MACHINE LEARNING AND PARKINSONS DATA

Aashaka Desai, Jack Maloy, Matthew Miller

INTRODUCTION

The ability to predict if someone has a disease given the values of various attributes is something that can greatly impact people's lives and society as a whole. Catching illnesses and diseases early is essential in ensuring the best possible outcome. Pouring over lots of data and trying to learn and then predict outcomes is exactly where machine learning can come into play. We are trying to apply this idea to Parkinson's data - a disease that 60,000 Americans are diagnosed with every year.

Parkinson's Disease is a degenerative neurological disorder affecting the central nervous system, thereby affecting movement and speech. It is often marked with tremors, muscular rigidity and slow imprecise movement. It is a result of degeneration of cells in the substantia nigra, which produce dopamine - a neurotransmitter essential for initiating movement. Some treatments allow dopamine to be synthesized from ingested compounds like levodopa and carbidopa. These treatments are most effective in the early stages of the disease but early diagnosis of Parkinson's is incredibly difficult. The motor symptoms that mark its presence can be similar to symptoms of other neurological disorders. We hope to use machine learning to aid early diagnosis and treatment of Parkinson's.

Another characteristic symptom of Parkinson's is vocal changes - a breathy, hoarse and soft voice that is result of change in CNS control of vocal folds. The fundamental frequency of a voice, closely related to vocal pitch, is caused by vibration of vocal folds. Previous research has shown that the variability of fundamental frequency is lower in patients with Parkinson's Disease than the control groups. Gender and current treatment plan (levodopa) can also affect the variability of fundamental frequency. The dataset we used for this project consists of data from 31 people, 23 with Parkinson's. The columns consists of different voice measures like average vocal fundamental frequency, minimum vocal fundamental frequency, maximum vocal fundamental frequency and health status (1 for Parkinson's, 0 for healthy). The rows corresponded to different voice recordings - each participant gave approximately 6 recordings.

With this dataset, we are trying to uncover the exact relationship between these various voice measures and a diagnosis of Parkinson's Disease. We use Gaussian Naive Bayes classifier, K-Nearest-Neighbor classifier and Decision Tree classifier as algorithms, outputting the accuracy for each - trying to find the one that is the most efficient and accurate.

METHODOLOGY

Since we have a single set of Parkinson's data, we will be splitting the dataset up into a training portion and a testing portion, where 80% of the data will be used for training and 20% will be used for testing. The url for the dataset is taken as input, and then various machine learning algorithms are run with their accuracies printed to the screen. Afterward, we will compare accuracies of the various algorithms, along with their runtimes. We will use a weighted decision matrix to come to a conclusion.

RESULTS

First we will use Figure 3 to determine which value of K is best for this dataset when using the K Nearest Neighbor (KNN) algorithm, and will use that K in our further analysis. We will analyze our results through the accuracy of the algorithm on the testing portion of the data as well as the amount of time it takes for each of these algorithms. Finally, we will combine both of these metrics and use a weighted decision matrix to see which of these algorithms performs best on this parkinsons dataset, with our given weights.

When reviewing Figure 3, you can see that the highest accuracy is achieved for the K value of 3 and 5. We chose to precede using the value of K = 3. All K's tested had a negligible difference in running time and thus that was not be taken into consideration during the analysis. So, for now on, the KNN algorithm will be run with a K value equal of 3.

We can see in Figure 1, that the K Nearest Neighbor algorithm and the Decision Tree algorithm have the same accuracy, with the Naive Bayes algorithm having significantly less accuracy. When analyzing the average times of each algorithm, by viewing Figure 2, we see that the Naive Bayes is the fastest on average. The Decision Tree algorithm was the next fastest, beating out the KNN algorithm, with which it was tied in accuracy. To be more precise and we

will now perform a weighted decision matrix to choose which algorithm is 'best' when taking both average time and accuracy into consideration. The one with the largest total will be the algorithm that is chosen as the best.

Algorithm	Average Time	Time Weight	Accuracy	Accuracy Weight	***Total***
Naive Bayes	0.0029969	-10	0.76923	2	***1.509***
KNN	0.0042180	-10	0.92308	2	***1.804***
Decision Tree	0.0032596	-10	0.92308	2	***1.812***

Because the time is such a small number, we chose a larger weight so that it would affect the outcome a reasonable amount. The total is calculated by summing each attribute by its weight, i.e. $Total = (AvgTime * -10) + (Acc * 2)$. A smaller runtime indicated better performance, and thus should be rewarded in our calculation. This is why the weight for time must be negative. Adding smaller negative numbers will make the total score larger than adding larger negative numbers.

From this chart we can see that the Decision Tree Algorithm produced the largest output in the weighted decision matrix, making it the chosen output with the weights that we have selected. Although the Naive Bayes algorithm was the fastest, it was also the least accurate, and since we valued accuracy more by virtue of being a larger number that dominated the summation, it's total was last. The choice of the decision matrix makes sense, as KNN and Decision Tree both have the same accuracy, but Decision Tree is slightly faster on average. Therefore it should have the larger total in the decision matrix, which it does.

FUTURE WORK

Some shortcomings to our current method is that one specific dataset is used. Adding the ability to generalize the input would make our program more powerful and be able to be used in a wider context than it is right now. Something else we could do in the future would be to use

cross-validation to test our algorithms instead of splitting them into a training set and a testing set. This would more closely track the accuracy of prediction on more generalized data.

FIGURES

Figure 1: Accuracy Comparison on Testing Data

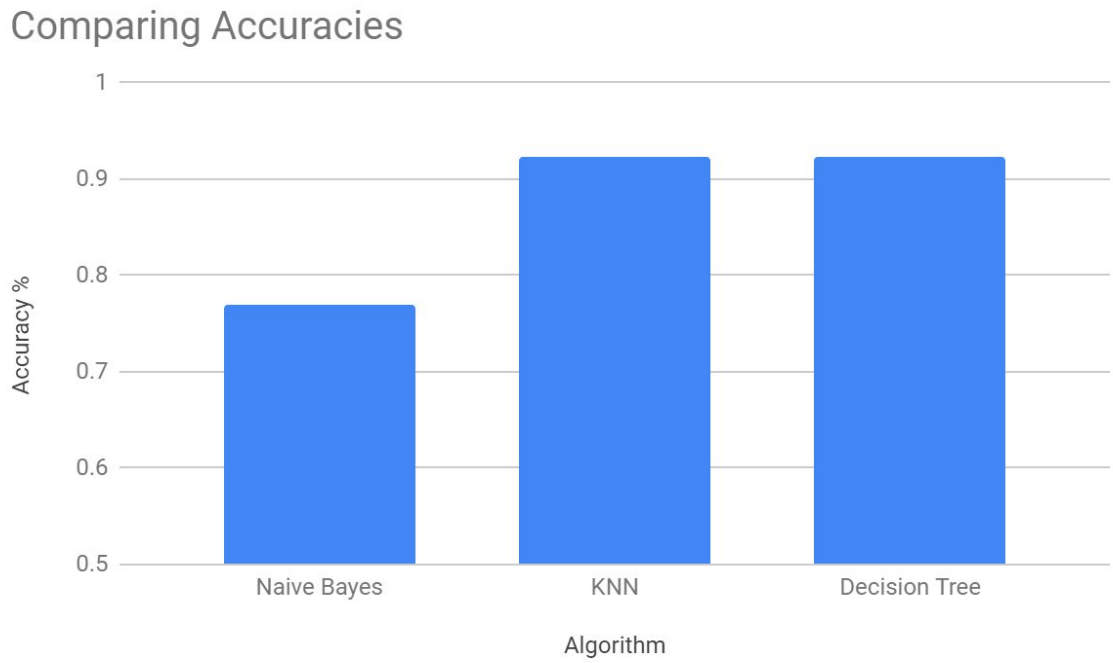


Figure 2: Average Run Times of each algorithm

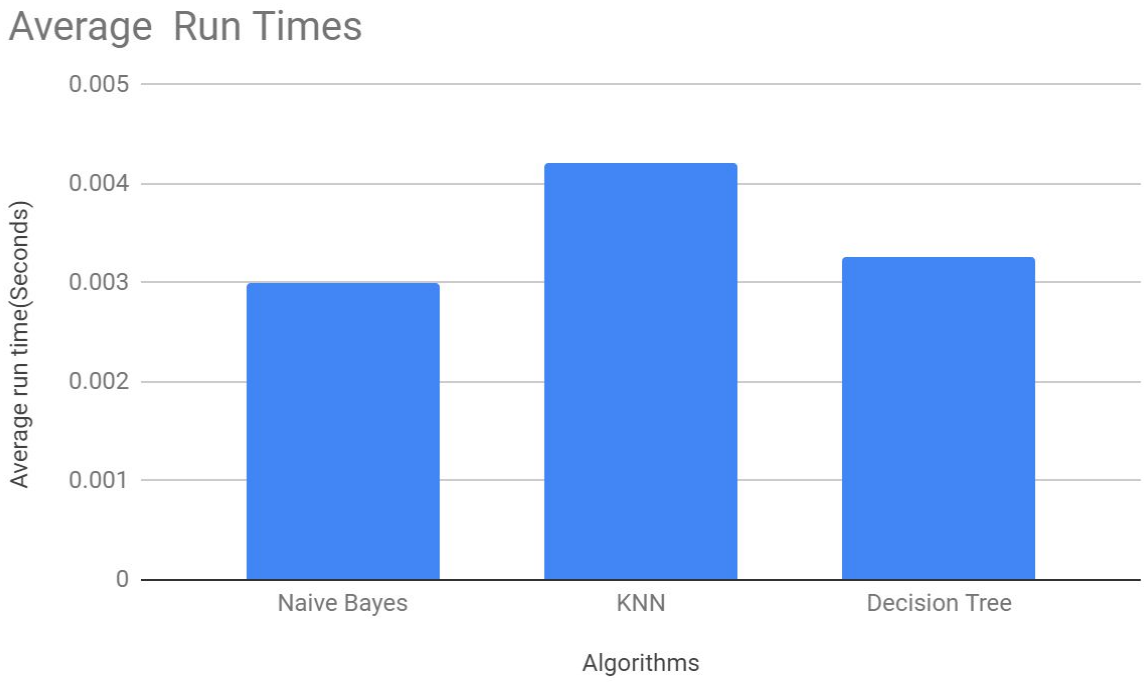
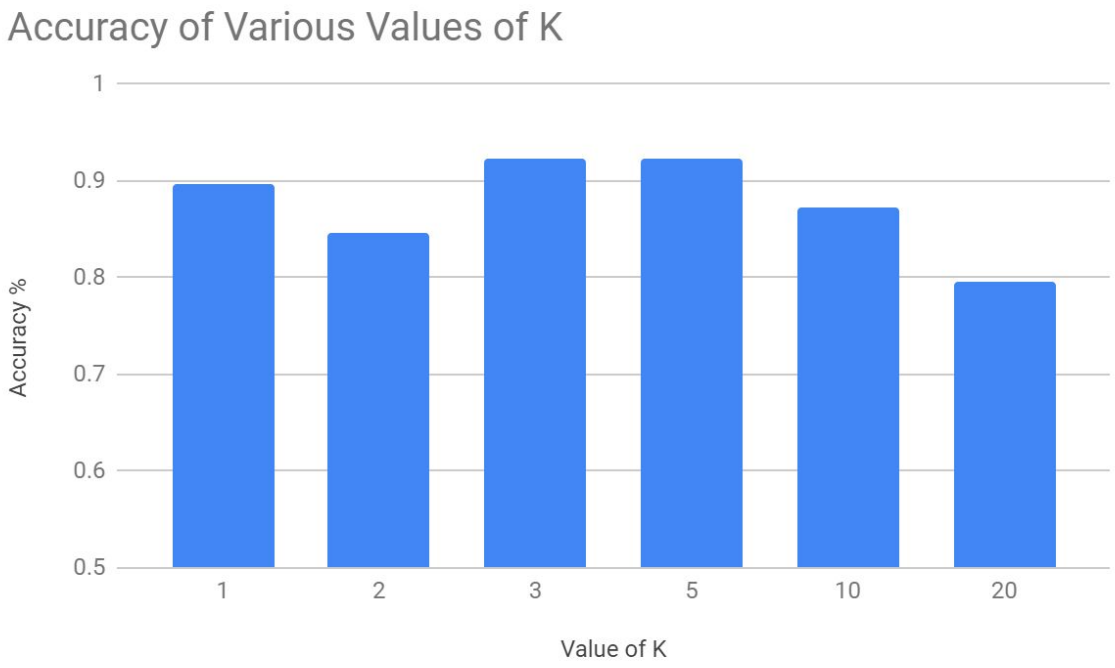


Figure 3: Accuracy of Various Values of K



BIBLIOGRAPHY

Bowen, Leah K et al. “Effects of Parkinson's Disease on Fundamental Frequency Variability in Running Speech.” *Journal of medical speech-language pathology* vol. 21,3 (2013): 235-244.

V.S Sriram, Tarigoppula & Venkateswara Rao, M & V Satya Narayana, G & Pandu Ranga Vital, T & Dowluru, Kaladhar SVGK. (2013). Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms. IJEIT. 3. 212-215.

Nooritawati Md. Tahir and Hany Hazfiza Manap, 2012. Parkinson Disease Gait Classification based on Machine Learning Approach. *Journal of Applied Sciences, 12: 180-185.*

“Diagnosis - Early Symptoms & Early Diagnosis.” *ParkinsonsDisease.net*, parkinsonsdisease.net/diagnosis/early-symptoms-signs/.

UCI Machine Learning Repository: Parkinsons Data Set, archive.ics.uci.edu/ml/datasets/parkinsons.

“Statistics.” *Parkinso n's Foundation*, 28 Mar. 2019, parkinson.org/Understanding-Parkinsons/Statistics.