# Mini Project #4 - Predictive Analytics

David Strube
dstrube3@gatech.edu

***Abstract*—**In this exploratory challenge for Mini Project #4 for CS 6440 - Intro to Health Informatics, predictive analytical methods shall be applied to a selected dataset. Working with this dataset as the model, a predictive based solution was built utilizing the Decision Tree algorithm in the scikit-learn library (Pedregosa F, Varoquaux G, Gramfort A, et al, 1970) using Python and Google Colab. This solution should be both powerful and generic enough to be applied to any dataset of sufficiently valid and reliable data.

## 1 PROBLEM STATEMENT

The chosen Problem is how to predict whether an instance of breast cancer is malignant or benign given a range of available factors.

### 1.1 Topic Area

The chosen Topic Area is Other.

## 2 APPROACH

The approach used for analyzing this dataset was to employ a DecisionTreeClassifier algorithm. This approach was chosen because the nature of the data was that of a binary set of targets (malignant or benign) rather than a spectrum of targets, like a real number line of expected outcomes, in which case a regression algorithm would have been more appropriate than a classifier. Furthermore, a DecisionTree was chosen because it was among the easier set of approaches to understand and explain.

### 2.1 Dataset

The chosen Dataset is the 1995 University of Wisconsin dataset of breast cancer (Wolberg WH, Street WN, Mangasarian OL, 1995). The function to invoke this dataset comes built-in with scikit-learn, so the dataset was relatively easy to use. It has 569 rows (also known as “instances”) and 30 columns (also known as “features”). Alternatively, much larger data sources of different natures that were being considered include two from CDC and two from OpenML. The sources from CDC that were considered were “Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System” (CDC, 2021) and “Behavioral Risk Factor Data: Tobacco Use

(2011 to present)" (CDC, 2020). These sources were not selected because unfortunately they were not correctly formatted for the data processing logic available in scikit-learn. The sources from OpenML considered were "hypothyroid" (Vanschoren, 2014) and "open_payments" (Vanschoren, 2020) . These sources were not selected because unfortunately some unparseable errors were thrown when I tried to apply some of the DecisionTree algorithms to them.

### 2.2 Model

The best Model was achieved by running the training data through 3 functions: one for exploring the decision tree without pruning, one with pre-pruning, and one with post-pruning. The optimal model was found to be a decision tree classifier that uses pre-pruning. Details to follow in Analysis and Outcome.

### 2.3 Algorithm(s) & Libraries

The 3 functions that I wrote to explore the decision tree were: `decisionTree`, `decisionTreePrePruning`, and `decisionTreePostPruning`. The function `decisionTree` only took in the training data and an optional parameter for maximum cross validation size. It returned the optimal settings found by cycling through the range of criterion available and a range of cross validation sizes; these settings would be passed to the other functions. The function `decisionTreePrePruning` explored a range of maximum depth sizes from 2 to 20. The function `decisionTreePostPruning` explored a range of Cost-Complexity Pruning (`ccp_alpha`) sizes from 1 to 10. Each function would graph its findings in a learning curve line chart and a confusion matrix (to display the True-Positive, True-Negative, False-Positive, and False-Negative scores) before exiting.

The libraries used were `matplotlib`, `mpl_toolkits`, `numpy`, and `sklearn`. I also used Python's built-in `time` library to display how long each function took to complete.

## 3 ANALYSIS

### 3.1 Notebook

The output from the `decisionTree` function shows the optimal cross validation size between 1 and 100 is 93 with the best criterion being *gini*. This is confirmed by the confusion matrix which shows 91% correlation between true and predicted target of 0, and a 96% correlation between true and predicted target of 1. The learning curve levels off with a score just above 92% for the cross validation score between 300 and 400 training examples. The performance of the model had similar results with fit times leveling off around 93%.

The output from the `decisionTreePrePruning` function shows the optimal maximum depth size between 2 and 20 is 5. This is confirmed by the confusion matrix which shows 91% correlation between true and predicted target of 0, and a 97% correlation between true and predicted target of 1. The learning curve shows the training score and the cross validation score starting to converge above 400 training examples.

The output from the `decisionTreePostPruning` function shows the optimal *ccp_alpha* value between 0 and 10 is 0. In other words, post pruning doesn't help at all in the search for the best model.

**4 OUTCOME**

The optimal model was found to be one that uses pre-pruning with a cross-validation size of 93, a criterion of *gini*, and a maximum depth size of 5. These settings yielded a predictive accuracy score of approximately 96.6%.

Note, as discussed on this post in Ed, I was not able to attach the Jupyter Notebook to this document, so it will be submitted separately: https://edstem.org/us/courses/3716/discussion/344491?answer=799588

## 5 REFERENCES


1. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html. Published January 1, 1970. Accessed March 31, 2021.
2. Wolberg, WH, Street, WN, Mangasarian, OL. UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic). Published November, 1995. Accessed March 31, 2021.
3. Nutrition, Physical Activity, & Obesity - Behavioral Risk Factor Surveillance. CDC. https://chronicdata.cdc.gov/Nutrition-Physical-Activity-and-Obesity/Nutrition-Physical-Activity-and-Obesity-Behavioral/hn4x-zwk7. Updated January 29, 2021. Accessed March 31, 2021.
4. Behavioral Risk Factor Data: Tobacco Use (2011 to present). CDC. https://chronicdata.cdc.gov/Survey-Data/Behavioral-Risk-Factor-Data-Tobacco-Use-2011-to-pr/wsas-xwh5. Updated August 13, 2020. Accessed March 31, 2021.
5. Vanschoren J. hypothyroid. OpenML. https://www.openml.org/d/1000. Uploaded October 14, 2014. Accessed March 31, 2021.
6. Vanschoren J. open_payments. OpenML. https://www.openml.org/d/42738. Uploaded November 26, 2020. Accessed March 31, 2021.