National College of Ireland

# An Investigation into Cervical cancer using ensemble learning approach

MSc Research Project
Data Analytics

Rishab Raina
x16132335

School of Computing
National College of Ireland

Supervisor: Cristian Rusu

**National College of Ireland**
**Project Submission Sheet – 2017/2018**
**School of Computing**

| | |
|---|---|
| **Student Name:** | Rishab Raina |
| **Student ID:** | x16132335 |
| **Programme:** | Data Analytics |
| **Year:** | 2017 |
| **Module:** | MSc Research Project |
| **Lecturer:** | Cristian Rusu |
| **Submission Due Date:** | 11/12/2017 |
| **Project Title:** | An Investigation into Cervical cancer using ensemble learning approach |
| **Word Count:** | 6157 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

| | |
|---|---|
| **Signature:** | |
| **Date:** | 11th December 2017 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS:**

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Contents

# An Investigation into Cervical cancer using ensemble learning approach

Rishab Raina

MSc Research Project in Data Analytics

11th December 2017

**Abstract**

Cervical cancer is one of the most daunting disease after breast cancer known to the medical world, it can be fatal if it is not detected at early stages and can lead to death. Diagnosing cervical cancer can be done using pap smear tests which has to be done on a weekly basis making it a tiring process for some and due to the technological limitations in some developing nation the disease can go undetectable which leads to more deaths. This paper studies the different attributes which can cause cervical cancer. This is done by studying patients medical history without using the data for pap smear tests as the aim here was to use ensemble learning as an aid to detect the cause of cervical cancer. In this paper, five different machine learning techniques are used, and an ensemble model is presented and evaluated.

## 1 Introduction

Cervical cancer is the second most dangerous disease after Brest Cancer (Keating (2017)) known to the medical world as it remains incurable at later stages. Lots of recent development have been made to increase the detection rate of the disease by using a machine learning algorithm for the image analysis of the pap smear tests. Meanwhile, various technological advancements in the field of machine learning is bringing us closer day by day to find the cure for the disease which is incurable. Furthermore, machine learning techniques are used nowadays to predict the true outcome of any disease such as cancer, these techniques are also used in the fields such as banking, weather forecasting, etc. The results attained by machine learning techniques are credible, and new studies are done every day to improve the models accuracy and to provide in-depth analysis of a complex data. Machine learning is a trending topic and is used by scientists/biologists to know the facts before they have taken place.

The cervical cancer is caused when there is an abnormal growth of the cells in the cervix which can be caused by many different attributes in the dataset. This report will try to give insights about the different attributes which can influence the disease. The

report from NCC (MedicineNet (2017)) states that the death rate is 85% in the developing nations alone. This statistic is daunting as the developed nations like the USA also face difficulties for the diagnosis of cervical cancer as the disease is treatable if detected at initial stages (Milton (2017)). According to the current technological limitations, a weekly check-up is required to detect any abnormality in the cell growth to detect cervical cancer, and expertise in the field of cervical cancer is also required to analyze the pap smear tests. Most of the researchers have tried to use the data of pap smear test to analyze them using machine learning algorithms so that domain expertise will not be required to study pap-smear tests and the death rate could be decreased. The goal of this research is to study the medical history of the patient to detect the likelihood of having cancer and to analyze different attributes which are most likely to cause cervical cancer. By using machine learning techniques, the above goal is achieved. Around 300 cases of cervical cancer are seen every year in Ireland. ( MedicineNet (2017)). One of the interesting fact about the cervical cancer is that the survival rate increases to 80% if the cancer is diagnosed at the initial stage. As at initial stage the disease only requires medications to be treated, but at later stages it requires radiotherapy and the can be untreatable at later stages which can lead to death. In this research, an ensemble model is created using a very simplified R code using five different machine learning algorithms and the result is presented in terms of specificity and sensitivity to give better insights of the model, and the ensemble model has also been compared against the different single model.

The paper flow is as follows, In section 2 all the related work in this field has been summarized, Section 3 explains the methodology used for the research, section 4 is based on the implementation, section 5 is the evaluation of the research and paper concludes in Section 6.

# 2 Related Work

## 2.1 Intoduction

An extensive amount of studies has been made to predict cervical cancer at initial stages which also includes using machine learning algorithm to classify pap smear images, cervix images, and genomics. However, most of the studies made a focus on classifying the data taken from pap smear test, and cervical image screening which can be a problem for some people as going to medical center is must for doing the relevant test, and this is done on a weekly basis making it a time-consuming task for some of the population. Figure 1 tries to show the problem relating to the screening process of cervical cancer. This paper aims to take out the contribution of different attributes towards Cervical Cancer using patients medical history. Section 2.2 tells us about the previous work done in the field of Cervical cancer. The gap in these paper in accordance with cervical cancer has been identified and discussed in the section 2.3

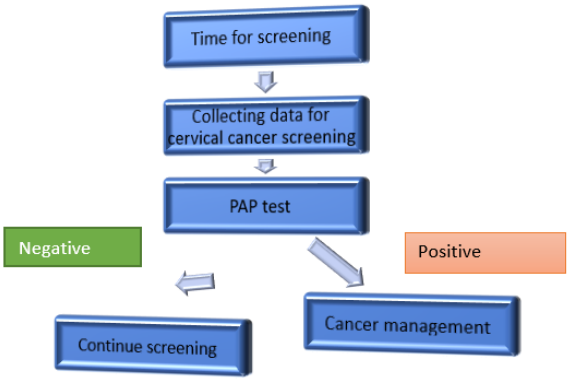


Figure 1: Problem

## 2.2 Literature Review of Classification models on Cervical Cancer

A classification was performed by (Chang et al. (2013)) to predict the recurrence of cancer in patients by using machine learning algorithms on their previous health records. The techniques chosen by the author was Decision Tree C 5.0 and MARS (Multivariate Adaptive Regression Splines). In this paper, the author has taken twelve attributes from the patients medical record such as the age of the patient, tumor type and so forth. The accuracy of the model for Decision tree C5.0 and MARS were 92.44% & 86.60% respectively. The best model according to this paper in terms of chosen performance metric, i.e., Accuracy was Decision Tree C 5.0. This paper concludes by saying that the model created should be validated by a domain expertise in the field of cervical cancer for its effectiveness.

The method such as CART(Classification and regression tree), RFT (Random forest tree) RFT with k-means have been performed by the author (Vidya and Nasira (2016)) on the dataset taken from NCBI (National center for Biotechnology Information). All the models were implemented using MATLAB. The accuracy achieved was 83.87%, 93.54% & 96.77% for the CART, RFT and RFT with k means respectively. The splitting criteria used in the paper involves GINI index which has not been properly explained.

Image classification technique has also been used for predicting cervical cancer using pap smear tests images. This was done by the author (Devi et al. (2016)) using SVM (Support Vector Machines), fuzzy based techniques and texture classification. The aim here was to differentiate cancerous cell from the normal ones by analyzing the images from pap smear tests. SVM model for the image classification achieved the accuracy of 71% in the paper. However, the paper is unclear as there are no images in the paper which is the sole aim of the paper i.e. image analysis. The author could have explained the paper in a better way.

A novel idea was presented by the author (Detection (2016)) to use MRI (Magnetic Resonance Images) to detect cervical cancer SVM (Support vector machine) is used on transform features to detect GLCM, Contour. Pixels are represented by the GLCM (grey

level cooccurrence matrix) as per the author. Image enhancement with feature extraction is done by Contourlet  Gabor respectively. MATLAB was used for implementing the model with an accuracy of 83% as presented in the paper.

The author (Sharma (2016)) has used WEKA (Cs.waikato.ac.nz. (2017)) to handle large data set. A decision tree model was built using WEKA with an accuracy of 37.8% to predict cancer. The accuracy achieved for the model is low which will be used to detect cancer. The author has also used 10-fold cross-validation to deal with the class imbalance.

A simple approach has also been used to predict cancer by using three different techniques namely Decision tree J48, MLP (Multilayer perceptron) and Naive Bayes. This was also implemented in WEKA (Cs.waikato.ac.nz. (2017)) by the author (Latha et al. (2014)). The accuracy of the decision tree was 93.03% with a sensitivity and specificity of 70% and 90% respectively. The aim here was to predict the number of stages in cervical cancer.

An analysis of feature selection technique was performed by the author (Ashok and Aruna (2016)). The aim of the paper was to know which feature selection technique is the among the other four analyzed. The techniques analyzed were Mutual Information, Sequential Forward Search, Sequential Floating Forward Search and Random Subset feature selection. The accuracy achieved was 98% for the sequential floating forward search determining the best feature selection method among other four.

A deep learning technique is used to analyze an existing framework of medical records with patients having cervical cancer. The aim here was to increase the efficiency of the framework as the existing one suffered from a very low recall value and a very high false negative rate/specificity. Convolutional neural nets were used on images to extract the information from them, and then the author (Xu et al. (2016)) combined the extracted data with the non-image data for training the neural net. This end to end deep learning technique achieved the recall of 87.83% and 90% for the true negative rate respectively.

A database of patients medical record was created by (Sarwar (2016)). The aim of the author was to use the database for the screening of Pap smear images to bring down the death rate. For achieving this task, the author has used 15 machine learning techniques to automate the screening process fully from start to end. The 15 different techniques used in the paper were Random Sub Space, Random forest, Naive Bayes..., Multiclass classifier., Random Committee., END., Bagging., artificial neural network., Radial Basis Function., FC..., Decision Table., J48., PART. and Decorate. The author tested the algorithm for two class and multi-class with an accuracy of 93% and 73% respectively.

The aim of the author (Lisboa (2006)) was to use the artificial neural network for increasing the benefits during the tests of cervical cancer. The author has used artificial neural networks to increase the benefits from the clinical trials, and the results were credible, 21 out of 28 patients showed a sign of improvement with three patients having no change/no improvement. The author has used the ANN technique for image analysis and also described the benefits of ANN in the field of cancer.

Different validation techniques with different machine learning models were used to study

cervical cancer. The author (Kourou et al. (2015)) has used different k-fold techniques such as 10-fold validation, 20-fold validation and so on. SVM (Support Vector machine) model trained with 10-folds gave the most accuracy i.e. 95%. The goal of the paper was to make one understand the benefits of different validation techniques and how they benefit when used accordingly with the right machine learning algorithm. All the models were successfully evaluated in the paper.

The author (Sarwar et al. (2015)) has proposed a hybrid ensemble technique which improves the predictive performance of the model for the classification of pap smear images in cervical cancer testing. The idea here is to classify the set of images and decide whether the patient is developing the disease or not. The author has analyzed the samples of pap smear images using hybrid ensemble technique with an accuracy of 78% for 7- class problem. The results were achieved by writing a MATLAB code. The results are beneficial as they are helping to analyze the images taken from pap smear test without the human intervention.

### 2.3 Identified Gaps and Conclusion

All the research made on the cervical cancer are based on the classification of pap-smear tests or a classification among normal, abnormal and cancerous cells. Some researchers have used machine learning techniques such as decision tress and ANN to predict the disease at initial stages. However, we do not know which factors are contributing to the disease. Can drinking alcohol cause cervical cancer? Or The number of smokes/year or /day? Or Is it some other STD which is contributing towards cervical cancer? Or Can we have a better predictive model? This research will try to answer the questions raised, which is done by analyzing the patients medical history. The collected data will be analyzed using an ensemble model created in a R code using a data frame which contains data from different models created, and the contribution of different attributes towards Cervical cancer will be determined.

## 3 Methodology

CRISP-DM (Cross Industry Standard Process for Data Mining) is the backbone of this research (Locke (2017)).CRISP-DM here is used to understand the problem and make the process easier by solving each puzzle in various steps. CRISP-DM has five stages in the modified approach namely Problem understanding, Data collection and preparation, Making the model, Evaluation of the result and deployment. In this research, the CRISP-DM model is modified as per the need for this study.

Figure 2 shows us the modified CRISP-DM approach. The first step was to understand the downsides of the current methods used to predict cervical cancer. The death rate due to the cervical cancer is growing every year, and the patient still needs to go to the doctor every week for the checkup for a proper diagnostic. Researchers have used

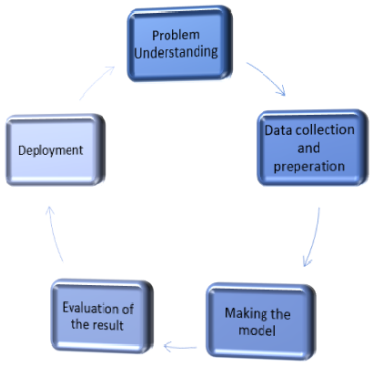


Figure 2: Modified CRISP-DM

that data for image classification so that a domain expertise is not required in analyzing the test. This approach is beneficial in some way, but it beats the purpose of having an advanced algorithm which can be trained to predict cervical cancer without having the data for pap smear test, thus saving time. The second stage was to acquire the data and clean it.The dataset is taken from UCI machine learning repository(Kelwin Fernandes and Fernandes. (n.d.)). The dataset has 27 predictor variables which were cleaned by analyzing the weight to information gain ratio in R. Some of the attributes whose contribution was less towards the target variable were removed and were not deemed important for this research. Though some of the attributes having lower contribution were kept for the analysis as they were important as per the research perspective. There were 835 records of the patients. The dataset contained null values, and those were replaced by 0 - (meaning false) so that results are not effected. Machine learning approach can only give us better results if the data is prepared properly and all the features are selected carefully.

Some of the attributes from the data set include HPV, Smokes/ day, Smokes/year, and different type of STD's are present in the data which may be responsible for causing the virus. All the missing values have been eradicated using excel to increase the performance of the model and for not giving misleading results. The third step was to make five different machine learning models in R and combine them using an ensemble approach using R code. The five model which were used in this research is random forest, C4.5, GBM, neural networks and KNN. These all models were combined by their predictor table using XGBoost as the classifier.

All the results have been evaluated and compared in terms of the kappa value achieved, the accuracy of the model, Sensitivity and the specificity of the model. The different performance metric is needed to represent a different set of models in different domains, for this research Sensitivity and the specificity of the model has been more focused in context to this research.

## 3.1 Proposed Solution

A different idea of combining five different algorithms have been done using the R programming. The predict table from all the five different algorithms was taken to build a

data frame so that new data can be given to a classifier algorithm in this case XGBoost. All the data gathered was then boosted and the performance achieved was credible. The important features which can influence cervical cancer have also been presented in the Feature importance plot in section 5.2. The code was also optimized into one R script and has been explained further in section 3.1.1

### 3.1.1 Code Optimization

The R code used in this research have been optimized in one R script. The machine learning code for each of the algorithm was huge initially and had to be written in separate R scripts making it complex for implementing an ensemble model. In this research, we present an ensemble model which was built using the caret package in one R script making it much smaller in size and more comfortable to understand than before.

# 4 Implementation

## 4.1 Data Preparation

Figure 3 shows the original data which was collected from the UCI machine learning repository (Kelwin Fernandes and Fernandes. (n.d.)). The dataset had 36 attributes out of which only 23 attributes were taken as per the weight by information gain ratio in R. Some of the attributes having low contribution were also kept as per the research perspective as they seemed necessary as per the other researches conducted in the past. The target variable is Biopsy which is binomial i.e. 0 & 1. The data has been acquired,

| | Age | Number of | First sexua | Num of pr | Smokes | Smokes (y | Smokes (p | Hormonal | Hormonal | IUD | IUD (years | STDs | STDs (num | STDs:cond | STDs:cervi | STDs:vagin | STDs:vulvc | STDs:syphi | STDs:pelvi | STDs:genit | STDs:molli | STDs:AIDS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 18 | 4 | 15 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 15 | 1 | 14 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 34 | 1 | 15 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 52 | 5 | 16 | 4 | 1 | 37 | 37 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 46 | 3 | 21 | 4 | 0 | 0 | 0 | 1 | 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 42 | 3 | 23 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 51 | 3 | 17 | 6 | 1 | 34 | 3.4 | 0 | 0 | 1 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Figure 3: Cervical cancer Data

and all the knowledge related to this field has been gained through related work which tells us all the work which has been done in this domain. That data set had few null values as the patients didn't want to disclose it. The values have been replaced by 0 as per the data source. Data transformation has been performed in excel to ensure that the model gives the best outcome.

In this research, we are proposing an ensemble model which is combined using XGBoost classifier by using the package caret in R. There are five different algorithms used for ensemble model namely random forest, GBM, C4.5, neural net, KNN. All these models have been compared against the ensemble model in the following section. Feature selection has been performed using weight by information gain ratio. Features like HPV, Smokes/year, etc. are essential as per the weight by information ratio implemented in R.

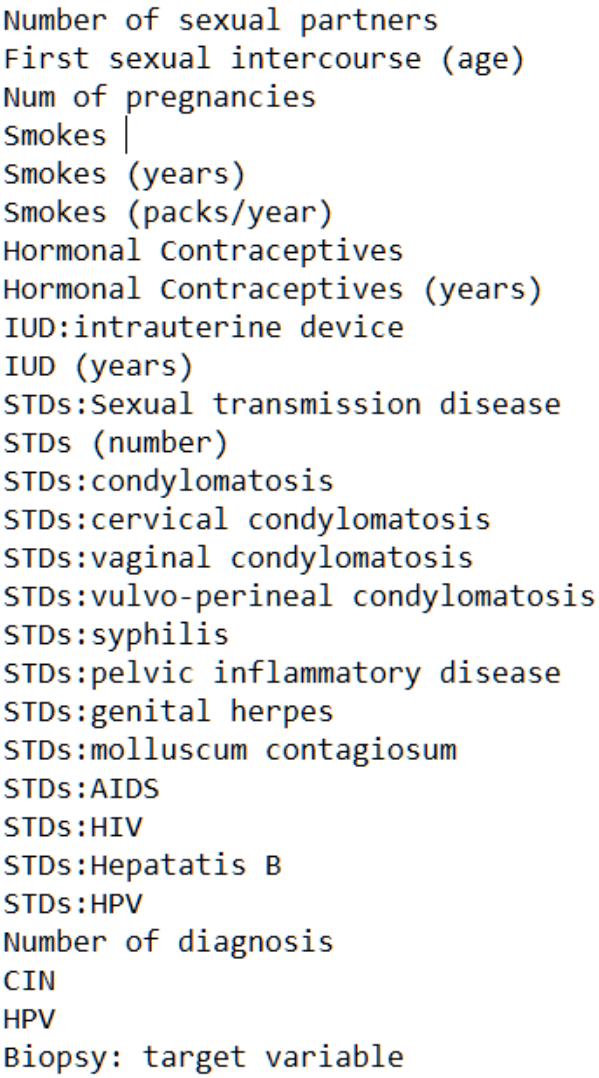


Figure 4: Data attributes

Figure 4 shows us the attributes in our data set. All these attributes have been analyzed using R, and the attribute which has the most influence over the target variable has been presented using variable importance plot in a random forest. The aim of this research is to provide the attributes which have the most effect on cervical cancer and also to increase the overall efficiency of the model. To train the models, we are using 70% of the data as a training set and 30% of the data as a testing set. The splitting of the data has been performed using R code.

```
> weights <- information.gain(Biopsy~., cc)
> print(weights)
                                    attr_importance
Number.of.sexual.partners                0.00000000
First.sexual.intercourse                 0.00000000
Num.of.pregnancies                       0.06589180
Smokes                                   0.00000000
Smokes..years.                           0.00000000
Smokes..packs.year.                      0.00000000
Hormonal.Contraceptives                  0.00000000
Hormonal.Contraceptives..years.          0.02988800
IUD                                      0.00000000
IUD..years.                              0.02346489
STDs                                     0.00000000
STDs..number.                            0.02645493
STDs.condylomatosis                      0.00000000
STDs.cervical.condylomatosis             0.00000000
STDs.vaginal.condylomatosis              0.00000000
STDs.vulvo.perineal.condylomatosis       0.00000000
STDs.syphilis                            0.00000000
STDs.pelvic.inflammatory.disease         0.00000000
STDs.genital.herpes                      0.00000000
STDs.molluscum.contagiosum               0.00000000
STDs.AIDS                                0.00000000
STDs.HIV                                 0.00000000
STDs.Hepatitis.B                         0.00000000
STDs.HPV                                 0.00000000
STDs..Number.of.diagnosis                0.00000000
CIN                                      0.00000000
HPV                                      0.05225539
```

Figure 5: Information by weight ratio

Figure 5 represents the information by weight ratio. The weight by information ratio is taken out using R code and shows the weights of different attributes towards our target variable biopsy. The variables with low importance can be removed but are kept so that you can determine if there is any change in the importance after the machine learning algorithm has been performed.

## 4.2 Machine learning models

Classification of the data set has been performed using R.The assessment of these models is primarily done by splitting data into 70-30. Sensitivity & Specificity is used in this research to give a broader perspective about the dataset. Different models trained for making an ensemble model are as follows :

1. Random Forest :- Random forests are one of the most popular classification techniques used by the researchers worldwide as random forest makes many trees for the single data input and then the voting is done to predict the class. It is the most effective way to deal with larger data set according to (Stat.berkeley.edu (2017)). As per the (Stat.berkeley.edu (2017)) Random Forests can handle missing value without deleting a variable. There is no pruning in Random Forests. Random forest also gives a variable importance plot which was of utmost importance for this research to determine the variables which may effect the cervical cancer.

2. Artificial Neural Networks :- The process of Artificial neural network is based on how the human brain works. Similar to the human brain there is a network of neurons which are responsible for making the model learn. Artificial neural network are generally described as the black box of machine learning algorithm as the process of working is very difficult to understand. The advantage of neural network is that they modify themselves overtime while learning new insights about the data each

time a neuron sees a pattern. Each variable is taken based on its node weight and given importance in the formation of neural network The variables which have higher weight are given more importance in the formation of the neural network. In this research hidden layer value was set to 1 as it gave the best accuracy as seen in the Figure 6.

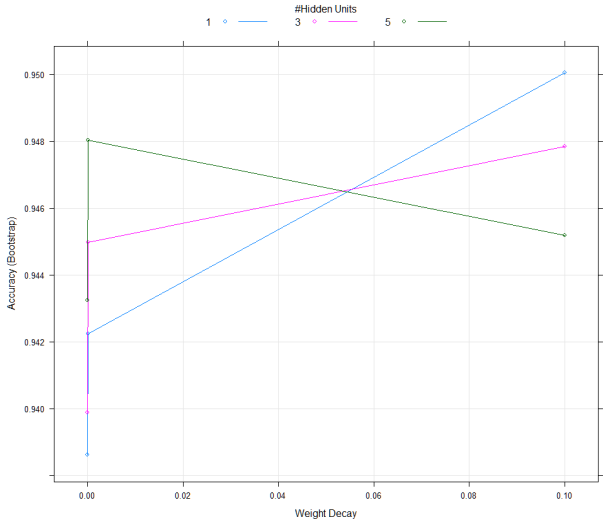


Figure 6: Hidden layers

3. Decision tree C 4.5:- Decision tree C4.5/J48 are often used for the classification. Decision tree is one of the oldest machine learning technique.(Lucidchart (2017)) The decision tree form a tree-like structure to make decisions and predict possible outcomes. Decision trees are easy to build and relatively easy to understand comparatively to Artificial neural network and Support vector machine In this research classification was performed using a decision tree. The model used for training was made under caret package in R using method as J48.

4. GBM:- GBM classifies the weak learners in the set of trees and boosts them. It is a very powerful technique build over decision trees, and it contains various function to determine the weak learner additive model.(Brownlee (2017)). In other words, we can call GBM as an ensemble model of a weak set of trees. Classification and regression problems can be handled by GBM with the basic idea of generalizing the model using a loss function (Natekin and Knoll (2017)). In this research, all the default parameters have been used

5. KNN:- K- Nearest Neighbour algorithm is a non-parametric technique which is used to classify data based on the distance function stored by the algorithm on the training Data Set. Here K stands for the number of closest training data sets which is used to determine the class of it. In this research, we have taken the value of K as 9 which was automatically decided by the model. The value of k as 9 was the best for doing the implementation as shown in the Figure 7.

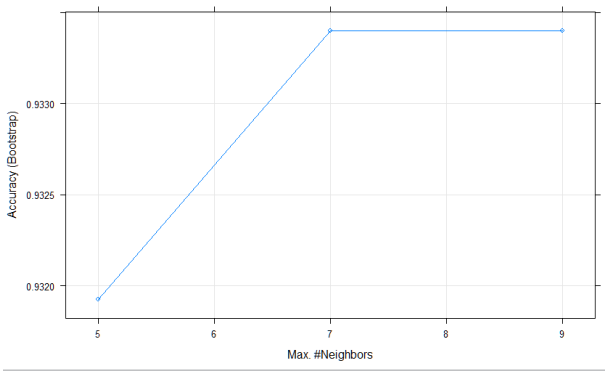


Figure 7: Kmax selection

6. Ensemble model:- An ensemble model was built using the R code in Rstudio. The research proposes an idea of combining five different machine learning algorithm to make a better model. The model was trained using the predictions table taken out of five various machine learning algorithms. These predictions tables were then combined in a data frame with the class variable, and the XGBoost was used to make an ensemble model. XGBoost is a step forward from gradient boosted trees and is now used widely for boosting.

# 5 Evaluation

## 5.1 Performance Evaluation

The performance evaluation of the model has been done in accordance to this report.The data was split into the ratios of 70:30 to make a training set to train the various models and the testing set was use to evaluate the model. In this research Sensitivity and Specificity has been more focused as the problem is that the data is related to medical world and requires more focus on true positive rate and true negative rate. If the true positive rate/sensitivity will be higher there will be less false negatives which is of the utmost important here in this research and same goes for the true negative rate which will mean there are very less false positive. Further performance metric can also be calculated using True positive/negative & False positive/negative from the configuration matrix. Kappa has also been chosen for this research as that statistics provides the agreement rate between the expected and predicted values. Also output from the random forest is presented to show the influence of the different attributes on over target variable. Results are presented in a tabular form and also 3D-bar charts are used for giving the graphical representation of the result.

## 5.2 Feature importance

In the Figure 8, a graphical representation of different attributes has been shown. This importance plot is build using random forest and according to the plot we can interpret that the HPV has influenced the cervical cancer. The second and third most important variable effecting the cervical cancer is the age of the first sexual intercourse and the hormonal contraceptive/year. The HIV has the least effect on the cervical cancer.

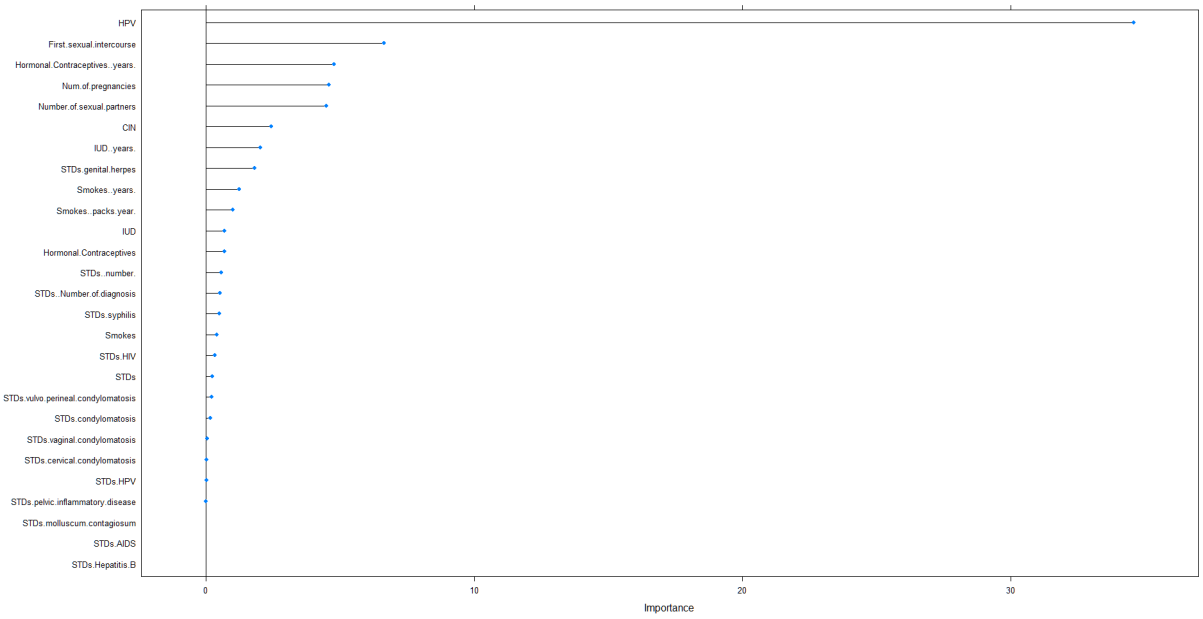


Figure 8: Feature importance

## 5.3 Accuracy, Specificity & Sensitivity

The figure 10 shows the comparison of the different machine learning models RF, GBM ,C4.5, ANN & KNN against the ensemble model implemented using XGBoost. In this reseacrch our focus is on sensitivity and specificity values because accuracy can be a misleading factor when it comes to the medical data as the goal is to know the false negative outcomes which can be determined using specificity. The ensemble model has the best overall results with the accuracy of 97%, sensitivity/true positive rate of 84% and specificity/true negative rate of 97%.The statistical equation to take out the Accuracy, Sensitivity and specificity are as follows.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{1}$$

$$Sensitivity = TP/(TP + FN) \tag{2}$$

$$Specificity = FP/(FP + TN) \tag{3}$$

*Where TP:- True positive*
*TN:- True Negative*
*FP:- False Positive*
*FN:- False Negative*

```
> cmatrix

results6   0   1
       0 237   5
       1   3  12
```

| TN | FN |
|---|---|
| FP | TP |

Figure 9: Confusion Matrix Ensemble model

The figure 9 shows us the confusion matrix of the ensemble model.

Since the sensitivity of the ensemble model is 84% Type II error will is extremely less which is beneficial in this case as this research deals with medical data, making it sensitive towards Type II error. The results prove that the ensemble model is the best fit for the data. The graphical representation of the table has been presented in the Figure 11.

| . | RF | GBM | C4.5 | ANN | KNN | Ensemble Model |
|---|---|---|---|---|---|---|
| Accuracy | 96.00% | 94.00% | 95% | 95.00% | 95% | 97% |
| Sensitivity | 83.00% | 55.00% | 83.00% | 72% | 61% | 84% |
| Specificity | 97.00% | 97.00% | 96.00% | 97% | 98% | 97% |

Figure 10: Performance comparison

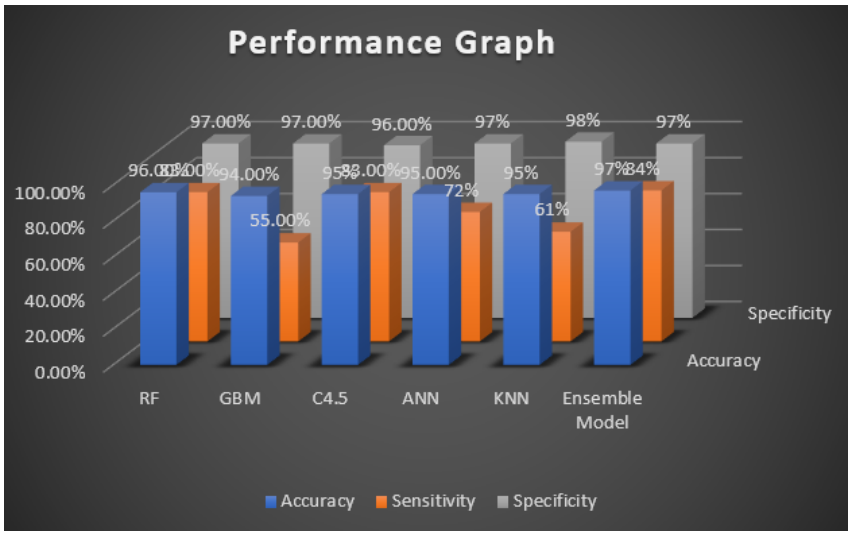


Figure 11: Performance Plot

The plot in Figure 12 shows the True positive rate across different models. The true positive rate is of utmost importance in this paper as already discussed in 5.3. GBM has the lowest sensitivity when compared to other models.The graph proves that the ensemble model is better for doing research related to medicine.

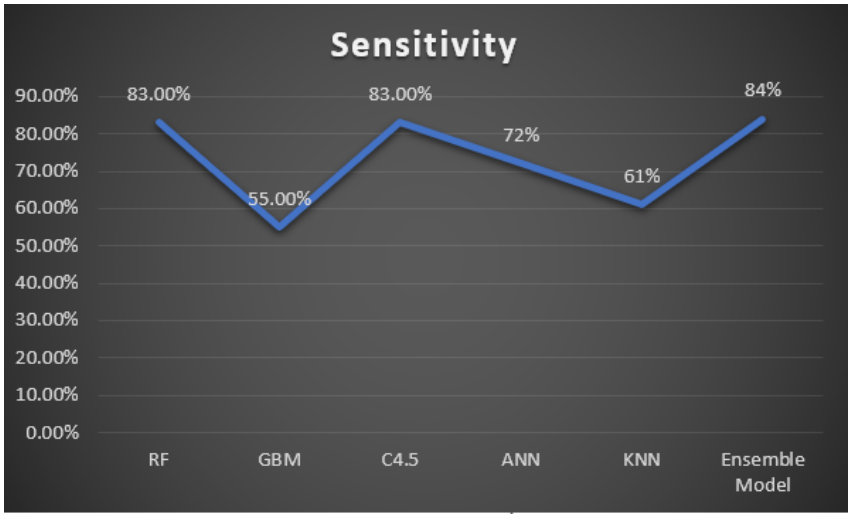


Figure 12: True Positive Rate

## 5.4 Kappa Statistics

In the figure 13 the kappa value of the models have been shown. The kappa value gives us the agreement rate between the expected and the predicted outcome. 0-0.20 means poor agreement,0.21-0.40 means fair agreement,0.41-0.60 means moderate agreement, 0.61-0.80 means good agreement and 0.81-1 means excellent agreement. In this research there was a good agreement between the expected vs predicted values with the overall percentage of 77 for the ensemble model.

| . | RF | GBM | C4.5 | ANN | KNN | Ensemble model |
|---|---|---|---|---|---|---|
| Kappa | 0.72 | 0.57 | 0.7 | 0.65 | 0.64 | 0.77 |

Figure 13: Kappa comparison

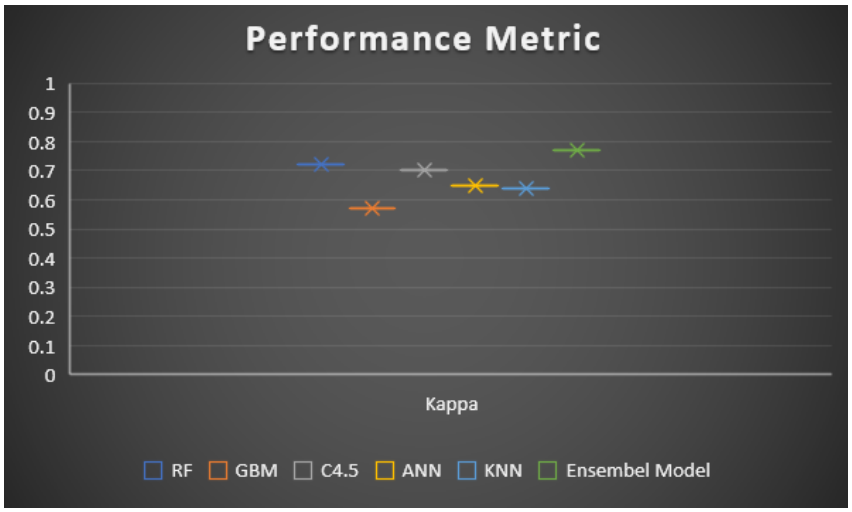


Figure 14: Performance Plot

## 5.5 Comparison with other methods

This aim of this research was to create an ensemble model so that the sensitivity can be increased. According to the Literature, the other authors who have already worked on ensemble model used the tool WEKA(Sarwar et al. (2015)) and have achieved the accuracy of 78.7% compared to 97% produced by the model in this paper which was implemented using R code.
The decision tree J48 model implemented by (Latha et al. (2014)) has achieved the accuracy of 93.03% with the specificity and sensitivity of 90% & 70% respectively. The model implemented in this research has outperformed the model presented in the literature with a difference of 3% regarding accuracy and the sensitivity achieved in this case was 83% compared to 70% making it even a better model for researching medicine. The model in this research has decreased the TYPE II error considerably.
The Random Forest algorithm implemented by (Vidya and Nasira (2016)) has achieved 93.54% accuracy which is lower when compared to the model presented in this report with the accuracy of 96%.
According to the comparison with literature, it has been proved that the ensemble model performed in this research has outperformed all the models in the literature.

## 5.6 ROC Curve

The ROC curve shown in the figure 15 is for the ensemble model. The AUC(Area under the curve) value achieved by the model was 0.86 which tells that the model implemented was a good fit.

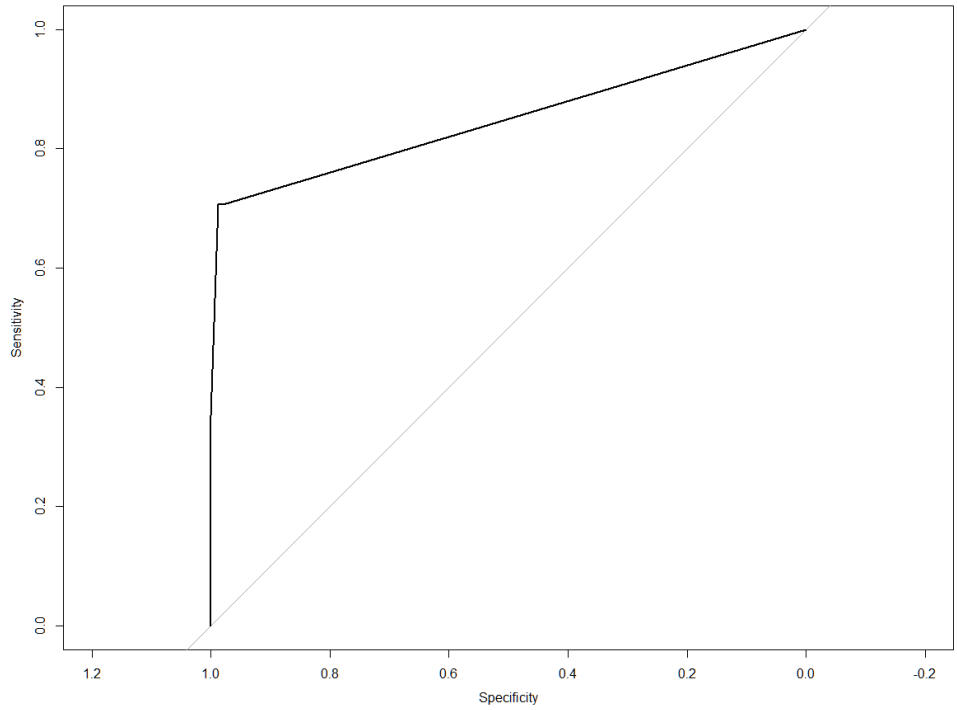


Figure 15: ROC Curve

# 6 Conclusion and Future Work

An ensemble model has been implemented in this research successfully and it has been proven that it outperforms all the model when compared to the Literature. The objective here was to find out the variables influencing the Cervical cancer which has been clearly shown in the section 5.2. The model implemented has decreases the value of type II error as compared to the previous model referenced in this research. The ensemble model proves to be highly accurate among all the model with the accuracy of 97%. In the future a large data set can be acquired to make the model more efficient, however finding a data related to cervical cancer is a toughest part as the privacy of the patients are at stake.

Future studies can include working on end to end deep learning models which can greatly improve the sensitivity of the model and the model created here can be applied on other diseases related to the medical world so that the performance can be verified.

# Acknowledgement

I would like to express my sincere gratitude to my supervisor Prof. Cristian Rusu. for his continuous support in this thesis, for his patience, motivation, and immense knowledge. His guidance helped me in all the time...

# References

Ashok, B. and Aruna, P. (2016). Comparison of Feature selection methods for diagnosis of cervical cancer using SVM classifier, **6**(1): 94–99.

Brownlee, J. (2017). A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning - Machine Learning Mastery.
**URL:** *http://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/*

Chang, C.-C., Cheng, S.-L., Lu, C.-J. and Liao, K.-H. (2013). Prediction of Recurrence in Patients with Cervical Cancer Using MARS and Classification, *International Journal of Machine Learning and Computing* **3**(1): 75–78.
**URL:** *http://www.ijmlc.org/index.php?m=content&c=index&a=show&catid=35&id=273*

Cs.waikato.ac.nz. (2017). Data mining with open source machine learning software in java$_2$017.
**URL:***https://www.cs.waikato.ac.nz/ml/weka/*

Detection, C. C. (2016). Cervical Cancer Detection and Prevention, *Cancer Detection and Prevention* **9**(2): 663–671.

Devi, M. A., Ravi, S. and Punitha, J. V. S. (2016). Detection of Cervical Cancer using the Image Classification Algorithms, **9**(3): 1589–1602.

Keating, M. (2017). Cervical Cancer - Marie Keating Foundation. [online] Marie Keating Foundation.
**URL:** *http://www.mariekeating.ie/cancer-information/cervical-cancer/*

Kelwin Fernandes, J. S. C. and Fernandes., J. (n.d.). Transfer learning with partial observability applied to cervical cancer screening.' iberian conference on pattern recognition and image analysis.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction, *Computational and Structural Biotechnology Journal* **13**: 8–17.
**URL:** *http://dx.doi.org/10.1016/j.csbj.2014.11.005*

Latha, D. S., Lakshmi, P. V. and Fathima, S. (2014). Staging Prediction in Cervical Cancer Patients A Machine Learning Approach, **2**(2): 14–23.

Lisboa, P. J. G. (2006). The use of artificial neural networks in decision support in cancer: a systematic review, *Neural Netw.* **19**: 408–415.

Locke, S. (2017). CRISP-DM and why you should know about it.
**URL:** *https://itsalocke.com/crisp-dm/*

Lucidchart (2017). what is a decision tree diagram.
**URL:** *https://www.lucidchart.com/pages/decision-tree*

MedicineNet (2017). Genital Warts (HPV) Picture Image on MedicineNet.com.
**URL:** *http://www.medicinenet.com/image-collection/genital*$_w$*arts*$_h$*pv*$_p$*icture/picture.htm*

Milton, S. (2017). International Cervical Cancer - NCCC.
**URL:** *http://www.nccc-online.org/about-nccc/international-cervical-cancer/*

Natekin, A. and Knoll, A. (2017). Gradient boosting machines, a tutorial.
**URL:** *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3885826/*

Sarwar, A. (2016). Cervical Cancer : Open Access Artificial Intelligence Based Semi-automated Screening of Cervical Cancer Using a Primary Training Database, **1**(1): 1–10.

Sarwar, A., Sharma, V. and Gupta, R. (2015). Hybrid ensemble learning technique for screening of cervical cancer using Papanicolaou smear image analysis, *Personalized Medicine Universe* **4**(March): 54–62.
**URL:** *http://linkinghub.elsevier.com/retrieve/pii/S2186495015000024*

Sharma, S. (2016). Harnessing the Power of Decision Tree approach in Machine Learning for Cervical Cancer Stage Prediction using See5 and SIPINA, **2**(2): 1176–1182.

Stat.berkeley.edu (2017). classification description$_2$017.
**URL:***https://www.cs.waikato.ac.nz/ml/weka/*

Vidya, R. and Nasira, G. M. (2016). Prediction of cervical cancer using hybrid induction technique: A solution for human hereditary disease patterns, *Indian Journal of Science and Technology* **9**(30): 1–10.

Xu, T., Zhang, H., Huang, X., Zhang, S. and Metaxas, D. N. (2016). Multimodal deep learning for cervical dysplasia diagnosis, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9901 LNCS**: 115–123.