com/visualize-machine-learning-data-python-pandas/