**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**

To Do in Final Write-up

**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 diabetes mellitus, I have witnessed how prevention and 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%. Jobeda et al. (2021) [2] compared the performance of seven ML algorithms on the Pima Indian Diabetes (PID) dataset. From their research, LR and Support Vector Machine (SVM) work well on diabetes prediction, the Neural Network (NN) with two hidden layers provided 88.6% accuracy. Except SVM, Jingyu et al. (2020) [3] trained Naïve Bayes classifier and LightGBM on a dataset of 520 diabetic patients and potential diabetic patients aged 16-90. Based on their research, SVM has the highest accuracy rate of 96.54%, while Naïve Bayes achieved 93.27% and LightGMB is only 88.46%. 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 LR. Song et al. (2021) reported that the overall diabetes prediction accuracy ranged from 66.7% to 99.4%. This research noted that ensemble models, such as RF and gradient boosting, tended to perform better than any single algorithm. Hasan et al. (2021) and Singh et al. (2020) reviewed several studies in diabetes prediction using machine learning models, accuracy rate ranging from 70% to 99%. Like Song, they also noted that the ensemble models help the performance improvement. Swapna et al. (2018) [11] discussed the diabetes prediction using deep learning architecture. The research employed long short-term memory (LSTM) and Convolutional Neural Network (CNN) for extracting complex temporal dynamic features, and then passed the features into SVM for classification. The performance improved 0.03% and 0.06% in CNN and CNN-LSTM compared to the ones without SVM. The deep learning system reached an accuracy rate of 95.7%.

**Dataset and Features**

In the proposal, we mentioned 3 potential datasets. After analyzing each one, this research will focus on Pima Indians Diabetes (PID). This dataset contains 778 records, with 8 medical information, and 1 column to identify if the patient is diagnosed with diabetes. Medical information includes 1) The number of pregnancies the patients has had, 2) Plasma glucose concentration 2 hours in an oral glucose tolerance test, 3) Diastolic blood pressure, 4) Triceps skin fold thickness, 5) 2-hour serum insulin, 6) Body mass index (BMI), 7) Diabetes pedigree function and 8) Patient’s age.

**Methods**

1. Classification:
   1. Logistic Regression – Gradient Descent & Newton Method
   2. Generative Learning Algorithm – to compare with Logistic Regression
   3. Naïve Bayes
   4. Support Vector Machine
   5. Deep Learning of Neural Network (Convolutional)
   6. Imbalanced Data – Ensemble Models
      1. Gradient Boosting
      2. Random Forest
2. Feature Selection:
   1. Principal Component Analysis
   2. K-means – to remove incorrectly clustered data

**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. Confusion matrix.

**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.
12. Bokhare, Anuja and Vandan Raj, N. 2023 International Conference for Advancement in Technology (ICONAT) Advancement in Technology (ICONAT), 2023 International Conference for. :1-5 Jan, 2023.
13. T.M. Alam, et al., Informatics in medicine unlocked a model for early prediction of diabetes, Inform. Med. Unlocked 16 (2019) 100204.
14. Patil BM. Hybrid prediction model for Type-2 diabetic patients. Expert Syst Appl 2010;37:8102–8.