Northeastern University

ASSIGNMENT FRONT SHEET

Course Name: **ALY6050 20906 Intro to Enterprise Analytics**

Professor Name: **Dr. Christopher Johnson**

Student Name: **Dong Quoc Tuong (Lukas)**

Student Class: **Fall 2019 CPS** **Term: B. 2020**

Module 6: Final Project

Completion Date: Feb 27t **Due Time:12:00am**

Statement of Authorship

I confirm that this work is my own. Additionally, I confirm that no part of this coursework, except where clearly quoted and referenced, has been copied from material belonging to any other person e.g. from a book, handout, another student. I am aware that it is a breach of Northeastern University's regulations to copy the work of another without clear acknowledgement and that attempting to do so renders me liable to disciplinary procedures. To this effect, I have uploaded my work onto Turnitin and have ensured that I have made any relevant corrections to my work prior to submission.

☒ **Tick here** to confirm that your paper version is identical to the version submitted through Turnitin

Introduction

In America, there are 11,000 new cases of invasive cervical cancers being found annually. Despite the fact that the number of new cases are on the downward trend over the past decade, it still kills about 4,000 women in America and 300,000 women globally.(Fontham et al., 2020) The sooner one discovers it, the more chances they have to survive. Therefore, in order to effectively eliminate such diseases, we need to create a Machine Learning algorithm that can detect the cancer as soon as possible. As a result, for the final project, I will be analyzing to analyze the dataset: Cervical cancer (Risk Factors) from UCI. ("UCI Machine Learning Repository: Cervical cancer (Risk Factors) Data Set," 2017) The dataset was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela. I am planning to do a mix of EDA, Feature Importance Analysis with Boruta and Classification prediction models

Data Preparation

We will start our project by loading the necessary libraries and reading the dataset. There are 8 libraries that we are going to use, 1 for visualization, 3 for data manipulation, 1 for feature importance analysis and 3 for the prediction models.

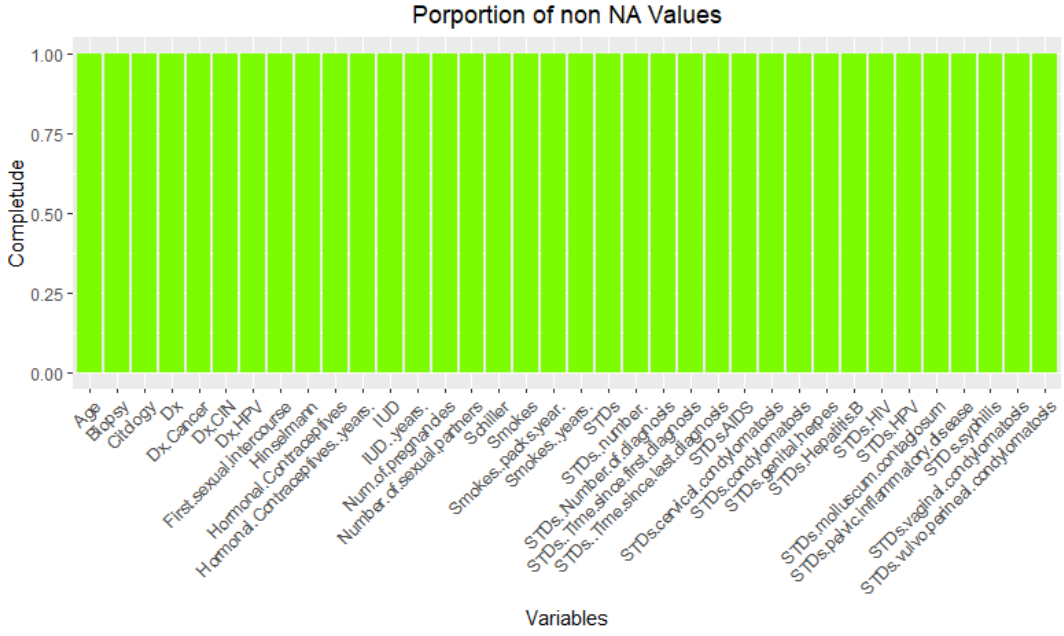
```
# Load Llibraires, data and surface analysis
library(ggplot2)      #Data Visulization
library(dplyr)        #Data Manipulation
library(caret)        #Streamline the model training process
library(Boruta)       #Feature Importance Analysis
library(caTools)      #To split the data
library(e1071)        #SVM Model
library(rpart)        #Decision Tree Model
library(randomForest) #Random Forest
```

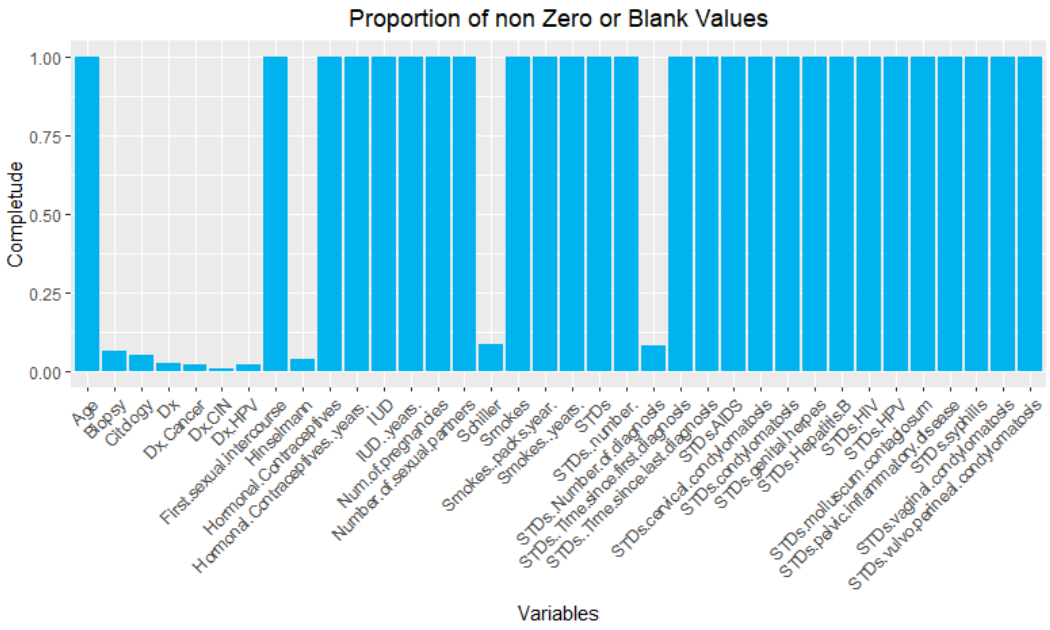
Cervical cancer has 858 inputs with 36 variables ranging from demographic information, habits to historic medical records. Some variables have missing data as patients refused to answer due

to privacy concerns. It is noteworthy that many variables were interpreted as factors, caused by the "?" populated as placeholders for the missing values

```
> #  Explore the data
> glimpse(cervical_cancer)
Rows: 858
Columns: 36
$ Age                                 <int> 18, 15, 34, 52, 46, ...
$ Number.of.sexual.partners           <chr> "4.0", "1.0", "1.0",...
$ First.sexual.intercourse            <chr> "15.0", "14.0", "?",...
$ Num.of.pregnancies                  <chr> "1.0", "1.0", "1.0",...
$ Smokes                              <chr> "0.0", "0.0", "0.0",...
$ Smokes..years.                      <chr> "0.0", "0.0", "0.0",...
$ Smokes..packs.year.                 <chr> "0.0", "0.0", "0.0",...
```

So, before going any further, I will check the integrity of the data, since it can hide the actual summary statistics that we are looking for. There is a wide array of functions to indicate Missing values such as NAs, zeros, negative vales or blank strings, so it is easier to create a function that deal specifically with the NAs and then plot the completude (non missingness) of the data





Data Manipulation

After we know the data a little bit better, it is much better to take care of the missing values to enable a closer exploratory data analysis. First we create a function to identify all the columns that need repair, then a function to fix the missing values. The reason why x=”-1” is to allow numeric operations to be performed on them after “?” caused them all to be factors. “-1” is also a good place holder since it distinctively shows there is something with that input. Then we apply the two functions onto the columns and we can establish an attribute that represents the cervical cancer later on.

In addition, we create a correlation map to reject unimportant variables. After that we need to cut off the correlation of anything combos that are less than 0.7. The final result will look like this, indicating that in general, these created highly correlated combos of variables.

```
print(highlyCorrelated)
[1]  4 11 13  8  9 10  5 12  6  3  7  1  2
```

The last four columns (“Hinselmann”,“Schiller”,“Citology”,“Biopsy”) indicate the results of the cervical exams. Positive result does not mean that the patient suffers from Cervical cancer, but

the likelihood increase the more positives a patient receive. Therefore I would create a Variable called “CervicalCancer” that is created according to such formula:

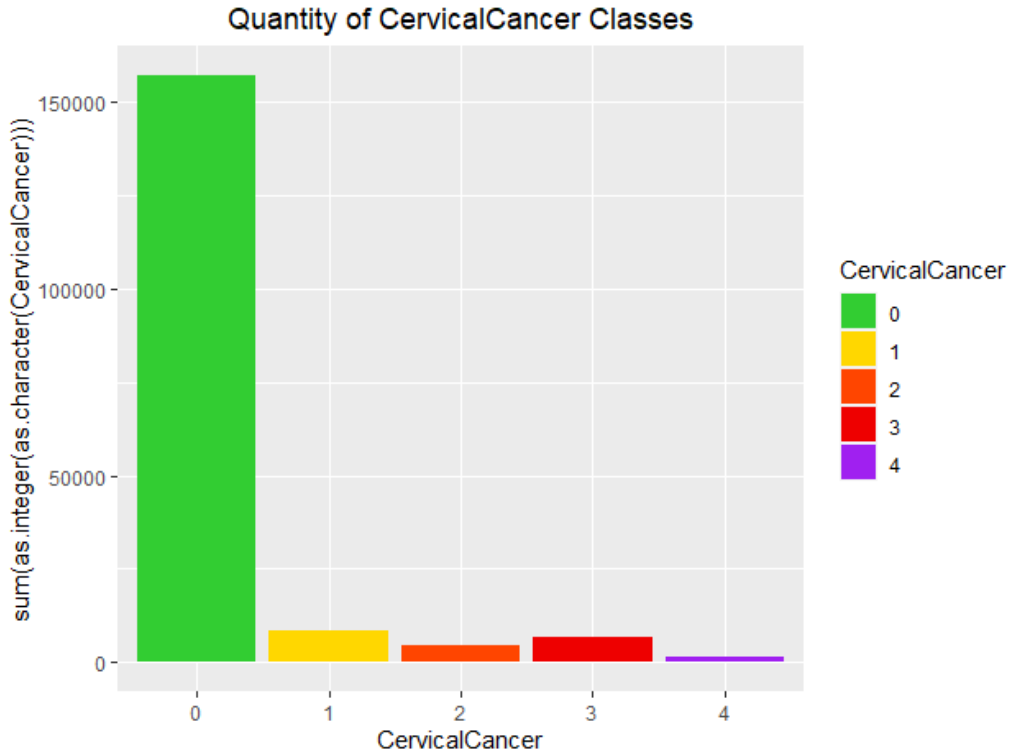
$$\mathbf{CervicalCancer = Hinselmann + Schiller + Citology + Biopsy}$$

```
#Create target varibles to represent the cervical cancer
cervical_cancer$CervicalCancer = cervical_cancer$Hinselmann + cervical_cancer$Schiller
                                 + cervical_cancer$Citology + cervical_cancer$Biopsy
                                          #We plus all of these columns together becau
cervical_cancer$CervicalCancer = factor(cervical_cancer$CervicalCancer
                                        , levels=c("0","1","2","3","4"))
```

Exploratory Data Analysis (EDA)

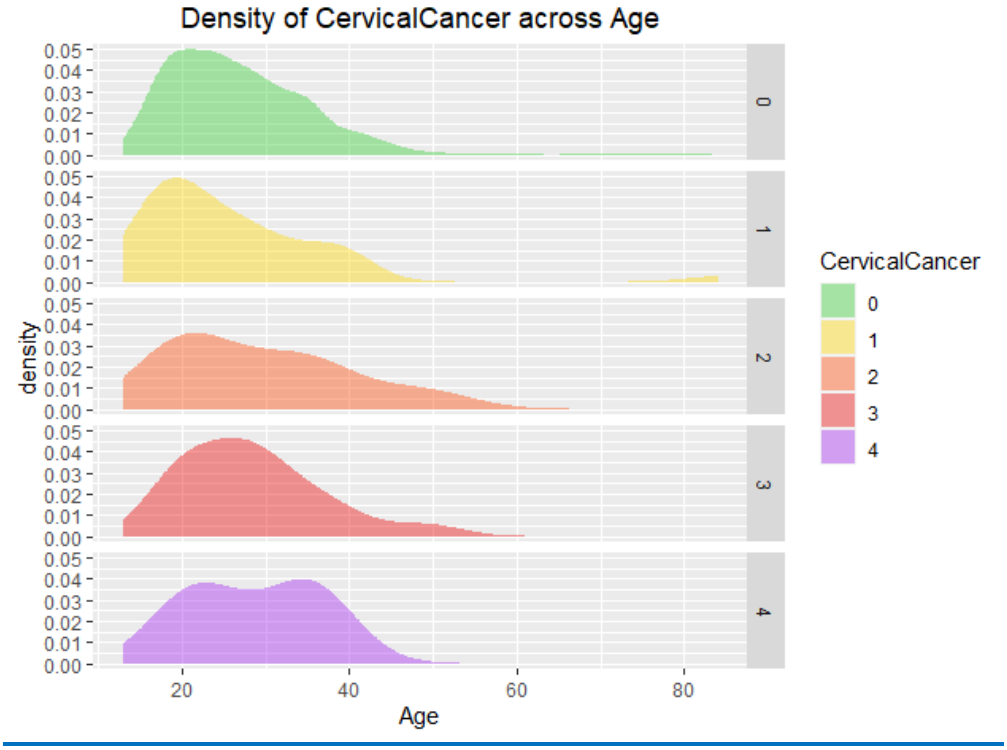
After creating the “CervicalCancer” column, we can see that approximately 90% of the patients do not have any symptom and only 1% actually show serious condition by showing all 4 of them.

```
   0    1    2    3    4
0.88 0.05 0.03 0.04 0.01
>
```

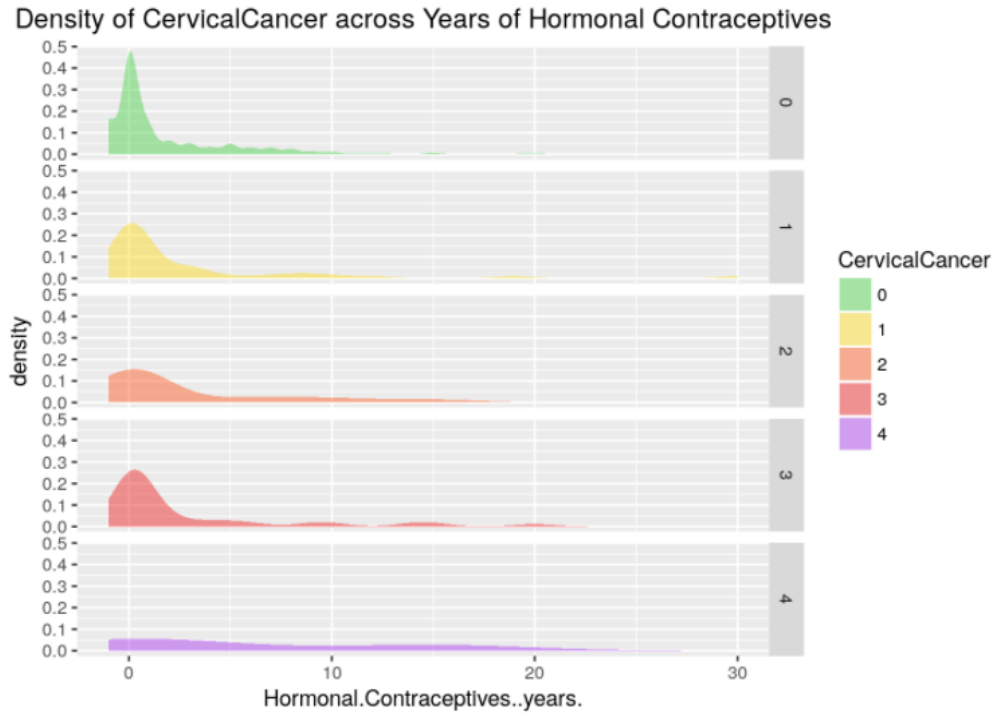


Thus, we can say that the accuracy of a baseline model, predicting everybody will not have cancer would have a high accuracy of 88%

When we look at the age of the patients in the density plot, there is a notable desolation of the peak in every density plot, indicating that the correlation between " Age" and "CervicalCancer" is strong



Additionally, when one look at the "Hormonal contraceptives years" and "CervicalCancer", there is also a decreasing height of the peaks and the right skewedness of the density plot also suggest a strong correlation also. While we could potentially do this for every single one of them, it is much better to carry out a much faster approach to find the features that most influence "CervicalCancer" like Boruta

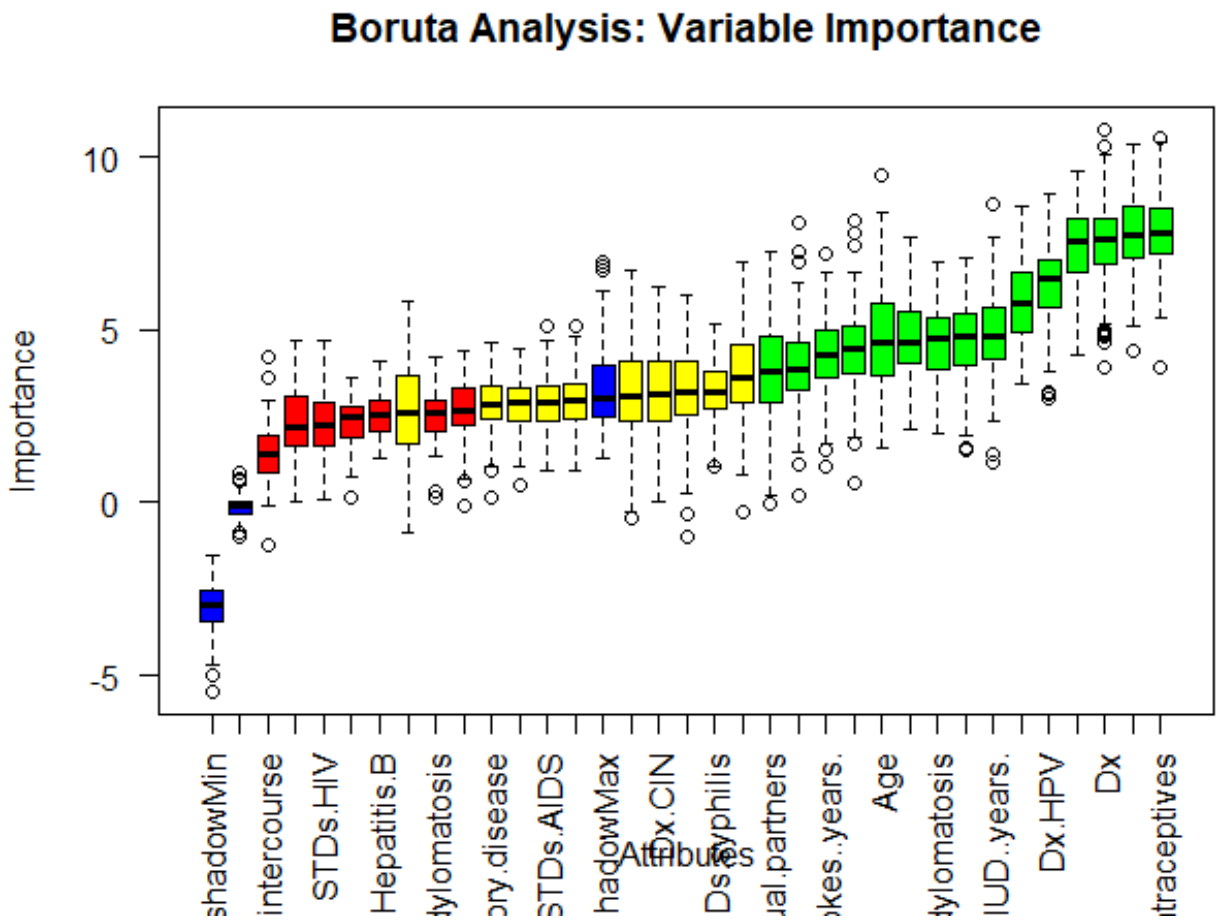


Feature Importance Analysis with Boruta

Boruta method assessed the feature importance by performing shuffling of predictors' values and joining them with the original predictors and construct a random forest on the merged dataset. After that, we will make comparison between the original variables with the randomized variables to measure their importance and choose the one with the higher importance ratio than that of the randomized variables. .(Deepanshu Bhalla, 2017) One often consider Robuta over other feature selection packages in R thanks to its various advantage:

- Suitable for classification and regression analysis because it takes into account multi-variable relationships
- A enhancement on random forest variable importance measure and can handle interplay between variables as well as changing nature of random forest importance measure

In this exercise, the Boruto process starts with creating a cop of original dataset but remove the medical results columns. Then we set the seed and use Boruta() function. The plot is as followed



We confirmed 15, eliminated 7 and 10 are put as tentative.

```
Boruta performed 199 iterations in 2.57951 mins.
 15 attributes confirmed important: Age, Dx, Dx.Cancer, Dx.HPV,
Hormonal.Contraceptives and 10 more;
 7 attributes confirmed unimportant: First.sexual.intercourse,
Smokes, STDs.cervical.condylomatosis, STDs.genital.herpes,
STDs.Hepatitis.B and 2 more;
 10 tentative attributes left: Dx.CIN, Num.of.pregnancies, STDs,
STDs..Number.of.diagnosis, STDs..Time.since.last.diagnosis and 5
more;
```

And here are the variables that are confirmed to make a difference to the dependent variables

```
> getSelectedAttributes(final.boruta, withTentative = F)
 [1] "Age"
 [2] "Number.of.sexual.partners"
 [3] "Smokes..years."
 [4] "Smokes..packs.year."
 [5] "Hormonal.Contraceptives"
 [6] "Hormonal.Contraceptives..years."
 [7] "IUD"
 [8] "IUD..years."
 [9] "STDs..number."
[10] "STDs.condylomatosis"
[11] "STDs.vulvo.perineal.condylomatosis"
[12] "STDs.pelvic.inflammatory.disease"
[13] "STDs.HPV"
[14] "STDs..Number.of.diagnosis"
[15] "STDs..Time.since.first.diagnosis"
[16] "STDs..Time.since.last.diagnosis"
[17] "Dx.Cancer"
[18] "Dx.CIN"
[19] "Dx.HPV"
[20] "Dx"
```

The data frame of the final result derived from Boruta can be found in the R file

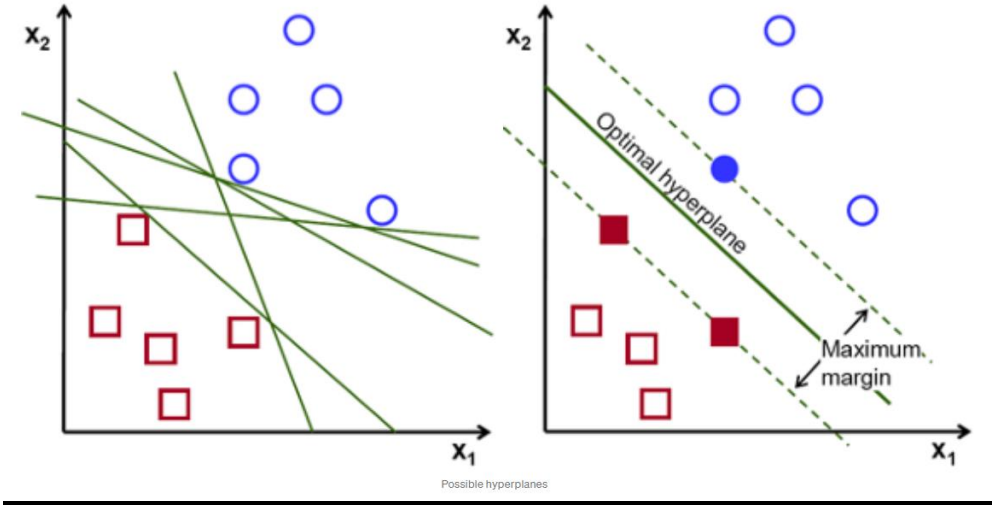
Prediction Models

We will create a new variable “C_cervical_cancer” with all the aforementioned “Confirmed” variables and the “CervicalCancer”. The we assign the values to yes or no for potential for cancers. Since this is a very critical type of cancer and given to the small sample of potential cancer patient, it is better that we take all symptoms (1-4) seriously instead of leaving anything out and assign them with 1, the rest is 0. Then we encode the target features as factors and split the dataset into the training and test set of 0.75/0.25 ratio for predictions.

- **Support Vector Machine**

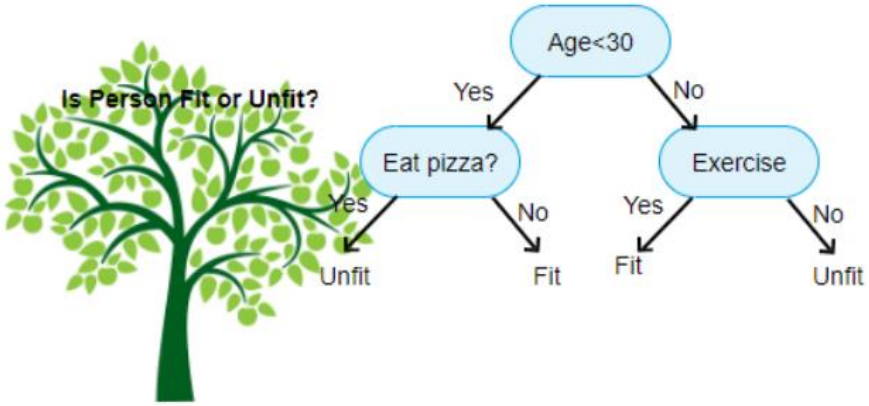
Support Vector Machine (SVM) aims at finding a hyperplane in an N-dimensional space (N- the number of features that distinctly categorizes the data points. There are numerous possible hyperplanes to classify them but we want to find a plane with the maximum margin (distance between data points of both classes). Once achieved, future data points can perform

reinforcement and classify with more confidence , evidently shown in the picture below.(Rohith Gandhi, 2018). In this project, we choose type as “C-classification” abd kernel as “linear”



- **Decision Tree**

Decision Tree model makes judgments by breaking the inputs into smaller decisions. Decision tree is strength lies in the fact that its process has understandable rules and conduct classification task regardless of computational power. It works well with both continuous and categorical variables to provide the researcher a clear indication which independent variables are indispensible. Regardless, it is not recommended to use continuous attributes too often, can produce errors and sometimes energy consuming even on a small number of training examples. Decision Tree is used in tasks like animal pictures classification.



- **Random Forest**

This is an advanced version of Decision tree, except the questions that are posed include some randomness. The algorithm creates a “forest”, an ensembles of mini decision trees through the “bagging method. Then all the trees are merged together to fet a more accurate and stable prediction. Unfortunately, Random forest often over-fits when dealing with noisy data. (Kirill Fuchs, 2017). In this project we choose the number of tree is 500

- **Models comparison and Precision vs. Recall trade off**

After running all the models we can build the Confusion Matrix tables as below.

```
> confusionMatrix(test_set[, 21], y_pred1)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 187   2
         1  25   1

               Accuracy : 0.8744
                 95% CI : (0.8226, 0.9156)
    No Information Rate : 0.986
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0451

 Mcnemar's Test P-Value : 2.297e-05

            Sensitivity : 0.88208
            Specificity : 0.33333
         Pos Pred Value : 0.98942
         Neg Pred Value : 0.03846
             Prevalence : 0.98605
         Detection Rate : 0.86977
   Detection Prevalence : 0.87907
      Balanced Accuracy : 0.60770

       'Positive' Class : 0
```

```
> confusionMatrix(test_set[, 21], y_pred2)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 186   3
         1  26   0

               Accuracy : 0.8651
                 95% CI : (0.8121, 0.9078)
    No Information Rate : 0.986
    P-Value [Acc > NIR] : 1

                  Kappa : -0.0257

 Mcnemar's Test P-Value : 4.402e-05

            Sensitivity : 0.8774
            Specificity : 0.0000
         Pos Pred Value : 0.9841
         Neg Pred Value : 0.0000
             Prevalence : 0.9860
         Detection Rate : 0.8651
   Detection Prevalence : 0.8791
      Balanced Accuracy : 0.4387

       'Positive' Class : 0
```

```
> confusionMatrix(test_set[, 21], y_pred3)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 188   1
         1  25   1

               Accuracy : 0.8791
                 95% CI : (0.8278, 0.9195)
    No Information Rate : 0.9907
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0551

 Mcnemar's Test P-Value : 6.462e-06

            Sensitivity : 0.88263
            Specificity : 0.50000
         Pos Pred Value : 0.99471
         Neg Pred Value : 0.03846
             Prevalence : 0.99070
         Detection Rate : 0.87442
   Detection Prevalence : 0.87907
      Balanced Accuracy : 0.69131

       'Positive' Class : 0
```

We will use the precision and recall as metrics for our models' performance assessment. Precision (specificity) is about ensuring the model's accuracy, indicating the fraction of relevant examples among the retrieved example. Recall (sensitivity), on the other hand, is about model's ability to get the right instances. Precision and Recall range from 0-1 and the higher they are the better. It is noteworthy that there is a trade-off, meaning that when precision rises, recall fails and vice versa. Researchers often use an abstract metric like F1 score, which is the combination of both to make the decision. However, in such dangerous cases like identifying Cervical cancer patient, Recall is more important than precision because the cost of wrong treatment might be high, but not at high as the opportunity cost of wrongly passing up a potential cancer patient. Luckily, I would choose Random forest method due to the fact that it has the highest recall rate (0.8826) and precision rate (0.5) compared to the rest. (Samuel Hillis & Sara Hoormann, 2016)

Conclusion

To sum up, Cervical cancer is one of the most life threatening diseases out there which is responsible for thousands of deaths per year. But implementing the Random Forest method like above, I believe we can save a lot of people in the future.

References

Deepanshu Bhalla. (2017, August 5). Feature Selection : Select Important Variables with Boruta Package. Retrieved February 22, 2021, from https://www.listendata.com/2017/05/feature-selection-boruta-package.html

Fontham, E. T. H., Wolf, A. M. D., Church, T. R., Etzioni, R., Flowers, C. R., Herzig, A., … Smith, R. A. (2020). Cervical cancer screening for individuals at average risk: 2020 guideline update from the American Cancer Society. *CA: A Cancer Journal for Clinicians*, *70*(5), 321–346. https://doi.org/10.3322/caac.21628

Kirill Fuchs. (2017). Machine Learning: Classification Models | by Kirill Fuchs | Fuzz | Medium. Retrieved February 22, 2021, from https://medium.com/fuzz/machine-learning-classification-models-3040f71e2529

Rohith Gandhi. (2018, June 7). Support Vector Machine — Introduction to Machine Learning Algorithms. Retrieved February 22, 2021, from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Samuel Hillis, & Sara Hoormann. (2016, January 16). Precision and Recall: Understanding the Trade-Off. Retrieved February 22, 2021, from https://medium.com/opex-analytics/why-you-need-to-understand-the-trade-off-between-precision-and-recall-525a33919942

UCI Machine Learning Repository: Cervical cancer (Risk Factors) Data Set. (2017). Retrieved February 22, 2021, from https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29

Inspired by :

Divya Bongouni's Kaggle paper: https://www.kaggle.com/divyabongouni/key-cervical-cancer-predictions-with-boruta