Diabetes Mellitus Prediction Using Machine Learning Algorithms

Xinxie Wu, xinxiewu@stanford.edu

Predicting

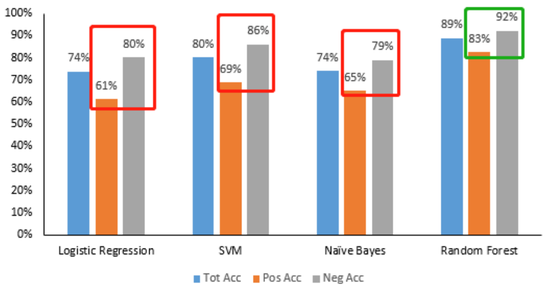
Diabetes Mellitus (DM) is a chronic disease with high blood sugar and doesn't have a permanent cure; hence early detection is required.
This research applies a **2-part methodology** into diabetes prediction, based on PID dataset.
In **baseline** approach, we train Logistic Regression, Naïve Bayes & SVM, with **total accuracy** of ~75% but **imbalanced gap**, 20%, between positive and negative.
In **improvement** approach, Random Forest shrinks the imbalance gap down to <10%; PCA/K-means refined the dataset and LR's retraining results in 95% accuracy; Multiple Layers' Neural Network is trained by changing hyperparameters and gets 89% accuracy, convolutional layer is added, and the accuracy is improved to 90.91%.

Baseline & Random Forest

In baseline, **SVM (80.09%)** performs the best, followed by Naïve Bayes (74.03%) and Logistic Regression (73.59%), all with negative-positive gap 15-20%.

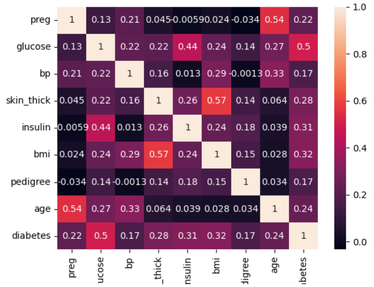
Although both **small-size dataset** and **normalized features** bring a higher expectation on NB than LR, but gap is small, which can be explained by **some features' high correlation**.

1,000 Random Forests are trained, the best-performance one is picked. Negative-positive gap shrinks to less than 10%.



Dataset & Features: Pima Indian Diabetes (PID)

Atttibute	Description	Missing Count	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

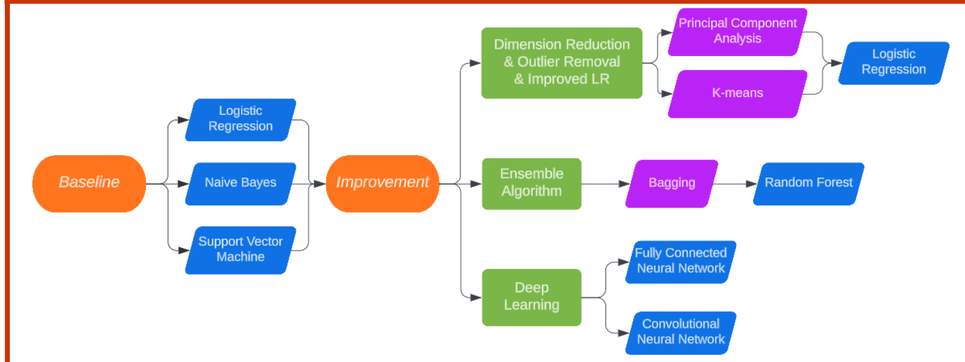


PID has 768 observations in total, 268 (34.9%) as diabetes. PID includes 8 numeric attributes covering the personal health status and medical examination results. All features are used and further selected by PCA & K-means.

EDA:

1. Missing Value – Median
2. Normalization
3. Correlation Matrix – Glucose 0.5
4. Training vs Testing: 70% / 30%

2-Part Methodology: Models & Workflow



Principal Component Analysis (7)

$$\underset{\phi_{11},\phi_{12},\ldots,\phi_{p1}}{\text{maximize}} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2$$

LR re-training determines the optimal number as 7

K-means (2)

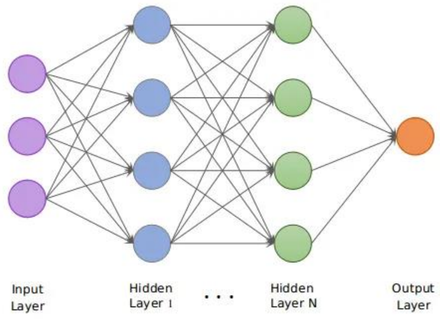
$$J = \sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}\left\|x^{(i)} - \mu_j\right\|^2$$

Neural Network

Dataset is split into training (80%), validation (10%) and testing (10%).

Neural networks, with different number of hidden layers (2, 3, 4) and neurons, are trained; best-performance one in validation is picked and test.

Convolutional layer is also added and tested.



Results & Discussion

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%
(PCA-7 & KM-574) LR	173	34.10%	95.95%	89.83%	99.12%	98.15%	89.83%	93.81%
MLP: 2L (88, 50) - Valid	77	32.47%	83.12%	84.00%	82.69%	70.00%	84.00%	76.36%
MLP: 3L (64, 32, 40) - Valid	77	32.47%	83.12%	80.00%	84.62%	71.43%	80.00%	75.47%
MLP: 4L (5, 3, 4, 2) - Valid	77	32.47%	84.42%	92.00%	80.77%	69.70%	92.00%	79.31%
MLP: 4L (5, 3, 4, 2) - Test	77	38.96%	89.61%	90.00%	89.36%	84.38%	90.00%	87.10%
MLP: 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%

Results & Discussion:

1. Baseline models show total accuracy ~75%, but 20% negative-positive gap;
2. Random Forest help to shrink the negative-positive gap down to <10%;
3. PCA returns an optimal number as 7, 2-cluster K-means removes 194 outliers. Improved LR reaches accuracy 95.95%;
4. 4-layers network performs best with 89.61%; 90.91% if convolutional layers are added.

Future Work

For the future work, **k-fold cross-validation** is under consideration since our research focused on 7/3 dataset split. Also, **neural networks** with more different number of layers/neurons 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 by the **size of the dataset**.

References

[1] Sharma, T., & Shah, M. (2021). A comprehensive review of machine learning techniques on diabetes detection.
[2] Aishwarya Mujumdar, and Dr. Vaidehi V. Diabetes Prediction using Machine Learning Algorithms, 2019.
[3] Tejas N. Joshi, and Prof. Pramila M. Chawan. Diabetes Prediction using Machine Learning Techniques, 2018.
[4] Jobeda Jamal Khanam, and Simon Y. Foo. A comparison of machine learning algorithms for diabetes prediction, 2021.
[5] [3] Jingyu Xue et al (2020). Research on Diabetes Prediction Method Based on Machine Learning.