Diabetes Mellitus Prediction Using Machine Learning Algorithms

Project Category: Life Sciences
Project Mentor: Hong Liu

Name: Xinxie Wu
SUNet ID: xinxiewu
Department of Computer Science
Stanford University
xinxiewu@stanford.edu

Key Information to include

- No external collaborators as of this milestone manuscript, 05/05/2023
- Not sharing this project with any other class/group

Motivation / Introduction

Diabetes Mellitus (DM) is a chronic disease that affects the body's ability to convert food into energy and is the 7th leading cause of death in the United States. DM can be classified into three main types: type 1, type 2, and gestational diabetes. According to the Centers for Disease Control and Prevention (CDC), the number of adults diagnosed with diabetes has more than doubled in the last 20 years. Currently, over 37 million US adults have diabetes, and 1 in 5 of them are unaware of their condition.

Early diagnosis and prevention are essential in managing the disease and reducing its complications. However, DM is a complex disease with various interdependencies on human body's different organs, making it challenging for medical practitioners to detect and diagnose it early. Machine learning models, based on patients' medical data, have the potential to aid in the early detection and prediction of DM.

This research aims to investigate the effectiveness of various machine learning algorithms for predicting diabetes using the Pima Indian Diabetes (PID) dataset. Logistic Regression (LR), Support Vector Machines (SVM), and Naïve Bayes (NB) algorithms will serve as baseline for comparison. Principle Component Analysis (PCA) and k-means will be employed for feature selection, followed by the retraining of LR using the newly selected features. The study will investigate whether the performance of LR is improved by the implementation of these techniques. In addition, Neural Network (NN) algorithm will be executed to further compare the results. Finally, ensemble models, such as Random Forest (RF), will be used to assess the effectiveness of the algorithms in addressing the imbalanced nature of the PID dataset.

Related Work / Literature Review

Various studies have explored the accuracy and performance of different algorithms in diabetes prediction. Sharma et al. (2021) [1] evaluated several algorithms, including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbors (KNN) on the PID dataset, and found RF to achieve the highest accuracy of 83.6%. Jobeda et al. (2021) [2] compared seven ML algorithms on

the PID dataset as well and found LR and Support Vector Machine (SVM) to work well, with a two hidden layers Neural Network (NN) achieving 88.6% accuracy. Jingyu et al. (2020) [3] trained Naïve Bayes classifier and LightGBM on a dataset of 520 diabetic patients and found SVM to have the highest accuracy rate of 96.54%, followed by Naïve Bayes at 93.27% and LightGMB at 88.46%. Mujumdar et al. (2019) [5] achieved 96% accuracy using LR by including external factors. Ensemble algorithms, such as RF and Gradient Boosting (GB), have been found the potential to outperform individual algorithms in diabetes prediction (Song et al. (2021) [7]). Swapna et al. (2018) [11] applied deep learning architecture with long short-term memory (LSTM) and Convolutional Neural Network (CNN) for feature extracting, and SVM for classification, achieving an accuracy rate of 95.7%. The performance improved 0.03% and 0.06% in CNN and CNN-LSTM compared to the ones without SVM.

Dataset and Features

The Pima Indian Diabetes (PID) dataset, sourced from the UCI Machine Learning Repository [15] and originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), comprises of health and medical examination data of 768 female patients, who are at least 21 years old, from Arizona, USA population who were examined for diabetes. This dataset is imbalanced, with 268 records (34.9%) identified as diabetic patients, while the remaining 500 (65.1%) are non-diabetic. Aside from the diabetes identifier (output in this research), PID contains 8 numeric attributes (input in this research), which describe the personal health status and medical examination results. **Table 1**, in Appendix, provides a detailed overview of the attributes and their respective statistics.

Although the PID dataset does not contain any missing values, some variables (such as Glucose and Diastolic Blood Pressure) have recorded values of 0, which is not reasonable and thus defined as the missing value in our research. As data quality is a crucial aspect of the research, we need to address the issue of missing values. Based on domain knowledge, the values of these 8 attributes are expected to be related to whether a patient is diabetic. Therefore, in this research, we assigned values based on the diabetes identifier. Specifically, the median value of each variable with missing values was assigned by diabetes status. If the median value was 0, the mean value was used instead.

To examine the relationship between different variables, this research calculated Pearson's correlation coefficients which is between -1 and 1. **Figure 1**, in Appendix, shows the correlation matrix in heatmap. Based on **Figure 1**, we found that Glucose is highly related to Diabetes, with the Pearson coefficient as 0.5, followed by BMI (0.32), which makes sense in medical practice.

Normalization is a technique used to transform data to a common scale, which helps to reduce runtime complexity and improve model performance. In this research, the data is normalized by subtracting the mean of each feature and a division by the standard deviation. This way, each feature has a mean of 0 and a standard deviation of 1.

$$Value_{New} = \frac{Value_{Old} - Mean}{Std}$$

Methods

1. **Baseline Algorithm**: In this research, Logistic Regression (LR), Naïve Bayes (NB) and Support Vector Machine (SVM) serve as the baseline. As of the milestone, baseline has been completed;
2. **Unsupervised Learning Algorithm**: This research plans to use Principal Component Analysis (PCA) and k-means for selecting features and removing outlier/extreme data points (those incorrectly clustered data). A retraining of LR based on the newly selected variables will be performed, and the results will be used to compare with the original ones;
3. **Deep Learning**: Neural Network will be built to see if any help performance improvement;
4. **Ensemble Algorithms**: As noted, PID is an imbalanced dataset, so ensemble algorithm, such as Random Forest, will be implemented and the results will be used to compare;
5. **Others**: If time allowed, we would try k-fold cross validation.

Preliminary Experiments, Results and Discussion

For the baseline models, the PID dataset was split into 70% for training and 30% for testing. This research has completed three baseline algorithms: Logistic Regression (LR), Naïve Bayes (NB) and Support Vector Machine (SVM). Performance evaluation is based on the confusion matrix, details provided in **Table 2**.

Algorithm	Test Size	TN	FP	FN	TP	Total Accuracy	Positive Accuracy	Negative Accuracy
LR	231	123	31	28	49	74.46%	63.64%	79.87%
NB	231	119	29	32	51	73.59%	61.45%	80.41%
SVM	231	126	29	25	51	76.62%	67.11%	81.29%

Table 2: Preliminary Results

Based on **Table 2**, we found that SVM performs best among the three baseline algorithms, followed by LR. Generative algorithm with strong assumptions (NB) only achieved 73.59% accuracy rate. Although the overall accuracy is ~75%, the positive accuracy (~65%) is much lower than the negative ones (~80%), which means the dataset is imbalanced. **The overall performance improvement and the imbalance issue will be discussed further in the next step**.

Next Steps / Conclusion and Future Work

1. Implement unsupervised learning algorithms mentioned in Method, **PCA & k-means**;
2. Implement **Neural Network** algorithm and explore the number of hidden layers which achieves the highest prediction accuracy rate;
3. Implement **Random Forest** to see if the performance of **imbalanced dataset** improves;
4. If time allowed, try **k-fold cross validation**.

Contributions

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

Appendix

Atttibute	Description	Missing	Mean	Std	Min	25%	Median	75%	Max
Preg	Number of pregnancy	0	3.85	3.37	0.00	1.00	3.00	6.00	17.00
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	5	120.89	31.97	0.00	99.00	117.00	140.25	199.00
BP	Diastolic blood pressure (mm Hg)	35	69.11	19.36	0.00	62.00	72.00	80.00	122.00
SkinThickness	Triceps skin fold thickness (mm)	227	20.54	15.95	0.00	0.00	23.00	32.00	99.00
Insulin	2-hour serum insulin (µIU/mL)	374	79.80	115.24	0.00	0.00	30.50	127.25	846.00
BMI	Body mass index (kg/m^2)	11	31.99	7.88	0.00	27.30	32.00	36.60	67.10
DPF	Diabetes pedigree function	0	0.47	0.33	0.08	0.24	0.37	0.63	2.42
Age	Age (years)	0	33.24	11.76	21.00	24.00	29.00	41.00	81.00

Table 1: Attributes of PID dataset

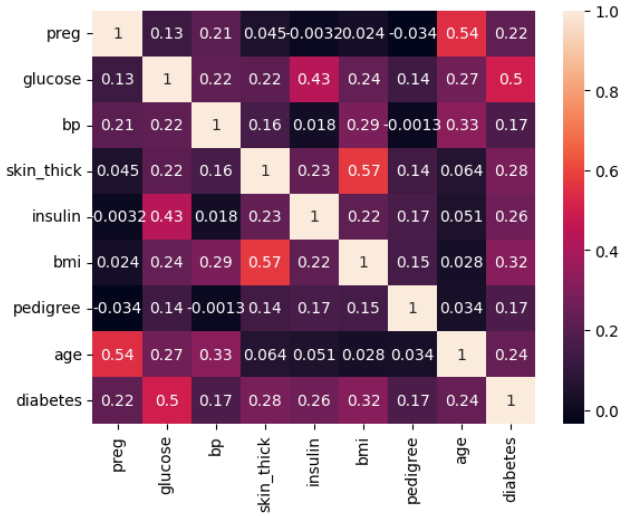

Figure 1: Correlation Matrix

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.

[15] http://archive.ics.uci.edu/ml/datasets/PimaþIndiansþDiabetes.