Machine Learning University of Central Arkansas, Conway AR Dr. Olcay Kursun

**Parkinson’s Disease Prediction & Classification**

***Kima Brown2***

2  Computer Science Department UCA, *kbrown52@cub.uca.edu*

**Abstract:** Parkinson’s Disease poses a serious risk to mostly older adults and in very rare cases is diagnosed in people under the age of 60. In the United States alone 60,000 new case are diagnosed each year and 4 to 6.5 million world-wide in developing countries all over the world. The use of machine learning and data mining techniques has revolutionized the whole process of predicting the presence of PD in gait stride with wearable foot devices. In this work, we evaluate and investigate selected classification algorithms and clustering using Waikato Environment for Knowledge Analysis (WEKA for short) by inputting data from a study conducted by Physio Bank a biomedical community. WEKA is a data mining software primarily used for academic and research purpose and is opened sourced. The algorithms tested are Naïve Bayes, SMO, Multilayer Perceptron, Random SubSpace, Decision Table, J48, Random Forest and also clustered Simple K Means. Experimental results show that Random Forest proves to be the best algorithm with highest accuracy of 99.9752%.

**Keywords:** Parkinson’s Disease, Data mining, Prediction, Classification, WEKA

**1. Introduction**

Data mining is sometimes called data or knowledge discovery, is the process of analyzing data from multiple perspectives and summarizing it into useful information. It is the process of exploring new patterns from large datasets involving methods from statistics, machine learning and artificial intelligence and also database management. Data mining approach is actually part of the knowledge discovery process. Data mining is used today by companies with a strong consumer focus in various domains such as medical, healthcare, higher education, telecommunication etc. There are several data mining functions such as Association Rules, Classification, Concept Descriptions, Prediction, Clustering and Sequence discovery to find the useful patterns

**2. Causes of Parkinson’s Disease**

Parkinson's disease (PD) belongs to a group of conditions called motor system disorders, which are the result of the loss of dopamine-producing brain cells. The four primary symptoms of PD are tremor, or trembling in hands, arms, legs, jaw, and face; stiffness of the limbs and trunk; slowness of movement; and postural instability, or impaired balance and coordination. Gait disorders are considerable cause of falls in patients with neurological diseases. Grasping these disorders allows prevention and better awareness into underlying diseases. Mobile and wearable sensor technologies can serve as affordable and practical tools to assess the severity of neurological conditions. As these symptoms become more pronounced, patients may have difficulty walking, talking, or completing other simple tasks.

**3. Methodology**

This study is based on data taken from PhysioBank study and it was run against seven algorithms to predict the likelihood of a PD patient gait stride verse a healthy control group subject. PD usually affects people over the age of 60. In some people the disease progresses more quickly than in others. As the disease progresses, the shaking, or tremor, which affects most people with PD may begin to interfere with daily activities. There are currently no blood or laboratory tests that have been proven to help in diagnosing PD. Therefore, the diagnosis is based on medical history and a neurological examination. The disease can be difficult to diagnose accurately. Doctors may sometimes request brain scans or laboratory tests to rule out other diseases. In the study, we are reviewing a look at a patience’s swing stride as it relates to the control group without the disease to help determine if the patience with PD show a significant difference in walking ability. In this study, we analyze gait-related measurement signals on the study of Parkinson’s disease (PD). The dataset is obtained from PhysioBank, a digital recording archive bank of physiologic signals related data for use by the biomedical research community. PD affects approximately 1 million Americans (estimates range between 4 to 6.5 million people worldwide) and about 1% of older adults. The database contains 93 patients with PD and 73 healthy controls. The database includes the vertical ground reaction force records of subjects as they walked at their usual, self-selected pace for approximately 2 minutes on level ground. Underneath each foot were 8 sensors that measure force as a function of time sampled at 100 Hz. We apply deep learning techniques (LSTM, auto-encoders, transfer learning) to classify subjects as healthy or with Parkinsonism. These tools can help automate monitoring the debilitating gait symptoms of the patients with neurological diseases and thus help predict patients with severe gait disturbances that make them prone to falls. Minimizing the cost and the weight of the sensors to be worn makes the approach more practical for the patients. The prognosis: at present, there is no cure for PD, but a variety of medications provide dramatic relief from the symptoms. In some cases, surgery may be appropriate if the disease doesn't respond to drugs. A therapy called deep brain stimulation (DBS) has now been approved by the U.S. Food and Drug Administration.

**4. Algorithm Description**

1. **Naive Bayes:** Naive Bayes is an algorithm that uses Bayes theorem to classify data. Bayes theorem uses probability theory to classify the data. As more data is introduced the probability of an event is adjusted.
2. **SMO:** Sequential Minimal Optimization, or SMO, is an algorithm that solves quadratic programming problems that arise during training. It’s widely used for support vector machines and implemented by the LIBSVM tool.
3. **Multilayer Perceptron:** MLP entails at least three layers of nodes: an input layer, a hidden layer and an output layer. Excluding the input nodes each node is a neuron that inhabits a nonlinear activation function. It can differentiate data that is not linearly detachable.

1. **Random SubSpace:** Random SubSpace is an ensemble learning method that reduces correlation between estimators in a training set by training them on random samples. This method is similar to bagging except for the random sampling.
2. **Decision Table:** Are succinct visual examples for specifying what actions to implement depending on said conditions. They are algorithms whose output is a set of occurrences. The data expressed in decision tables can also be interpreted as decision trees and in programming a if-then-else or switch-case statements.
3. **J48:** J48 is a decision tree algorithm that takes training data and builds a predictive model. This model is mapped to a tree structure to achieve classification with minimal numbers of decisions.
4. **Random Forest:** Random Forest is a decision tree model that constructs sets of random decision trees. These trees allow for different applied variables to be studied.
5. **Simple K-Means:** This clustering algorithm has a loose relation to k-nearest neighbor classifier popularly used in machine learning technique for classification that is often misrepresented for k-means because of the name. Relating the 1-nearest neighbor classifier to the cluster centers obtained by k-means classifies new data into present clusters. This process is branded as nearest centroid classifier or Rocchio algorithm.

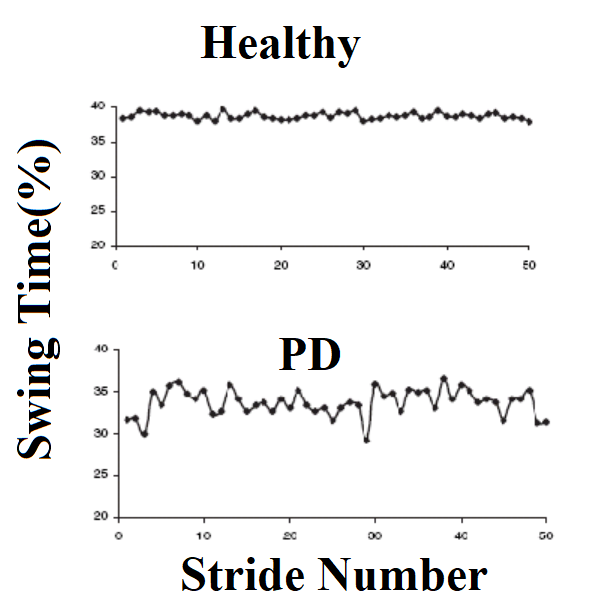
**5. Data Description**

This dataset combines one instance of a PD patient and a control group subject total left foot and total right foot gait stride with 8 sensors under each food recorded for two minutes of walking time. Each PD and control group subject has 12119 rows or (a.k.a. seconds) totaling 24238 rows of data. I also added a flag of ‘Yes’ for PD patient and ‘No’ for Control group subject. It’s import to note we are only comparing one instance of PD two minute walk and one instance of control group subject two minute walk and not all 93 PD patients and not all 73 control group subjects. There are three attributes with characteristics of real, integers and floating point numbers. This dataset is derived from a 11/14/2004 study and was conducted on men only.

1. **Total\_Left\_Foot**: Time recorded from 0 to 12119 seconds roughly two minutes as a participant walked a normal pace for two mines on level ground from sensors on his left foot.

1. **Total\_Right\_Foot**: Time recorded from 0 to 12119 seconds roughly two minutes as a participant walked a normal pace for two mines on level ground from sensors on his right foot.
2. **PD-Yes-Control-No:** This is basically a flag to separate the PD patients from the control group subjects. Yes means they have PD and No means they were a control group subject.

1. **Results of Study Data Processed in WEKA:** Results are based on if there was a noticeable difference in gait stride among PD patients verses control group subjects. This was determined by graphing the results on the below charts as well as running the data on algorithms in WEKA the results provides better predictions, accuracy and classification of the data**.**



**` WEKA**

WEKA is the tool used to carry out experiments and implementations. WEKA (Waikato Environment for Knowle0dge Analysis) is a data mining tool written in java and developed at Waikato. This open sourced data mining software is mainly used for academic and research purposes. The data file is in ARFF file format, data is inputted into the program known as attributes. Figure 1 is all the attributes used in the dataset.

**Interfaces**

Below are the interfaces after the dataset has been processed. The interfaces show a tree from the J48 instance and the performance output of each algorithm. The performance is evaluated based on three criteria, the prediction accuracy and the time taken to build the model.

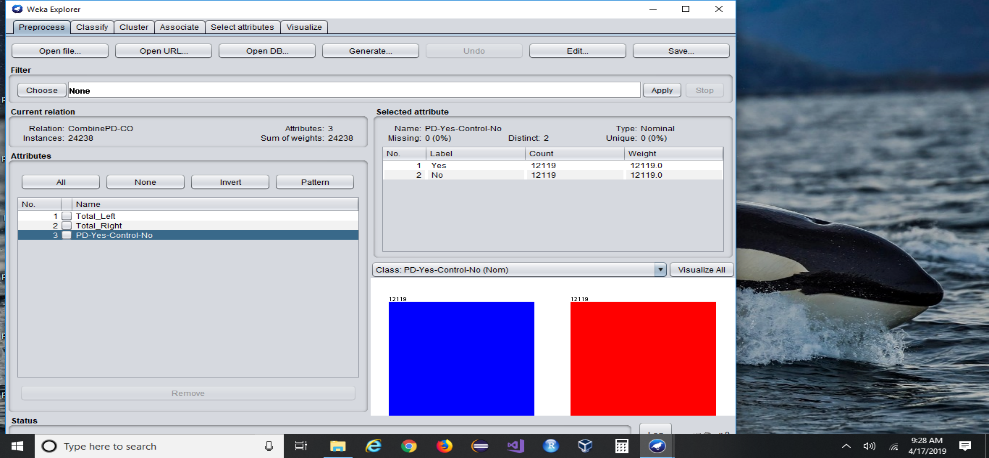


Fig 1a: Attributes used in data set

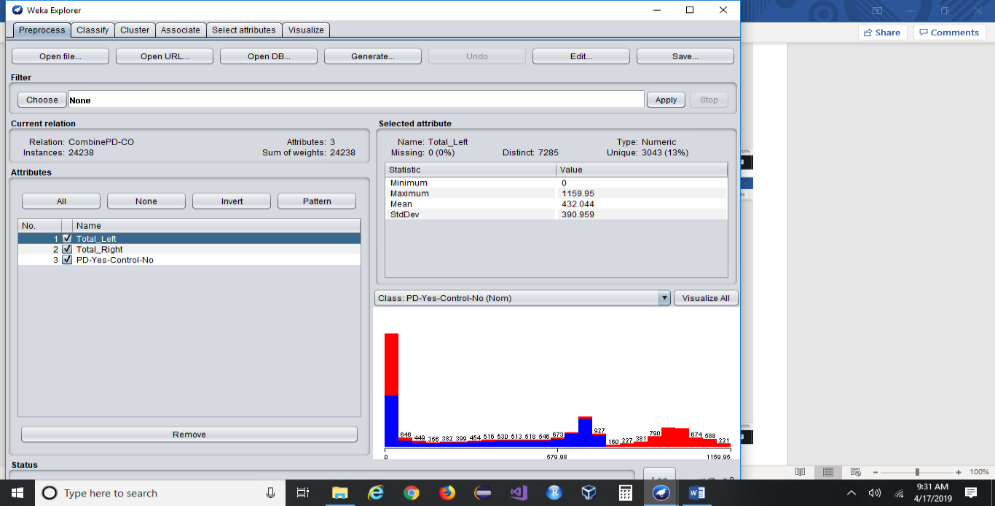


Fig 1b: Attributes used in data set all selected

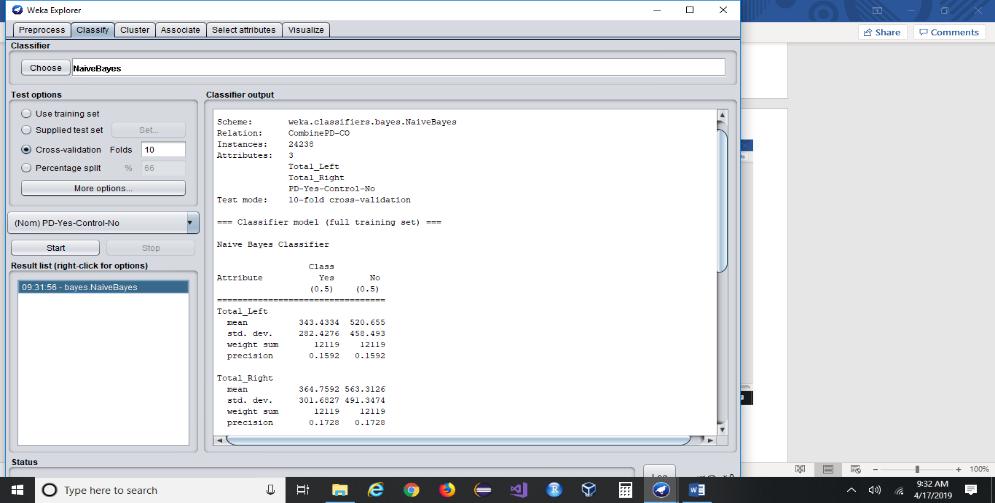


Fig 2a: Naive Bayes dataset output

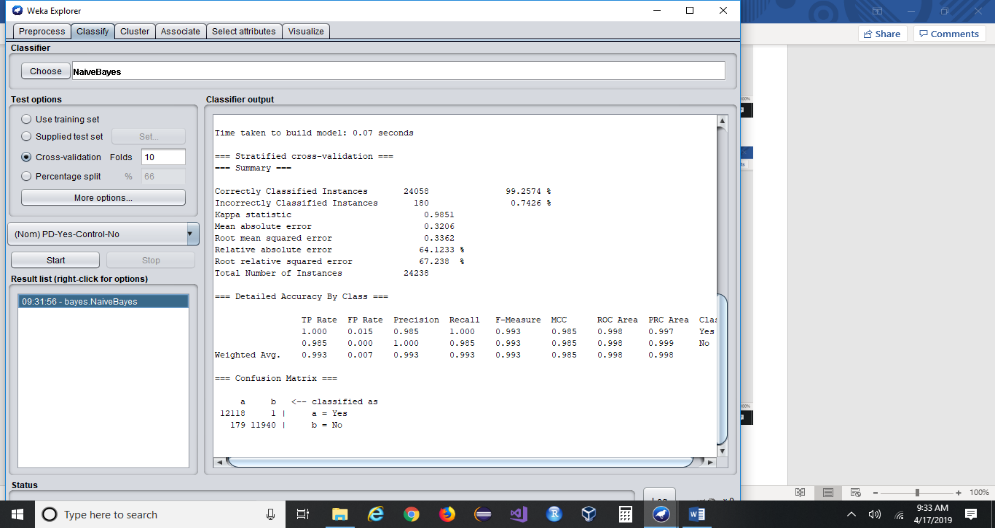


Fig 2b: Naive Bayes dataset output

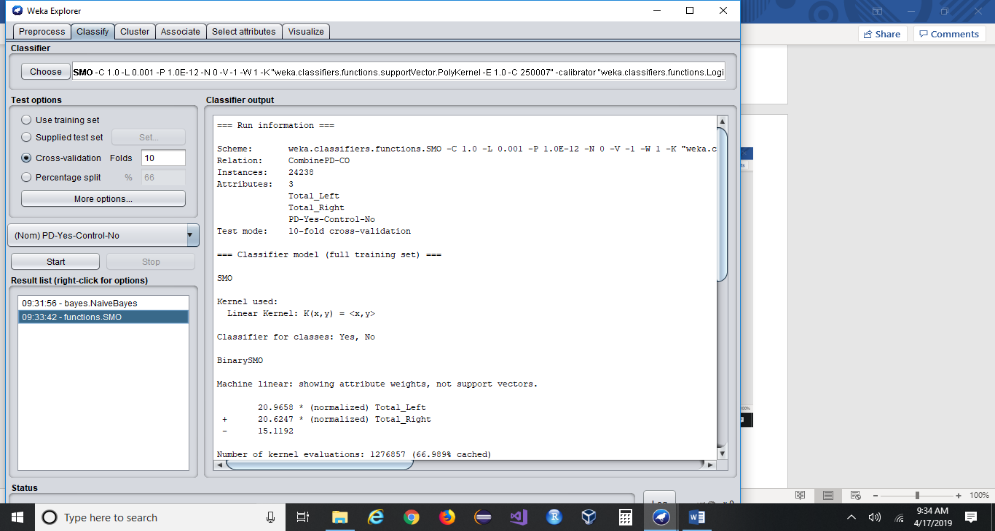


Fig 3a: SMO dataset output

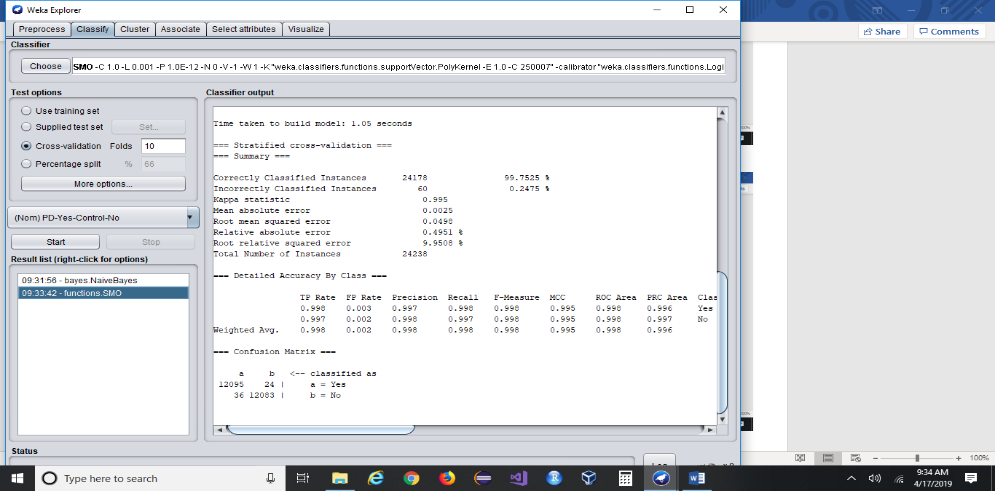


Fig 3b: SMO dataset output

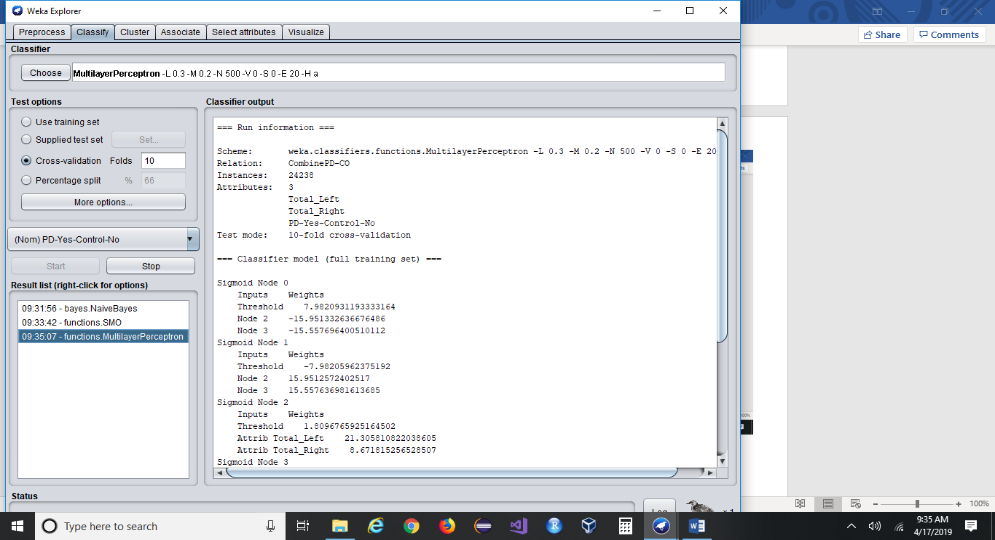


Fig 4a: Multilayer Perceptron dataset output

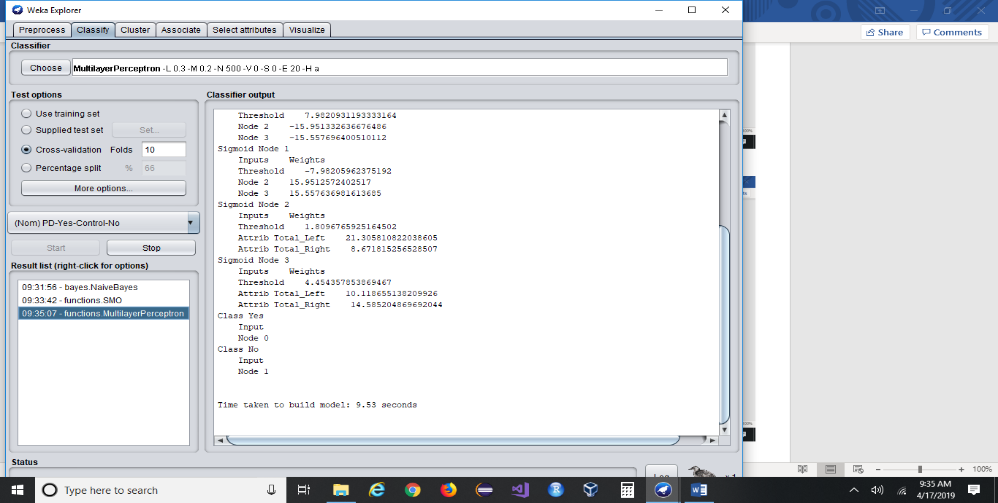


Fig 4b: Multilayer Perceptron dataset output

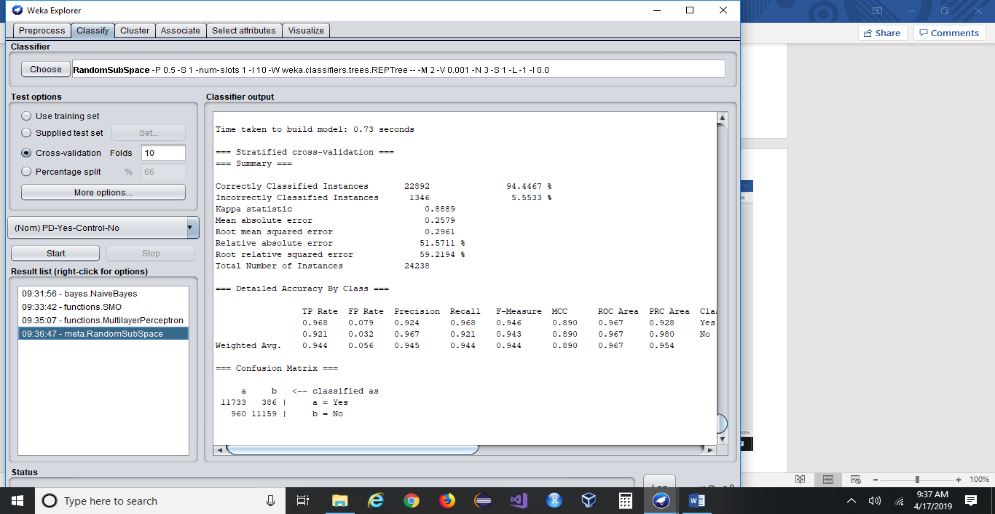


Fig 5: Random SubSpace dataset output

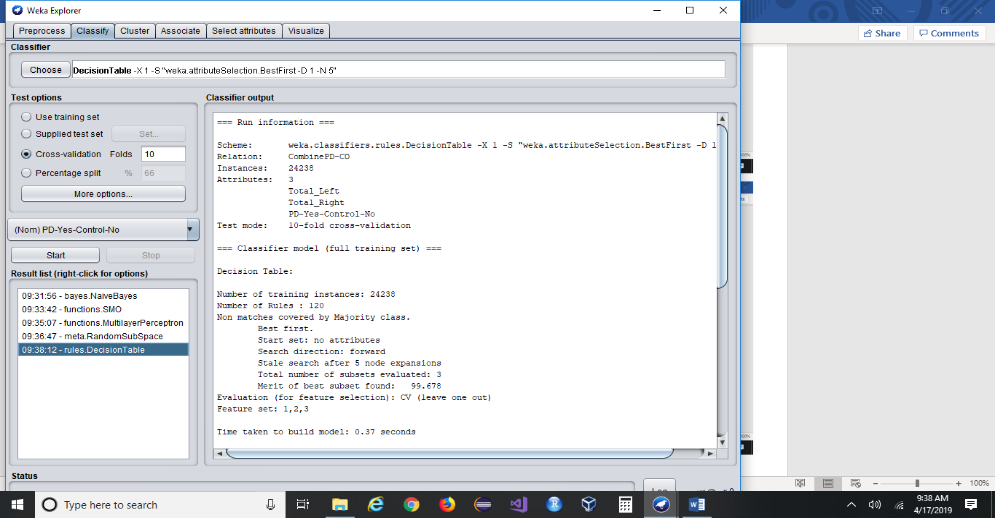


Fig 6a: Decision Table dataset output

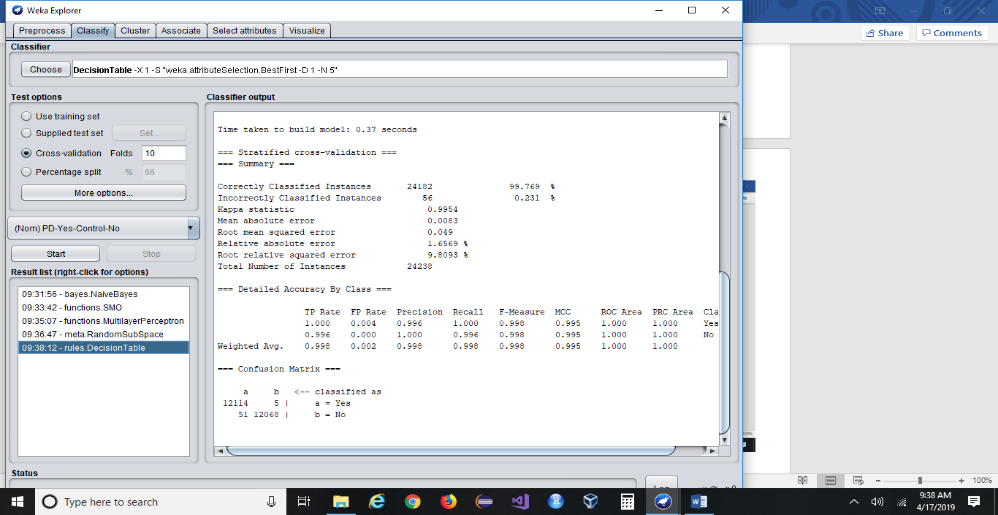


Fig 6b: Decision Table dataset output

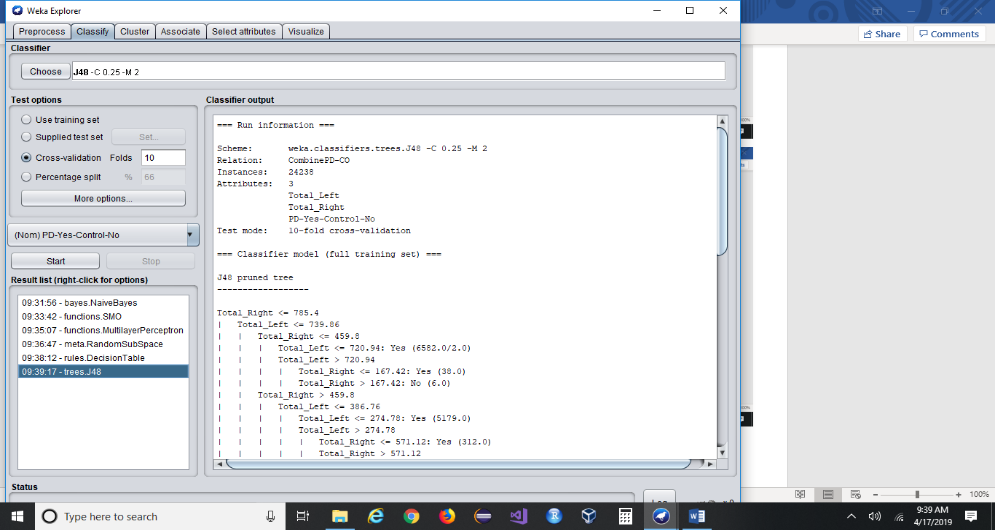


Fig 7a: J48 dataset output

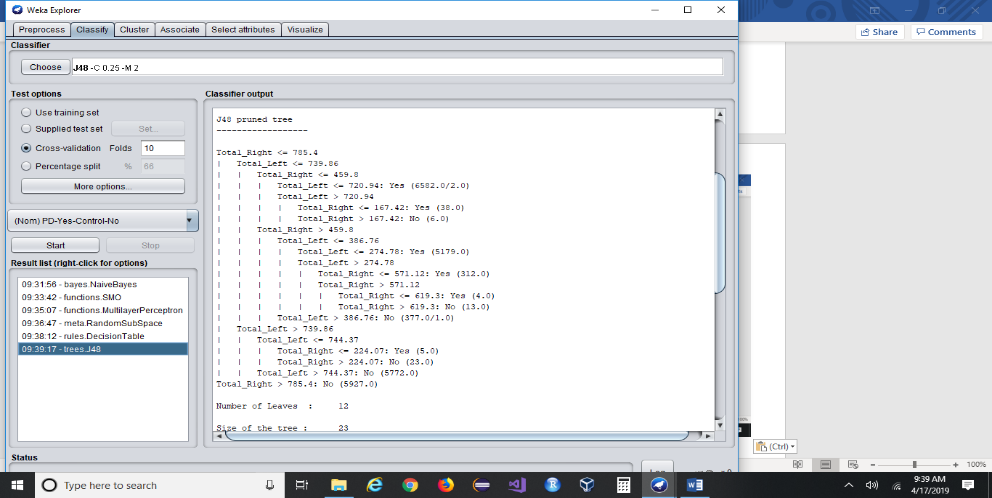


Fig 7b: Tree for J48

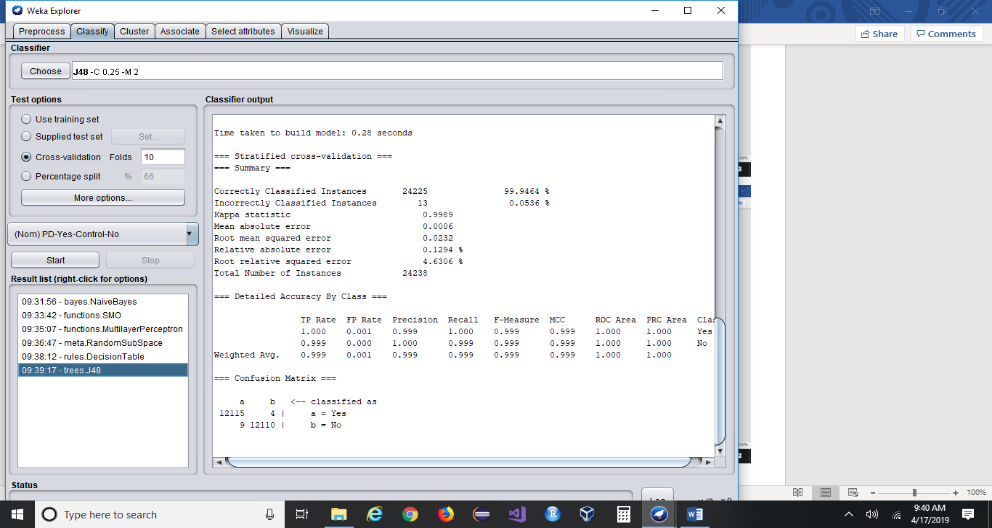


Fig 7c: J48 dataset output

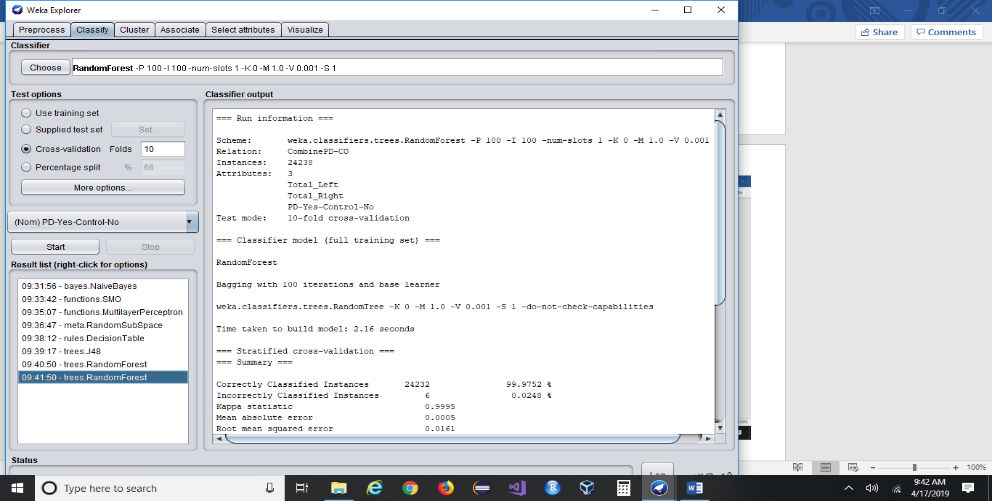


Fig 8a: Random Forest dataset output

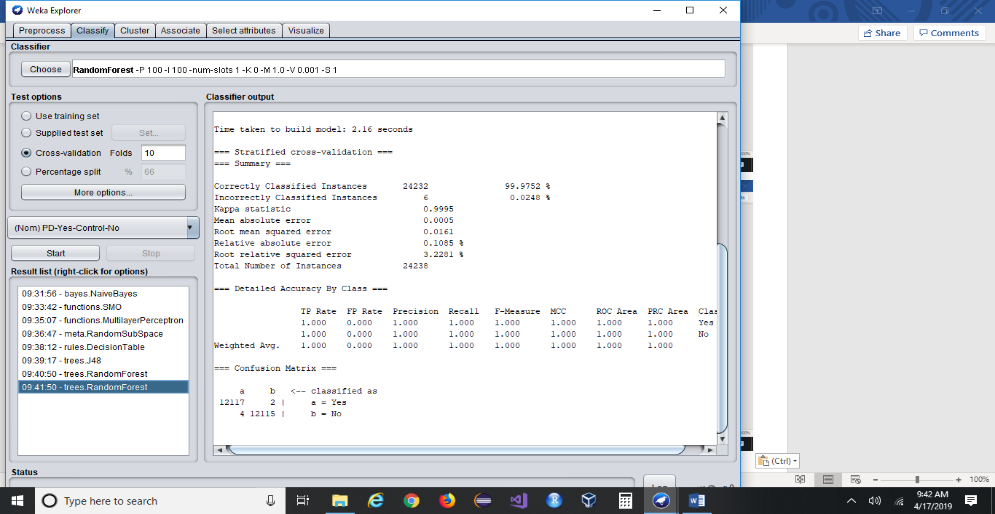


Fig 8b: Random Forest dataset output

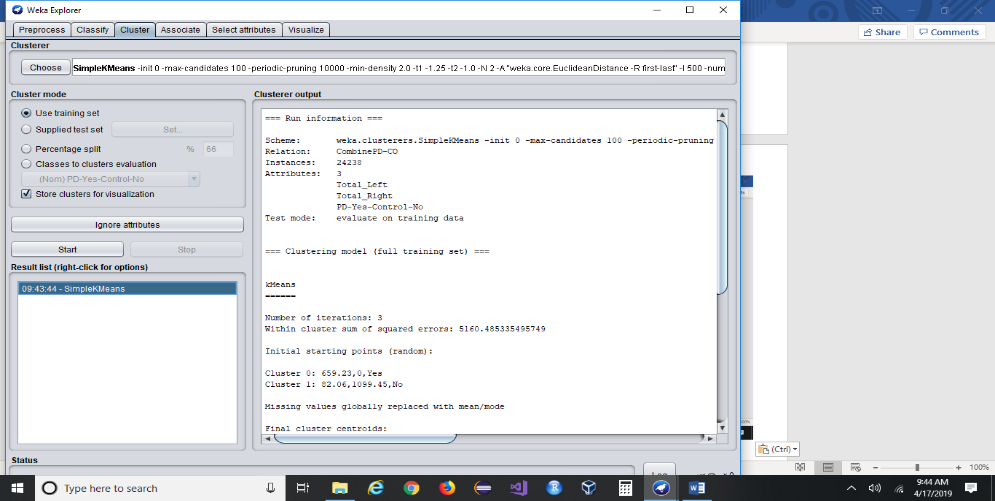


Fig 9a: Simple K Means dataset output

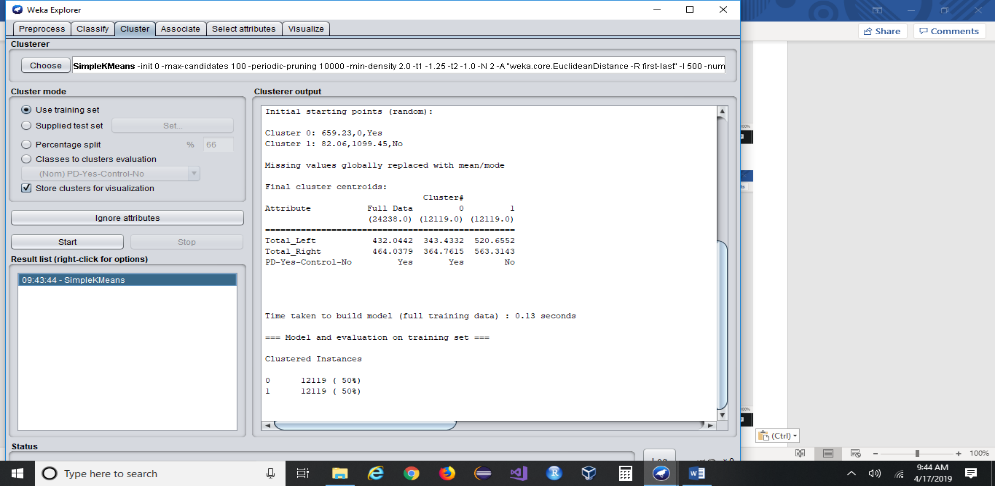


Fig 9b: Simple K Means dataset output

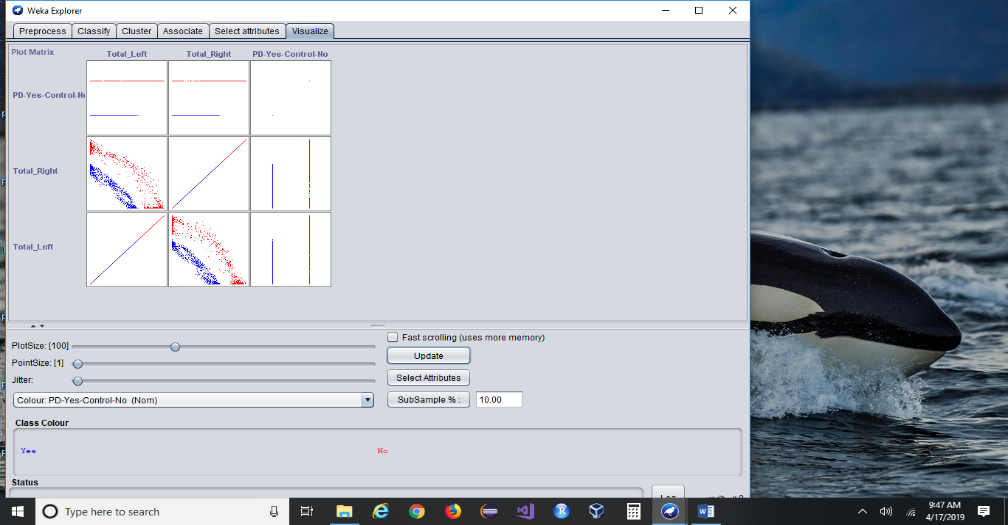


Fig 10a: Visualization of 10% of dataset displayed

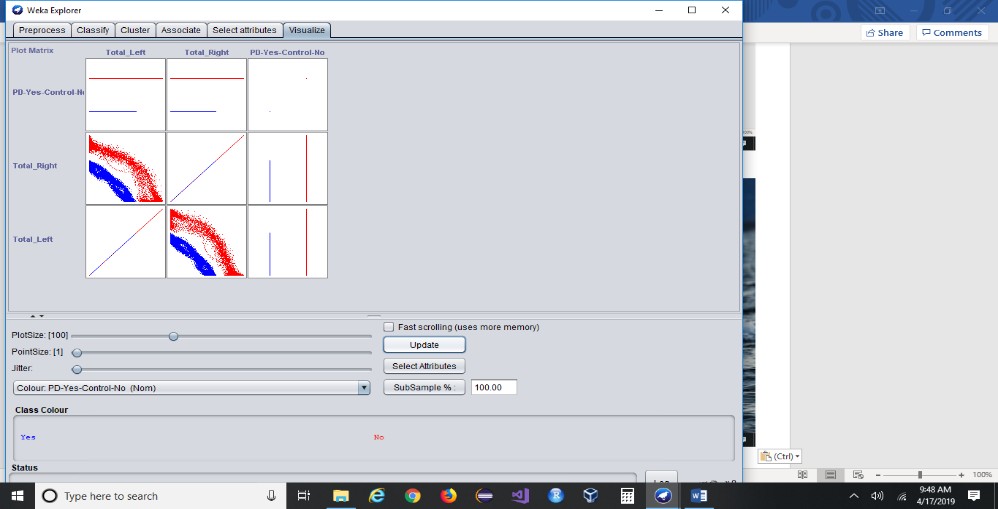
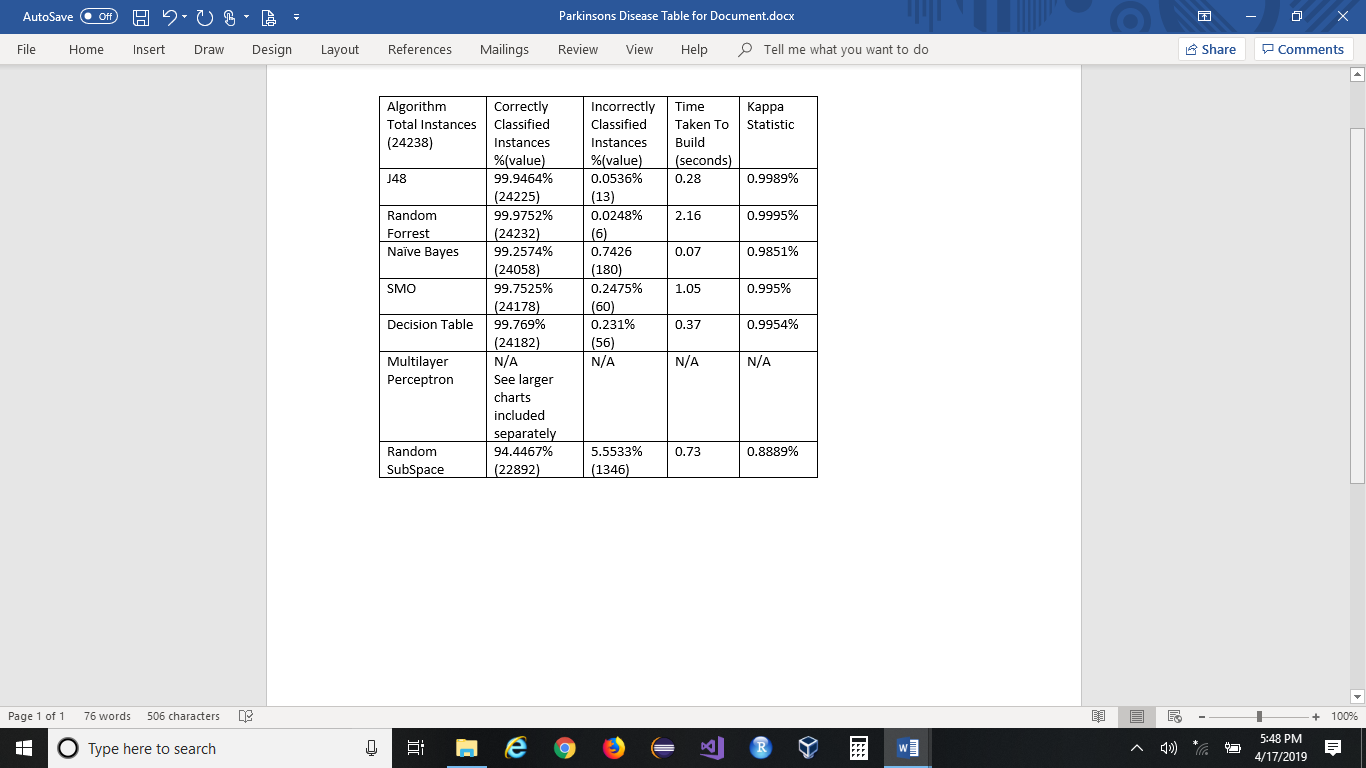


Fig 10b: Visualization 100% of dataset displayed

**7. Results**

With the WEKA results the Random Forest trees resulted in 1.00% accuracy across the board. With an average precision of 1.00, recall of 1.00, and a F-Measure of 1.00 (see larger screen prints enclosed in a separate file). A Confusion Matrix is used to calculate accuracy.



**8. Conclusion**

Based on this training set doctors will be able to use Random Forest to make treatment decisions. This dataset output will help doctors determine treatments with a 1.00% accuracy. This could help aid patients and help prevent falls with the proper diagnosis and medication treatment sooner than later. It will also save time and money by selecting treatments with the best possibility of predicting accurately.

**References**

1. **J48 source:**

<http://technobium.com/decision-trees-explained-using-weka/>

<http://ijarcsse.com/Before_August_2017/docs/papers/Volume_3/6_June2013/V3I6-0408.pdf>

<http://data-mining.business-intelligence.uoc.edu/home/j48-decision-tree>

2. **Random Forest source:**

<https://www.techopedia.com/definition/32935/random-forest>

<https://en.m.wikipedia.org/wiki/Random_forest>

<https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm>

<https://www.analyticsvidhya.com/blog/2018/10/interpret-random-forest-model-machine-learning-programmers/>

3. **Naïve Bayes source:**

<https://www.techopedia.com/definition/32335/naive-bayes>

<https://en.m.wikipedia.org/wiki/Naive_Bayes_classifier>

<http://it-definition.com/developer-naive-bayes-classifier.php>

<http://dataaspirant.com/2017/02/06/naive-bayes-classifier-machine-learning/>

4. **SMO source:**

<https://en.m.wikipedia.org/wiki/Sequential_minimal_optimization>

<http://cs229.stanford.edu/materials/smo.pdf>

5. **Random Subspace source:**

<https://en.m.wikipedia.org/wiki/Random_subspace_method>

<http://euler.stat.yale.edu/Seminars/2003-04/seminars/Bauer.pdf>

6. **WEKA Retrieved from**

http://www.WEKA.com/