**CS6220 Data Mining Techniques - Final Project**

Team Members: Paul Cruz, Chih-Ming Sun, Haocheng Yang

GitHub Repo: <https://github.com/Paulrkcruz/CS6220-Final_Project-Diabetes>

**Prediction of Diabetes using Data Mining Techniques**

**Abstract**

The remarkable advances in biotechnology and biomedical sciences have led to a significant production of data, such as high throughput genetic sequencing data and clinical information, generated from large Electronic Health Records (EHRs). To this end, the application of machine learning and data mining methods in the biomedical sciences is presently, more than ever before, vital and indispensable in our efforts to transform all available information into valuable knowledge. Diabetes is defined as a group of metabolic disorders exerting significant pressure on human health across the world. Extensive amounts of research in all aspects of diabetes have led to the generation of enormous amounts of data. The aim of this study was to ascertain if diabetes is either predictable or preventable to the highest accuracy by employing data mining techniques to a publicly available dataset from the University of California, Irvine. Therefore, three machine learning classification models namely Random Forrest, Gradient Boosting, and AdaBoost. The performance of each model was evaluated on various measures like Precision, receiver operating characteristic curve (ROC), and the correlative coefficient. Accuracy is measured over correctly and incorrectly classified instances. By focusing on two features of the data found to be the most significant, Glucose levels and BMI, application and analysis of selected models allowed for proper analysis. The results obtained show that the Random Forrest Classifier outperforms with the highest accuracy of 79.00% comparatively against the other two models. These results are verified using Receiver Operating Characteristic (ROC) curves in a proper and systematic manner. As a result, the model was shown to be very useful for the prediction of Diabetes.

1. **Introduction**

Diabetes is a well-known disease affecting individuals all over the world. According to the WHO (World Health Association), an estimated 422 million people worldwide currently have diabetes with the majority living in low-and middle-income countries. In addition, 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 and are expected to grow. But can diabetes be prevented?

Recent research utilizing data science via data mining techniques including predictive modeling using machine learning is inspiring confidence in the medical world. Classification strategies are broadly used in the medical field for classifying data into different classes according to individual classifiers. Diabetes, being a disease that affects the human body’s ability to produce the hormone insulin, which in turn causes metabolism of carbohydrates abnormal while at the same time causing elevated levels of glucose in the blood. Generally, an individual with diabetes suffers from high blood sugar, with intensified hunger, frequent urination, being caused by this abnormal level of sugar in the blood. Though there are many factors that may contribute to a diabetes diagnosis, such as heredity, weight, and height, the major factor is blood sugar concentration. The early identification of abnormal glucose levels may be the only remedy to further complications and therefore a more serious diagnosis. Diabetes is a serious health matter during which the measure of sugar substance cannot be controlled. Thus, analysis of the levels of sugar in the blood and how it may compare to other compounds measured is paramount in predictive modeling.

The problem in the diagnosis of diabetes is that many times the diagnosis could have been much sooner, we need a way to diagnose early. In this work we aim to show how this can be done using the machine learning algorithmic approaches of Data mining. We focus on over 70 datasets recorded on patients with diabetes ranging from several weeks to several months, containing readings of glucose, insulin, and patient lifestyle data. Decision Tree machine learning classification algorithms are used and evaluated on the University of Irvine dataset with the aim to predict diabetes in a patient. Therefore, experimental performance of many of the most commonly used predictive modelling algorithms and methods will be employed with the intent to solve this problem.

1. **Methodology**

Archived file diabetes-data.tar.z which contains 70 sets of data recorded on diabetes patients (several weeks' to months' worth of glucose, insulin, and lifestyle data per patient + a description of the problem domain) is extracted and processed and merged as a CSV file.

1. **Code**
2. **Results**

1. **Discussion**

The results obtained

1. **Future Work**

One of the most important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in finding a system which results in the prediction of diabetes through data science via predictive modelling algorithms. Though this work may not be the final solution to the prediction of diabetes, it serves as an example of the power data science may have on the prediction of other diseases as well. During this work, three machine learning classification algorithms are chosen and evaluated on various measures. Experimental modelling is then performed on the University of California, Irvine, diabetes dataset. The results determine the adequacy of the system with an achieved accuracy of 79.00% using the Random Forrest classification algorithm. In the future, this system with the use of machine learning classification algorithms may be able to be used to predict or diagnose other diseases as well. The work therefor be extended and improved for the automation of diabetes analysis using the methods described here as well as other machine learning algorithms.

**References**

Huang, R. (2021). Prediction of Pima Indians Diabetes with Machine Learning Algorithms. *UCLA*. ProQuest ID: Huang\_ucla\_0031N\_19508. Merritt ID: ark:/13030/m5md4q60. Retrieved from https://escholarship.org/uc/item/6rh07945

Altıntaş, Ergin. “UCI Diabetes Data Set.” *Kaggle*, 1 May 2020, www.kaggle.com/ealtintas/uci-machine-learning-repository-diabetes-data-set.

World Health Organization (WHO). “Diabetes.” *Kaggle*, 2 April 2021, <https://www.who.int/health-topics/diabetes#tab=tab_1>.

***What’s the format of the final report?***

Although you are not expected to write a full-fledged research paper, this exercise helps you understand the structure of the components thereof. The project report might have the following seven main sections -

a. Abstract - This is a brief summary not exceeding 500 - 750 words describing the problem, the solution method, and the results.

b. Introduction – Approximately one to two pages long and contains (i) Statement of the problem you are trying to solve  
(ii) Why is it important to solve this problem?

(iii) Background information and bit of literature survey to present what’s already known about this problem.

c. Methodology - What’s your solution, how do you propose to solve the identified problem?

d. Code - Brief explanation of the code.

e. Results – here you just state the results as you see them. E.g., In a fictitious study you were studying a data collected over a period of 50 years of height of children aged 12 years. You can state the trends shown in the data, like in all countries the average height of 12-year-olds seem to be increasing by 3 mm every decade, except in countries C and D. In these two countries it seemed to decrease by 2 mm in the 1980s.

f. Discussion - interpretation of the results - what do the numbers really mean, do they make sense. E.g., In the fictitious study we used to illustrate the results section, Does it make sense that in countries C and D the average height decreased. Probably, there was a famine in Country C and Country D had an epidemic that affected the health of children in the 1980s.

g. Future Work - Is this study conclusive or does it lead to some future work? h. Conclusion - what conclusions can you draw.

i. References – Research papers, articles, and Internet resources referred to in the rest of the report.

<https://www.kaggle.com/uciml/pima-indians-diabetes-database>

<https://escholarship.org/uc/item/6rh07945>