Final Assignment Report

CSE-0408 Summer 2021

Name:Rafi Ahmed ID:UG02-47-18-030
Department of Computer Science and Engineering
State University of Bangladesh (SUB)
Dhaka, Bangladesh
Email Address: rafiahmed362@gmail.com

***Abstract*—This paper introduced for KNN problem and Decision Tree problem.**

n

***Index Terms*—Languages: Python.**

I. Introduction

KNN Algoritham:-

KNN is a lazy learning, non-parametric algorithm. It uses data with several classes to predict the classification of the new sample point. KNN is non-parametric since it doesn't make any assumptions on the data being studied, the model is distributed from the data

Decision tree algorithm:-

In a decision tree, the algorithm starts with a root node of a tree then compares the value of different attributes and follows the next branch until it reaches the end leaf node. Deficiency of adenosine deaminase type 2 (DADA2) is an autosomal recessive systemic autoinflammatory disorder (SAID) described for the first time in 2014 [1, 2]. Both homozygous or compound heterozygous genotypes have been detected [3]. Although one mutation c.139G¿A;p.(Gly47Arg) is frequent, notably in the Georgian population, due to a founder effect, the disease seems to occur ubiquitously; indeed, patients with DADA2 have been identified in several countries [4].

II. Literature Review

KNN:-

Author's introduction: Zhongheng Zhang, MMed. Department of Critical Care Medicine, Jinhua Municipal Central Hospital, Jinhua Hospital of Zhejiang University. Dr. Zhongheng Zhang is a fellow physician of the Jinhua Municipal Central Hospital. He graduated from School of Medicine, Zhejiang University in 2009, receiving Master Degree. He has published more than 35 academic papers (science citation indexed) that have been cited for over 200 times. He has been appointed as reviewer for 10 journals, including Journal of Cardiovascular Medicine, Hemodialysis International, Journal of Translational Medicine, Critical Care, International Journal of Clinical Practice, Journal of Critical Care.

Decision tree algorithm:-

Angel Insua, Alberto Monje, Hom-Lay Wang, Marita Inglehart, Patient-Centered Perspectives and Understanding of Peri-Implantitis, Journal of Periodontology, 10.1902/jop.2017.160796, 88, 11, (1153-1162), (2017). Wiley Online Library Nima D. Sarmast,Howard H. Wang,Nikolaos K. Soldatos,Nikola Angelov,Samuel Dorn,Raymond Yukna,Vincent J. Iacono, First published: 01

III. Proposed Methodology

When we classified a data set including large number of unlabeled data, if only utilize the few training examples available, then we can't obtain a high accuracy classifier with inadequate training examples; if we want to obtain a classifier of high performance, then labeling the unlabeled data is necessary, but labeling vast unlabeled data wastes time and consumes strength. In this paper, we propose a novel method which uses SVM cooperated with KNN for classification based on semi supervised learning theory. The general model is depicted as above (See Figure 2). To begin with, we construct a weaker classifier SVM according to the few training examples available, then using the weaker SVM classifies the remaining large number of unlabeled data in the data set, picking out n examples belonging to each class around the decision boundary by calculating Euclidean distance in the feature space, because the examples located around the boundary are easy to be misclassified, but they are likely to be the support vectors, we call them boundary vectors, so picking out these boundary vectors whose labels are fuzzy labeled by the weaker classifier SVM. Secondly we recognize these boundary vectors as testing set while recognize initial training examples as training set, use KNN method to classify them and recognize the results as the labels for boundary vectors. In the end, we put these boundary vectors and their labels into initial training set to enlarge the number of the training examples then retrain a SVM, iteratively until the number of the training examples is m times of the whole data set. The experimental results on three UCI data sets indicate that the final classifier SVM has significant improvement on accuracy

IV. ADVANTAGES and DISADVANTAGE FOR KNN

ADVANTAGES:-

1.Simple to implement and intuitive to understand 2.Can learn non-linear decision boundaries when used for classfication and regression. Can came up with a highly flexible decision boundary adjusting the value of K. DISADVANTAGE:-

1.Performance Issue with large data-set: The time required to calculate the distance between the new point and each existing points is huge. Which then degrades the performance of the algorithm. 2.Value of K: It is really crucial to determine what value to assign to k. with different value of K you get different results

V. ADVANTAGES AND DISADVANTAGE FOR DECISION TREE

ADVANTAGES :-

1.When using Decision tree algorithm it is not necessary to normalize the data.

2.Decision tree algorithm implementation can be done without scaling the data as well.

3.When using Decision tree algorithm it is not necessary to impute the missing values.

4.The data pre-processing step for decision trees requires less code and analysis.

5.The data pre-processing step for decision trees requires less time

DISADVANTAGE:-

1.The mathematical calculation of decision tree mostly require more memory.

2.The mathematical calculation of decision tree mostly require more time.

3.The space and time complexity of decision tree model is relatively higher.

4.Decision tree model training time is relatively more as complexity is

VI. OUTPUTS

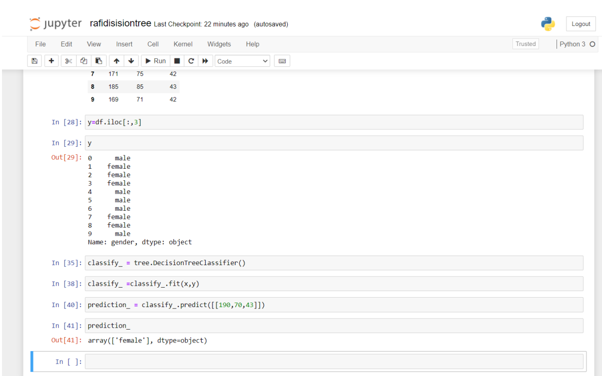

Fig. 1. Output of Decision tree.

VII. CONCLUSION AND FUTURE WORK

KNN:-

The article introduces some basic ideas underlying the kNN algorithm. The dataset should be prepared before running the knn() function in R. After prediction of outcome with kNN algorithm, the diagnostic performance of the model should be checked. Average accuracy is the most widely used statistic to reflect the performance kNN algorithm.

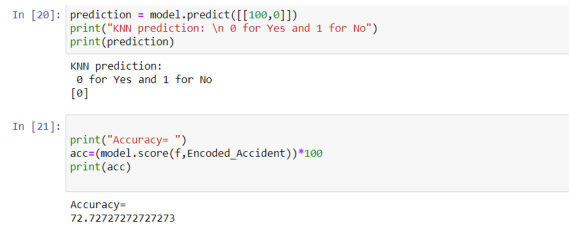

Fig. 2. Output of KNN.

Factors such as k value, distance calculation and choice of appropriate predictors all have significant impact on the model performance.

DECISION TREE:-

We report a large series of patients referred to us for genetic diagnosis of DADA2. We used information provided by the ordering clinicians to (1) describe the population with suspected DADA2, (2) compare our patients to those previously reported and (3) try to delineate prerequisites for a positive genetic diagnosis. We identified 13 patients carrying recessively inherited mutations in ADA2 that were predicted to be deleterious. Eight patients were compound heterozygous for mutations. Seven mutations were novel (4 missense variants, 2 predicted to affect mRNA splicing and 1 frameshift). Phenotypic manifestations included fever, vasculitis and neurological disorders. Prerequisites for quick and low-cost Sanger analysis included one typical cutaneous or neurological sign, one marker of inflammation (fever or elevated CRP level), and recurrent or chronic attacks in adults

ACKNOWLEDGMENT

I would like to thank my honourable **Khan Md. Hasib Sir** for his time, generosity and critical insights into this project.

REFERENCES

[1] [1] Zhou Q, Yang D, Ombrello AK, et al. Early-onset stroke and vasculopathy associated with mutations in ADA2. N Engl J Med. 2014;370:911–20. [2] Navon Elkan P, Pierce SB, Segel R, et al. Mutant adenosine deaminase 2 in a polyarteritis nodosa vasculopathy. N Engl J Med. 2014;370:921–31. [3] Oda H, Kastner DL. Genomics, biology, and human illness: advances in the monogenic autoinflammatory diseases. Rheum Dis Clin N Am. 2017;43:327–45. [4] Gonzalez Santiago TM, Zavialov A, Saarela J, et al. Dermatologic features of ADA2 deficiency in cutaneous polyarteritis nodosa. JAMA Dermatol. 2015;151:1230–4. [5] Short RD, Fukunaga K. The optimal distance measure for nearest neighbor classification. IEEE Transactions on Information Theory 1981;27:622-7. 10.1109/TIT.1981.1056403 [CrossRef] [Google Scholar] [6] Weinberger KQ, Saul LK. Distance metric learning for large margin nearest neighbor classification. The Journal of Machine Learning Research 2009;10:207-44. [Google Scholar] [7] Cost S, Salzberg S. A weighted nearest neighbor algorithm for learning with symbolic features. Machine Learning 1993;10:57-78. 10.1007/BF00993481 [CrossRef] [Google Scholar] X

VIII. CODE PART:

IX.

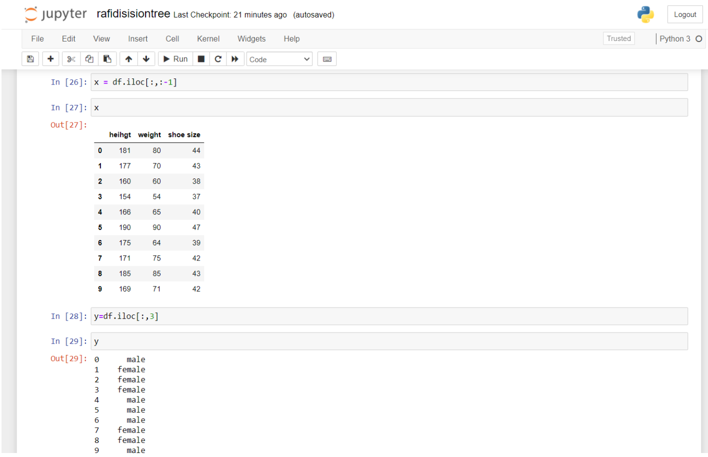

Fig. 3. 1.

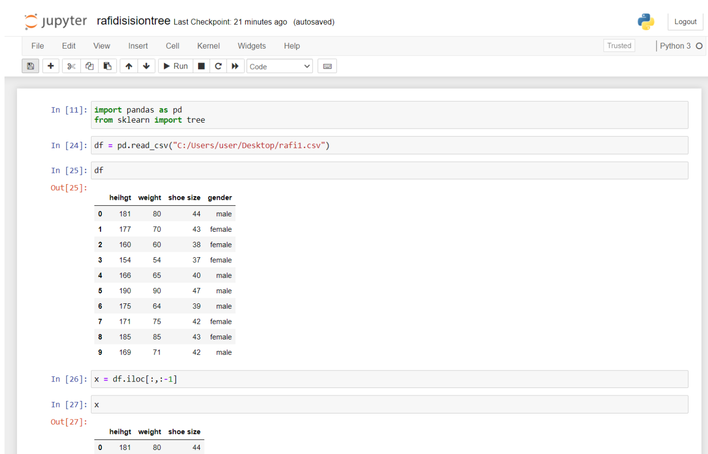

Fig. 4. 2.

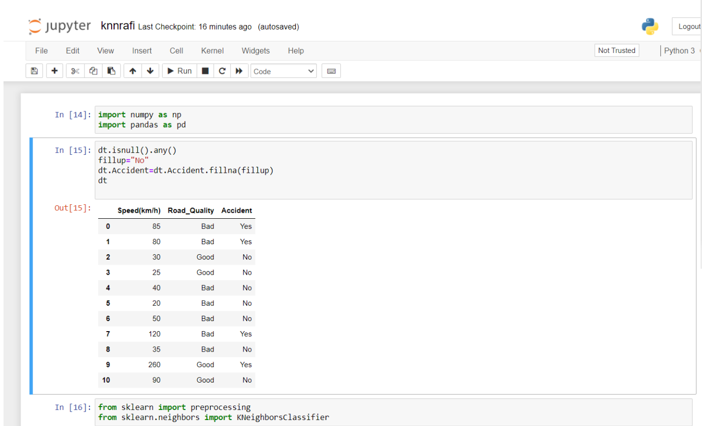

Fig. 5. 3.

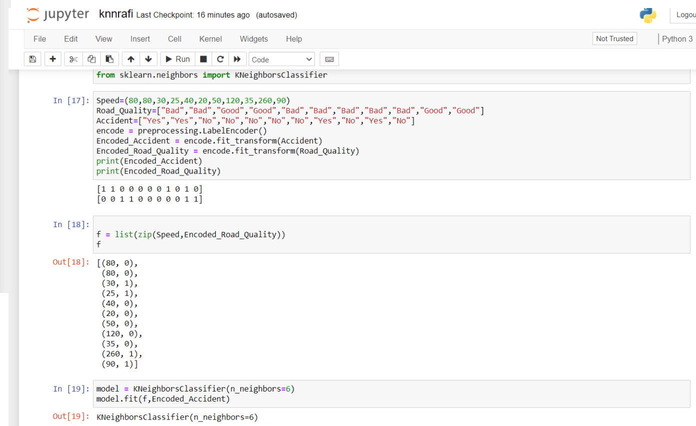

Fig. 6. 4.

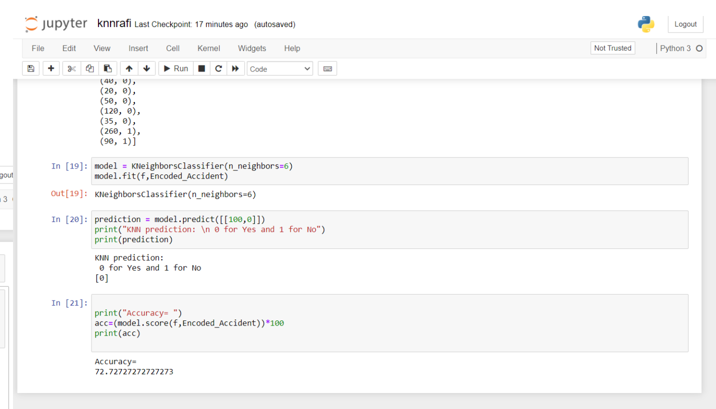

Fig. 7. 5.