Team #3

Emily Kabban

Vineel Busi

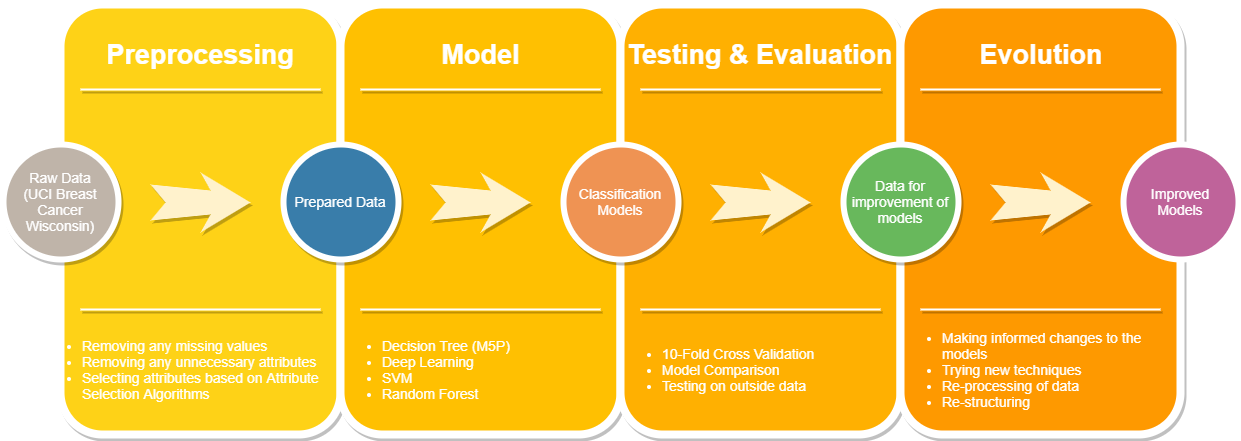
Derek Nkeng

Breast Cancer Malignancy Prediction

# Data Mining Problem

The goal of the project is to create working models for determining if a breast cancer anomaly is benign or malignant as a precursor or early prediction before lab tests are conducted. This will be based on physical and observable characteristics of the anomaly.

# System Design and Architecture

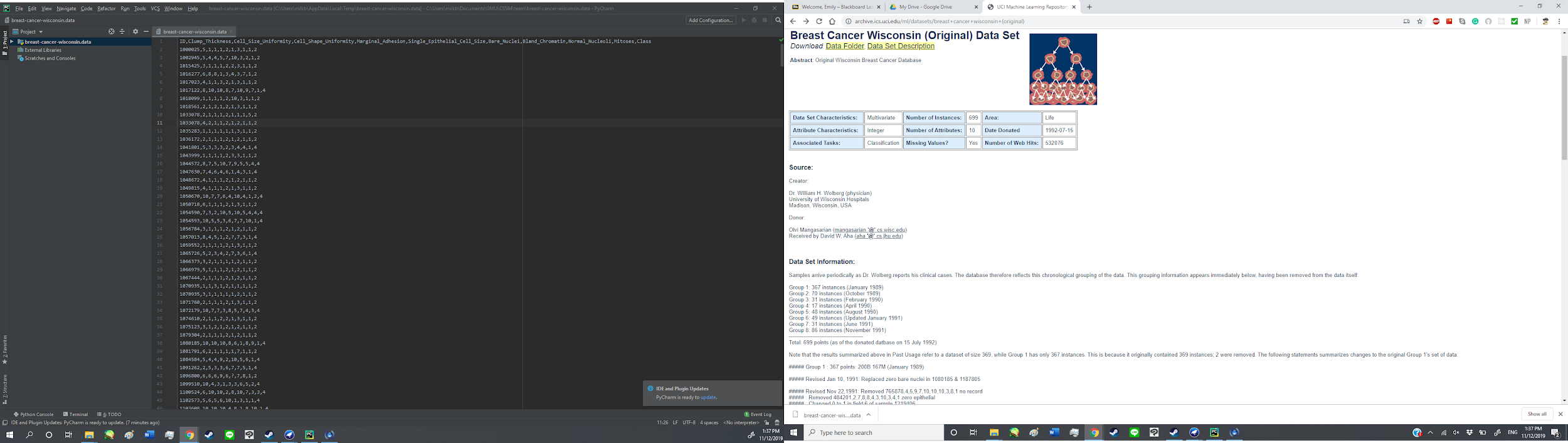


# Tools

Weka

# Data Capture

For this project we are using the following dataset which was provided for free use to the UCI machine learning repository in 1992. It features 10 quantitative attributes and 1 class attribute, separating the data into malignant and benign, and 699 instances.



1. Sample code number: id number

2. Clump Thickness: 1 - 10

3. Uniformity of Cell Size: 1 - 10

4. Uniformity of Cell Shape: 1 - 10

5. Marginal Adhesion: 1 - 10

6. Single Epithelial Cell Size: 1 - 10

7. Bare Nuclei: 1 - 10

8. Bland Chromatin: 1 - 10

9. Normal Nucleoli: 1 - 10

10. Mitoses: 1 - 10

11. Class: (2 for benign, 4 for malignant)

Citations:

1. O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

2. William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

3. O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.

4. K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).

<https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)>

## Preprocessing

## Data Cleaning

The UCI database page indicated that there were missing values present. It would appear that these have been corrected in updates to the repository seeing as how there are no missing values present in the accompanied CSV file.

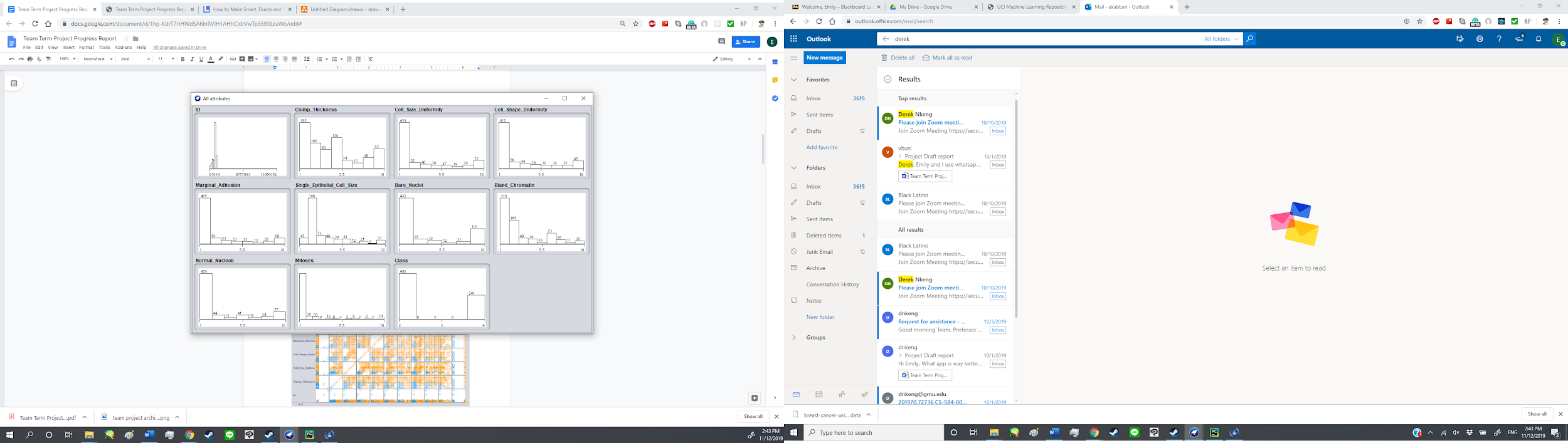
## Data Integration

All data is contained within a single file so no integration is necessary.

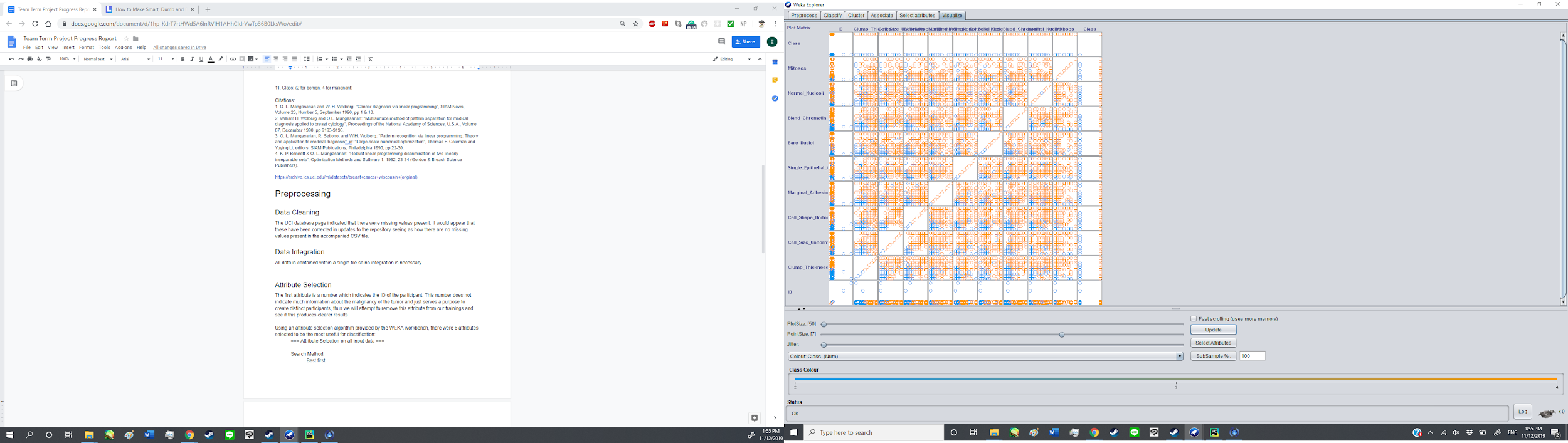
## Attribute Selection

The first attribute is a number which indicates the ID of the participant. This number does not indicate much information about the malignancy of the tumor and just serves a purpose to create distinct participants, thus we will attempt to remove this attribute from our trainings and see if this produces clearer results.

The spread of each attribute is visualized below. This may prove useful if we attempt to restructure the data qualitatively to broaden the data-mining methods and models available to us.



Upon visualizing the data, we have found that there are no clear correlations between attributes. This is shown in the figure below:



Using an attribute selection algorithm provided by the WEKA workbench, there were 6 attributes selected to be the most useful for classification:

=== Attribute Selection on all input data ===

Search Method:

Best first.

Start set: no attributes

Search direction: forward

Stale search after 5 node expansions

Total number of subsets evaluated: 65

Merit of best subset found: 0.907

Attribute Subset Evaluator (supervised, Class (numeric): 11 Class):

CFS Subset Evaluator

Including locally predictive attributes

Selected attributes: 2,3,4,7,8,9 : 6

Clump\_Thickness

Cell\_Size\_Uniformity

Cell\_Shape\_Uniformity

Bare\_Nuclei

Bland\_Chromatin

Normal\_Nucleoli

We will see how including only these attributes affects the accuracy of our models.

## Data Transformation

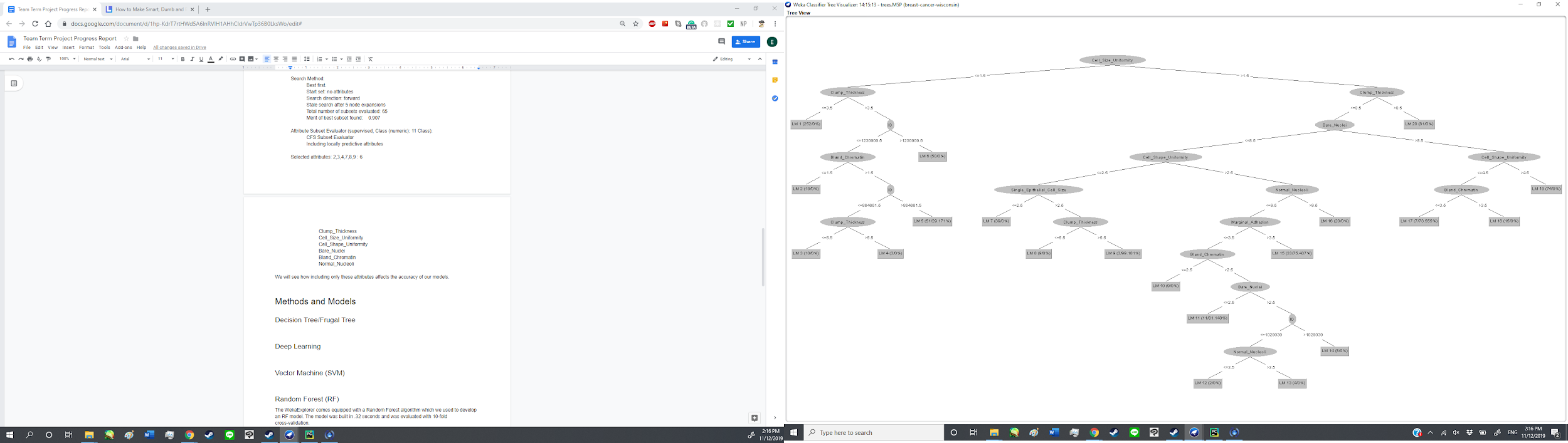
In order to explore more algorithm options and metrics for performance measurement we created a separate dataset in which the class labels were made into nominal values. Rather than 2 indicating benign and 4 indicating malignant, the labels were replaced with B for benign and M for malignant. This allows us to use tree algorithms which allow us to see the percentage of correctly identified instances.

# Methods and Models

## Decision Tree

## MP5

Using Weka, we were able to model an M5P tree of the following structure. It consists of 17 rules.



The cross-validation summary was as follows:

=== Cross-validation ===

=== Summary ===

Correlation coefficient 0.9254

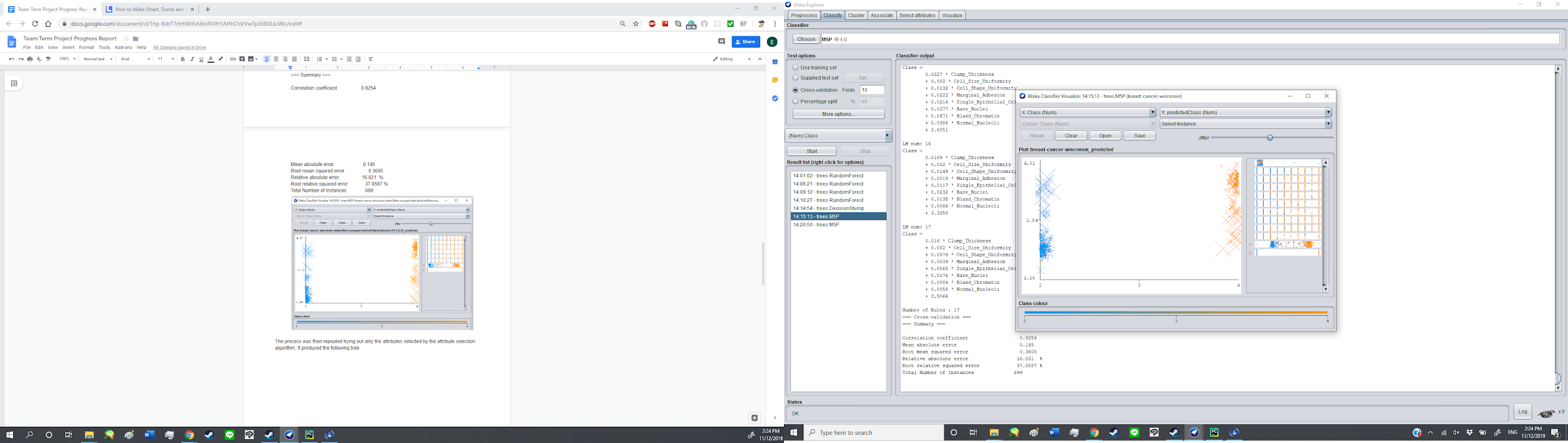
Mean absolute error 0.145

Root mean squared error 0.3605

Relative absolute error 16.021 %

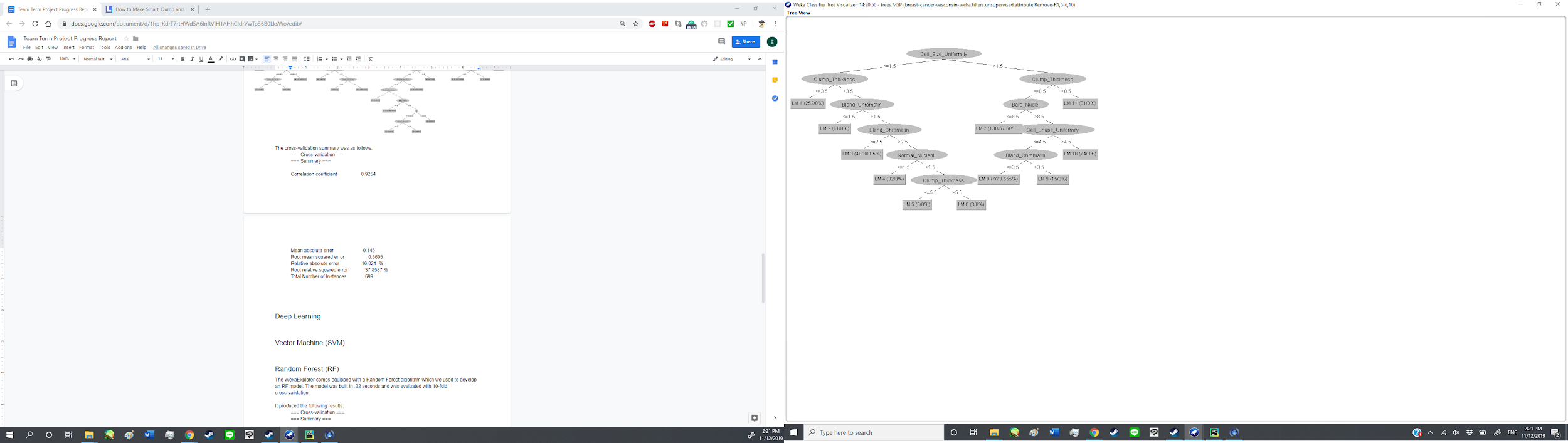
Root relative squared error 37.8587 %

Total Number of Instances 699



### MP5 With Attribute Selection

The process was then repeated trying out only the attributes selected by the attribute selection algorithm. It produced the following tree:



=== Cross-validation ===

=== Summary ===

Correlation coefficient 0.9261

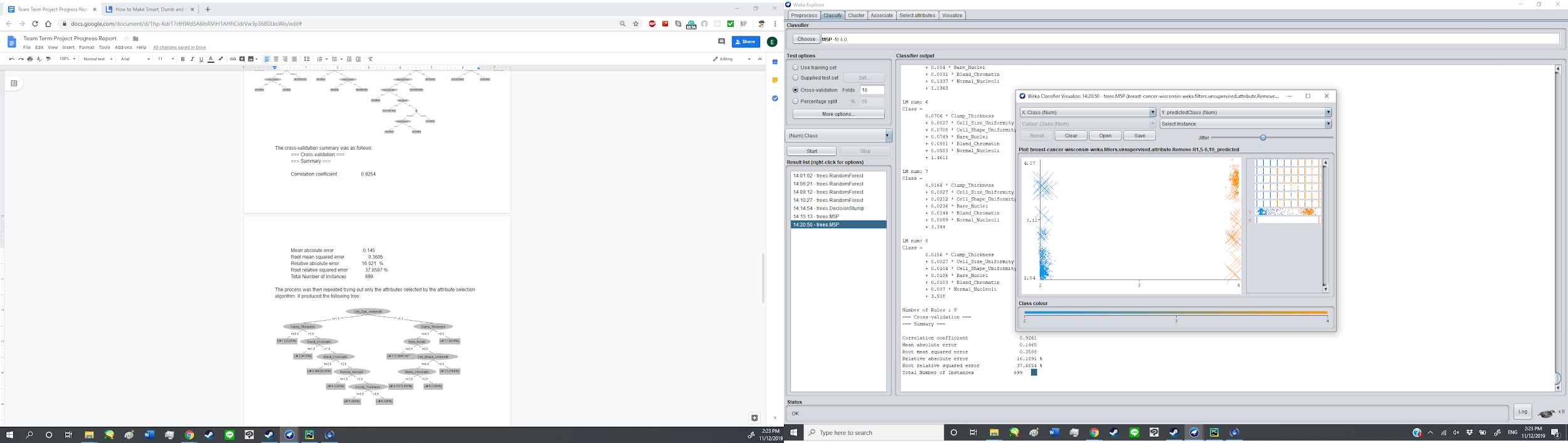
Mean absolute error 0.1465

Root mean squared error 0.3588

Relative absolute error 16.1891 %

Root relative squared error 37.6854 %

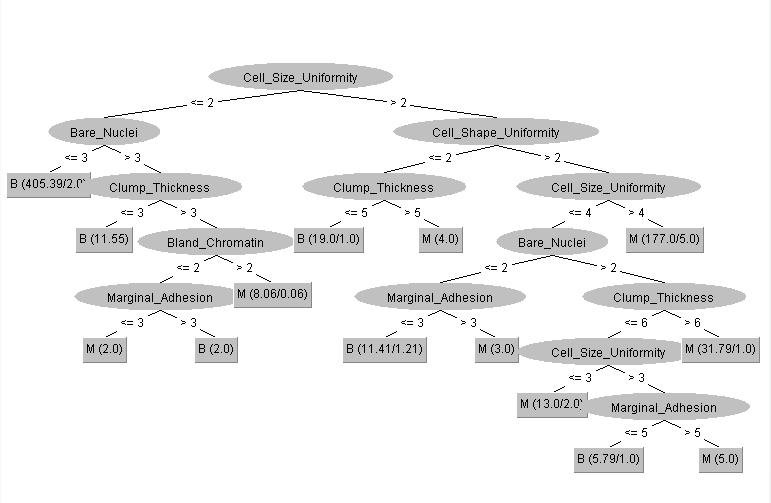
Total Number of Instances 699



### J48 (Nominal Class Labels)

In order to compare the performance of the decision tree methods using nominal classes, the transformed version of the dataset with nominal class labels was used to run the J48 algorithm. It was evaluated using 5-fold cross-validation. Unlike the original dataset, when running the attribute selection algorithm, this dataset featured 10 selected attributes which were used to create the tree. The only attribute omitted was the ID attribute which lacks meaning in this dataset besides to identify participants and thus would naturally not be selected.

The J48 algorithm produced the following tree structure:



=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 661 94.5637 %

Incorrectly Classified Instances 38 5.4363 %

Kappa statistic 0.8799

Mean absolute error 0.0694

Root mean squared error 0.2229

Relative absolute error 15.352 %

Root relative squared error 46.8927 %

Total Number of Instances 699

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.956 0.075 0.961 0.956 0.958 0.880 0.955 0.955 B

0.925 0.044 0.918 0.925 0.921 0.880 0.955 0.902 M

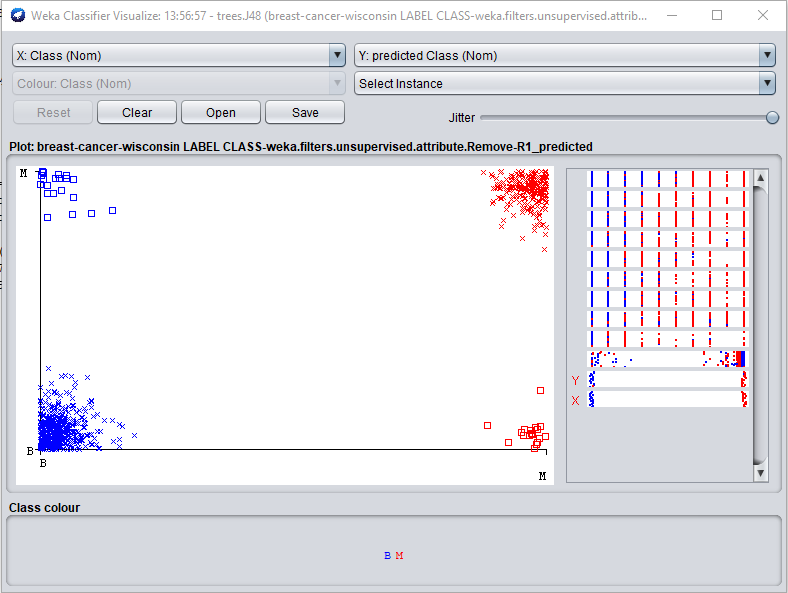
Weighted Avg. 0.946 0.064 0.946 0.946 0.946 0.880 0.955 0.937

=== Confusion Matrix ===

a b <-- classified as

438 20 | a = B

18 223 | b = M



The classifier error visualization above shows that a large majority of instances were correctly classified with crosses being correct and squares indicating an incorrectly classified instance. As shown in the summary, about 94.5% of instances were correctly identified with a mean absolute error of .0694 which is a notable improvement from the MP5 algorithm.

Mean absolute error for each fold:

(1) 0.029197

(2) 0.097409

(3) 0.050396

(4) 0.134703

(5) 0.092253

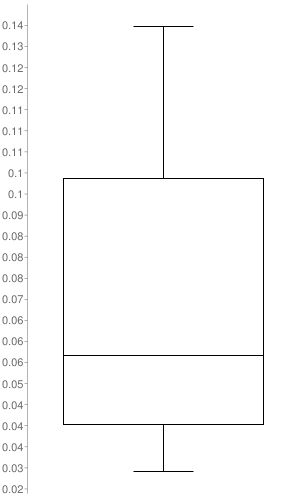
(6) 0.043126

(7) 0.055129

(8) 0.058102

(9) 0.031646

(10) 0.102374



## 

## Deep Learning

To be completed

## Vector Machine (SVM)

To be completed

## Random Forest (RF)

The WekaExplorer comes equipped with a Random Forest algorithm which we used to develop an RF model. The model was built in .32 seconds and was evaluated with 10-fold cross-validation.

It produced the following results:

=== Cross-validation ===

=== Summary ===

Correlation coefficient 0.937

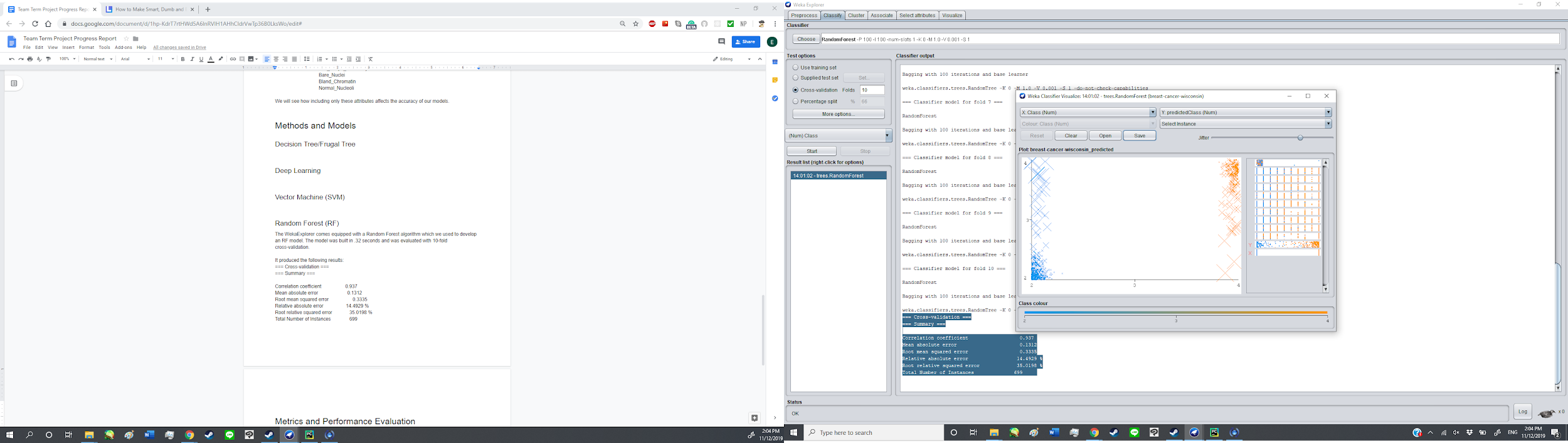
Mean absolute error 0.1312

Root mean squared error 0.3335

Relative absolute error 14.4929 %

Root relative squared error 35.0198 %

Total Number of Instances 699



We then repeated the process using only attributes selected with the attribute selection algorithm. It yielded similar results:

Correlation coefficient 0.9344

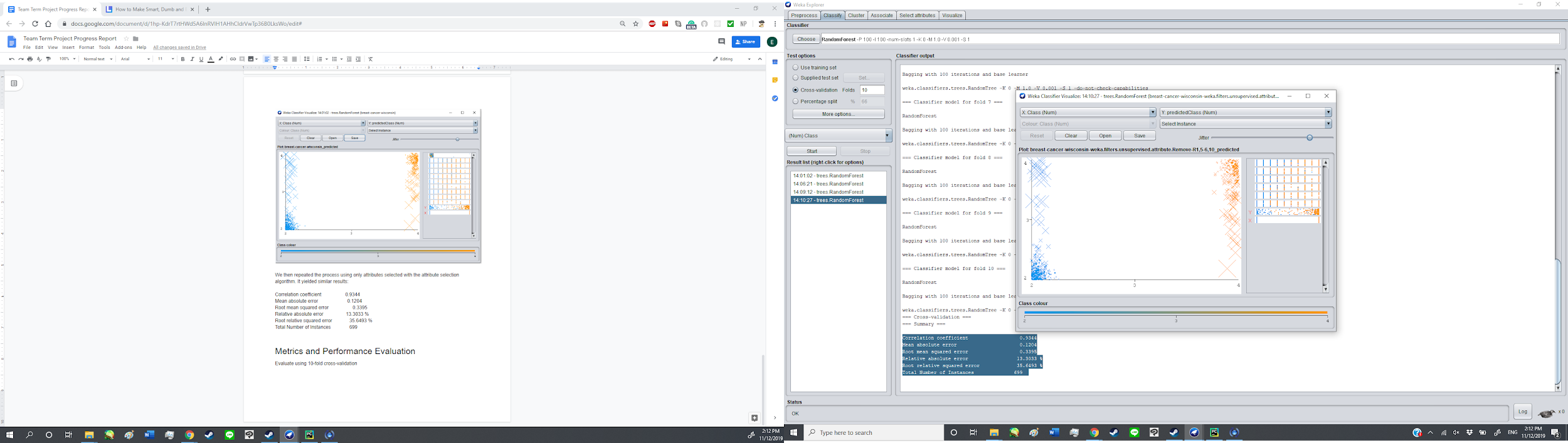
Mean absolute error 0.1204

Root mean squared error 0.3395

Relative absolute error 13.3033 %

Root relative squared error 35.6493 %

Total Number of Instances 699



# Metrics and Performance Evaluation

All models were evaluated using a 10-fold cross validation.

Once all models are built we will compare the performance of the models and do an in-depth analysis of their errors

We will need to note the success rate at which human subjects predict whether a tumor is benign of malignant to properly understand the quality of our models.

Still need to create box-plots.

# Summary of still to be done:

* Deep Learning Model
* SVM model
* Error analysis
* Box Plots
* Evolution of models
* We may also experiment with remodeling the data as qualitative attributes to see if it results in better results