**Diabetes Mellitus Prediction Using Machine Learning Algorithms**

**Project Category**: Life Sciences

**Project Mentor**: TBD

**Name**: Xinxie Wu

SUNet ID: xinxiewu

Department of Computer Science

Stanford University

[xinxiewu@stanford.edu](mailto:xinxiewu@stanford.edu)

**Abstract**

**Introduction / Motivation**

Diabetes Mellitus (DM) is a chronic (long-lasting) disease that affects how your body turns food into energy, the 7th leading cause of death in the United States. There are three main types of diabetes: type 1, type 2, and gestational diabetes.

According to Centers for Disease Control and Prevention (CDC), in the last 20 years, the number of adults diagnosed with diabetes has more than doubled. Currently more than 37 million US adults have diabetes, and even 1 in 5 of them don’t know they have it.

Working in the healthcare industry for 5 years, with 3 years in the area of diabetes mellitus, I have witnessed how the prevention and the early detection help pts out of diseases. However, early diagnosis/detection of DM is quite challenging for medical practitioners, since DM has a complex interdependence on various factors from human’s different organs. As a data scientist, I believe machine learning models, based on pts’ medical data, would help on the early identification/prediction of DM. Therefore, this application research will explore how the machine learning models would help in DM early prediction, and discuss the prediction accuracy among several models.

**Literature Review / Related Work**

Sharma et al. (2021) [1] discussed various machine learning algorithms for diabetes prediction, with an Indian population, including logistic regression (LR), decision tree (DT), random forest (RF), and k-nearest neighbors (KNN). Among these algorithms, RF algorithms achieved the highest accuracy of 83.6%. Mujumdar et al. (2019) [5] pointed out that the classification accuracy of diabetes prediction can be improved by including external factors, and achieved 96% accuracy using logistic regression.

**Dataset and Features**

**Diabetes 130-US hospitals for years 1999-2008 Data Set**: This dataset contains data on over 100,000 diabetic patients from 130 hospitals in the United States. It includes features such as patient demographics, lab test results, and hospital admission details.

<https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008>

**Methods**

1. Classification:
   1. Decision Tree
   2. Random Forest
   3. Neural Network
2. Feature Selection:
   1. Principal Component Analysis
   2. Minimum Redundancy Maximum Relevance

**Preliminary Experiments, Results and Discussion**

1. **Experiment**: This research is to develop several machine learning models (classification & deep learning), based on pts’ medical data. Also, features’ selection will be conducted, important features/variables will be discussed. Finally, we will compare models’ prediction accuracy;
2. **Evaluation**: Prediction accuracy is the top metric in this research. To achieve this, the dataset will be divided into training and testing sets, randomly.

**Conclusion and Future Work / Next Steps**

**Contributions**

As the only member in this project, Xinxie Wu is responsible for all parts of this research.

**References**

1. Sharma, T., & Shah, M. (2021). A comprehensive review of machine learning techniques on diabetes detection in 2021. International Journal of Computer Science and Mobile Computing, 10(4), 115-121.
2. Jobeda Jamal Khanam, and Simon Y. Foo. A comparison of machine learning algorithms for diabetes prediction, 2021.
3. Jingyu Xue et al 2020 J. Phys.: Conf. Ser. 1684 012062. Research on Diabetes Prediction Method Based on Machine Learning.
4. Shahabeddin Abhari, Sharareh R. Niakan Kalhori, Mehdi Ebrahimi, Hajar Hasannejadasl, and Ali Garavand. Artificial Intelligence Applications in Type 2 Diabetes Mellitus Care: Focus on Machine Learning Methods, 2019.
5. Aishwarya Mujumdar, and Dr. Vaidehi V. Diabetes Prediction using Machine Learning Algorithms, 2019.
6. Tejas N. Joshi, and Prof. Pramila M. Chawan. Diabetes Prediction using Machine Learning Techniques, 2018.
7. Song, I. U., Cho, H. J., Lee, H. W., & Kim, J. Y.. Predictive models for diabetes using machine learning techniques: A systematic review and meta-analysis. Journal of Medical Internet Research, 23(3), e23934. (2021).
8. Kaur, R., & Kaur, P. (2021). Diabetes prediction using machine learning techniques: A comprehensive review. International Journal of Computing and Digital Systems, 10(1), 18-25.
9. Hasan, M. S., Rahman, M. S., & Islam, M. S. (2021). Machine learning-based diabetes prediction: A review. Healthcare, 9(2), 157.
10. Singh, P., & Jaiswal, N. (2020). Diabetes prediction using machine learning: A review. International Journal of Advanced Computer Science and Applications, 11(1), 91-96.
11. Swapna G., Vinayakumar R., Soman K.P. (2018). Diabetes detection using deep learning algorithms.