Breast Cancer Meets Machine Learning

By: Samriddh Gupta

Class: CSC-493

Final Project Report

Prof. Mehdi Owrang

Introduction

The project focus on what we can do with the data when machine learning algorithms are used to find some relation or verify correlations that are already available among the variables. I get the data from the professor, and by using Weka, started applying different classification methods and functions on it, so that I could get different results. Then with some of the variables that I got, I made different models using R to implement functions such as decision tree, neural net, etc. The Codes which I used here are mostly written by me and uses multiple R packages.

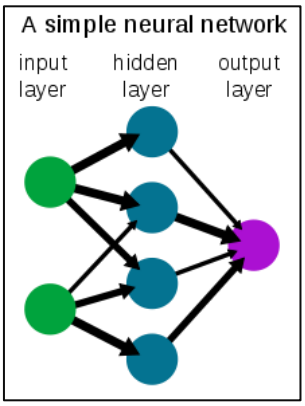
Data Collection and Cleaning

The data I got was vast and frankly too much to do for me in one semester. So, I decided that rather than taking the whole breast cancer data, I took the Triple-Negative Breast Cancer data. According to the article I read about Triple-Negative breast cancer, these are caused when patients test negative for estrogen receptors, progesterone receptors, and excess HER2 protein. These results mean the growth of the cancer is not fueled by the hormones estrogen and progesterone, or by the HER2 protein. So, triple-negative breast cancer does not respond to hormonal therapy medicines or medicines that target HER2 protein receptors. The focus area for mean in the data I used where the following variables: Age at the diagnose, Tumor Size, Tumor Extension, lymph nodes, survival months and life status (dead or alive). The data I get from here was then divided into two parts, the training data, and the testing data which was divided 80% and 20 % respectively. For one another study on the regression, I also investigated Ethnicity, Marital status NPI score etc. The results I got from there will be listed below.

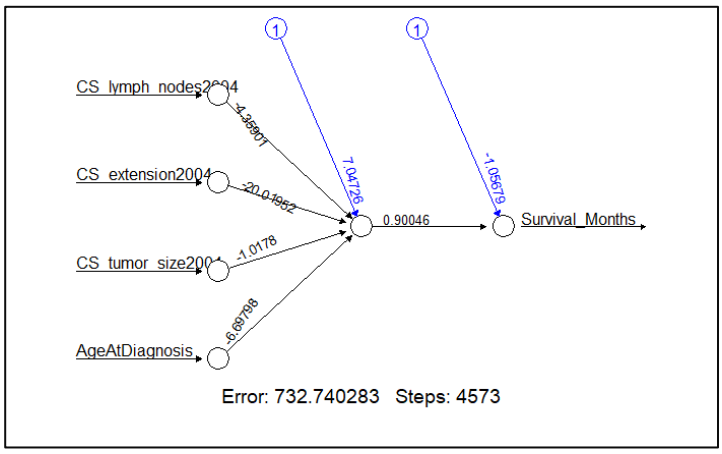
Machine learning Algorithm

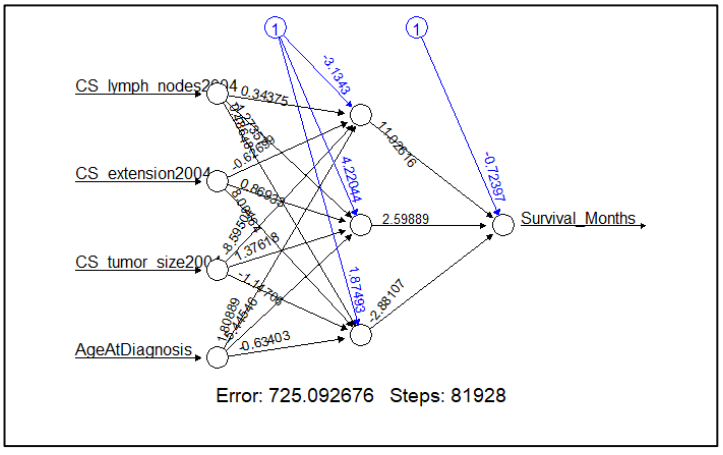
Neural Net

- **Theoretical Part:** A neural network is a network or circuit of neurons, or in a modern sense, an artificial neural network, composed of artificial neurons or nodes. Thus, a neural network is either a biological neural network, made up of real biological neurons or an artificial neural network, for solving artificial intelligence problems. The basic design of the neuron comes from the biological neuron where each neuron has a weight and that weight is applied in a form of mathematical function at the end of the node. In our case, we use a sigmoid function. Typically, a basic neural net has three layers: an input layer, a hidden layer, and the output layer. As the name suggests, the input layer's main functions are to provide variable data to the nodes and the output layer has the data giving the output of the model. The hidden layer is where most of the mathematical operations take place.



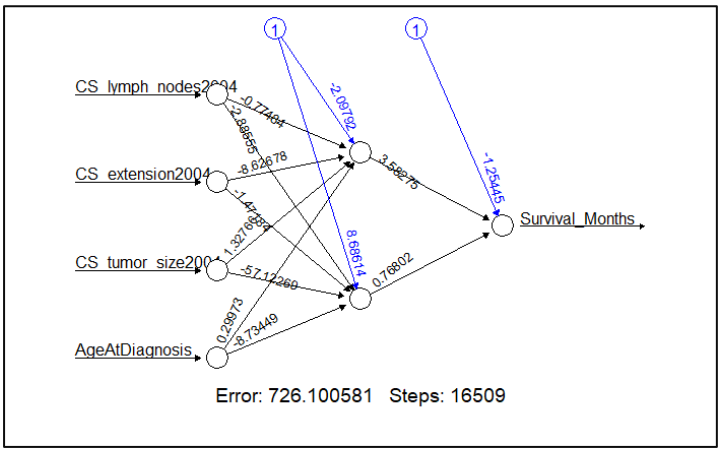
- **Project Part:** I used the model I created like this in survival analysis and whether a person will live or die. For the survival analysis, I did not get great results as the error I got from the testing data was very high here is the image of the neural net I made and their error for how many months will the patient survive after the diagnose of cancer.





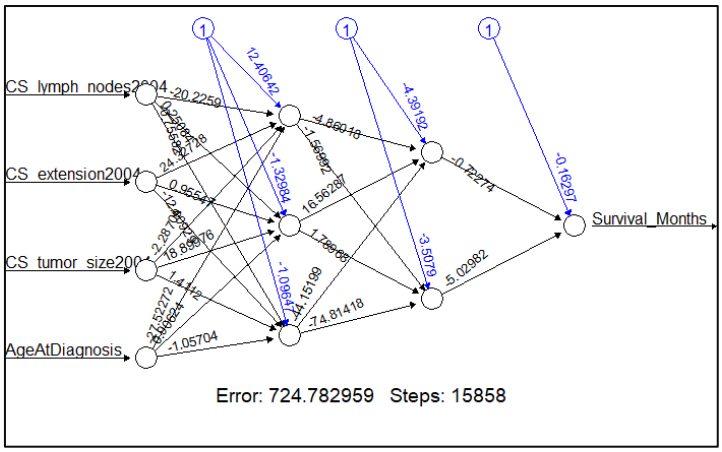
Error probability:

0.9873064



Error probability:

0.9909639



Error probability:

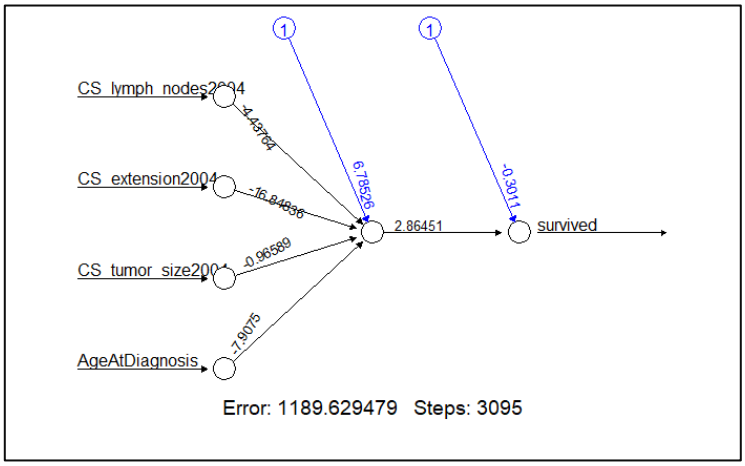
0.9898881

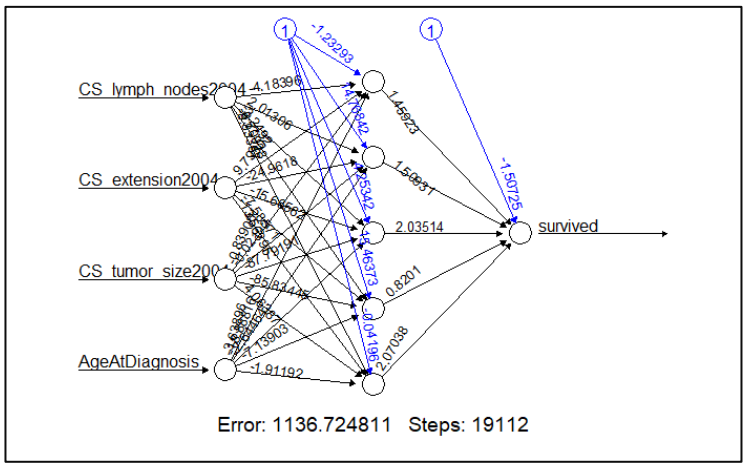
Error probability:

Gives an Error

As we can see from the error, they are quite high and that means that this model is not suitable for calculating. These results could be because of multiple issues, first is that the months were first to normalize between 0 and 1, then when we get the result they were expended back for getting the proper result, this could have caused the error as months will not match as they may differ even after rounding them up as power function is involved. Hence causing the problem in making confusion table and giving us such high errors. Other reason for the error could be that the variables I have chosen may not have many effects on the survival of months or there might

be more variable which I might have not included here which could have affected it. The second neural net I created was for, whether the person will live or die. Here are their neural net model and its error.



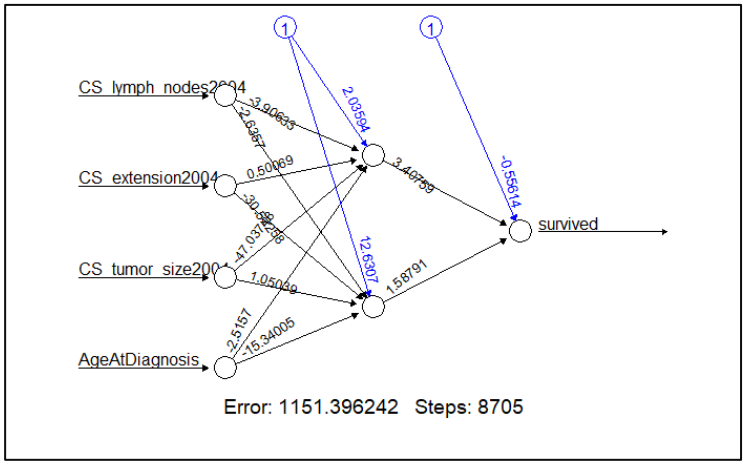


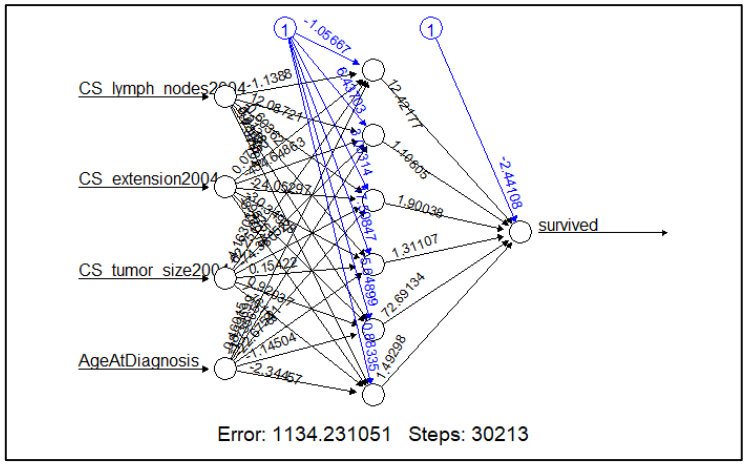
pred	0	1
0	292	147
1	847	3362

Error: 0.2138554

pred	0	1
0	268	131
1	871	3378

Error: 0.2155766





pred	0	1
0	292	147
1	847	3362

Error: 0.2138554

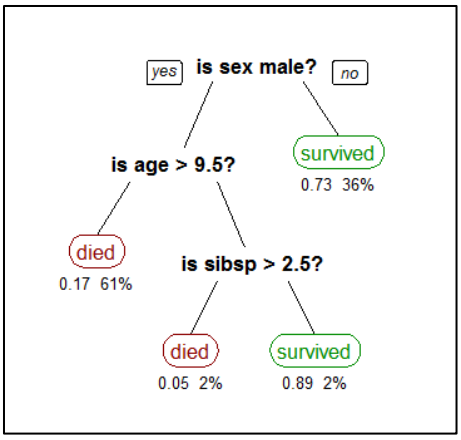
pred	0	1
0	252	124
1	887	3385

Error: 0.2175129

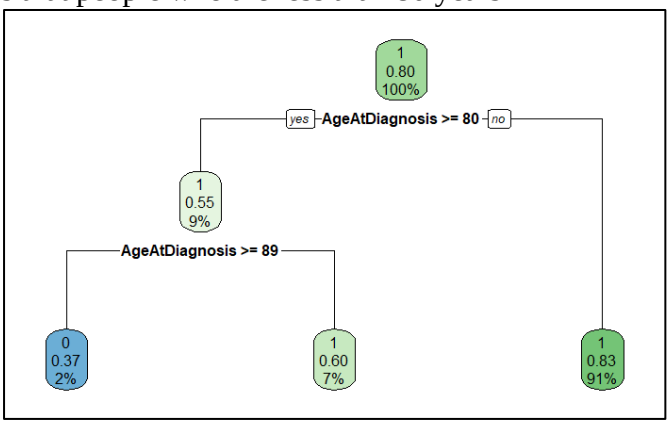
The error term is quite low for this model and we get a good model with a relatively high confusion matrix and a lower error value.

Decision Tree

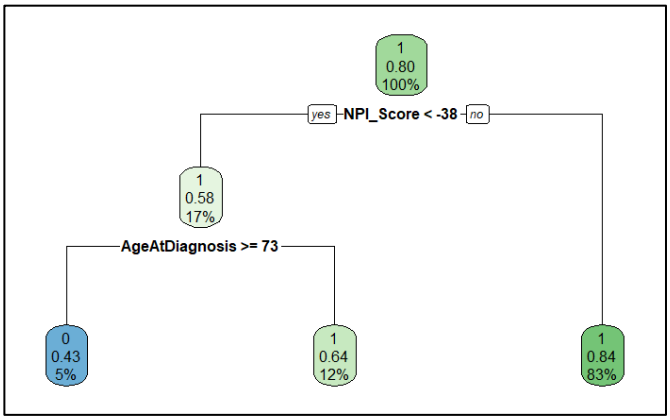
- **Theoretical Part:** A tree has many analogies in real life and turns out that it has influenced a wide area of machine learning, covering both classification and regression. In Decision Tree analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Usually, it is used when the data contain yes or no questions. It gives the percentage of the whole data that will lie in that category. What we can tell by decision tree is whether the data will fall in one category of in another.



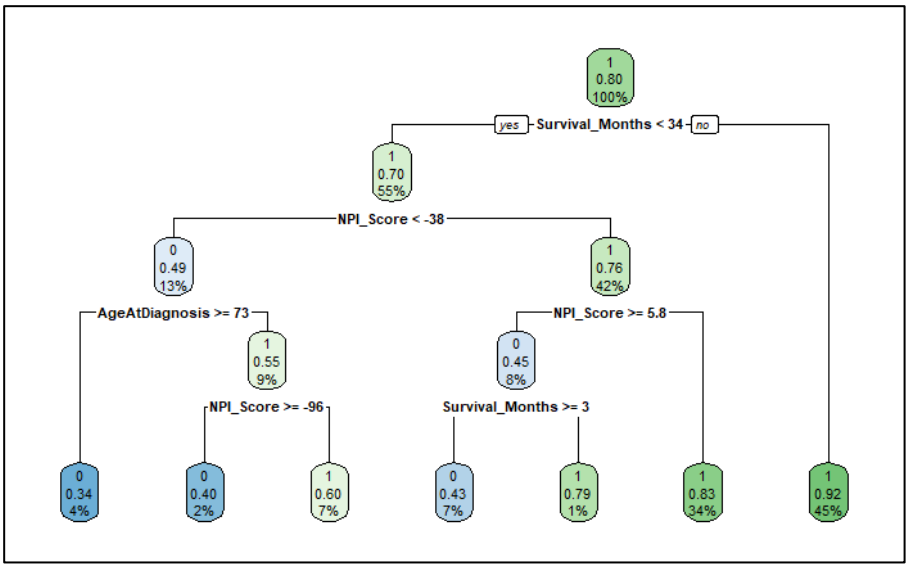
- **Project Part:** I made a couple of decision trees with multiple variables and will be discussing there here. The first decision tree I made was with the age. As you can see from the image. It displays one section of the age and gives the percentage of people who are alive on either side of the data. We can see from this that people who are less than 80 years survive every time. The people. It also tells us the percentage of people in that group like for example, in the first part of the node 80 percent of people survive at any given condition. People who are 89 years and above are bound to die with survival probably of only 37 percent and it accounts to be 2% of overall data. So, in simpler terms, we can say that people at and above 89 years will die with a probability of 37%. People between 89 and 80 years can survive with a 60% chance of survival and people below 80 can survive with an 83% chance of survival.



Here is the next decision tree with I am going to show have two variables in it. We have age and the other one is the NPI Score(Nottingham prognostic index). Over here we can explain the images as if your NPI score is less than -38 and your age at the time was 73 and greater, chance if your survival is 43 %. While the same condition of NPI score and age to be less than 73 chance of survival is 64 percent and if your NPI score is more than -38 then the chance of survival is 84%.

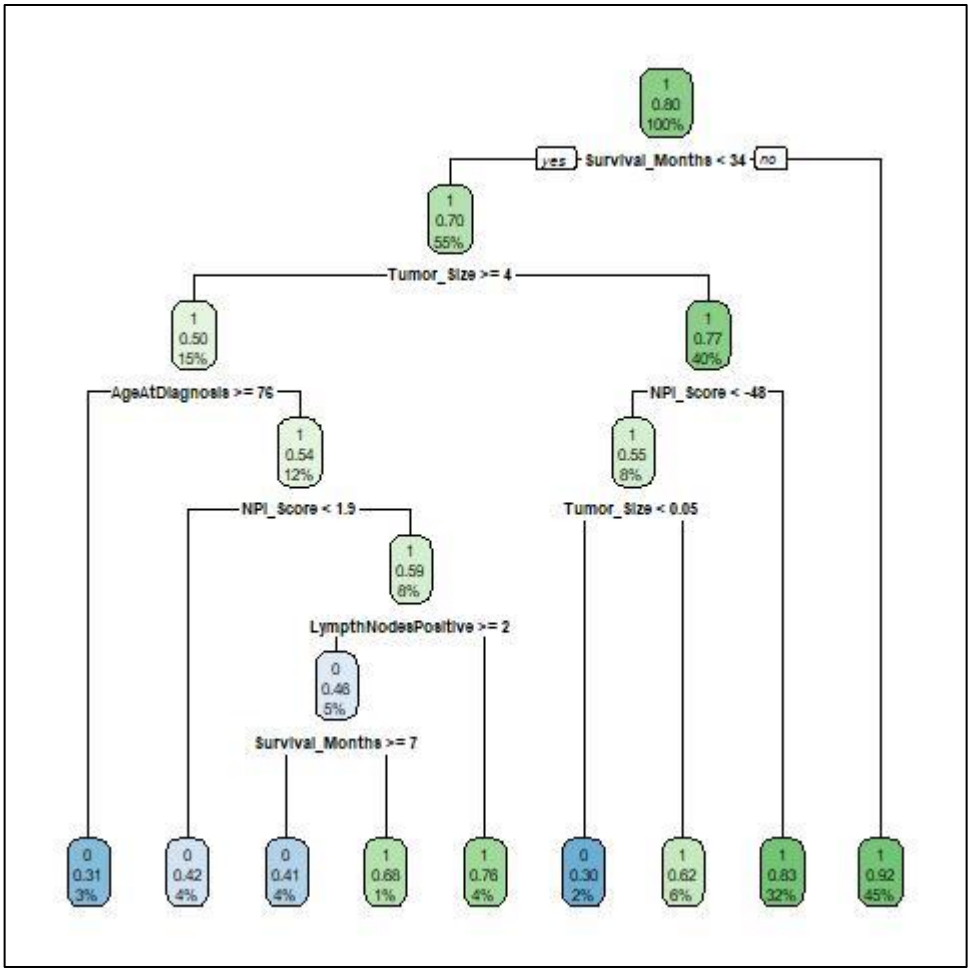


The next decision tree I decision tree that I will discuss have three variables in them. As you can see from the image we can further divide the data based on how many months the person survives. Let discuss each case from left to right. The first case can be explained as if a person survives for less than 34 months, have the NPI score of less than -38 and is 73 years and above at the time of detection then the chance of your survival is 34% while holding the same condition but if you were less than 73 years old but have NPI score of more than -96, then you have 40% chance of survival and if NPI score ranges between -96 and -38, then you have 60 % chance of survival. IF you survive for less than 34 months NPI score is more than -38 and more than 5.8 and have survived for more than 3 months then there is a 43% chance of survival and if you have less than 3 months then you 79% chance of survival. If your NPI score is less then 5.8 then you have an 83 % chance of survival and of you have survived for more than 34 months than you have a 92% chance of survival.



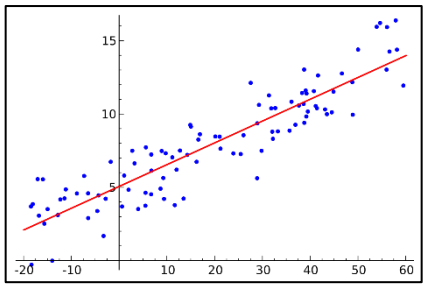
The last image that I have here has a total of five variables and it gives a detailed explanation of the cluster of an area where each plot can be added. So I will start explaining ending nodes from left to right. If survival months are less than 34 and tumor size more than 4 cm along with age at diagnosing was more than 76, the patients have a 31% chance of survival. With the same condition as earlier but if patients' age was less than 76 have

had an NPI score of less than 1.9, the patient will have a 42% chance of survival. If NPI score was more than 1.9 and have the same conditions then, along with Lympthnodespositives be more than 2 and patients have survived more than 7 months than chances of survival are 41%. If patients have survived less than 7 months the chance of survival is 69%. If Lympthnodespositives were less than 2 then patients have a 76% chance of survival. Now if Tumor size was less than 4 cm NPI score is less than -48 and further tumor size is less than 0.05, then there is a 30% chance of survival while if the tumor size is more than 0.05, then there is 62% chance of survival. If the NPI score was more than -48, then there is an 83% chance of survival.And finally, if the patients have survived for more than 34 months than there is a 92 % chance of survival.



Regression

- Theoretical Part: Regression is a statistical method used in finance, investing, and other disciplines that attempt to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables). This can be used to find what variables are corresponding to finding values and values are not. For this, I used seven variables to find any relation using the linear model.



- Project Part: The first thing I did was do a simple linear model by applying the formula and trying to get the result. I get the following equation.

Y= 28.54067 - 0.04512 * (X1) + 0.07564 * (X2) + 12.32351 * (X3) - 0.54310 * (X4) - 0.58474 * (X5) + 0.63392 * (X6) - 0.08749 * (X7)

Where Y= Survival Months, X1= Age at Diagnosis, X2= Tumor Size, X3= survived, X4= NPI Score, X5= Ethnicity Race Class, X6= Lymph Nodes Positive, X7= Marital Status Class.

Now if we see it like this, we can say that there is a lot of this which is favoring it and we are getting a really good model, but the problem with this is that when I calculated it R square value it comes out to be quite low(0.1107). This means that overall there is very little correlation between the Dependent variable and independent variable. Then when I tried to find whether the variables are necessary for the model and I got the following P-values.

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         28.540667   0.861246  33.139  < 2e-16 ***
AgeAtDiagnosis      -0.045123   0.009467  -4.766 1.89e-06 ***
Tumor_Size           0.075639   0.034337   2.203 0.027616 *
survived            12.323515   0.333547  36.947  < 2e-16 ***
NPI_Score           -0.543096   0.167048  -3.251 0.001151 **
Ethnicity_Race_Class -0.584738   0.092841  -6.298 3.06e-10 ***
LympthNodesPositive  0.633918   0.167405   3.787 0.000153 ***
Marital_Status_Class -0.087486   0.100611  -0.870 0.384559
```

We can see the variables have lower p-value and that means that those are not good for the models and those variables can be removed from the model, but even after removing those variables, I get the following p-value.

```
(Intercept)          34.384042   0.247029 139.190  < 2e-16 ***
Tumor_Size           -0.078691   0.010745  -7.324 2.49e-13 ***
NPI_Score             0.120377   0.003485  34.539  < 2e-16 ***
Marital_Status_Class -0.553376   0.100126  -5.527 3.30e-08 ***
```

The P-value without the other variables is small as well. So, we can say that the variables do not have any linear relation to the dependent variable.

Next, I tried something a little different. I tried to find If the Ethnicity or Marital status have on Survival months. I know that the variables do not affect them individually but, we can at least get which part of Ethnicity has more effect and which have less effect. Here are the p-values and results I got.

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    32.2863     0.1760 183.476  < 2e-16 ***
as.factor(Ethnicity_Race_Class)2  -3.1110     0.3744  -8.310  < 2e-16 ***
as.factor(Ethnicity_Race_Class)3  -0.9666     0.4079  -2.370 0.017814 *
as.factor(Ethnicity_Race_Class)4  -0.9209     2.8611  -0.322 0.747539
as.factor(Ethnicity_Race_Class)5  -1.5832     0.4804  -3.296 0.000984 ***
as.factor(Ethnicity_Race_Class)7  -2.5611     1.1101  -2.307 0.021052 *
```

As we can see from the tables, some of variables here can be use and here is result for the coefficients is given here.

```
Coefficients:
                     (Intercept)  as.factor(Ethnicity_Race_Class)2
                         32.2863                           -3.1110
as.factor(Ethnicity_Race_Class)3  as.factor(Ethnicity_Race_Class)4
                         -0.9666                           -0.9209
as.factor(Ethnicity_Race_Class)5  as.factor(Ethnicity_Race_Class)7
                         -1.5832                           -2.5611
```

From these coefficients, we can say that Ethnicity 2 which is Black Non-Spanish-Hispanic-Latino survives less than other Ethnicity, while Ethnicity 3 and 4 which are White people and Black people from South or Central American excluding Brazil are doing much better and survives more than other ethnicities.

Other Research

The Other research which I did was Getting the result I used for all the breast cancer data set. The problem I have with this was since the data was quite large, I was not able to do it in R and it took scientifically longer to calculate them here. Because of this I must leave this part blank and I was not able to do anything for this section and this will probably become work for the future.

Limitations

The limitation of the project was I think the data part of the project. The data which was provided to me though was very large, but in excel the data rows have a limit and it was not showing me any more data after a million records, so I was not able to see all the data in the project. Secondly, the NPI score was from 1 to 6 range, but we also have a negative result as well

as positive results more than 6 which means that different dictionary was used when data were combined which could have a case this problem. Another problem was the time, as the neural net took 30 minutes to compile the whole code and it was problematic. Rather than using my machine, we could use virtual computing with a stronger ram and faster computing power.

Conclusion

We did test for multiple models both in Weka as well as in R. There were several things we get and all that information I got; I have presented earlier. There are several other things as well that we can do, and new machine learning techniques can be applied here. But as a part of the project and constraints, I have got overall, I will end that here. I can say her thought that the models used for Whether a person is alive or not have great results from decision trees and we can divide all data set in many groups. The data gives us quite useful information and hope we can get more such information if someone took this project in the future.

Reference

- http://www.sthda.com/english/wiki/creating-and-saving-graphs-r-base-graphs
- https://www.datacamp.com/community/tutorials/decision-trees-R
- https://www.datacamp.com/community/tutorials/neural-network-models-r