BREAST CANCER DIAGNOSIS USING DIFFERENT MACHINE LEARNING TECHNIQUES

A Paper
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Souradip Roy

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Computer Science

May 2019

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Breast Cancer Diagnosis Using Different Machine Learning Techniques

By

Souradip Roy

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Simone Ludwig
Chair

Dr. Maria de los Angeles Alfonseca-Cubero

Dr. Jen Li

Approved:

May 20,2019
Date

Dr. Kenneth Magel
Department Chair

ABSTRACT

Cancer is one of the dangerous diseases which causes many deaths each year and breast cancer being one of them which is quite common among women. In today's time 12 percent of the women can develop breast cancer over her course of lifetime. There are two kinds of tumors that can be found in women, they are benign and malignant. The former is considered non-cancerous while the latter is deadly. In this work we applied different machine learning models and did a comparative study to see which one performs better in predicting unseen data to be benign or malignant. The dataset we have used is imbalanced, so we also experimented by improving the prediction of our models using oversampling technique on the minority class. We have calculated Accuracy, F1-scores, AUC and Confusion Matrix as our measures to evaluate and compare our models.

ACKNOWLEDGEMENTS

I will like to thank my research advisor Dr. Simone Ludwig and my parents and colleagues for trusting in me and who have supported me through my ups and downs in order to achieve my goal.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Breast cancer is one of the leading causes of death in women in many countries around the globe. It is a major requirement to reduce the threat of life from this disease at a very early stage. The detection of this disease at an early stage can increase the chance of a patient to live longer reducing the number of deaths per year. In these kinds of cases machine learning techniques can be useful which help in detection of a tumor to be benign (non-cancerous) or malignant (cancerous) without going into the long procedure of biopsy. If the identification of the tumor can happen at an early stage, then it is possible to start the treatment and prevent the disease from spreading further. There are various kind of machine learning task that can be helpful among which classification or predictive modeling is the most widely used one. The availability of large data sets in medical science and different kind of tools have made the machine learning tasks to be popular.

"We are living in the information age" is a popular saying, however, if the statement needs to be redefined it can be said that it is a data age. In the 21st century terabytes or petabytes of data are available online. This data is collective information of business, science and engineering, medicine and almost every other aspect of daily life. Few years back machine learning techniques were not of regular use because of limited availability of computational power but currently due to the high availability of the same machine learning is a regular practice. The data which is collected by different organizations helps these machine learning techniques to carry out the desired tasks. For example, when a person needs to gain some information, they generally put some search query in a search engine. This search information is stored as data and using a machine learning technique we can try to group people who query similar searches. There are two kinds of tasks that are carried out in machine learning which are clustering and classification. Clustering is the grouping of data having similar kind of hidden patterns. The classification task is categorizing

the data into different classes. The dataset contains two parts one is the features and the other is the class. Clustering tasks are performed on the data where the class is absent and classification on the data where it is present.

Machine Learning or Data Mining tasks are done in 3 stages which are data cleaning and data preprocessing, model training and model evaluation or testing. Data cleaning includes different kind of processes like excluding instances of missing values, correction of any corrupted data, removing of irrelevant data or columns from the data set. Data preprocessing technique is the process of converting the data to an understandable format. It is often common that real-world data has a wide range between the minimum and the maximum value of each feature. To reduce the range between the values there are lots of scaling techniques that can be used. Sometimes the feature set is very large, so to make the learning procedure easier and less time-consuming dimensionality reduction techniques are often used. The whole data cleaning and data preprocessing technique can be assumed as a single phase known as data handling in the machine learning process. In the preprocessing step there is another big part of data handling which is dividing the data into two parts namely training set and testing set. The training and testing data are generally in the ratio of 8:2.

The training data helps our model to learn about the pattern inside the data along with an instance falling into which class. The testing data tries to predict the class of each of the instances. Once the predicted class values are available we use a different model evaluation technique to see how well the model is performing on unseen data. The techniques we use other than accuracy to evaluate our model are confusion matrix, Area Under the curve, F1-score.

CHAPTER 2: RELATED WORK

There has a lot of research been performed on the diagnosis of breast cancer and most of the researches have high classification accuracies. A learning algorithm has reported an accuracy of 98.8% where a combination of logarithmic simulated annealing and perceptron algorithm was used [1]. There is another research done that used the fuzzy-GA method and it has obtained an accuracy of 97.36% [2]. It was mentioned that an accuracy of 98.10% has been obtained when a feed forward neural network rule extraction algorithm has been used [3]. A 10-fold cross validation technique on the C4.5 decision tree for classifying breast cancer has been used where it was able to gain an accuracy of 94.74% [4]. There is another accuracy of 96.8% reported where the research group used linear discrete analysis [5].

There is a study presented in [6] where an intelligent system had been developed using Support Vector Machine and Artificial Neural Network to automate breast cancer detection. Recently with the increase in data and the development of ML methods, it is seen that the methods have high classification reliability. An accuracy of 95.06% has been obtained using neuro-fuzzy technique [7]. There have been three different methods, learning vector quantization (LVQ), big LVQ, and artificial immune recognition system were used to detect the kind of cancer it is and reported accuracies were 96.7%, 96.8% and 97.2%, respectively [8].

CHAPTER 3: DATA SET

The dataset that is used in our experiments has been obtained from the University of California Irvine and is a publicly available dataset [9]. The dataset has a total of 11 columns which are 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. The class column defines the instance to be either benign or malignant. The id number does not contribute anything in the learning process for the training of our models so we rejected it. All the other columns are used in training our different models. The class is the dependent variable and all the other columns are the independent variables. All the independent variables have values ranging from 1 to 10 where 1 being the lowest and 10 being the highest. This dataset has been collected over 3 years starting from 1989 to 1991. The total number of samples are 699 in the dataset. The following table shows the formation of the whole dataset over a period of 3 years

Table 1: Dataset

Group	Instances	Time
1	367	January 1989
2	70	October 1989
3	31	February 1990
4	17	April 1990
5	48	August 1990
6	49	January 1991
7	31	June 1991
8	86	November 1991

It is seen at different times different number of instances have been added to the dataset and lastly in 1991 it was completed with a total number of 699 instances belonging to both benign

and malignant classes. The dataset has 458 samples which belong to the Benign class and 241 instances which are the Malignant ones. The dataset has 16 instances with missing values, so we rejected those 16 instances. The benign class is 65.5% of the dataset and the malignant is 34.5%. This is an imbalanced dataset where most of the instances fall into the Benign class. The imbalanced dataset problem can cause the models to predict the data as the major class.

Statistical Analysis [10] shows that there is a huge difference between the values of each of the features based on the instance belonging to benign or malignant. The dataset has been created periodically since this dataset belongs to the clinical cases of Dr. William H. Wolberg. Table 1 shows the number of instances added since January 1989 after the dataset started being built. There were two points which have been discarded from the first group as it was noticed that they are inconsistent. The data can be considered ‘noise-free’ [11] and has 16 missing values from 16 instances.

CHAPTER 4: APPROACH

This section will describe different kinds of experiment that were performed on this dataset. We did a comparative study using different binary classifiers to see which one performs better in detecting the class on unseen data. We have carried out the experiments on two datasets, one of them is the original or imbalanced dataset and the other is the oversampled which means after applying the SMOTE algorithm. We have used the following different measures, to determine the performance of our model, which are Accuracy, F1-Score, Area Under the Curve, Confusion Matrix.

SMOTE is an over-sampling technique where the minority class is over-sampled creating "synthetic" examples. In handwritten character recognition this technique proved quite useful [12]. The research group tried to create extra training data on real data using some operations. The synthetic examples are generated by operating in the feature space instead of the data space. The oversampling technique is done by taking each of the minority class samples and introducing synthetic examples along the line joining any of k minority class nearest neighbors. The k nearest neighbors are randomly chosen which depend upon the amount of over-sampling required. Our implementation uses 5 nearest neighbors in generating the synthetic data using the mentioned technique. The process of generating the synthetic data is the following: Difference between the sample which is currently considered and its nearest neighbor. The difference is then multiplied with a randomly chosen value between 0 and 1 and the obtained value to the feature vector is added which is being considered. The application of the SMOTE algorithm on the minority class forces the decision region of the minority class to be more general.

Our experiments include the use of different binary classifiers which are Naïve Bayes, K-Nearest Neighbors, Decision Tree, Dense Feed Forward Neural Network, Logistic Regression and Support Vector Machine.

The Naïve Bayes classifier is based on Bayes' theorem. Bayes' theorem calculates the posterior probability using the attribute values. It assumes that the values of a predictor or attribute on a given class is independent of the values of other predictors. This is called class conditional independence. The mathematical representation of the algorithm is:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

i) P(c|x) is the posterior probability of class given attribute.
ii) P(c) is the prior probability of class.
iii) P(x|c) is the likelihood which is the probability of predictor given class.
iv) P(x) is the prior probability of the predictor.

This experiment has been performed on both the original dataset and the oversampled one. The dataset has 9 attributes as mentioned in the Dataset section. These 9 attributes are the predictors in the dataset. The classifier assumes the presence of a feature in the dataset is independent or unrelated to any other feature. This model is very helpful for prediction on large datasets as well as Naïve Bayes is well known for outperforming other classifiers. There are various advantages and disadvantages of this algorithm. The algorithm is very easy and fast when it predicts on test data. The Naïve Bayes algorithm can be much better compared to other models like logistic regression. The Naïve Bayes algorithm can perform on less training data as well. The disadvantages are if there is any attribute which is not being observed in the training set then the algorithm will assign zero as its probability, however, the availability of a totally independent predictor in real life is nearly impossible.

The K-nearest neighbors classification algorithm is known as a lazy learning algorithm and is one of the simplest classification algorithms. It is also a non-parametric algorithm which means that the algorithm does not make any assumptions on the data distributions. KNN is useful for real world problems since most of the data does not maintain the theoretical assumptions made by models like linear regression. KNN can be used for both kinds of problems classification and

regression, but it is mostly used for classification tasks. The "K" in KNN refers to the number of nearest neighbors we want to consider during classification. The algorithm uses a voting procedure to predict sample data belonging to a class. Data is classified based on the plurality votes of its neighbors. KNN will predict data belonging to a class based on the maximum number of votes going to a class by the neighbors of the data point. This algorithm does not learn any model, instead it stores the whole training data which it uses as the representation. It makes prediction by calculating the similarity between an input sample and each training instances. There are few disadvantages of KNN. It needs to determine the value of K. Since it is distance-based learning it is very much unclear which type of distance needs to be used. It is all very unclear to understand which attribute to use to get the best result or all the attributes need to be used. The computation of this algorithm is quite high since we need to compute the distance of a data point to each of the training instances. For our experiment we have used the KNN algorithm to predict a breast cancer tumor to be benign or malignant. The algorithm has been performed on the original dataset as well as on the oversampled dataset.

The Logistic Regression model is our third model which performs the classification whether a tumor is benign or malignant. The logistic function is a Sigmoid Function that takes a value between 0 and 1. The expression of sigmoid function is:

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

The sigmoid function takes an "S" shape curve which can take any real-valued number and it can be mapped between 0 and 1. Logistic regression is very much like linear regression since the former uses an equation as its representation. There are three types of Logistic Regression which are Binary Logistic Regression, Multinomial Logistic Regression, Ordinal Logistic Regression.

Logistic Regression predicts the class for sample data based on a threshold. The estimated probability is classified into a class using the set threshold. The coefficients of the logistic regression is estimated from the training data which is done using the maximum-likelihood estimation. When a model predicts a value close to 1 for the default class, and very close to 0 for the other class. Let us consider a scenario where a tumor needs to be classified into benign or malignant. If linear regression is used there will be a threshold value based on which a tumor will be detected as one or the other. If the actual class is malignant and the predicted value is 0.4 whereas the threshold value is 0.5 then the data point will be classified as benign which is a serious concern during real time prediction. Logistic Regression overcomes this problem. The output of logistic regression strictly takes values either 0 or 1 as described above.

Decision tree is one of various other models we have used for our experiments. It can be used to solve both classification and regression problems. The main working procedure of the decision tree is to predict classes by learning decision rules inferred from the training data. There are two kinds of attribute selection measures we have used to create two kinds of decision tree for both original dataset and oversampled dataset which are information gain and Gini index. The root node is the attribute which has the highest values in both cases. The next step is to choose to split the training data into subsets in such a way that each subset contains data with the same value for an attribute. There are few assumptions that are made while creating the tree. The whole training set is considered as the root at the first step. Decision trees prefers categorical data while building the model. It is possible that continuous data attributes are present in the dataset so conversion of those data to discrete or categorical values are important prior to building the tree. The information gain of each attribute is obtained by calculating the difference between the entropy of the target values and the entropy of the attribute. The entropy is given by the formulation:

$$H(X) = -\Sigma p(x) log p(x)$$

The Gini index is a metric to measure how often a random chosen element would be incorrectly identified. The above statement means that an attribute with lower Gini index should be referred. The formulation of Gini index is:

$$Gini\ Index = 1 - \Sigma\ p_j{}^2$$

We used these two-attribute selection techniques to build two different decision trees to see which method performs better in classifying the dataset.

The other model we have used in classifying the dataset into benign or malignant is Dense Feed Forward Neural Network. Deep Learning provides a multi-layer approach to learn data representation. The architecture of a DNN will have the input layer followed by single or multiple hidden layer and output layer. Each layer comprises of multiple neurons, the number of neurons in the input layer will be the number of features present in the dataset. The number of neurons in the hidden layer and the number of hidden layers varies based on the output we are seeking. Learning rate is another parameter which needs to be set from the beginning of the training. The learning rate determines how fast a model will learn from the data set and besides the hidden layers and number of neurons the learning rate is another factor which helps the neural network to understand the data representation. The other factors that are associated while training a neural network are biases, weights and activation function. The weights are associated with each input and the weights provide the strength of connection between the neurons and the layers. The weights determine how influential a connection between neurons is over another. The biases add noise to the dataset helps during the non-linear learning of the neural network. The neural network has been trained on both the original dataset and the oversampled dataset.

The last model that has been used in the experiment is Support Vector Machine (SVM). It can be defined as given a labeled training data SVM outputs an optimal hyperplane which

categorizes new examples. Let us assume we have data points plotted in a two-dimensional plane. Now if we use the algorithm on these data points it will try to provide a function that will be able to separate these points based on the label provided to it. SVM does the same for n-dimensional data points. To separate data points there are many hyperplanes which are possible but the one with the maximum margin will be chosen. The maximum margin means the distance between the data points in both the classes should be maximum. The maximum margin helps to provide more confidence while classifying future data points. The number of hyperplanes depends on the number of features the data contains. If the number of features be n then the number of hyperplanes that will be generated is n.

The last experiment that has been executed is comparing all the classifiers over the balanced dataset. In the earlier experiments it has been noticed that the classifiers perform better on the oversampled dataset. This experiment has been performed to identify which classifier works best among all in classifying the benign or malignant cases. In this experiment we have passed the same test and train data to all the classifiers and calculated the metrics to detect which classifiers performs the best.

The performances measures are described next. Accuracy is one among many metrices to evaluate how well a model is performing. The performance can be evaluated based on the number of correct predictions the model performs. Informally we can say that accuracy is the fraction of the number of correct predictions among the total number of predictions. We can divide the total correct predictions in two parts for a binary classification task, one is the positive class and other is the negative class.

The Confusion Matrix is another matrix which helps to evaluate our models. Confusion matrix is a table layout that allows us to visualize the performance of our model. It is typically

used in the process of supervised machine learning tasks. The rows of the table indicate the number of instances that fall in the predicted class while the columns indicate the instance that belongs to the actual class or vice-versa. The reason to use a confusion matrix to evaluate models is because it becomes easy to understand how much misclassification our model is performing. The term misclassification refers to the process where the actual class is positive, but our model predicted it to be negative or conversely.

There is another model evaluation method which is known as F1-Score. In binary classification the F1-Score is a measure of the test's accuracy. There are two values named precision and recall which are involved in the calculation of F1-Score. Let us assume precision is denoted by p and recall by r then we call p the number of correct positive results divided by the number of all positive results the classifier predicted, and r is correct number of positive results divided by the number of all relevant samples. F1-Score is the harmonic average of the precision and recall where F1-Score tends to reach value 1.

CHAPTER 5: RESULTS

This section will describe the different results we achieved while using different models for classifying the dataset into benign and malignant.

Experiment 1

The first experiment is to apply the Naïve Bayes algorithm both on original and oversampled dataset, and Figure 1 and Figure 2 show the ROC curve for those datasets, respectively.

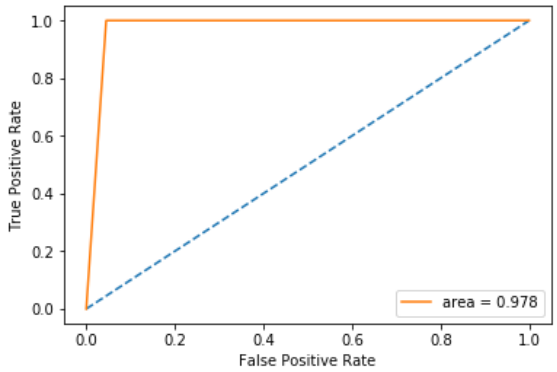

Figure 1: ROC Curve for Original Dataset using Naïve Bayes

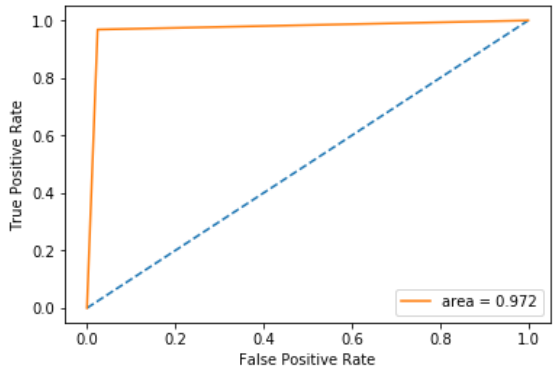

Figure 2: ROC Curve for Oversampled Dataset using Naïve Bayes

The area that the ROC curve covers is being noted on the lower right corner of both figures, which are 97.8 % and 97.2 %, respectively. The AUC is the area between the ROC curve and the x-axis.

Table 2: Measure Values for Naïve Bayes

Measure Names	**Original Dataset**	**Oversampled Dataset**
Training Accuracy	0.950	0.950
Testing Accuracy	0.970	0.970
AUC Score	0.978	0.972

Table 2 lists the accuracy values for both datasets and as well as the AUC score. The testing accuracies are 0.97 for both the original and oversampled dataset. It is the testing accuracy which determines the performance of the model.

Table 3: Confusion Matrix for Original Dataset

	Predicted Class **No**	Predicted Class **Yes**
Actual Class **No**	85	4
Actual Class **Yes**	0	48

Table 4: Confusion Matrix for Oversampled Dataset

	Predicted Class **No**	Predicted Class **Yes**
Actual Class **No**	81	2
Actual Class **Yes**	3	92

Table 3 and Table 4 show the confusion matrices for the original and oversampled dataset, respectively. It is seen from both the tables that the misclassification error is 4 and 5 for

the original and the oversampled dataset, respectively. The number of correct predictions are 133 for both the classes in the original dataset, and 173 for the oversampled dataset.

Table 5: F1-Score for Original Dataset

Label	Precision	Recall	F1-Score	Support
2	1.00	0.96	0.98	89
4	0.92	1.00	0.96	48
Micro average	0.97	0.97	0.97	137
Macro Average	0.96	0.98	0.97	137
Weighted Average	0.97	0.97	0.97	137

Table 6: F1-Score for Oversampled Dataset

Label	Precision	Recall	F1-Score	Support
2	0.96	0.98	0.97	83
4	0.98	0.97	0.97	95
Micro average	0.97	0.97	0.97	178
Macro Average	0.97	0.97	0.97	178
Weighted Average	0.97	0.97	0.97	178

Table 5 and Table 6 show the Micro, Macro and Weighted Average along with F1-Score for the original and oversampled data, respectively. The F1-Score in both tables for both classes are 0.98, 0.96 and 0.97. This means that the values for the false positives and the false negatives are very low, which indicates that the model is correctly classifying the correct class.

Experiment 2

The second model for our experiment is the K-nearest neighbors. Figure 3 and 4 are showing the results of our different measures which determines how well the KNN model is performing on our datasets.

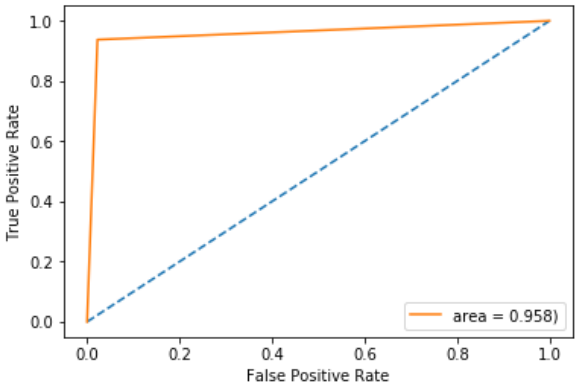

Figure 3: ROC Curve for Original Dataset using KNN

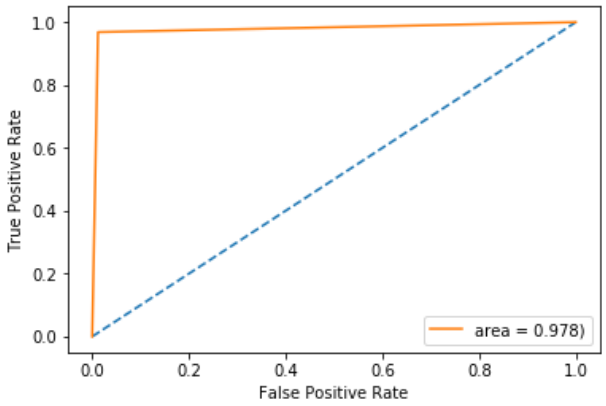

Figure 4: ROC Curve for Oversampled Dataset using KNN

The area that the ROC curve covers is being noted on the lower right corner of both figures, which are 95.8 % and 97.8 %, respectively.

Table 7: Measure Values for KNN

Measure Names	**Original Dataset**	**Oversampled Dataset**
Training Accuracy	0.963	0.977
Testing Accuracy	0.937	0.963
AUC Score	0.957	0.978

Table 7 lists the accuracy values for both datasets and as well as the AUC score. The testing accuracies are 0.937 and 0.963 for both the original and the oversampled dataset, respectively. It is the testing accuracy which determines the performance of the model.

Table 8: Confusion Matrix for Original Dataset

	Predicted Class **No**	Predicted Class **Yes**
Actual Class **No**	87	2
Actual Class **Yes**	3	45

Table 9: Confusion Matrix for Oversampled Dataset

	Predicted Class **No**	Predicted Class **Yes**
Actual Class **No**	82	1
Actual Class **Yes**	3	92

Table 3 and Table 4 show the confusion matrices for the original and oversampled dataset, respectively. It is seen from both tables that the misclassification error is 5 and 4 for the original and the oversampled dataset, respectively. The number of correct predictions is 133 for both the classes in the original dataset, and 173 for the oversampled dataset.

Table 10: F1-Score for Original Dataset

Label	Precision	Recall	F1-Score	Support
2	0.97	0.98	0.97	89
4	0.96	0.94	0.95	48
Micro average	0.96	0.96	0.96	137
Macro Average	0.96	0.96	0.96	137
Weighted Average	0.96	0.96	0.96	137

Table 11: F1-Score for Oversampled Dataset

Label	Precision	Recall	F1-Score	Support
2	0.96	0.99	0.98	83
4	0.99	0.97	0.98	95
Micro average	0.98	0.98	0.98	178
Macro Average	0.98	0.98	0.98	178
Weighted Average	0.98	0.98	0.98	178

Table 10 and Table 11 show the Micro, Macro and Weighted Average along with F1-Score for the original and oversampled data, respectively. The F1-Score on both tables for both classes are 0.97, 0.95 and 0.98, respectively. This means that the values for the false positives and the false negatives are very low, which indicates that the model is correctly classifying the correct class.

Experiment 3

The third model for our experiment is the Logistic Regression. Figures 5 and are showing the ROC results which determines how well the model is performing on our datasets.

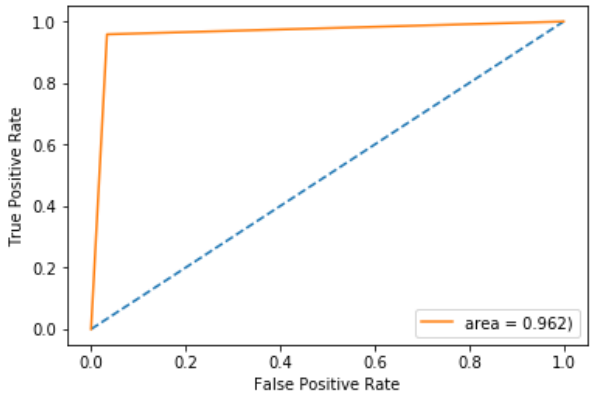

Figure 5: ROC Curve for Original Dataset using Logistic Regression

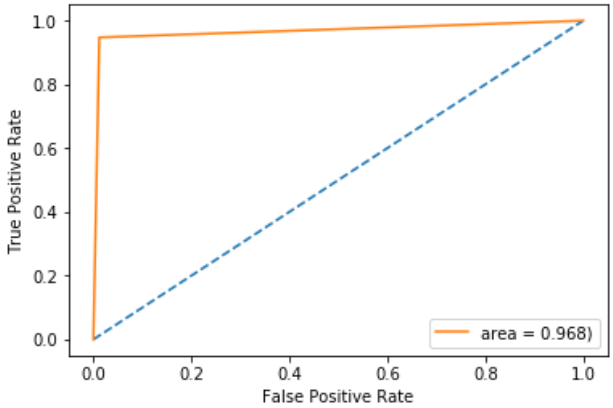

Figure 6: ROC Curve for Oversampled Dataset using Logistic Regression

The area that the ROC curve covers of both figures are 96.2 % and 96.8 %, respectively.

Table 12: Measure Values for Logistic Regression

Measure Names	**Original Dataset**	**Oversampled Dataset**
Training Accuracy	0.968	0.966
Testing Accuracy	0.96	0.966
AUC Score	0.962	0.967

Table 12 lists the accuracy values for both datasets and as well as the AUC score. The testing accuracies are 0.937 and 0.963 for both the original and oversampled dataset, respectively.

Table 13: Confusion Matrix for Original Dataset

	Predicted Class **No**	Predicted Class **Yes**
Actual Class **No**	86	3
Actual Class **Yes**	2	46

Table 14: Confusion Matrix for Oversampled Dataset

	Predicted Class **No**	Predicted Class **Yes**
Actual Class **No**	81	1
Actual Class **Yes**	5	90

Table 13 and Table 14 show the confusion matrices for the original and oversampled dataset, respectively. It is seen from both tables that the misclassification error is 5 and 6 for original and oversampled dataset, respectively. The number of correct predictions is 133 for both the classes in the original dataset, and 173 for the oversampled dataset.s

Table 15: F1-Score for Original Dataset

Label	Precision	Recall	F1-Score	Support
2	0.98	0.97	0.97	83
4	0.94	0.96	0.95	48
Micro average	0.96	0.96	0.96	137
Macro Average	0.96	0.96	0.96	137
Weighted Average	0.96	0.96	0.96	137

Table 16: F1-Score for Oversampled Dataset

Label	Precision	Recall	F1-Score	Support
2	0.96	0.99	0.96	83
4	0.99	0.95	0.97	95
Micro average	0.97	0.97	0.97	178
Macro Average	0.97	0.97	0.97	178
Weighted Average	0.97	0.97	0.97	178

Table 15 and Table 16 show the Micro, Macro and Weighted Average along with F1-Score for the original and oversampled data, respectively. The F1-Score for both tables for both classes are 0.97, 0.95 and 0.96, 0.97, respectively. This means that the values for the false positives the and false negatives are very low, which indicates that the model is correctly classifying the correct class.

Experiment 4

Our next model is the Decision Tree which has been calculated on two kinds of information gain, which are gini index and entropy. Figure 7 to 10 show the ROC results of our different measures, which determines how well the model is performing on our datasets.

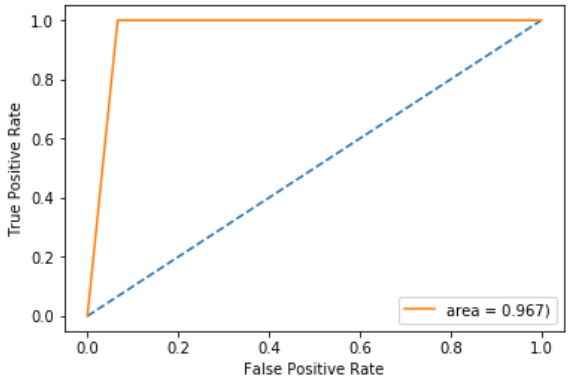

Figure 7: ROC Curve for Original Dataset using Decision Tree Gini Index

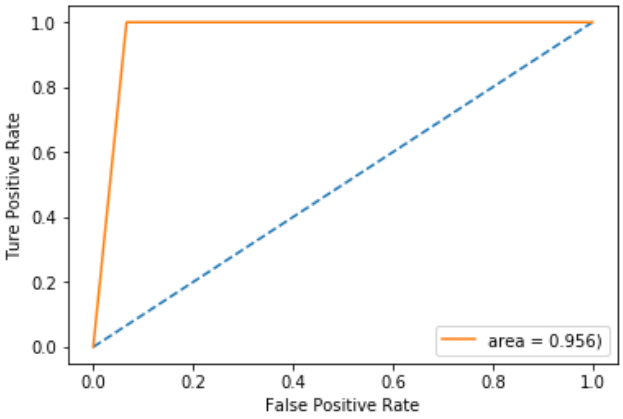

Figure 8: ROC Curve for Original Dataset using Decision Tree Entropy

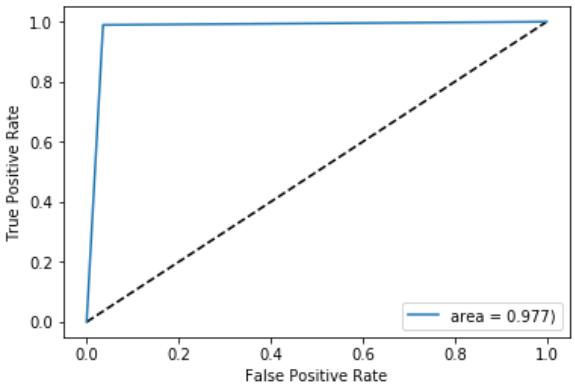

Figure 9: ROC Curve for Oversampled Dataset using Decision Tree Gini Index

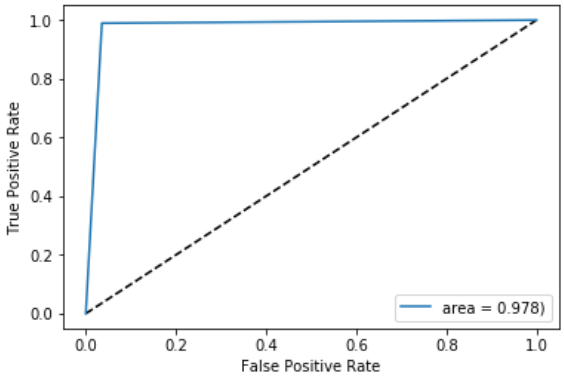

Figure 10: ROC Curve for Oversampled Dataset using Decision Tree Entropy

The area that the ROC curve covers of both figures and is 96.7 %, 95.6 %, 97.7 % and 97.8 %, respectively. In this experiment we obtained four ROC curves since we have used two types of information gains for both original and oversampled dataset.

Table 17: Measure Values for Decision Tree Gini Index

Measure Names	**Original Dataset**	**Oversampled Dataset**
Training Accuracy	0.93	0.939
Testing Accuracy	0.956	0.960
AUC Score	0.967	0.960

Table 17 lists the accuracy values for both datasets and as well as the AUC score. The testing accuracies are 0.956 and 0.960 for the original and oversampled dataset, respectively. It is the testing accuracy which determines the performance of the model.

Table 18: Measure Values for Decision Tree Entropy

Measure Names	**Original Dataset**	**Oversampled Dataset**
Training Accuracy	0.970	0.934
Testing Accuracy	0.948	0.971
AUC Score	0.950	0.972

Table 18 lists the accuracy values for both datasets and as well as the AUC score. The testing accuracies are 0.948 and 0.971 for the original and oversampled dataset, respectively. It is the testing accuracy which determines the performance of the model.

Table 19: Confusion Matrix for Oversampled Dataset using Gini Index

	Predicted Class **No**	Predicted Class **Yes**
Actual Class **No**	92	3
Actual Class **Yes**	4	79

Table 20: Confusion Matrix for Oversampled Dataset Using Entropy

	Predicted Class **No**	Predicted Class **Yes**
Actual Class **No**	91	4
Actual Class **Yes**	1	82

Table 19 and Table 20 show the confusion matrices for the oversampled dataset, for the gini index method and the entropy method, respectively. It is seen from both tables that the misclassification error is 7 and 5 for original and oversampled dataset, respectively. The number of correct predictions is 173 for both classes for both information gain methods.

Table 21: Confusion Matrix for Original Dataset using Entropy

	Predicted Class **No**	Predicted Class **Yes**
Actual Class **No**	85	5
Actual Class **Yes**	2	45

Table 22: Confusion Matrix for Original Dataset using Gini Index

	Predicted Class **No**	Predicted Class **Yes**
Actual Class **No**	85	5
Actual Class **Yes**	2	45

Table 21 and Table 22 show the confusion matrices for the original dataset, for the gini index method and the entropy method, respectively. It is seen from both tables that the misclassification error is 7 in both cases. The number of correct predictions is 130 for both classes of information gain methods.

Table 23: F1-Score for Oversampled Dataset using Gini Index

Label	Precision	Recall	F1-Score	Support
2	0.96	0.97	0.96	83
4	0.99	0.95	0.96	95
Micro average	0.96	0.96	0.96	178
Macro Average	0.96	0.96	0.96	178
Weighted Average	0.96	0.96	0.96	178

Table 24: F1-Score for Oversampled Dataset using Entropy

Label	Precision	Recall	F1-Score	Support
2	0.99	0.96	0.97	83
4	0.95	0.99	0.97	95
Micro average	0.97	0.97	0.97	178
Macro Average	0.97	0.97	0.97	178
Weighted Average	0.97	0.97	0.97	178

Table 23 and Table 24 show the Micro, Macro and Weighted Average along with F1-Score for the original and oversampled data, respectively. The F1-Score for both tables and for both classes are 0.96 and 0.97, respectively. This means that the values for the false positives and the false negatives are very low, which indicates that the model is correctly classifying the correct class.

Table 25: F1-Score for Original Dataset using Gini Index

Label	Precision	Recall	F1-Score	Support
2	0.98	0.94	0.96	90
4	0.90	0.96	0.93	47
Micro average	0.94	0.95	0.95	137
Macro Average	0.95	0.95	0.94	137
Weighted Average	0.94	0.95	0.95	137

Table 26: F1-Score for Original Dataset using Entropy

Label	Precision	Recall	F1-Score	Support
2	0.98	0.94	0.96	90
4	0.90	0.96	0.93	47
Micro average	0.94	0.95	0.95	137
Macro Average	0.95	0.95	0.94	137
Weighted Average	0.94	0.95	0.95	137

Table 25 and Table 26 show the Micro, Macro and Weighted Average along with F1-Score for the original and oversampled data, respectively. The F1-Score for both tables for both classes are 0.96, 0.93 and 0.96, 0.93, respectively. This means that the values for the false positives and the false negatives are very low, which indicates that the model is correctly classifying the correct class.

Experiment 5

The next model is the Dense Feed Forward Neural Network. Figure 11 and 12 show the results of the ROC measure which determines how well the Decision Tree model is performing on our datasets.

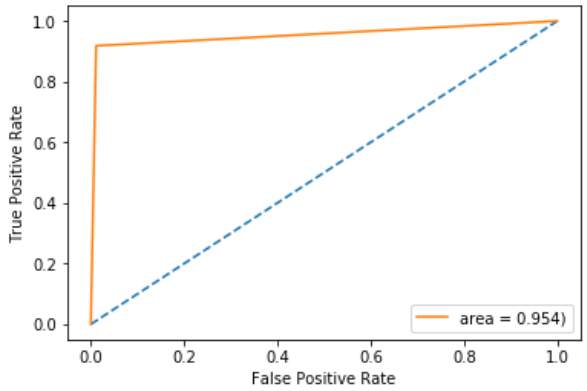

Figure 11: ROC Curve for Original Dataset using Dense Feed Forward Neural Network

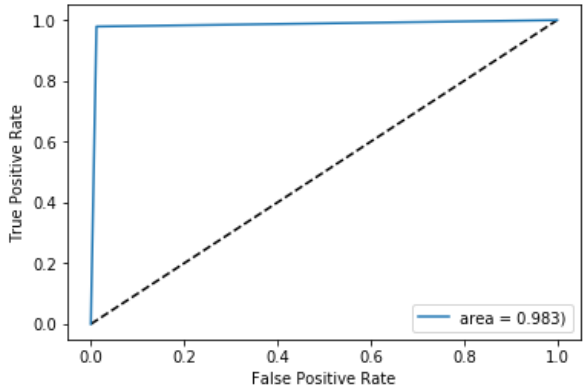

Figure 12: ROC Curve for Oversampled Dataset using Dense Feed Forward Neural Network

The area that the ROC curve for both figures are 95.4 % and 98.3 %, respectively.

Table 27: Measure Values for Dense Feed Forward Neural Network

Measure Names	**Original Dataset**	**Oversampled Dataset**
Training Accuracy	0.963	0.971
Testing Accuracy	0.96	0.983
AUC Score	0.953	0.983

Table 27 lists the accuracy values for both datasets and as well as the AUC score. The testing accuracies are 0.96 and 0.983 for both the original and oversampled dataset, respectively. It is the testing accuracy which determines the performance of the model.

Table 28: Confusion Matrix for Original Dataset

	Predicted Class **No**	Predicted Class **Yes**
Actual Class **No**	87	1
Actual Class **Yes**	4	45

Table 29: Confusion Matrix for Oversampled Dataset

	Predicted Class **No**	Predicted Class **Yes**
Actual Class **No**	82	1
Actual Class **Yes**	2	93

Table 28 and Table 29 show the confusion matrices for the original and oversampled dataset, respectively. It is seen from both tables that the misclassification error is 5 and 3 for original and oversampled dataset, respectively. The number of correct predictions is 133 for both classes of the original dataset and 175 for the oversampled dataset.

Table 30: F1-Score for Original Dataset

Label	Precision	Recall	F1-Score	Support
2	0.96	0.99	0.97	88
4	0.98	0.92	0.95	49
Micro average	0.96	0.96	0.96	137
Macro Average	0.97	0.95	0.96	137
Weighted Average	0.96	0.96	0.96	137

Table 31: F1-Score for Oversampled Dataset

Label	Precision	Recall	F1-Score	Support
2	0.98	0.99	0.98	83
4	0.99	0.98	0.98	95
Micro average	0.98	0.98	0.98	178
Macro Average	0.98	0.98	0.98	178
Weighted Average	0.98	0.98	0.98	178

Table 30 and Table 31 show the Micro, Macro and Weighted Average along with F1-Score for the original and oversampled data, respectively. The F1-Score on both tables for both classes are 0.97, 0.95 and 0.98, respectively. This means that the values for the false positives and the false negatives are very low, which indicates that the model is correctly classifying the correct class.

Experiment 6

The sixth model for our experiments is the SVM. Figure 13 and 14 show the results of the ROC measure which determines how well the SVM model is performing on our datasets.

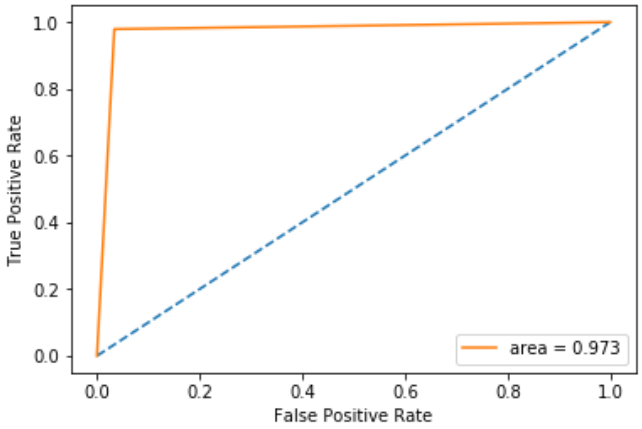

Figure 13: ROC Curve for Original Dataset using SVM

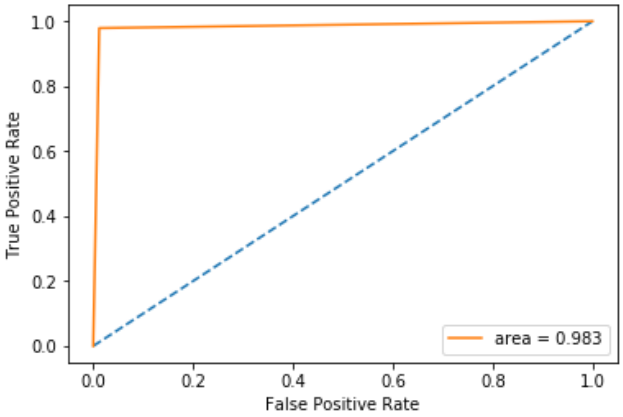

Figure 14: ROC Curve for Oversampled Dataset using SVM

The area that the ROC curve is 97.3 % and 98.3 % in Figure 13 and 14, respectively.

Table 32: Measure Values for SVM

Measure Names	**Original Dataset**	**Oversampled Dataset**
Training Accuracy	0.970	0.964
Testing Accuracy	0.970	0.983
AUC Score	0.972	0.983

Table 32 lists the accuracy values for both datasets and as well as the AUC score. The testing accuracies are 0.970 and 0.983 for the original and oversampled dataset, respectively. It is the testing accuracy which determines the performance of the model.

Table 33: Confusion Matrix for Original Dataset

	Predicted Class **No**	Predicted Class **Yes**
Actual Class **No**	86	3
Actual Class **Yes**	1	47

Table 34: Confusion Matrix for Oversampled Dataset

	Predicted Class **No**	Predicted Class **Yes**
Actual Class **No**	87	1
Actual Class **Yes**	2	93

Table 33 and Table 34 show the confusion matrices for the original and oversampled dataset, respectively. It is seen from both tables that the misclassification error is 4 and 3 for the original and the oversampled dataset, respectively. The number of correct predictions is 133 for both classes for the original dataset, and 180 for the oversampled dataset.

Table 35: F1-Score for Original Dataset

Label	Precision	Recall	F1-Score	Support
2	0.99	0.97	0.98	89
4	0.94	0.98	0.96	48
Micro average	0.97	0.97	0.97	137
Macro Average	0.96	0.97	0.97	137
Weighted Average	0.97	0.97	0.97	137

Table 36: F1-Score for Oversampled Dataset

Label	Precision	Recall	F1-Score	Support
2	0.98	0.99	0.98	83
4	0.99	0.98	0.98	95
Micro average	0.98	0.98	0.98	178
Macro Average	0.98	0.98	0.98	178
Weighted Average	0.98	0.98	0.98	178

Table 35 and Table 36 show the Micro, Macro and Weighted Average along with F1-Score for the original and oversampled data, respectively. The F1-Score for both tables for both classes is 0.98, 0.96 and 0.98, respectively. This means that the values for the false positives and the false negatives are very low, which indicates that the model is correctly classifying the correct class.

Experiment 7

The last experiment that has been carried out uses all the classifiers on the oversampled dataset. Figure 15 to 21 show the results of the ROC measure, which determines how well the all models are performing on our datasets.

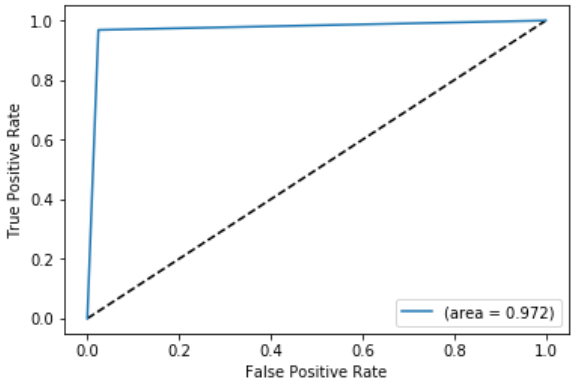

Figure 15: ROC Curve for Oversampled Dataset using Decision Tree Gini Index

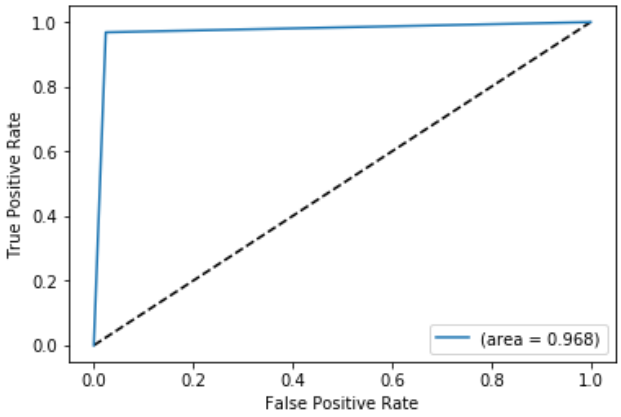

Figure 16: ROC Curve for Oversampled Dataset using Decision Tree using Entropy

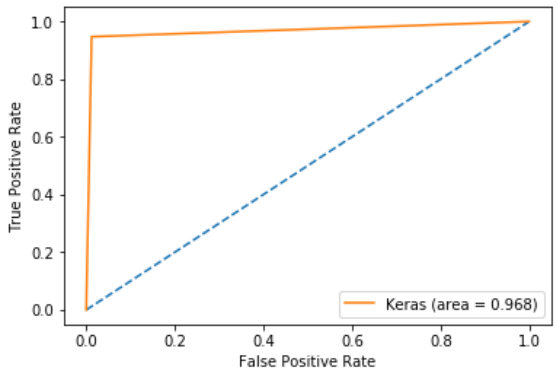

Figure 17: ROC Curve for Oversampled Dataset using Logistic Regression

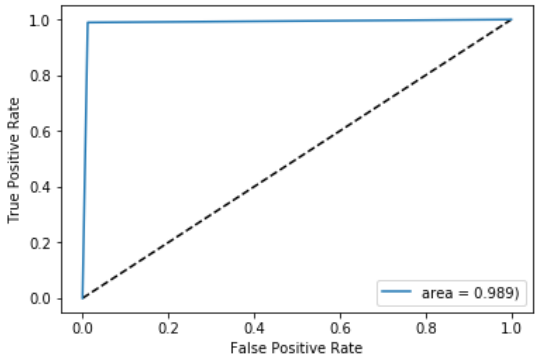

Figure 18: ROC Curve for Oversampled Dataset using Dense Feed Forward Neural Network

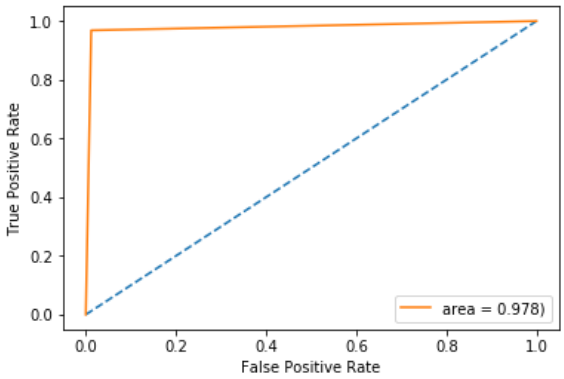

Figure 19: ROC Curve for Oversampled Dataset using K-Nearest Neighbors

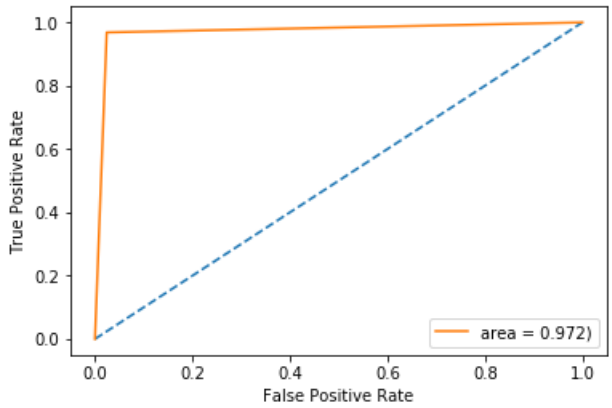

Figure 20: ROC Curve for Oversampled Dataset using Naïve Bayes classifier

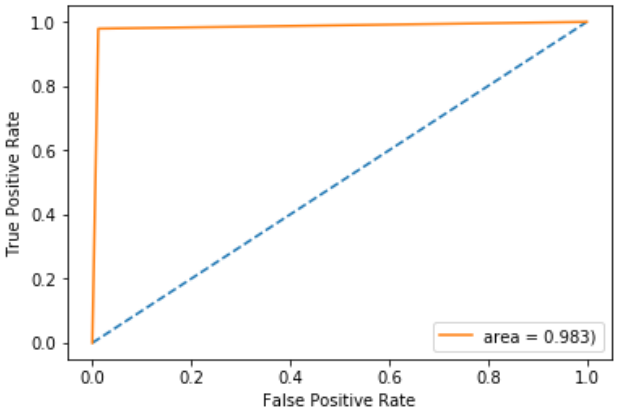

Figure 21: ROC Curve on Oversampled Dataset using SVM

The area that the ROC curves is 97.2 %, 96.8 %, 96.8 %, 98.9 %, 97.8 %, 97.2 % and 98.3 % in Figure 15 to 20, respectively. These AUC values are calculated on all models using only the oversampled dataset to verify which model obtains the best performance.

Table 37: Measure Values for all the classifiers

Measure Name	**Gini Index Decision Tree**	**Support Vector Machine**	**Naïve Bayes**	**K-Nearest Neighbors**	**Dense Feed Forward Neural Network**	**Logistic Regression**	**Entropy Decision Tree**
Training Accuracy	0.925	0.980	0.959	0.963	0.974	0.957	0.925
Testing Accuracy	0.971	0.983	0.971	0.977	0.988	0.966	0.966
AUC Score	0.972	0.983	0.972	0.978	0.988	0.967	0.968

Table 37 lists the accuracy values for both datasets and as well as the AUC score. The testing accuracies are 0.971, 0.983, 0.971, 0.977, 0.988, 0.966 and 0.966 on both the oversampled dataset. It is the testing accuracy which determines the performance of the model. It is clearly seen that the Dense Feed forward method shows the best performance in classifying the cancer data.

CHAPTER 6: CONCLUSION

In these above experiments it is seen that a model when trained with the oversampled dataset the model does a better job in classifying tasks. In the last part of the experiment where all the models are tested on the same test data it is noticed that the Feed Forward Neural Network and the Support Vector Machine are the two models that reaches an accuracy of 98%. The Feed Forward Neural Network outperforms the Support Vector Machine by 0.002% accuracy.

The future work for this implementation includes using images of breast cancers which will give the model a more detailed view about the kind of cancer. The use of images with the dataset which have been used in this work can be used to implement an intelligent system. The use of different feature values along with the images of the breast cancer will provide the machine learning algorithm a more detailed view about the cancer. This work can be extended by reducing the feature set to the minimal and selecting only those features that contribute the most in detecting the instance to be benign or malignant.

REFERENCES

[1] Albrecht, A. A., Lappas, G., Vinterbo, S. A., Wong, C. K., & Ohno-Machado, L. (2002). Two applications of the LSA machine. In Proceedings of the 9th international conference on neural information processing.

[2] Pena-Reyes, C. A., & Sipper, M. (1999). A fuzzy-genetic approach to breast cancer diagnosis. Artificial Intelligence in Medicine.

[3] Setiono, R. (2000). Generating concise and accurate classification rules for breast cancer diagnosis. Artificial Intelligence in Medicine.

[4] Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. Journal of Artificial Intelligence Research.

[5] Ster, B., & Dobnikar, A. (1996). Neural networks in medical diagnosis:Comparison with other methods. In Proceedings of the international conference on engineering applications of neural networks.

[6] Mehmet Faith Akay, (2008). Support vector Machines combined with feature selection for breast cancer diagnosis. Expert Systems with Applications.

[7] Nauck, D., & Kruse, R. (1999). Obtaining interpretable fuzzy classification rules from medical data. Artificial Intelligence in Medicine.

[8] Goodman, D. E., Boggess, L., & Watkins, A. (2002). Artificial immune system classification of multiple-class problems. In Proceedings of the artificial neural networks in engineering.

[9] Wolberg, W.H., Street, W.N., & Mangasarian, O.L (1995). Wisconsin Breast Cancer Dataset.

[10] William H. Wolberg, O. L. Mangasarian (1990). Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology. PNAS - Proceeding of the National Academy of Sciences.

[11] William H . Wolberg,(2001). Sparsity Through Automated Rejection. University College London.

[12] Ha, T. M., & Bunke, H. (1997). Off-line, Handwritten Numeral Recognition by Perturbation Method. Pattern Analysis and Machine Intelligence.

[13] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall & W. Philip Kegelmeyer (2002). SMOTE: Synthetic Minority Oversampling Technique. Journal of Artificial Intelligence Research.

[14] Shweta Kharya. (2012). Using data Mining Techniques for Diagnosis and Prognosis of Cancer Disease. International Journal of Computer Science, Engineering and Information Technology.