# The Accuracy of KNN, decision tree, random forest, SVM, neural network, naive Bayes classifier and PLA for early Prediction of Diabetes

Hsieh Cheng-Han

October 2022

## abstract

This work compares the accuracy of some classifiers for early the prediction of diabetes. More specifically, the research compares the accuracy of k-nearest neighbors (KNN) algorithm, decision tree, random forest, support vector machine (SVM), neural network, naive Bayes classifier, and perceptron learning algorithm (PLA) on the prediction of diabetes, which the dataset is collected with eight features, times of pregnancy, concentration of glucose in blood, blood pressure, skin thickness, concentration of insulin in blood, body mass index (BMI), the value of diabetes pedigree function and age.

The result show that XXX is the most accurate on the prediction of diabetes.

## 1 Introduction

Diabetes is a chronic disease which may cause many complications. There're lots of reason that can put a person at the highly risk of having diabetes, such as age, obesity, lack of exercies, and more on. So many reasons interweave together making the manual prediction on diabetes is nearly impossible. However, lots of works [1] [2] [3] show that it is possible to have high accuracy by using machine learning techniques, such as random forest, K-means clustering, neural network, and so on.

By collecting the essential data of human body, prediction of diabetes can be turn into classification problem. Imagine that an individual case with essential data is a point in hyperspace, if it is closer to the cluster having diabetes, this case is more likely to have diabetes in the future, otherwise, this case is more likely healthy. But there are lots of machine learning techniques born to solve classification problem, it remains a problem that which technique having the hightest accuracy on the prediction of diabetes.

To find out which techniques is more suitable to predict diabetes, this work examines the diagnosis of diabetes using KNN algorithm, decision tree, random forest, SVM, neural network naive Bayes classifier, and PLA.

## 2 Related works

**k-nearest neighbors (KNN) classification algorithm:** The KNN classification algorithm is a supervised learning method which is first developed by Fix and Hodges [4]. The idea of KNN is based on the idiom, "birds of a feather flock together". By picking the $k$-nearest neighbors of a data point, the unkonwn class label can be determined. Lots of works [5] [6] [7] show the fact that KNN performs well for prediction of diabetes disease.

**decision tree:** Unlike KNN uses distance to determine the outcome, decision tree uses a sequence of deicsion that maximize the information gain, which can distinguish the class label of data as much as possible, to determine the outcome. Many works [8] [9] have applied deicsion tree method and gain a good accuracy. The advantage of decision tree is fast, easy to implement, and the decision is clear. But the disadvantage is that it is very likely overfitting and the structure of tree will become more complex with the more the class labels. To solve this problem, the

following techniques is developed:

**random forst:** Instead of a signle decision tree, random forest use lots of decision trees, which form a "forest". The decision trees are constructed by random subset of dataset. The key differs random forest from decision tree is that while decision trees consider all the possible feature splits, random forests only select a subset of those features, which reduce the risk of overfitting, bias, and overall variance. In [10] [11], random forest shows that it can greatly reduce the problem of over-fitting of the single decision tree, and gain an ever higher accuracy.

**support vector machine (SVM): neural network: naive Bayes classifier: perceptron learning algorithm (PLA):**

# 3 KNN

**The rough process of KNN is described as follow: suppose that there is a dataset which contain $N$ data point, denoted as $(X_i, Y_i)$ where $X_i$ is the features of the i-th individuals data and $Y_i$ is the class label of it. Now a data with unknown class label is given, denoted $(X, Y)$. By a preset distance function $d(P, Q)$, ordering the dataset as $(X_{(1)}, Y_{(1)}), (X_{(2)}, Y_{(2)}), \cdots, (X_{(N)}, Y_{(N)})$ where $d(X_{(1)}, X) \leq d(X_{(2)}, X) \leq \cdots \leq d(X_{(N)}, X)$. Pick the $k$-first class labels to determine the unknow class label, $Y$.**

# 4 decision tree

**Given a dataset, $D$, which contains $N$ data and class label, the construction of a decision tree can be described as below: suppose there are $M$ candidate decisions, denoted $f_i$. A decision can separate dataset $D$ into $m$ kinds, denote $D'_j$. The decision tree will adopt $\max_f G(D, f) = I(D) - \sum_{j=1}^{m} \frac{N'_j}{N} I(D'_j)$ as node decision, and then recursively construct the tree until the data in separated dataset have the same class label. When a data with unknown class label comes, a decision tree determines recursively by the decision node until the leaf node. The function of calculating information, $I$, can be various from implementation. The most famous two information function is entropy, and gini impurity. The formula of information entropy is $I_H(X) = -\sum_x p(x) \log_2 p(x)$ and gini impurity is $I_G(X) = 1 - \sum_x p(x)^2$.**

# 5 random forest

**Given the fact that a signle decision tree can be easily over-fitting, a technique called "random forest" is developed. By constructing $m$ decision trees with randomized subsets of training dataset, the different decision trees forms a "forest". The process that random forest determine a data with unknown class label can be outlined as follow: suppose for a coming data, $c_i$ decision trees in a random forest classify it as class $i$. Then the random forest will classify the data as class $\arg\max_i c_i$.**

# 6 SVM

# 7 neural network

# 8 Naive Bayes classifier

# 9 PLA

# 10 Experiment result

# 11 Conclusion

# References

**[1] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, 2019, 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920300557**

[2] T. Mahboob Alam, M. A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T. Imtiaz Baig, A. Hussain, M. A. Malik, M. M. Raza, S. Ibrar, and Z. Abbas, "A model for early prediction of diabetes," *Informatics in Medicine Unlocked*, vol. 16, p. 100204, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914819300176

[3] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Frontiers in Genetics*, vol. 9, 2018. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fgene.2018.00515

[4] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *International Statistical Review / Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989. [Online]. Available: http://www.jstor.org/stable/1403797

[5] M. NirmalaDevi, S. A. alias Balamurugan, and U. V. Swathi, "An amalgam knn to predict diabetes mellitus," in *2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN)*, 2013, pp. 691–695.

[6] D. Shetty, K. Rit, S. Shaikh, and N. Patil, "Diabetes disease prediction using data mining," in *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2017, pp. 1–5.

[7] V. Vijayan and A. Ravikumar, "Study of data mining algorithms for prediction and diagnosis of diabetes mellitus," *International journal of computer applications*, vol. 95, no. 17, 2014.

[8] A. A. Al Jarullah, "Decision tree discovery for the diagnosis of type ii diabetes," in *2011 International Conference on Innovations in Information Technology*, 2011, pp. 303–307.

[9] W. Chen, S. Chen, H. Zhang, and T. Wu, "A hybrid prediction model for type 2 diabetes using k-means and decision tree," in *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 2017, pp. 386–390.

[10] W. Xu, J. Zhang, Q. Zhang, and X. Wei, "Risk prediction of type ii diabetes based on random forest model," in *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, 2017, pp. 382–386.

[11] P. Palimkar, R. N. Shaw, and A. Ghosh, "Machine learning technique to prognosis diabetes disease: Random forest classifier approach," in *Advanced Computing and Intelligent Technologies*, M. Bianchini, V. Piuri, S. Das, and R. N. Shaw, Eds. Singapore: Springer Singapore, 2022, pp. 219–244.