SYNOPSYS CHAMPIONSHIP 2017 PROJECT ABSTRACT

THE ABSTRACT IS A REQUIRED PART OF YOUR PROJECT

Bring your abstract with you to project check-in on Wednesday, March 22, 2017.

PROJECT NUMBER: 112-H22-C1______ STUDENT NAME(s): Michelle Li________________________

You should bring at least ten copies of your abstract with you when you come to the Championship. One copy should remain on display with your project during the Championship. You will want others to give to the judges. Your abstract should be written after you finish your research and experimentation and should include:

- Your project title, the full name(s) of all team members, and your school (all centered).
- The purpose of your project
- Your hypothesis or evaluation criteria
- A brief statement about the procedures and equipment you used
- Your results (analysis of data)
- Your conclusions

Type or print neatly using 10- or 12- point black type. Single space throughout.
Center your project title, your name(s), and school.
You must use this form. Your abstract should be less than 500 words, and it should fit within the lines on this form.

A Machine Learning Approach For the Diagnosis of Parkinson's Disease via Speech Analysis

Michelle Li, Evergreen Valley High School, San Jose, CA

Parkinson's Disease is the second most prevalent form of dementia, affecting over 10 million people worldwide. Unfortunately, no single, reliable methods of diagnosis exist, resulting in decreased time to treat or slow the disease. This motivates a machine learning approach via speech analysis to provide a more reliable means of early diagnosis. The purpose of this experiment is to perform a comparative analysis of various classes of machine learning algorithms to identify the best models predicative of Parkinson's. It should be reasonable to create a model achieving at least 90% accuracy and a Matthews Correlation Coefficient (MCC) of at least 0.9, which would provide a significant improvement over current methods of diagnosis.

The following equipment were used in this project: a laptop equipped with Python 3.6, a code editor, and the University of Oxford / National Center for Voice and Speech dataset for Parkinson's. My procedure trained and validated each of the following models using 10-fold cross validation: Logistic Regression, Linear Discriminant Analysis, k Nearest Neighbors, Decision Tree, Multilayer Perceptron (MLP) Neural Network, Naive Bayes, and Gradient Boost. For each of these models, I used two versions of the Oxford Dataset: the raw dataset and a scaled version. I analyzed the accuracy and MCC of these models for 3 different train-test splits: 80-20, 75-25, and 70-30.

The two best performing models, k Nearest Neighbors and the MLP Neural Network, both produced a validation accuracy, sensitivity, specificity, ROC, and F1-score of 0.98 (98%), and a MCC of 0.94 (1.0 being a perfect classifier). These models tended to perform better on the rescaled dataset than on the raw dataset, and achieved the best results with a 75-25 train-test split. Most of the models I evaluated tended to underfit, due in part to the 10-fold cross validation training method I used to prevent overfitting.

Overall, my results show that KNN and MLP NN produce very robust, promising results that far exceeded my engineering goal and most literature on this subject. This suggests that a machine learning model can be implemented to significantly improve diagnosis methods of Parkinson's Disease. Not only is my machine learning method of diagnosis more reliable and robust, it is also cheaper (requiring only features of the patient's voice) to implement. These results are significant because millions of Parkinson's patients would benefit from a more reliable method of diagnosis.