2020-01-01

Using Machine Learning On An Imbalanced Cancer Dataset

James Ekow Arthur
University of Texas at El Paso

USING MACHINE LEARNING ON AN IMBALANCED CANCER DATASET

JAMES ERNEST EKOW ARTHUR

Master's Program in Mathematical Sciences

APPROVED:

Maria C Mariani, Ph.D., Chair

Guthrie Joe, Ph.D.

Sarkodie-Gyan, Ph.D.

Stephen Crities, Ph.D.

Dean of the Graduate School

USING MACHINE LEARNING ON AN IMBALANCED CANCER DATASET

by

JAMES ERNEST EKOW ARTHUR

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Mathematical Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

August 2020

Acknowledgements

I would like to express my deep-felt gratitude to my advisor, Dr. Christiana Mariani Maria of the Mathematical Science Department at The University of Texas at El Paso, for her advice, encouragement, enduring patience and constant support.

I also wish to thank the other members of my committee, Dr. Guthrie Joe of the Mathematics Department and Dr. Sarkodie -Gyan of the Engineering Department, at The University of Texas at El Paso.

Additionally, I want to thank The University of Texas at El Paso Mathematical Science Department professors and staff for all their hard work and dedication, providing me the means to complete my degree and prepare for a career as a Mathematical scientist. This includes (but certainly is not limited to) the following individuals:

Dr. Osei Tweneboah

> He made it possible for me to have many wonderful experiences as a student. He made my learning experience a wonderful one and guided me through this work .

Dr. Bhuiyan, Md Al Masum He guided me through the simulations .

Abstract

With an estimated 1.4 million cancer diagnosis worldwide and the increasing death of cancer patients. It is prudent to investigate methods, approaches and smarter ways of predicting and diagnosing of cancer so that a holistic techniques can be used to curb or reduce false predictions , increase exact predictions and also meticulos prognosis information .

Can a feasible technique be developed for the general problem of prognosis and diagnosis of cancer be developed ?

We will show here that this problem of cancer prognosis and diagnosis can be efficiently tackled with the aid of machine learning techniques and the best, feasible and efficient technique can be used to reduce this cancer menace.

Cancer has been characterized as a heterogeneous disease consisting of many different subtypes. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. The importance of classifying cancer patients into high or low risk groups has led many research teams, from the biomedical and the bioinformatics field, to study the application of machine learning (ML) methods. Therefore, these techniques have been utilized as an aim to model the progression and 16 treatment of cancerous conditions. In addition, the ability of ML tools to detect key features from complex datasets reveals their importance. A variety of these techniques, including Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Decision Trees (DTs) have been widely applied in cancer research for the development of predictive models, resulting ineffective and accurate decision making. Even though it is evident that the use of Machine Learning methods can improve our understanding of cancer progression, an appropriate level of validation is needed in order for these methods to be considered in the everyday clinical practice. In this work, what we present are view of recent ML approaches employed in the modeling of

cancer prognosis and diagnosis.

Table of Contents

List of Figures

Chapter 1

Introduction

Machine learning is not new to cancer research. Artificial neural networks (ANNs) and decision trees (DTs) have been used in cancer detection and diagnosis for nearly 20 years (Simes 1985; Maclin et al. 1991; Ciccheti 1992). Today machine learning methods are being used in a wide range of applications ranging from detecting and classifying tumors via X-ray and CRT images (Petricoin and Liotta 2004; Bocchi et al. 2004) to the classification of malignancies from proteomic and genomic (microarray) assays (Zhou et al. 2004; Dettling 2004; Wang et al. 2005). According to the latest PubMed statistics, more than 1500 papers have been published on the subject of machine learning and cancer. However, the vast majority of these papers are concerned with using machine learning methods to identify, classify, detect, or distinguish tumors and other malignancies. In other words machine learning has been used primarily as an aid to cancer diagnosis and detection (McCarthy et al. 2004). It has only been relatively recently that cancer researchers have attempted to apply machine learning towards cancer prediction and prognosis. As a consequence the body of literature in the field of machine learning and cancer prediction/prognosis is relatively small.

The fundamental goals of cancer prediction and prognosis are distinct from the goals of cancer detection and diagnosis. In cancer prediction/prognosis one is concerned with three predictive foci

1. the prediction of cancer susceptibility (i.e. risk assessment)
2. the prediction of cancer recurrence and
3. the prediction of cancer survivability

In the first case, one is trying to predict the likelihood of developing a type of cancer prior to the occurrence of the disease. In the second case one is trying to predict the likelihood of redeveloping cancer after to the apparent resolution of the disease. In the third case one is trying to predict an outcome (life expectancy, survivability, progression, tumor-drug sensitivity) after the diagnosis of the disease. In the latter two situations the success of the prognostic prediction is obviously dependent, in part, on the success or quality of the diagnosis. However a disease prognosis can only come after a medical diagnosis and a prognostic prediction must take into account more than just a simple diagnosis (Hagerty et al. 2005).

Indeed, a cancer prognosis typically involves multiple physicians from different specialties using different subsets of biomarkers and multiple clinical factors, including the age and general health of the patient, the location and type of cancer, as well as the grade and size of the tumor (Fielding et al. 1992; Cochran 1997; Burke et al. 2005). Typically histological (cell-based), clinical (patient-based) and demographic (population-based) information must all be carefully integrated by the attending physician to come up with a reasonable prognosis. Even for the most skilled clinician, this is not easy to do. Similar challenges also exist for both physicians and patients alike when it comes to the issues of cancer prevention and cancer susceptibility prediction. Family history, age, diet, weight (obesity), high-risk habits (smoking, heavy drinking), and exposure to environmental carcinogens (UV radiation, radon, asbestos, PCBs) all play a role in predicting an individual's risk for developing cancer (Leenhouts 1999; Bach et al. 2003; Gascon et al. 2004; Claus 2001; Domchek et al. 2003). Unfortunately these conventional "macro-scale" clinical, environmental and behavioral parameters generally do not provide enough information to make robust predictions or prognoses. Ideally what is needed is some very specific molecular details about either the tumor or the patient's own genetic make-up (Colozza et al. 2005).

With the rapid development of genomic (DNA sequencing, microarrays), proteomic (protein chips, tissue arrays, immuno-histology) and imaging (fMRI, PET, micro-CT) technologies, this kind of molecular-scale information about patients or tumors can now be readily ac-

quired. Molecular biomarkers, such as somatic mutations in certain genes (p53, BRCA1, BRCA2), the appearance or expression of certain tumor proteins (MUC1, HER2, PSA) or the chemical environment of the tumor (anoxic, hypoxic) have been shown to serve as very powerful prognostic or predictive indicators (Piccart et al. 2001; Duffy 2001; Baldus et al. 2004). More recently, combinations or patterns of multiple molecular biomarkers have been found to be even more predictive than single component tests or readouts (Savage and Gascoyne 2004; Petricoin and Liotta 2004; Duffy 2005; Vendrell et al. 2005) If these molecular patterns are combined with macro-scale clinical data (tumor type, hereditary aspects, risk factors), the robustness and accuracy of cancer prognoses and predictions improves even more. However, as the number of parameters we measure grows, so too does the challenge of trying to make sense of all this information.

In the past, our dependency on macro-scale information (tumor, patient, population, and environmental data) generally kept the numbers of variables small enough so that standard statistical methods or even a physician's own intuition could be used to predict cancer risks and outcomes. However, with today's high-throughput diagnostic and imaging technologies we now find ourselves overwhelmed with dozens or even hundreds of molecular, cellular and clinical parameters. In these situations, human intuition and standard statistics don't generally work. Instead we must increasingly rely on non-traditional, intensively computational approaches such as machine learning. The use of computers (and machine learning) in disease prediction and prognosis is part of a growing trend towards personalized, predictive medicine (Weston and Hood 2004). This movement towards predictive medicine is important, not only for patients (in terms of lifestyle and quality-of-life decisions) but also for physicians (in making treatment decisions) as well as health economists and policy planners (in implementing large scale cancer prevention or cancer treatment policies).

Given the growing importance of predictive medicine and the growing reliance on machine learning to make predictions, we believed it would be of interest to conduct a detailed review of published studies employing machine learning methods in cancer prediction and prognosis. The intent is to identify key trends with respect to the types of machine learning

methods being used, the types of training data being integrated, the kinds of endpoint predictions being made, the types of cancers being studied and the overall performance of these methods in predicting cancer susceptibility or patient outcomes. Interestingly, when referring to cancer prediction and prognosis we found that most studies were concerned with three "predictive" foci or clinical endpoints: 1) the prediction of cancer susceptibility (i.e. risk assessment); 2) the prediction of cancer recurrence and 3) the prediction of cancer survivability. We also found that almost all predictions are made using just four types of input data: genomic data (SNPs, mutations, microarrays), proteomic data (specific protein biomarkers, 2D gel data, mass spectral analyses), clinical data (histology, tumor staging, tumor size, age, weight, risk behavior, etc.) or combinations of these three. In comparing and evaluating the existing studies a number of general trends were noted and a number of common problems detected. Some of the more obvious trends include a rapidly growing use of machine learning methods in cancer prediction and prognosis (Figure 1), a growing reliance on protein markers and microarray data, a trend towards using mixed (proteomic + clinical) data, a strong bias towards applications in prostate and breast cancer, and an unexpected dependency on older technologies such as artificial neural networks (ANNs). Among the more commonly noted problems was an imbalance of predictive events with parameters (too few events, too many parameters), overtraining, and a lack of external validation or testing. Nevertheless, among the better designed and better validated studies it was clear that machine learning methods, relative to simple statistical methods, could substantially (15–25) improve the accuracy of cancer susceptibility and cancer outcome prediction. In other words, machine learning has an important role to play in cancer prediction and prognosis.

1.1 Background of the Study

Over the past decades, a continuous evolution related to cancer research has been performed. Scientists applied different methods, such as screening in early stage, in order

to find types of cancer before they cause symptoms. Moreover, they have developed new strategies for the early prediction of cancer treatment outcome. With the advent of new technologies in the field of medicine, large amounts of cancer data have been collected and are available to the medical research community. However, the accurate prediction of a disease outcome is one of the most interesting and challenging tasks for physicians. As a result, ML methods have become a popular tool for medical researchers. These techniques can discover and identify patterns and relationships between them, from complex datasets, while they are able to effectively predict future outcomes of a cancer type. Given the significance of personalized medicine and the growing trend on the application of ML techniques. In these studies prognostic and predictive features are considered which may be independent of a certain treatment or are integrated in order to guide therapy for cancer patients, respectively. In addition, we discuss the types of ML methods being used, the types of data they integrate, the overall performance of each proposed scheme while we also discuss their pros and cons. An obvious trend in the proposed works includes the integration of mixed data, such as clinical and genomic . However, a common problem that we noticed in several works is the lack of external validation or testing regarding the predictive performance of their models. It is clear that the application of ML methods could improve the accuracy of cancer susceptibility, recurrence and survival prediction. Based on , the accuracy of cancer prediction outcome has significantly improved by 15–20 percent the last years, with the application of ML techniques.Several studies have been reported in the literature and are based on different strategies that could enable the early cancer diagnosis and prognosis. Specifically, these studies describe approaches related to the profiling of circulating miRNAs that have been proven a promising class for cancer detection and identification. However, these methods suffer from low sensitivity regarding their use in screening at early stages and their difficulty to discriminate benign from malignant tumors. Various aspects regarding the prediction of cancer outcome based on gene expression signatures are discussed . These studies list the potential as well as the limitations of microarrays for the prediction of cancer outcome. Even though gene signatures could significantly improve our ability

for prognosis in cancer patients, poor progress has been made for their application in the clinics. However, before gene expression profiling can be used in clinical practice, studies with larger data samples and more adequate validation are needed. In the present work only studies that employed ML techniques for modeling cancer diagnosis and prognosis are presented.

1.2 Problem Statement

Every year, Pathologists diagnose 14 million new patients with cancer around the world. That's millions of people who'll face years of uncertainty. Pathologists have been performing cancer diagnoses and prognoses for decades. Most pathologists have a 96–98 percent success rate for diagnosing cancer. They're pretty good at that part. The problem comes in the next part. According to the Oslo University Hospital, the accuracy of prognoses is only 60 percent for pathologists. A prognosis is the part of a biopsy that comes after cancer has been diagnosed, it is predicting the development of the disease. It's time for the next step to be taken in pathology.

Machine Learning (ML) is one of the core branches of Artificial Intelligence. It's a system which takes in data, finds patterns, trains itself using the data and outputs an outcome. So what makes a machine better than a trained professional? ML has key advantages over Pathologists. Firstly, machines can work much faster than humans. A biopsy usually takes a Pathologist 10 days. A computer can do thousands of biopsies in a matter of seconds. Machines can do something which humans aren't that good at. They can repeat themselves thousands of times without getting exhausted. After every iteration, the machine repeats the process to do it better. Humans do it too, we call it practice. While practice may make perfect, no amount of practice can put a human even close to the computational speed of a computer. Another advantage is the great accuracy of machines. With the advent of the Internet of Things technology, there is so much data out in the world that humans can't possibly go through it all. That's where machines help us. They can do work faster than

us and make accurate computations and find patterns in data. That's why they are called computers. There is the need to compare the Machine learning techniques used for the prognosis and diagnosis of cancer so that the best optimum method is used on the data for the best outcome.

1.3 Objectives of the Study

The objectives of the study includes the following;

- To Introduce the machine learning techniques for the Study.
- To perform Data analysis of the Data .
- To perform simulations and show outputs of the machine learning techniques and models generated
- To perform simulations to determine the effect of varying the models.
- Comparison of the models.

1.4 Methodology

In this study machine learning techniques namely Logistic regression , Random Forest ,Support Vector Machine, Linear Support Vector Machine . Support Vector Machine (Radial Basis kernel "Gaussian") Support Vector Machine Radial (Non-Linear Kernel). SVM Radial (Non-Linear Kernel) with grids.

We are going to use the Breast cancer data set for this project . The data had ten (10) attributes ranging from id number , Clump Thickness , Uniformity of Cell Size , Uniformity of Cell Shape , Marginal Adhesion , Single Epithelial Cell Size , Bare Nuclei , Bland Chromatin, Normal Nucleoli , Mitoses, Class: for benign, malignant

The goal of this project is to predict the binary class of breast Cancer present, which represents whether or not a patient has heart disease (0, or 1). We proposed seven supervised ML techniques to compare their predictive ability of heart disease. These techniques help us to make order the variables in terms of importance that play roles in causing heart disease. The results suggest that the ML techniques are effective in classifying patients into "YES" and "NO" Breast Cancer present.

So the diagnosis, prognosis, and preventing heart disease are very important. Good data-driven systems for predicting heart disease can improve the research and prevention process, making sure that more people can live healthy lives.
Several ML models will be trained to accurately predict whether a sample patient has been diagnosed with Breast Cancer by training it on the dataset mentioned above. We will randomly split the data into training set (50% for building a predictive model) and test set (25% for evaluating the model) and used set.seed for reproducibility. Finally, We will predict the test data based on the key variables and compute the prediction accuracy using the Precision / Recall that validates the fitted model with the data.

1.5 Significance of the study

This study is a step towards understanding and appreciating the usage of machine learning in cancer prognosis and diagnosis . A clear understanding of cancer predictions is highlighted and the approaches and ideas can be used to tackle other health related ailments in future .

1.6 Organisation of Thesis

This study is organized into five chapters and outlined as follows

- **Introduction** This chapter presents a general introduction to the study with a background to the study, the problem statement, objectives, methodology and the significance of the study.
- **Literature Review**: In this chapter various literatures with relation to Machine Learning, Classification techniques, are presented
- **Methodology**: This chapter discusses various methods adopted for the study with a focus on definitions as well as a case study.

- **Model Formulation and Classification Analysis**: In this chapter, the model is formulated for each technique and analysis of each technique is explained.
- **Analysis and Simulations:** In this chapter, further analysis are discussed with simulations .
- **Conclusion and Recommendations**: This chapter concludes the entire study and lays out some recommendations for future studies.

Chapter 2

Literature Review

In this chapter we provide a brief introduction to machine learning and the techniques used in this study.

Machine Learning (ML) is one of the core branches of Artificial Intelligence. It's a system which takes in data, finds patterns, trains itself using the data and outputs an outcome. So what makes a machine better than a trained professional? ML has key advantages over Pathologists. Firstly, machines can work much faster than humans. A biopsy usually takes a Pathologist 10 days. A computer can do thousands of biopsies in a matter of seconds. Machines can do something which humans aren't that good at. They can repeat themselves thousands of times without getting exhausted. After every iteration, the machine repeats the process to do it better. Humans do it too, we call it practice. While practice may make perfect, no amount of practice can put a human even close to the computational speed of a computer. Another advantage is the great accuracy of machines. With the advent of the Internet of Things technology, there is so much data out in the world that humans can't possibly go through it all. That's where machines help us. They can do work faster than us and make accurate computations and find patterns in data. That's why they're called computers.

Machines are by nature not intelligent. Initially, machines were designed to perform specific tasks, such as running on the railway, controlling the traffic flow, digging deep holes, traveling into the space, and shooting at moving objects. Machines do their tasks much faster with a higher level of precision compared to humans. They have made our lives easy and smooth. The fundamental difference between humans and machines in performing their work is intelligence. The human brain receives data gathered by the five senses: vision,

hearing, smell, taste, and tactility. These gathered data are sent to the human brain via the neural system for perception and taking action. In the perception process, the data is organized, recognized by comparing it to previous experiences that were stored in the memory, and interpreted. Accordingly, the brain takes the decision and directs the body parts to react against that action. At the end of the experience, it might be stored in the memory for future benefits. A machine cannot deal with the gathered data in an intelligent way. It does not have the ability to analyze data on its own for classification, benefit from previous experiences, and store the new experiences to the memory units; that is, machines do not learn from experience. Although machines are expected to do mechanical jobs much faster than humans. In the street we interact with other machines through a common language, recognize dangers and the ways to avoid them, decide about a disease from its symptoms and laboratory tests, recognize the face of the criminal, and so on. The challenge is to make dumb machines learn to cope correctly with such situations. Because machines have been originally created to help humans in their daily lives, it is necessary for the machines to think, understand to solve problems, and take suitable decisions akin to humans. In other words, we need smart machines. In fact, the term smart machine is symbolic to machine learning success stories and its future targets. The question of whether a machine can think was first asked by the British mathematician Alan Turing in 1955, which was the start of the artificial intelligence history. He was the one who proposed a test to measure the performance of a machine in terms of intelligence. Computers are machines that follow programming instructions to accomplish the required tasks and help us in solving problems. Our brain is similar to a CPU that solves problems for us. Suppose that we want to find the smallest number in a list of unordered numbers. We can perform this job easily. Different persons can have different methods to do the same job. In other words, different persons can use different algorithms to perform the same task. These methods or algorithms are basically a sequence of instructions that are executed to reach from one state to another in order to produce output from input. If there are different algorithms that can perform the same task, then one is right in questioning which algorithm is better. For example, if two

programs are made based on two different algorithms to find the smallest number in an unordered list, then for the same list of unordered number (or same set of input) and on the same machine, one measure of efficiency can be speed or quickness of program and another can be minimum memory usage. Thus, time and space are the usual measures to test the efficiency of an algorithm. In some situations, time and space can be interrelated, that is, the reduction in memory usage leading to fast execution of the algorithm. For example, an efficient algorithm enabling a program to handle full input data in cache memory will also consequently allow faster execution of program.

2.0.1 Artificial Intelligence and Machine Learning

Machine learning is a branch of artificial intelligence that aims at enabling machines to perform their jobs skillfully by using intelligent software. The statistical learning methods constitute the backbone of intelligent software that is used to develop machine intelligence. Because machine learning algorithms require data to learn, the discipline must have connection with the discipline of database. Similarly, there are familiar terms such as Knowledge Discovery from Data (KDD), data mining, and pattern recognition. One wonders how to view the big picture in which such connection is illustrated.

2.0.2 Applications Of Machine Learning

For the purpose of this study. We will limiting ourselve to using machine learning for the prognosis and diagnosis of cancer. However , there are other applications of machine learning

Machine learning has proven itself to be the answer to many realworld challenges, but there are still a number of problems for which machine learning breakthrough is required. The need was felt by the cofounder and ex-chairman of Microsoft, Bill Gates, and was translated into the following wordings on one occasion .The following are some applications of machine learning

- Automatic Recognition of Handwritten Postal Codes

  Today, in order to communicate, we use a variety of digital devices. However, the postal services still exist, helping us send our mails, gifts, and important documents to the required destination. The way machine learning has benefited this sector can be understood by citing the example of the US Postal Service. The US Postal Service was able to exploit the potentials of machine learning in the 1960s when they successfully used machines to automatically read the city/state/ZIP code line of typed addresses to sort letters. Optical character recognition (OCR) technology was able to correctly interpret the postal address using machine learning algorithm. The images that consist of typed, handwritten or printed content of text are readable for humans. In order to make such text content readable for machines, Optical character recognition technology is used.

- Language identification.

  There exists no longer a hidden supposition that the language of the document to be processed is already known. If the language is identified wrongly, it means that we should expect a poor performance from the OCR technology. The OCR technology is one of the applications of pattern recognition, a branch of machine learning. The focus of pattern recognition is to recognize pattern and regularities in data. The data can be text, speech, and/or image. The OCR example is the one in which input data is in the form of an image.

- Computer-Aided Diagnosis.

  Another example of the application of pattern recognition using image data , computer-aided diagnosis Pattern recognition algorithms used in computer-aided diagnosis can assist doctors in interpreting medical images in a relatively short period. Medical images from different medical tests such as X-rays, MRI, and ultrasound are the sources of data describing a patient's condition.

The responsibility of a radiologist is to analyze and evaluate the output of these medical tests that are in the form of a digital image. The short time constraint requires that the radiologist be assisted by machine. Computer-aided diagnosis uses pattern recognition techniques from machine learning to identify suspicious structures in the image. How does an algorithm catch suspicious structure? Supervised learning is done to perform this task. Few thousand labeled images are given to the machine learning algorithm, such as Bayesian classifier, artificial neural network,radial basis function network, and support vector machine. The resulting classifier is expected to classify new medical images correctly. Mistakes in diagnosis by the machine learning algorithm can bring disaster for a family. The fault can cause damage to a person in monetary terms and it can risk his/her life, too.

The following are two importance of other classifiers :

- Suppose our classifier detects breast cancer in a patient who actually had no such disease. The results obtained by the classifier will create harmful psychological conditions for the patient. In order to confirm the result of the classifier, further tests can result in monetary losses for the patient.

- Suppose our classifier does not detect breast cancer in patient who actually has such a disease. This will lead to wrong medical treatment and can threaten the life of the patient in near or far future. In order to avoid such mistakes, the complete substitution of doctor with technology is not recommended. The role of technology should be supportive. It should be the doctor(generally a radiologist) who must take the responsibility of the final interpretation of medical image.

Computer-aided diagnosis is assisting medical doctors or radiologists in the diagnosis of a number of health problems. Few examples are as follows: Pathological brain detection, Breast cancer ,Lung cancer ,Colon cancer , Prostate cancer ,Bone metastases ,Coronary artery disease Congenital heart defect Alzheimer's disease

Chapter 3

Classification Techniques

In this chapter we will introduce all the classification techniques and needed definitions

3.1 Random Forest

The random forest technique is a type of additive model that predicts the data by combining decisions from a sequence of base models. It reduces the variance by avoiding overfitting of the model. The class of base models can be expressed as follows

$$g(x) = f_0(x) + f_1(x) + f(x) \tag{3.1}$$

where the final model g is the sum of simple base models fi. We define each base classifier as a simple decision tree. So it is an ensemble technique that considers multiple learning algorithms to obtain best predictive model.
At this point, all the base models or trees are made independently using a different subsample of the data. Once we have a new generated training set, we divide it randomly into two parts. The two-third samples are used to build a tree and the one-third samples are used to obtain the predictions of trees. We take the majority vote of these one-third predictions as the predicted value for the data point and then we estimate the error.
We now present the algorithm of random forest that is used in this study:

- We first take a random sample of size N with replacement from the data.
- Take a random sample without replacement of the predictors.

- Construct a split by using predictors selected in step 2.
- Repeat steps 2 and 3 for each subsequent split until the tree is as large as desired.
- Drop the out-of-bag data down the tree. We then store the class as- signed to each observation along with each observation's predictor val- ues.
- Repeat steps 1-5 a large number of times.
- For each observation in the dataset, we count the number of trees that it is classified in one category over the number of trees.
- Assign each observation to a final category by a majority vote over the set of trees. Thus, if 51 percent of the time over a large number of trees a given observation is classified as a 1, becomes its classification.

So the random forest include three main tuning parameters such as node size, number of trees (ntree), and number of predictors sampled (mtry) for splitting. To build a best predictive model, we estimate the best tuning pa- rameters and important variables using mean decrease accuracy (MDA) and mean decrease Gini (MDG) indices. The MDA determines the importance of a variable by measuring the change in prediction accuracy, when the values of the variable are randomly permuted compared to the original observations. However, the MDG index is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest.

3.2 Logistic Regression

Logistic Regression is a powerful classification algorithm that is used to predict the probability of a categorical variable. We assume that the predictors$x_k$are independent of each other, so the model has

little or no multi- collinearity. We express the model as:

$$logit(p) = \beta_0 + \beta_1 1x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k, \tag{3.2}$$

$$L(\beta_0 + \beta_1 1x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k) = \prod_{i=1}^{b} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i} \tag{3.3}$$

We now seek to estimate the parameters $\beta_0, \beta_1, \beta_2, \beta_3, ..., \beta_k$ that maximize the likelihood function $L(\beta_0 + \beta_1 1x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k)$

At this point, it is mathematically convenient to maximize the logarithm of likelihood function as follows:

$$l(\beta) = \sum_{i=1}^{N} (y_i \beta^T x_i) - log(1 + e\beta^T x_i) \tag{3.4}$$

We then use the regularization technique to obtain a parsimonious model with important features from the original model. The regularization tech- nique penalizes the magnitude of coefficients of features to minimize the error between predicted and actual observations. In this study, we use the L1 regularization is used.

3.2.1 L1 regularization

We use the lasso-regularization technique by adding a L1 penalty term in Eq. (4). It forces the sum of the absolute value of the regression coefficients to be less than a fixed value. This is due to the fact that the tuning parameter makes certain coefficients to be set to zero, effectively by choosing a simpler model that does not include those coefficients [21]. So we maximize the penalized versions as follows:

$$l_\lambda(\beta) = \sum_{j=1}^{N} (y_i \beta^T x_i - log(1 + e^{\beta^T} x_i)) - \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.5}$$

where λ is a tuning parameter that controls the strength of penalty term. So we select λ

in a way that the resulting model minimizes the out of sample error.

3.3 Support Vector Machine

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection. The advantages of support vector machines are

Effective in high dimensional spaces. Still effective in cases where number of dimensions is greater than the number of samples. Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient. Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial. SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation

3.3.1 Linear Support vector Machine

Linear support vector machines (svms) have become popular for solving classification tasks due to their fast and simple online application to large scale data sets. Note that the Linear SVM also implements an alternative multi-class strategy, that is the multi-class SVM formulated by Crammer and Singer , by using the option multiclass crammersinger'. In practice, one vs rest classification is usually preferred, since the results are mostly similar, but the runtime is significantly less.

For Linear SVM the attributes coefficient and intercept have the shape n classes and n features respectively. Each row of the coefficients corresponds to one of the n classes classifiers and similar for the intercepts, in the order of the "one" class.

3.3.2 Support Vector Machine Radial

Radial kernel support vector machine is a good approach when the data is not linearly separable. The idea behind generating non-linear decision boundaries is that we need to do some nonlinear transformations which transforms them into a higher dimensional space. We do this non-linear transformation using the Kernel trick. Now the performance of SVM is influenced by the values of the tuning parameters. Now there are 2 Tuning parameters in the SVM i.e the regularization parameter . We can implement cross-validation to find the best values of both these tuning parameters which affect our classifier's . Another way of finding the best value for these hyper-parameters is by using certain optimization techniques such as Bayesian Optimization.

3.3.3 Support Vector Machine Radial With Kernel

Radial kernel support vector machine is a good approach when the data is not linearly separable.The idea behind generating non linear decision boundaries is that we need to do some non linear transformations on the features X_i which transforms them to a higher dimension space.We do this non linear transformation using the Kernel trick.Now there are 2 hyper parameters in the SVM i.e the regularization parameter 'c' and γ.We can implement cross validation to find the best values of both these tuning parameters which affect our classifier's $C(X)$ performance. Another way of finding the best value for these hyper-parameters are by using certain optimization techniques such as Bayesian Optimization.

3.3.4 Support vector Machine Radial with grids

Support vector machines and other models employing the kernel trick do not scale well to large numbers of training samples or large numbers of features in the input space, several approximations to the RBF kernel (and similar kernels) have been introduced.Typically, these take the form of a function z that maps a single vector to a vector of higher dimensionality, approximating the kernel:

3.4 Imbalanced Datasets

Unbalanced datasets are prevalent in a multitude of fields and sectors, and of course, this includes the health industry and the financial services. The challenge appears when machine learning algorithms try to identify these rare cases in rather big datasets. Due to the disparity of classes in the variables, the algorithm tends to categorize into the class with more instances, the majority class, while at the same time giving the false sense of a highly accurate model. Both the inability to predict rare events, the minority class, and the misleading accuracy detracts from the predictive models we build. In simple terms, an unbalanced dataset is one in which the target variable has more observations in one specific class than the others.

3.4.1 Why checking this Imbalanced Datasets

Models trained on unbalanced datasets often have poor results when they have to generalize (predict a class or classify unseen observations). Despite the algorithm you choose, some models will be more susceptible to unbalanced data than others. Ultimately, this means you will not end up with a good model, and the reasons include: The algorithm receives significantly more examples from one class, prompting it to be biased towards that particular class. It does not learn what makes the other class "different" and fails to understand the underlying patterns that allow us to distinguish classes.

3.4.2 How to check Imbalanced Datasets

- Change your approach . Instead of building a classifier, sometimes it is beneficial to change your approach and the scope ; one option would be to analyze your data from the 'anomaly detection' point of view. Then, you can apply from 'One Class SVM' to 'Local Outlier Factor (LOF)' algorithms.

- Collect more data from the minority class. This option appears trivial, but it solves

the problem when it is applicable.

- Resample the dataset . Apart from using different evaluation criteria, one can also work on getting different dataset. Two approaches to make a balanced dataset out of an imbalanced one are under-sampling and over-sampling.. For the purpose of this thesis we will use this approach to curb the imbalanced dataset.

- Use penalized models . Many algorithms have their own penalized version. Usually, algorithms treat all misclassifications the same, so the idea is to penalize misclassifications from the minority class more than the majority. Mistakes made during training carry an additional cost (that is why they are called cost-sensitive classifiers), but in theory, these penalties help the model improve the attention given to the minority class. Sometimes, the penalties are called weights. Achieving the correct matrix of penalties can be difficult and sometimes does not do much to improve results, so try several schemas until you find the one that works best for your situation.

Chapter 4

Background Of Data

Breast cancer is a huge killer among women worldwide. An estimated 500,000 women died in 2018 alone. Take about 9 and a half football fields and fill it end-to-end with women. That's how about how many people died from breast cancer in a year. That is very bad. The dataset used in this study is publicly available and was created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA. The data was collated periodically under his supervision. The data had ten (10) attributes ranging from id number , Clump Thickness , Uniformity of Cell Size , Uniformity of Cell Shape , Marginal Adhesion , Single Epithelial Cell Size , Bare Nuclei , Bland Chromatin, Normal Nucleoli , Mitoses, Class: for benign, malignant

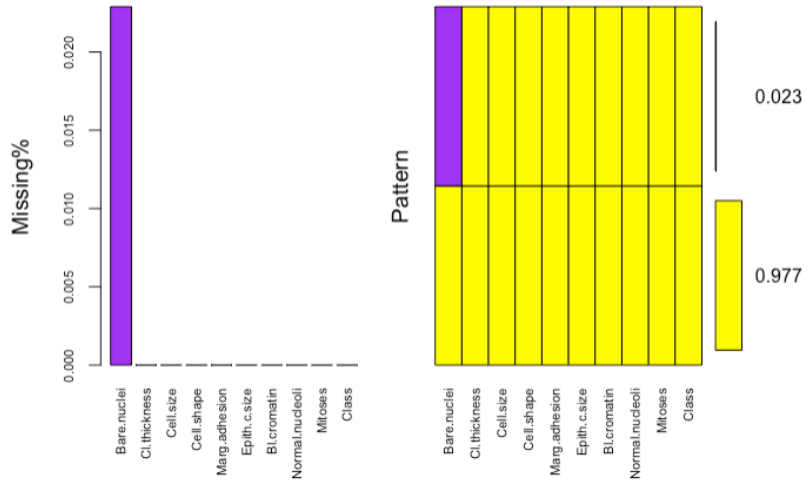
4.1 Analysis of the Data

This analysis aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the breast cancer is benign or malignant. To achieve this we used machine learning classification methods to fit a function that can predict the discrete class of new input.

4.1.1 Exploratory Data Analysis

The dataset was found to have 699 rows and 11 columns. 'Diagnosis' is the column that we will use to predict if the cancer is malignant or benign. M = malignant or B = benign. The data was then visualized . Being able to see that data is very important. It helps us

Figure 4.1: Visualized Dataset



understand what the data is saying and it makes it easier to interpret .

From the visualised data we note that the only variable with missing data with a percentage of 0.023 is Bare.nuclei. When the balance check was done on the data it was noted that the data had 241 benign variable value and 458 malignant variable value. This is a classical example of an unbalanced dataset. The dataset was balanced by using Both Over Sampling and Under sampling. The balance check was then noted to be 340 benign variable value and 359 variable value for malignant.The percentage ratio of the balanced data was concluded as 48.6 percent for benign and 51.4 for malignant.

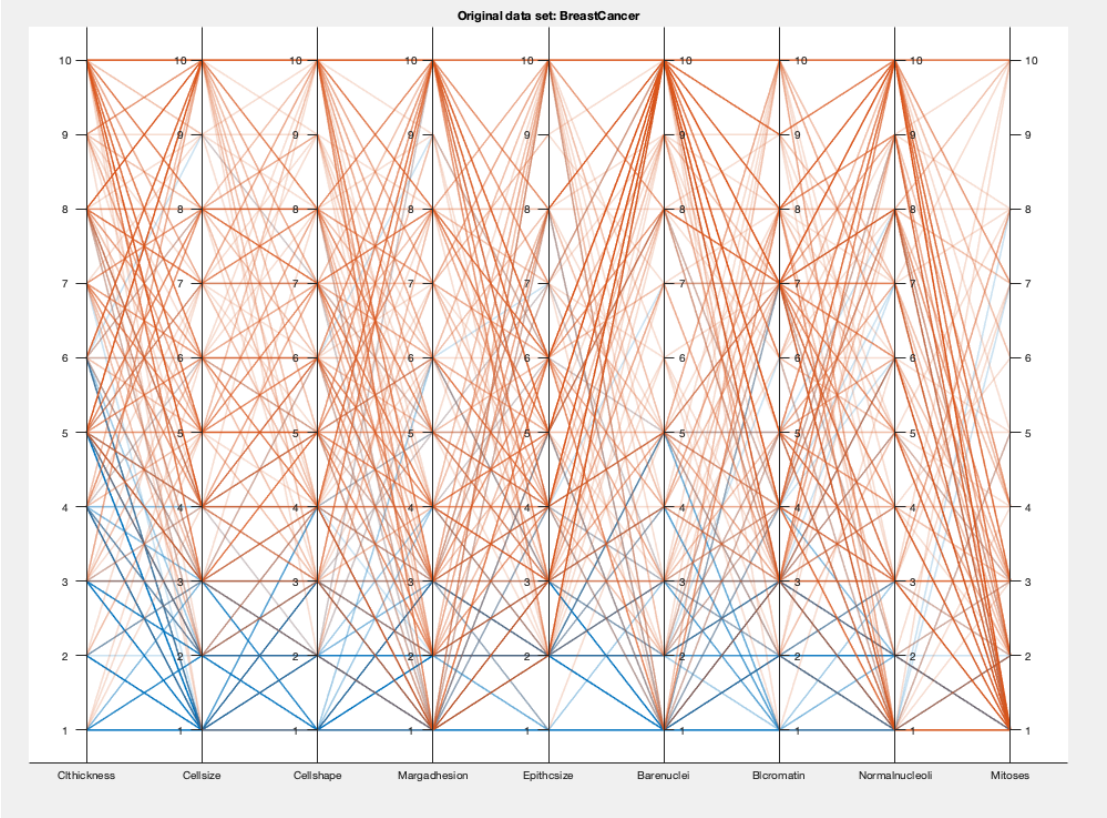


Figure 4.2: Visual representation of the raw dataset by range

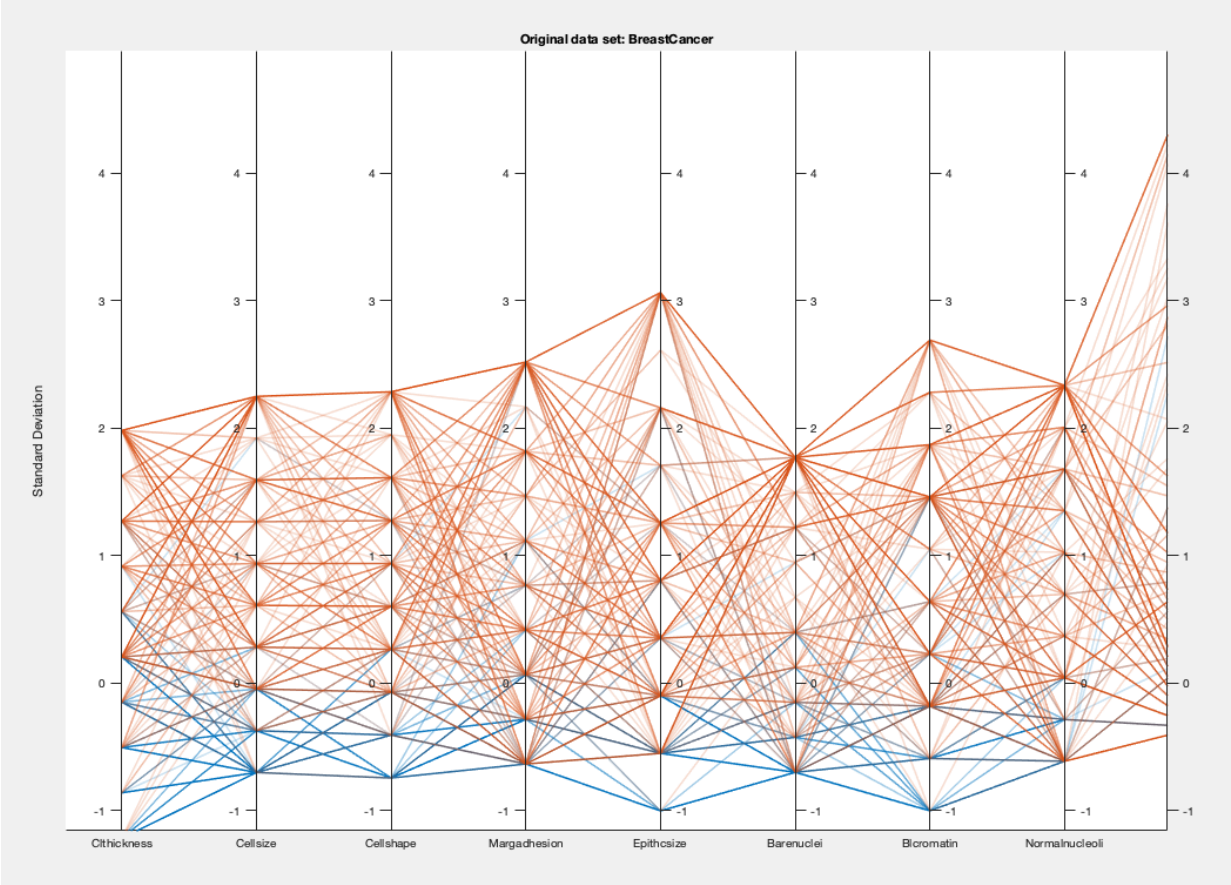


Figure 4.3: Visual representation of the dataset by standardizing

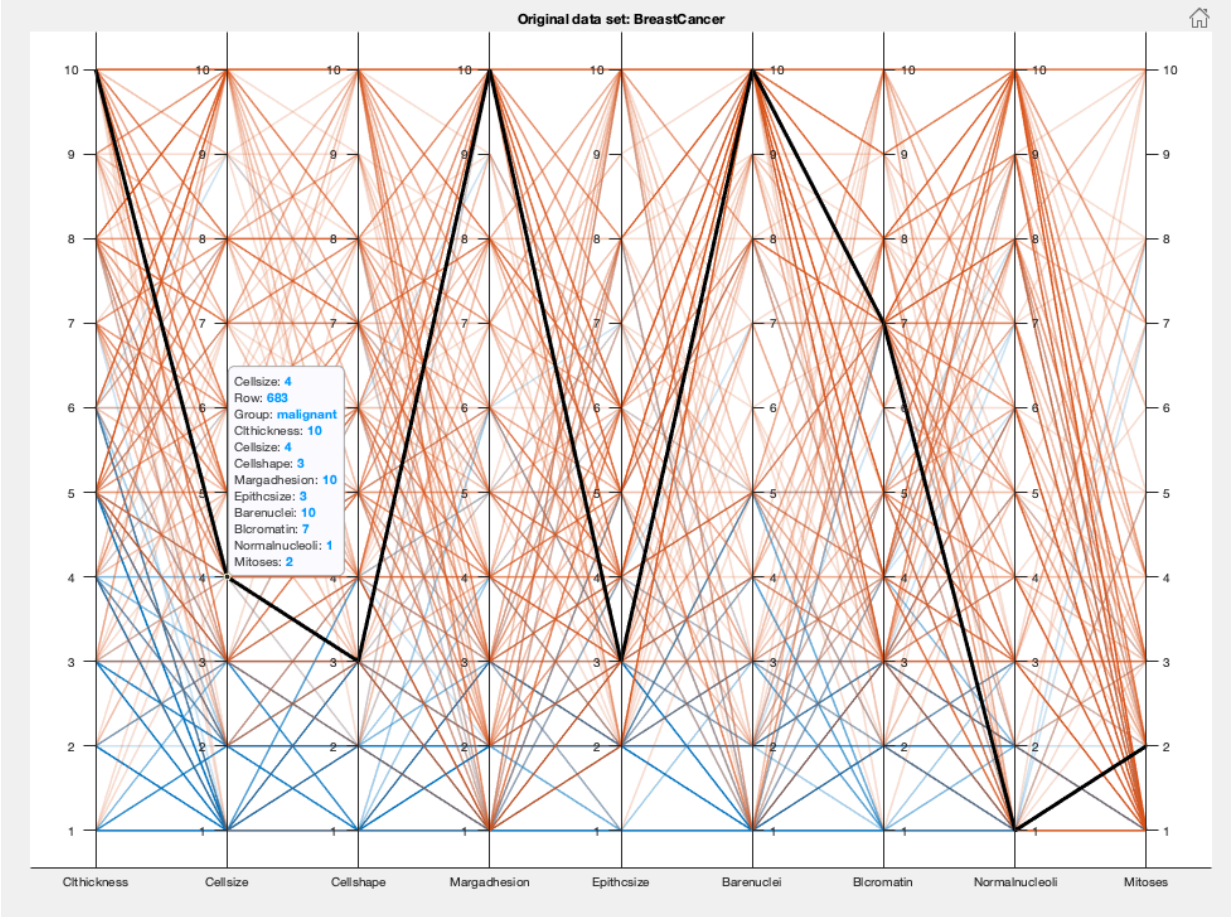


Figure 4.4: Visual representatation of the Cell size

```
      names             pvaules
 [1,] "Cl.thickness"    "2.84683486682793e-74"
 [2,] "Cell.size"       "1.21459962945422e-108"
 [3,] "Cell.shape"      "1.26448507714893e-111"
 [4,] "Marg.adhesion"   "2.48305460877158e-75"
 [5,] "Epith.c.size"    "1.7861705267743e-99"
 [6,] "Bare.nuclei"     "2.01210338144642e-100"
 [7,] "Bl.cromatin"     "8.1233923409114e-96"
 [8,] "Normal.nucleoli" "7.25211702645263e-84"
 [9,] "Mitoses"         "4.37960685248513e-36"
[10,] "0"               "0"
     Association among Class and other predictors Using Fisher test
      names             pvaules
 [1,] "Cl.thickness"    "9.99990000099999e-06"
 [2,] "Cell.size"       "9.99990000099999e-06"
 [3,] "Cell.shape"      "9.99990000099999e-06"
 [4,] "Marg.adhesion"   "9.99990000099999e-06"
 [5,] "Epith.c.size"    "9.99990000099999e-06"
 [6,] "Bare.nuclei"     "9.99990000099999e-06"
 [7,] "Bl.cromatin"     "9.99990000099999e-06"
 [8,] "Normal.nucleoli" "9.99990000099999e-06"
 [9,] "Mitoses"         "9.99990000099999e-06"
[10,] "0"               "0"
```

Figure 4.5: P Values

4.1.2 Exploring the Association between the Class and the Bare Nuclei

From the p-values of the fisher's exact test above, there is an association between the class and the other variables.

4.1.3 Data Splitting

The dataset was splitted into three parts. Namely Training dataset, Testing Data set and Validation Dataset. The data we use is usually split into 3 datasets.The training data set has a known output and the model learns on this data so it can be generalized to other data later on. The test data set is used to test the model's predictions.

Model Selection

In most researches we tend to use a bunch of different machine learning algorithms but at a high level they can all be classified into two groups:

- Supervised Learning
- Unsupervised learning

Supervised learning is a system whereby both input and desired output data are provided. They are labeled for classification to provide a learning basis for any future data processing. Supervised learning problems can be grouped into either regression problems or classification problems.Hence we used supervised learning

Regression Problem

A regression problem is when the output variable is a continuous value like a salary or weight.

Unsupervised Learning

Unsupervised learning is the algorithm that uses information that is neither classified nor labeled and it allows the algorithm to act on the information by its self.

The following classification Techniques were considered for this study

- Logistic Regression
- Random Forest
- Support Vector Machine (Linear, Non Linear, With Grids)
- Artificial Neural Networks

4.2 Results And Discussions

4.2.1 Analysis Of Logistic Regression

In a logistic regression, our main aim and purpose is to classify the response variable y, which is coded as 0 and 1, from high-dimensional explanatory variables. In general, in logistic regression, the response variable y is a Bernoulli random variable. Confusion Matrix is one of the way to measure the performance of a classification problem where the output can be of two or more type of classes The confusion matrix has actual and predicted and furthermore, both the dimensions have True Positives (TP)", True Negatives (TN)", False Positives (FP)", False Negatives (FN)". Explanation of the terms associated with confusion matrix are as follows;

- True Positives(TP) is the case when both actual class and predicted class of data point is 1.
- True Negatives(TN)

  It is the case when both actual class and predicted class of data point is 0.
- False Positives(FP)

  It is the case when actual class of data point is 0 and predicted class of data point is

Accuracy

It is the most common performance metric for classification procedures. It may be defined as the number of correct predictions made as a ratio of all predictions made. Accuracy measures the correct prediction of the classifier compared to the overall data points. It however does not give us the best picture of the cost of misclassification or unbalanced testing data set. A typical classification method should maximize the accuracy.

Precision

Precision is the number of correct positives returned by our ML model. Precision is how consistent results are when measurements are repeated

Recall or Sensitivity

Recall or sensitivity is the number of positives returned by our ML model.

Specificity

Specificity is the number of negatives returned by our ML model.

F1 Score F1 score is the weighted average of the precision and recall. The best value of F1 would be 1 and worst would be 0.

AUC(Area Under ROC Curve) AUC (Area Under Curve)-ROC (Receiver Operating Characteristic) is a performance metric, based on varying threshold values, for classification problems. ROC is a probability curve and AUC measure the separability. AUC-ROC metric will tell us about the capability of model in distinguishing the classes. Higher the AUC, better the model. AUC - ROC curve is a performance measure for classification problem at various thresholds settings. Often, we think that precision and recall both indicate accuracy of the model. While that is somewhat true, there is a deeper, distinct meaning of each of these terms. Precision means the percentage of your results which are relevant. On the other hand, recall refers to the percentage of total relevant results correctly classified by your algorithm. Precision and recall are two extremely important model evaluation metrics. While precision refers to the percentage of your results which are relevant, recall refers to the percentage of total relevant results correctly classified by your algorithm. Unfortunately, it is not possible to maximize both these metrics at the same time, as one comes at the cost of another. For simplicity, there is another metric available, called F-1 score, which is a harmonic mean of precision and recall. For problems where both precision and recall are important, one can select a model which maximizes this F-1 score. For other problems, a trade-off is needed, and a decision has to be made whether to maximize precision, or recall. The perfect test will have a PRC that passes through the upper right corner (corresponding to 100 percent precision and 100 percent recall). Generally you can

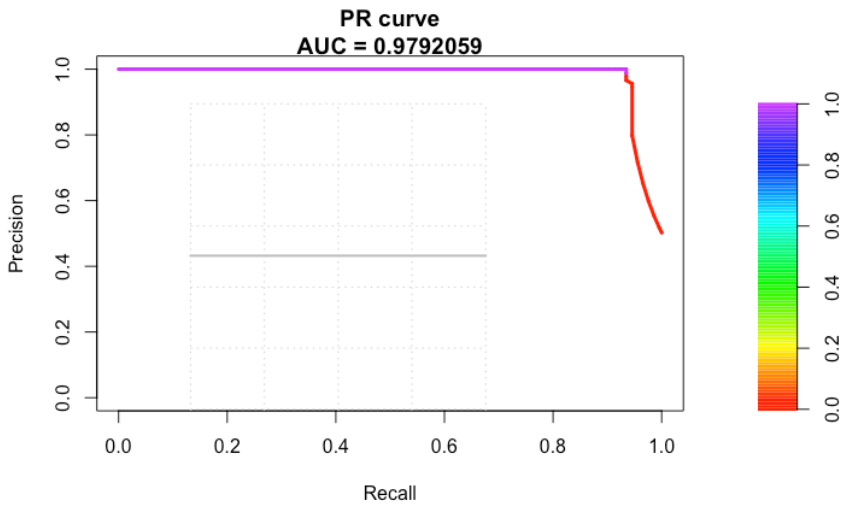
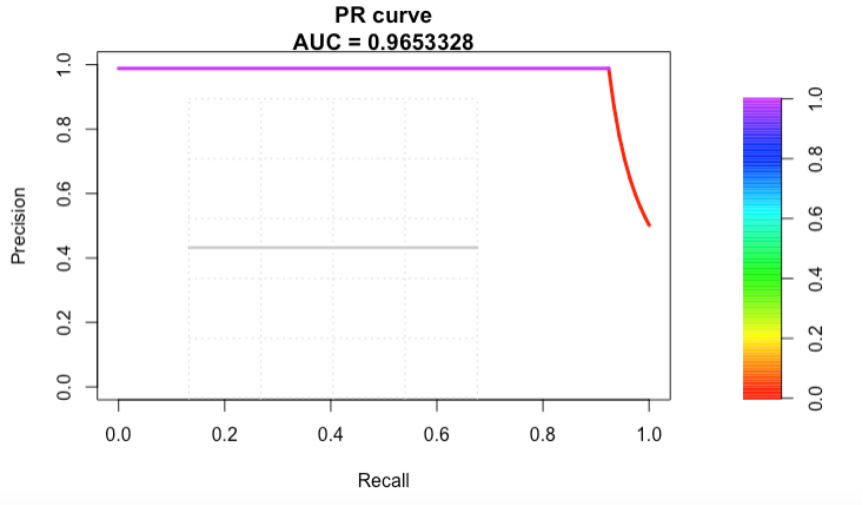


Figure 4.6: Precision/ Recall Logistic Regression

say that the closer a PRC is to the upper right corner, the better the test is. The logistic Regression had a AUC of 0.9792059.

Figure 4.7: SVM Linear



4.2.2 Analysis of Support Vector Machine (Linear)

When considering classification problems, the downfall of linear models is that ultimately, the decision boundary is a straight line, plane or hyperplane with coefficients equal to the models weights / parameters and thus can only classify data which is linearly separable, which could be a big limitation when working on more complex analytics problems. You might be thinking how the SVM which is a linear model can fit a linear classifier to non linear data. Intuitively with a simple linear regression model we may manually engineer x, x^2, x^3 features to attempt to achieve a fit to a non linear set of data points.We notice that the SVM - Linear had a AUC of 0.965

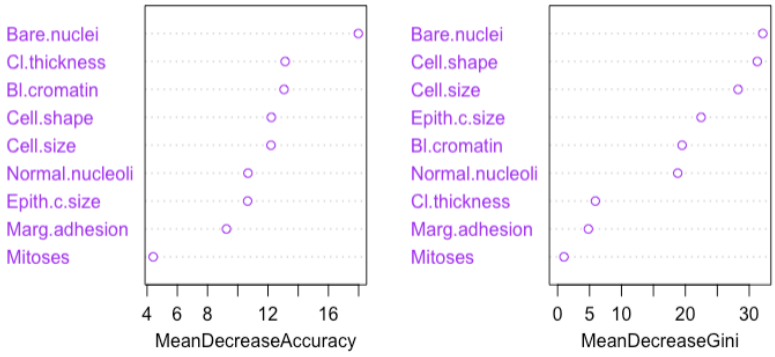


Figure 4.8: Important variable Check

4.2.3 Analysis Of Random Forest

Random forest is considered one of the most important machine learning algorithms. This is due to their relatively good accuracy, robustness and ease of use. The reason why random forests and other ensemble methods are excellent models for some data science tasks is that, they don't require as much pre-processing compare to other methods and can work well on both categorical and numerical input data. A simple decision tree isn't very robust, but random forest which runs many decision trees and aggregate their outputs for prediction produces a very robust, high-performing model and can even control over-fitting.

Feature importance will basically explain which features are more important in training of model. Sometimes training model only on these features will prove better results comparatively

Random forest is a commonly used model in machine learning, and is often referred to as a black box model. In many cases, it outperforms many of its parametric equivalents, and is less computationally intensive to boot.Using above visualizing methods we can understand and make others understand the model and its training

From this two graph we can conclude that the variables Bare nuclei and Cell shape are the important variables.

Per the Precision Recall curve of the Random Forest we notice that the AUC is 0.992.

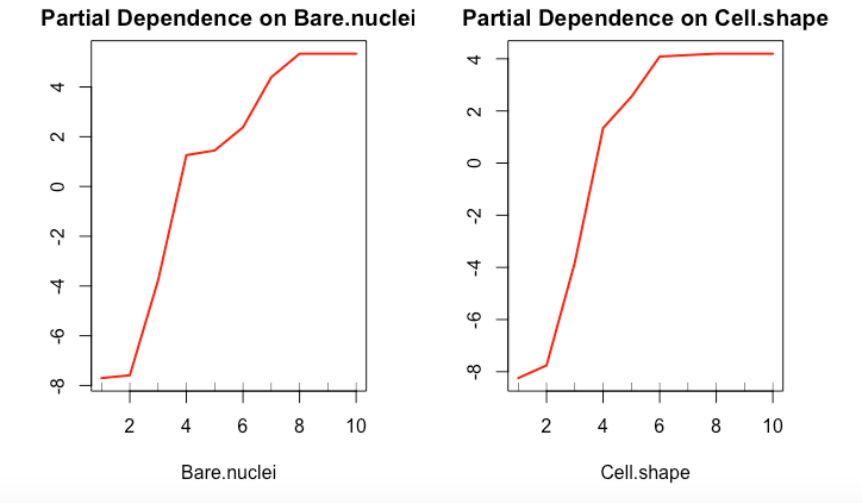


Figure 4.9: Partial Dependence

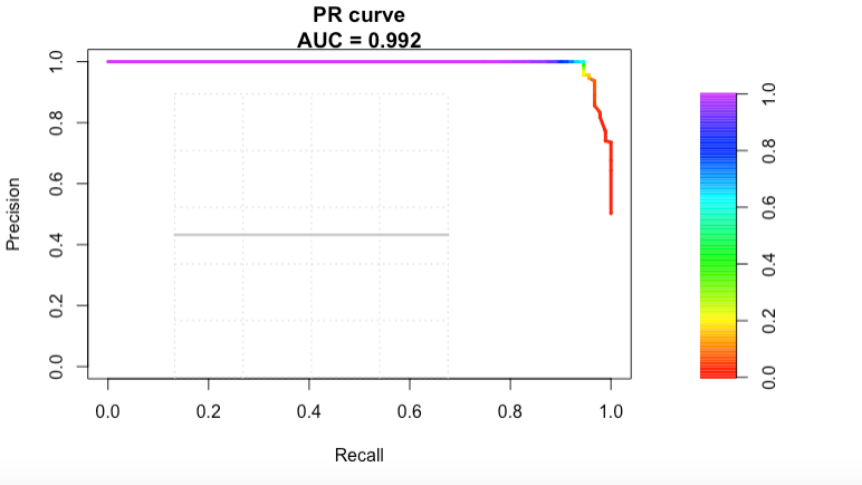


Figure 4.10: Precision/Recall Random Forest

4.2.4 Analysis of Support Vector Machine Non Linear Kernel

The implementation is so similar to linear or simple SVM. The difference is to select any kernel function like RBF(gaussian), polynomial, sigmoid and etc instead of a linear and 1-degree model. Kernel function takes it's inputs vectors in the original space and returns the dot product of the vectors in the feature space and this is called kernel functioning.

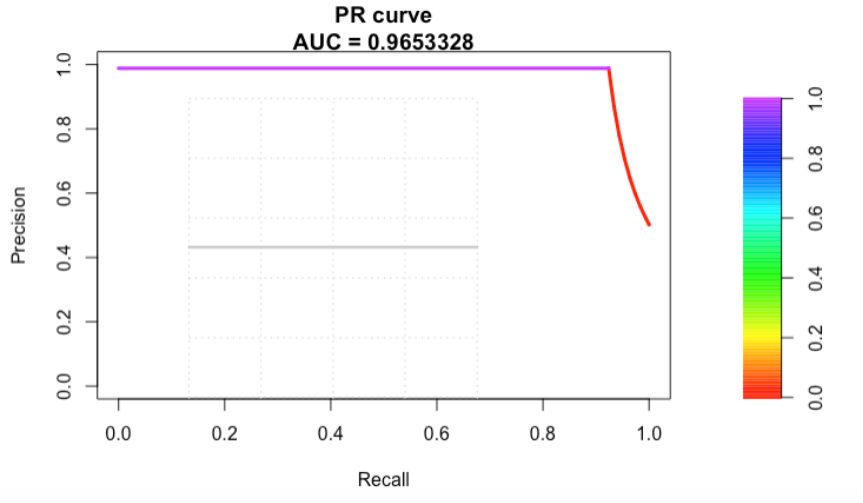


Figure 4.11: SVM Non Linear

4.2.5 Analysis Of Support Vector Machine Radial Non linear

A radial basis function (rbf) is equivalent to mapping the data into an infinite dimensional Hilbert space, and so we cannot illustrate the radial basis function concretely, as we do to a quadratic kernel. A radial basis function allows you to have features that pick out circles (hyperspheres) . The decision boundaries become much more complex as multiple such features interact. A string kernel lets you have features that are character subsequences of terms. All of these are straightforward notions which have also been used in many other places under different names.The PR of SVM non linear is 0.9763394

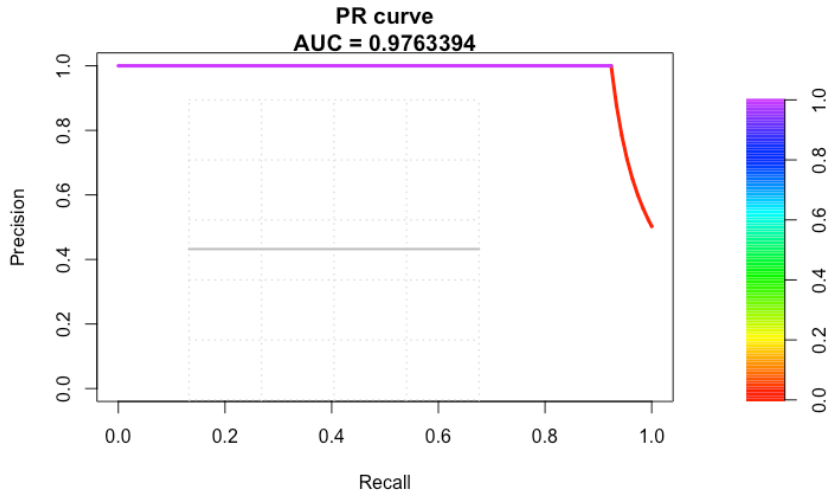


Figure 4.12: SVM Radial Non linear

4.2.6 Artificial Neural Network

An artificial neural network is an interconnected group of nodes, inspired by a simplification of neurons in a brain. Here, each circular node represents an artificial neuron and an arrow represents a connection from the output of one artificial neuron to the input of another. The Precision Recall of the ANN is 0.974

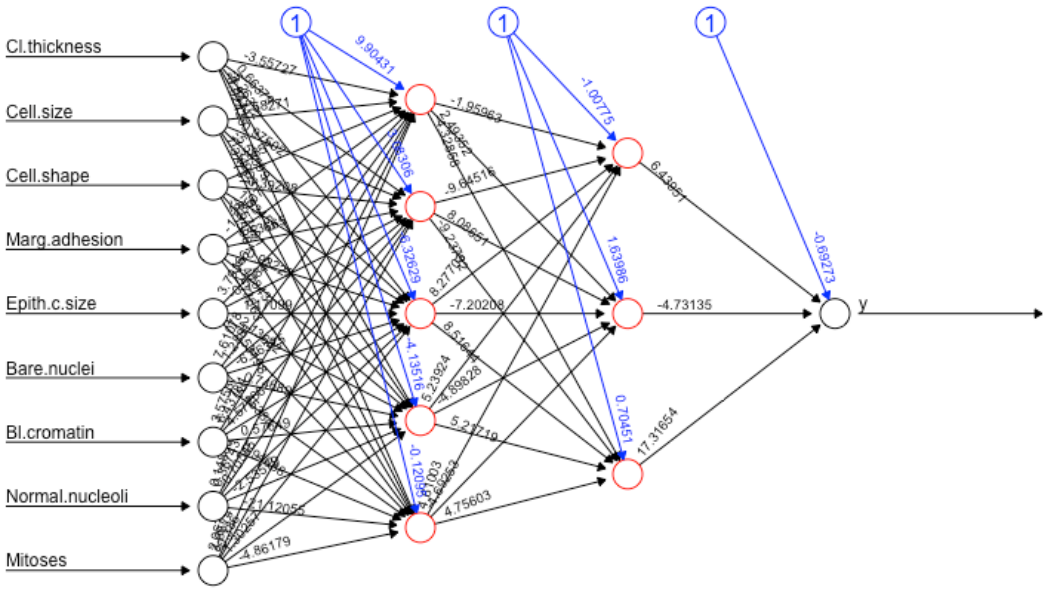


Figure 4.13: The visualized ANN model for the dataset

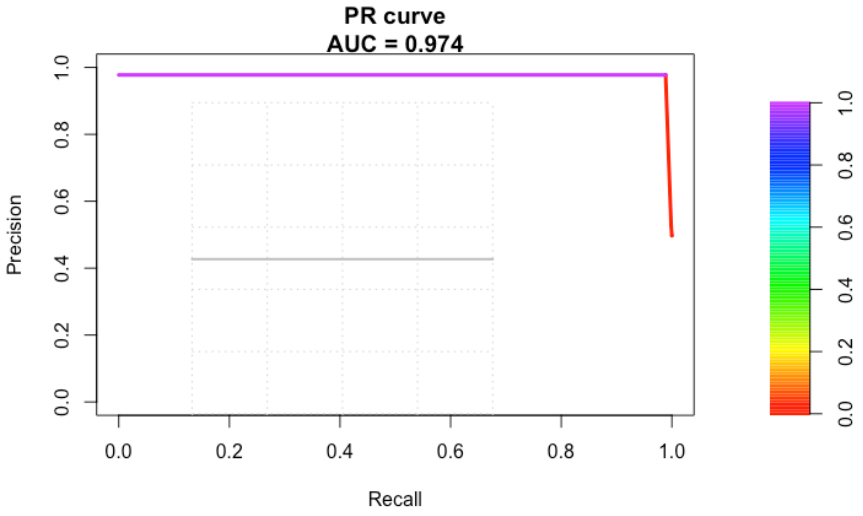


Figure 4.14: ANN

4.3 Summary of Analysis

Classification Technique	Precision Recall AUC Value
Logistic Regression	0.992336
Random Forest	0.9921113
Support Vector Machine (Linear)	0.9653328
Support Vector Machine (Radial Non Linear)	0.9763394
Support Vector Machine (Radial With Grids)	0.9276
Artificial Neural Network	0.974

We notice from the table above that the Logistic Regression gave the best PR AUC value

4.4 Classifications of Important Variables

The important variables of the Datasets was considered . It was noted that the important variables were Cell shape, Marg adhesion , BL cromatin, Normal nuclleoli.

Below are the Classification technique Results when only inportant variables are considered

4.5 Summary Of Important Variable Analysis

Classification Technique	Precision Recall AUC Value
Logistic Regression	0.99
Random Forest	0.995
Support Vector Machine (Linear)	0.935
Support Vector Machine (Radial Non Linear)	0.974
Support Vector Machine (Radial With Grids)	0.928
Artificial Neural Network	0.831

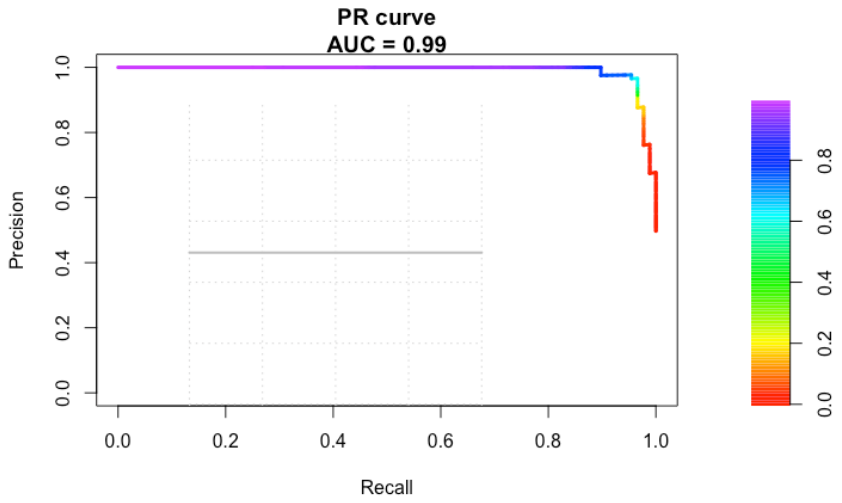


Figure 4.15: PR Imp.Var Logistic Regression

4.6 Summary Of Analysis

From the tables above we note that Logistic Regression had the best PR AUC . And also we could clearly see that SVM Radial with grids had the worse PR AUC value. But however when the important values for the datasets were considered it was not the case. Random Forest had the best PR AUC and ANN had the worse PR AUC when only the important variables were considered.

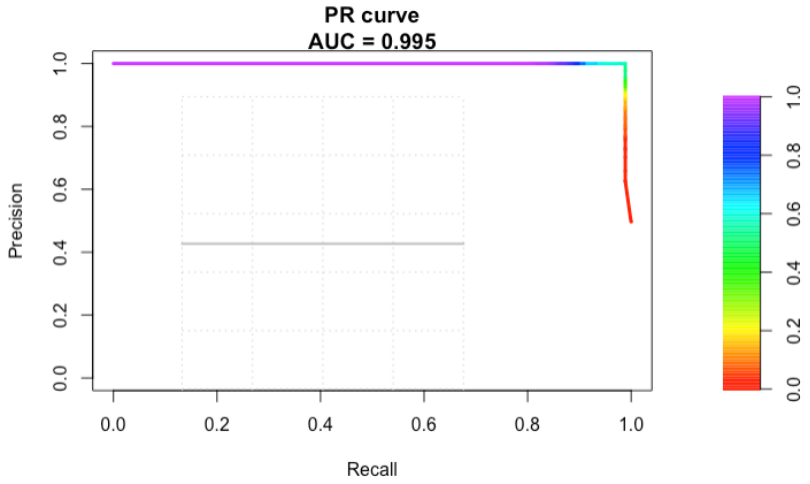


Figure 4.16: PR Imp.Var Random Forest

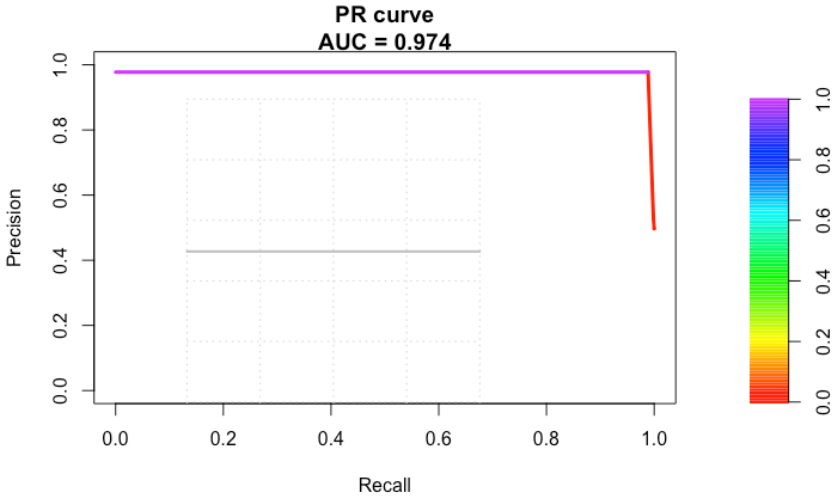


Figure 4.17: Imp.Var Non linear Kernel

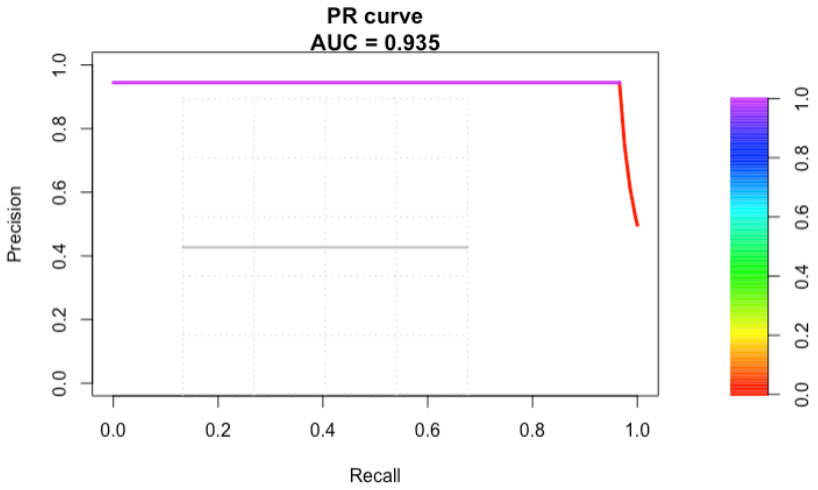


Figure 4.18: Imp.Var SVM Linear Kernel

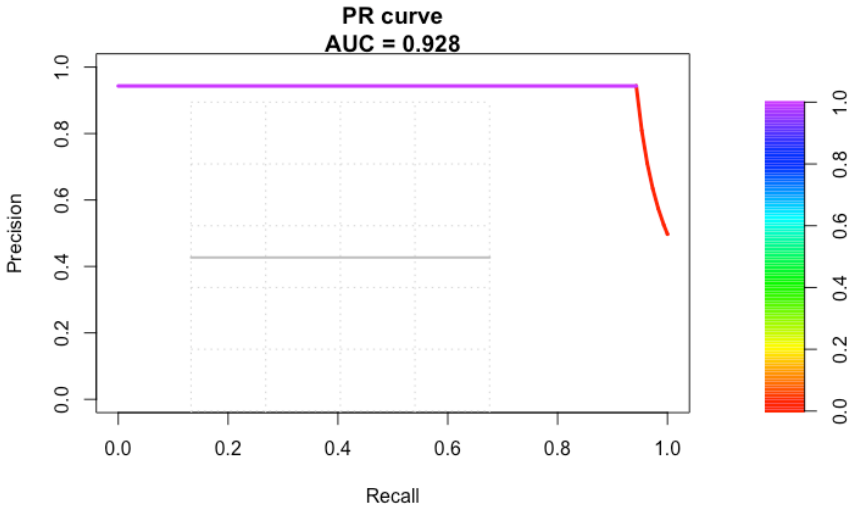


Figure 4.19: IMp.Var SVM Radial Grid

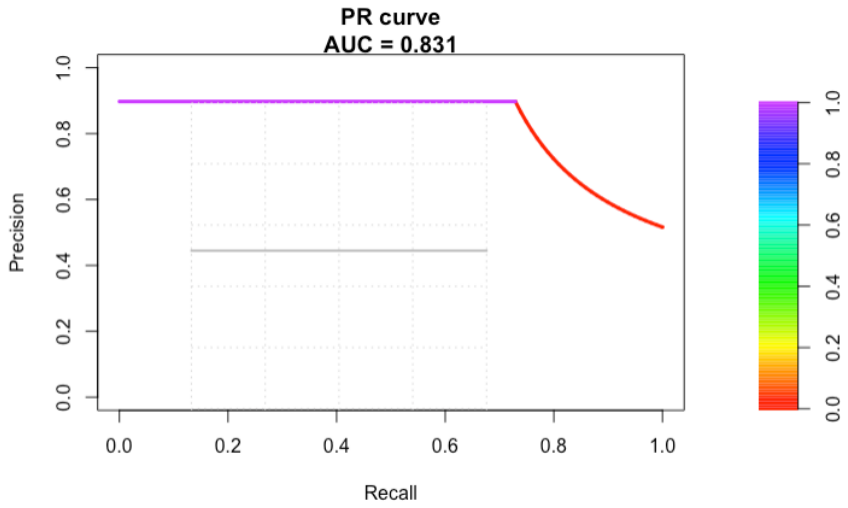


Figure 4.20: Imp.Var ANN

Chapter 5

Concluding Remarks

In this work we used six machine learning models namely; Logistic Regression, Random Forest, Support Vector Machine (Non Linear), Support Vector Machine (Linear), Support Vector Machine (Radial with grids) and Artificial Neural Network for the classification of the imbalanced Machine learning Dataset. Logistic Regrression and Random forest were noted to give the best AUC values with Logistic regression being the best of the two. In the case where one want to use only the important variables, it is highly recommended to do so only when you have a lots of attributes or variables. In the case of this study, when the important variables were considered it was noted that Random Forest had the best AUC value.

5.1 Future Work and Recommendations

- A bigger dataset with lots of attributes will be easier to work with.
- The Imortant Variable criteria can be used to make most predictions faster and easier.
- For an unbalanced dataset The Precision Recall is recommended.

References

[1] M. C. Mariani , O. K. Tweneboah,M. A M Bhuiyan, (2019) *Statistical Data Mining Algorithms for the prognosis of Diabetis and Autism*503, 304-321. Statistical Mechanics and its Applications, 503, 304-321.

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2675494/

[3] https://www.researchgate.net/publication/221345963$_{L}inear_{S}upport_{V}ector_{M}achines$

[4] https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/

Curriculum Vitae

James Arthur was born on April 1988. He graduated from University Of Mines and Technology, Ghana, in 2011. In the fall of 2018, he entered the Graduate School of The University of Texas at El Paso. While pursuing a master's degree in Mathematical Science,he worked as a Teaching Assistant.he was a member of African Students Organization.

Permanent Address: 910 N Oregon St Apt 2 El Paso, Texas 79902