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

**Abstract**

Diabetes Mellitus (DM) is a metabolic disease characterized by high blood sugar and doesn’t have a permanent cure; hence early detection is required. This research proposed three baseline algorithms and then discussed three methods of improvement, based on an imbalanced dataset of 768 observations, Pima Indian Diabetes (PID). For the baseline models, SVM (80.09%) outperforms Logistic Regression (73.59%) and Naïve Bayes (74.03%), but all with 15-20% performance gap between positive and negative data points. For improvement, Random Forest achieves 88.74% accuracy and shrinks the positive-negative gap down to less than 10%; PCA, with 7 principal components, is picked, and 194 outliers are removed by k-means algorithm. Improved Logistic Regression is re-trained on the refined dataset and reaches 95.95% prediction accuracy; Convolutional Neural Network is also discussed in this research, achieving an accuracy of 90.91%.

**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**

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.

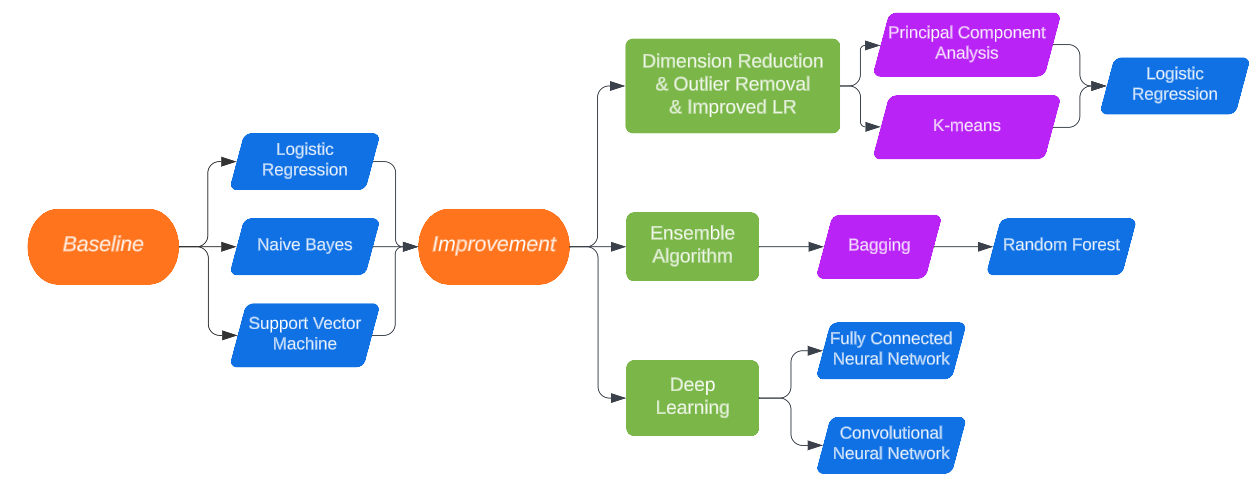
To examine the relationship between different variables, this research calculated Pearson's correlation coefficients which is between -1 and 1. **Figure 1**, in Appendix, shows the correlation matrix in heatmap. Based on **Figure 1**, we found that Glucose is highly related to Diabetes, with the Pearson 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.

In this research, the dataset is split into 70% training and 30% testing. For the deep learning part, the dataset is split into 80% training, 10% validation and 10% testing.

**Methods**

**Figure 2** describes a comprehensive two-part methodology adopted in this research: Baseline & Improvement. In the baseline approach, we evaluated the performance of Logistic Regression (LR), Naïve Bayes (NB), and Support Vector Machine (SVM). Subsequently, the improvement approach incorporates three distinct strategies aimed at enhancing predictive performance. The first one encompasses dimensionality reduction employing Principal Component Analysis (PCA) for effective feature space reduction, followed by an outlier removal through K-means clustering. Subsequent LR retraining on the refined dataset is to explore the performance improvement. Additionally, since the PID dataset is imbalanced as analyzed above, this research explores the efficacy of Random Forest (RF), one of the ensemble algorithms. Finally, we examined the suitability of Neural Network (NN) for diabetes prediction. In this part, Fully-connected Neural Network (FNN) was built and tested; also, convolutional layers (CNN) was added to FNN and analyze if any performance improvement.



**Figure 2**: Learning Algorithms

***Logistic Regression (LR), Naïve Bayes (NB) and Support Vector Machine (SVM)***

LR is a linear model that estimates the probability of a binary outcome. Compared with LR, NB is a generative algorithm that applies Bayes’ theorem with the assumption of independence among features which are assumed to follow the normal distribution. SVM aims to find an optimal hyperplane that maximally separates the data points, by mapping the data into higher-dimensional feature space with kernel functions.

***Principal Component Analysis (PCA) and K-means***

As a dimensionality reduction technique, PCA identifies the principal components, which are linear combinations of the original features that capture the maximum variance. K-means is a clustering algorithm which partitions a dataset into a specified number of clusters, by minimizing the within-cluster sum of squared distances, which involves an iteration of two steps: assign data to the nearest centroid and update the centroid based on the assigned data points.

***Ensemble Algorithm, Bagging, and Random Forest (RF)***

Since the PID dataset is imbalanced, this research applied ensemble algorithms. Ensemble Algorithm is a technique that combines multiple individual models to improve overall prediction performance. Bagging is an ensemble method which generates multiple subsets of the original dataset through bootstrapping, trains individual models on each subset, and combine via voting or averaging. RF is an ensemble algorithm based on bagging by employing decision trees as the basis. Randomness and individual models’ combination helps prediction of imbalanced dataset.

***Deep Learning and Convolutional Neural Network (CNN)***

Convolutional neural network (CNN) is an improvised variant of multilayer perceptron. CNN is generally made up of an input, an output layer, and many hidden layers. The hidden layers of a CNN typically are made up of convolutional, pooling, and fully connected layers.

**Experiments, Results and Discussion**

Rigorous implementation of the above methodologies is realized using Python, scikit-learn and PyTorch libraries, ensuring meticulous preprocessing, model training, and hyperparameter optimization. The evaluation of results is conducted using the confusion matrix.

***Baseline and Random Forest***

LR, NB & SVM were applied in our baseline experiment, performance evaluation in **Table 2**. Based on **Table 2**, SVM outperforms the other two algorithms, achieving 80.09% total accuracy. Although both small-size dataset and normalized features bring a higher expectation on NB than LR, NB’s 74.03% total accuracy is only 0.44% higher than LR’s 73.59%, which can be explained by some features’ high correlation in **Figure 1**.

Except the difference among the baseline algorithms, the positive accuracy is ~15-20% lower than the negative accuracy for all the three models, which indicates the dataset’s imbalance. To improve on this, we built the model of RF. Due to RF’s randomness, 1,000 random models were trained and the best-performance one was picked. From **Table 2** and compared with LR, RF’s total/positive/negative accuracy improved 15.15%/21.25%/11.92%. Specifically, the performance difference between positive and negative shrunk to less than 10%, which indicates RF is a good way to deal with the dataset’s imbalance and helps to improve the prediction performance.



**Table 2**: Results of Baseline and Random Forest

***Unsupervised Learning and Improved Logistic Regression (LR)***

PCA was employed with a range of components values from 1 to 8 on the whole dataset, 768 observations. To pick the optimal number of principal components, improved LR was re-trained on each value and the best-performance one was picked, 7. In **Table 3**, we can see the LR performance with the selected 7 principal components improved 1.73%. Building up on this, we conducted k-means algorithm of 2 clusters. By setting the initial centroid as “auto”, we performed k-means 1,000 times and pick the highest-accuracy one. As a result of k-means, we removed 194 data points which were clustered incorrectly (those data points were considered as outliers/noise in this research). Combing the results of PCA and k-means, we re-trained an improved LR based on the refined dataset, performance in **Table 3**. We can see that the total/positive/negative accuracy improved 22.36%/28.58%/18.99%, with negative accuracy almost 100%. So, using PCA for dimensionality reduction and k-means for outlier removal improved the prediction accuracy.



**Table 3**: Results of Improved Logistic Regression

***Deep Learning and Neural Network***

Fully-connected Neural Network was trained by different number of the hidden layers (2, 3, 4) and neurons (4 values), 12 neural networks in total. Validation dataset was used for hyperparameters’ optimization, and the testing dataset was used for generalization ability check, evaluation metrics in **Table 4**. **Table 4** listed the best-validation-performance network, with different number of neurons, in each value of the hidden layers. Based on **Table 4**, the 4-hidden-layer network, with (5,3,4,2) neurons in each layer, shows the highest accuracy in both validation and testing dataset. The accuracy of 89.61% is also higher than baseline models’ performance.

Except the above one, this research added two convolutional (different parameters) and max pooling layers to a random one of the neural networks we built earlier (2-hidden-layer with (10,7) neurons). **Table 4** shows that, with convolutional layers’ parameters as (32, 64), we have the highest accuracy on the validation dataset. Applying this CNN to the testing set, 90.91% accuracy was achieved, which is 3.90% higher than the one without convolutional layers.



**Table 4**: Results of Deep Learning (L = Hidden Layers)

**Conclusion and Future Work**

Based on the above results of the PID dataset, we conclude that the baseline algorithms (LR, NB & SVM) provide ~75% total accuracy, but with the positive accuracy 15-20% lower than the negative one. The diabetes prediction accuracy was improved by three ways in this research: Improved LR by PCA & K-means, Random Forest & Neural Network. The improved LR, with the refined dataset by PCA & K-means, performs best, achieving 95.95%/89.83% of total/positive accuracy. The 2-layer CNN follows by reaching 90.91%/86.87%, while RF’s performance shows 88.74%/82.50% and shrinks the positive-negative performance gap to less than 10%.

For the future work, k-fold cross validation is considered to be applied since our research focused on 7/3 dataset split. Also, neural networks with more different layers/neurons options need to be trained and compare the performance. Finally, the dataset’s small size is another point we need to consider, such as if the accuracy would be impacted change by the size of the dataset.

**Appendix**



**Table 1**: Attributes of PID dataset

Chart

Description automatically generated with low confidence

**Figure 1**: Correlation Matrix

**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.
15. <http://archive.ics.uci.edu/ml/datasets/PimaþIndiansþDiabetes>.