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

Abstract

Diabetes Mellitus (DM) is a disease characterized by high blood sugar and early detection is needed. 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% accuracy; Convolutional Neural Network is trained and achieves 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. 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 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.

This research aims to investigate the effectiveness of various machine learning algorithms for predicting diabetes using the Pima Indian Diabetes (PID) dataset, all features. Logistic Regression (LR), Support Vector Machines (SVM), and Naïve Bayes (NB) algorithms serve as baseline. Principle Component Analysis (PCA) and k-means will be employed for feature selection, followed by LR retraining using the refined dataset. 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], comprises of health and medical examination data of 768 female patients 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** provides an overview of the attributes with respective statistics.

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

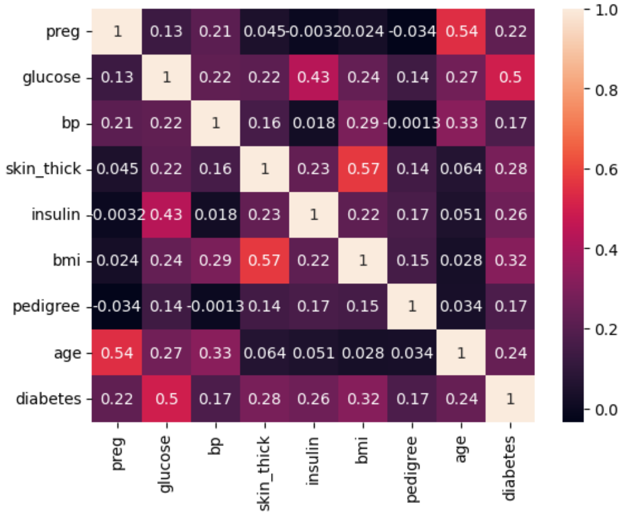
Table 1: Attributes of PID dataset

Although the PID dataset does not contain any missing values, some variables have recorded values of 0, which is not reasonable and thus defined as the missing value. 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, we assigned values based on the diabetes identifier. Specifically, the median value of each variable with missing values was assigned by diabetes status.

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.

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

To examine the relationship between different variables, this research calculated Pearson's correlation coefficients. **Figure 1** shows the correlation matrix in heatmap, where we found that Glucose is highly related to Diabetes, with the Pearson coefficient as 0.5, followed by BMI (0.32).



	preg	glucose	bp	skin_thick	insulin	bmi	pedigree	age	diabetes
preg	1	0.13	0.21	0.045	-0.0032	0.024	-0.034	0.54	0.22
glucose	0.13	1	0.22	0.22	0.43	0.24	0.14	0.27	0.5
bp	0.21	0.22	1	0.16	0.018	0.29	-0.0013	0.33	0.17
skin_thick	0.045	0.22	0.16	1	0.23	0.57	0.14	0.064	0.28
insulin	-0.0032	0.43	0.018	0.23	1	0.22	0.17	0.051	0.26
bmi	0.024	0.24	0.29	0.57	0.22	1	0.15	0.028	0.32
pedigree	-0.034	0.14	-0.0013	0.14	0.17	0.15	1	0.034	0.17
age	0.54	0.27	0.33	0.064	0.051	0.028	0.034	1	0.24
diabetes	0.22	0.5	0.17	0.28	0.26	0.32	0.17	0.24	1

Figure 1: Correlation Matrix

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 the two-part methodology in this research: Baseline & Improvement. In the baseline, we evaluated Logistic Regression (LR), Naïve Bayes (NB), and Support Vector Machine (SVM). The improvement incorporates three strategies. The first one encompasses dimensionality reduction employing Principal Component Analysis (PCA) and outliers' removal through K-means, followed by LR retraining on the refined dataset. Second, we explore the efficacy of Random Forest (RF) on imbalance. Finally, we examined the suitability of Neural Network (NN); convolutional layers were added and tested.

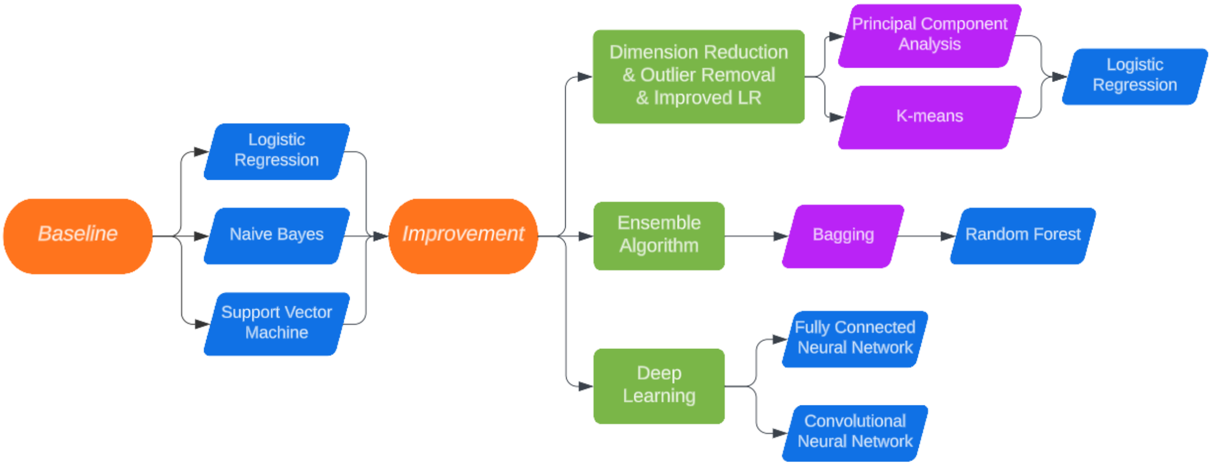


Figure 2: Learning Algorithms

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

LR estimates the probability of a binary outcome; NB is a generative algorithm that applies Bayes' theorem; SVM aims to find a hyperplane that maximally separates the data points using 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 partitions a dataset into a specified number of clusters, by minimizing the within-cluster sum of squared distances.

Ensemble Algorithm, Bagging, and Random Forest (RF)

Ensemble Algorithm is a technique that combines multiple individual models to improve overall prediction performance. Bagging is an ensemble method which generates multiple subsets through bootstrapping. 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. The hidden layers of a CNN typically are made up of convolutional, pooling, and fully connected layers.

Experiments, Results and Discussion

Rigorous implementation is realized using Python, scikit-learn and PyTorch, ensuring preprocessing, model training, and hyperparameter optimization; The evaluation is based on the confusion matrix.

Baseline and Random Forest

Performance of LR, NB & SVM is in **Table 2**. SVM (80.09%) outperforms the other two. Although both small-size dataset and normalized features bring a higher expectation on NB than LR, NB's 74.03% is only 0.44% higher than LR's 73.59%, which can be explained by some features' high correlation in **Figure 1**.

For baseline, the positive accuracy is ~15-20% lower than the negative one. To improve, we built 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%; The positive-negative gap shrunk to <10%. Therefore, RF is good to deal with the imbalance.

Algorithm	Test Size	Prevalence	Total Accuracy	Positive Accuracy	Negative Accuracy	Precision	Recall	F1-Score
Logistic Regression	231	34.63%	73.59%	61.25%	80.13%	62.03%	61.25%	61.64%
SVM	231	34.63%	80.09%	68.75%	86.09%	72.37%	68.75%	70.51%
Naïve Bayes	231	34.63%	74.03%	65.00%	78.81%	61.90%	65.00%	63.41%
Random Forest	231	34.63%	88.74%	82.50%	92.05%	84.62%	82.50%	83.54%

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. To pick the optimal number of principal components, LR was re-trained and compared, 7. 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 picked the highest-accuracy one. As a result, we removed 194 data points which were clustered incorrectly. We re-trained an improved LR based on the refined dataset, performance in **Table 3**. The total/positive/negative accuracy improved 22.36%/28.58%/18.99%. So, using PCA for dimensionality reduction and k-means for outlier removal improved the prediction accuracy.

Algorithm	Test Size	Prevalence	Total Accuracy	Positive Accuracy	Negative Accuracy	Precision	Recall	F1-Score
LR	231	34.63%	73.59%	61.25%	80.13%	62.03%	61.25%	61.64%
PCA-7 LR	231	34.63%	75.32%	60.00%	83.44%	65.75%	60.00%	62.75%
(PCA-7 & KM-574) LR	173	34.10%	95.95%	89.83%	99.12%	98.15%	89.83%	93.81%

Table 3: Results of Improved Logistic Regression

Deep Learning and Neural Network

Fully Connected Neural Network was trained by different number of hidden layers (2, 3, 4) and neurons (4 values), 12 neural networks in total. Validation set was used for hyperparameters' optimization, and optimal one was run on the testing set, evaluation in **Table 4**. 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.

We also added two convolutional (different parameters) and max pooling layers to a random neural network. **Table 4** shows that, with convolutional layers' parameters as (32, 64), we have the highest accuracy on the validation set. Applying this CNN to the testing set, 90.91% accuracy was achieved, which is 3.90% higher than the one without convolutional layers.

Algorithm	Test Size	Prevalence	Total Accuracy	Positive Accuracy	Negative Accuracy	Precision	Recall	F1-Score
2L (88, 50) - Valid	77	32.47%	83.12%	84.00%	82.69%	70.00%	84.00%	76.36%
2L (88, 50) - Test	77	38.96%	87.01%	80.00%	91.49%	85.71%	80.00%	82.76%
3L (64, 32, 40) - Valid	77	32.47%	83.12%	80.00%	84.62%	71.43%	80.00%	75.47%
3L (64, 32, 40) - Test	77	38.96%	85.71%	73.33%	93.62%	88.00%	73.33%	80.00%
4L (5, 3, 4, 2) - Valid	77	32.47%	84.42%	92.00%	80.77%	69.70%	92.00%	79.31%
4L (5, 3, 4, 2) - Test	77	38.96%	89.61%	90.00%	89.36%	84.38%	90.00%	87.10%
2L (10, 7) - Test	77	38.96%	87.01%	73.33%	95.74%	91.67%	73.33%	81.48%
2L (10, 7), CNN (8, 6) - Valid	77	32.47%	75.32%	80.00%	73.08%	58.82%	80.00%	67.80%
2L (10, 7), CNN (16, 32) - Valid	77	32.47%	70.13%	56.00%	76.92%	53.85%	56.00%	54.90%
2L (10, 7), CNN (32, 64) - Valid	77	38.96%	84.42%	76.67%	89.36%	82.14%	76.67%	79.31%
2L (10, 7), CNN (64, 88) - Valid	77	32.47%	76.62%	76.00%	76.92%	61.29%	76.00%	67.86%
2L (10, 7), CNN (32, 64) - Test	77	38.96%	90.91%	86.67%	93.62%	89.66%	86.67%	88.14%

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

Conclusion and Future Work

Based on this research, we conclude that the baseline algorithms (LR, NB & SVM) provide ~75% total accuracy, but with 15-20% positive-negative gap. The diabetes prediction accuracy was improved in three ways. 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 gap to <10%.

For the future work, k-fold cross validation will be applied since our research is 7/3 split. Also, neural networks with more different layers/neurons need to be compared. Finally, the dataset size is another concern, such as if the accuracy would be impacted change by the size of the dataset.

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.