Utilizing Routine Blood Analysis Parameters to Predict Breast Cancer

Moy, Cam F.

Rawl’s College of Business

Program - MS. Data Science

Table of Contents

[Table of Contents 2](#_Toc518483952)

[Table of Figures: 2](#_Toc518483953)

[Table of Tables: 3](#_Toc518483954)

[Introduction 3](#_Toc518483955)

[Background and statistics 3](#_Toc518483956)

[Data Set 4](#_Toc518483957)

[Age 4](#_Toc518483958)

[BMI 5](#_Toc518483959)

[Glucose 5](#_Toc518483960)

[Insulin 6](#_Toc518483961)

[HOMA 6](#_Toc518483962)

[Leptin 7](#_Toc518483963)

[Adiponectin 7](#_Toc518483964)

[Resistin 7](#_Toc518483965)

[MCP-1 8](#_Toc518483966)

[Hypothesis 8](#_Toc518483967)

[Statistical Analysis 8](#_Toc518483968)

[Data Exploration 9](#_Toc518483969)

[Data Distribution 9](#_Toc518483970)

[Correlation 11](#_Toc518483971)

[Balancing Dataset 12](#_Toc518483972)

[Normalizing the datasets 13](#_Toc518483973)

[Statistical Analysis 13](#_Toc518483974)

[K-Nearest Neighbor 13](#_Toc518483975)

[Logistic Regression 13](#_Toc518483976)

[Random Forest for Classification 13](#_Toc518483977)

[ROC Curves 14](#_Toc518483978)

[Conclusions 17](#_Toc518483979)

[Hypothesis Conclusion 17](#_Toc518483980)

[Ongoing Work 17](#_Toc518483981)

[References 19](#_Toc518483982)

# Table of Figures:

Figure 1 Relationships among variables 8

Figure 2 Age and BMI distributions 9

Figure 3 Glucose and insulin distributions. 9

Figure 4 HOMA and Leptin distributions 10

Figure 5 Adiponectin and resistin distributions. 10

Figure 6 MCP-1 distribution. 11

Figure 7 Correlation plot for all variables. 12

Figure 8 Normalization Function 13

Figure 9 ROC curve for KNN 15

Figure 10 ROC curve for Logistic Regression 16

Figure 11 ROC curve for Random Forest 17

# Table of Tables:

Table 1 Body Mass Index Parameters 5

Table 2 Ideal Blood Sugar Levels 6

Table 3 ROC curve color-code 14

Introduction

## Background and statistics

Egypt 1600 BC is the first record of this phenomena. Fast forward to ancient times where Hippocrates named this disease as the “traveling womb”. The traveling womb was the belief that a woman’s uterus could move throughout the body leaving health problems of every kind in its wake (Sara Suchy, 2012) . Today we know and recognize this disease as breast cancer. Breast cancer is a predominant illness which survival rate increases with an early diagnosis.

Breast cancer is the medical term given to the uncontrollable replication of cells know as, tumor(s) which affect both, men and women. When tumors are classified as malignant it indicates that the cells in question are abnormal and possess the capability to travel to other parts of the body. Breast cancer is the second most predominant cancer in American women; having 1 in 8 women developing it at some point in their life. For men, the odds are significantly lower affecting 1 in a 1000 men. The statistics for 2018 generated by the American Cancer society are as follow:

* They predict a total of 41, 400 deaths from breast cancer which 40, 920 will be women and 480 men (Cancer.Net, 2018).
* The diagnosis of 63, 960 cases of carcinoma in situ, the earliest form of breast cancer.
* Finally, they predict 266, 120 new diagnosis for non-initial phases of breast cancer.

Even though breast cancer survival is dependent on many factors; many researchers and medical practitioners agree that an early and correct diagnosis have a great positive impact in the survival rate of the patient.

## Data Set

The goal of this study is to correctly diagnose individuals that have breast cancer given the parameters and accumulated data from the UCI Machine Learning Repository “Breast Cancer Coimbra Data Set”. The reason for choosing this dataset lies within the parameters. Most breast cancer diagnosis research relies on tumor metrics. While tumor metrics are important to discover the phase in which the patient is as well as possible treatments, these measurements are usually taken after the patient has been diagnosed, while receiving treatment or during a mammography checkup. The “Breast Cancer Coimbra Data Set” provides more common measurements that every patient undergoes during a regular visit to a doctor or a routine blood analysis. Such measurements are:

* Age
* Body Mass Index (BMI)
* Glucose Levels
* Insulin Levels
* Homeostatic model assessment (HOMA)
* Leptin levels
* Adiponectin levels
* Adipocyte-Specific Secretory Factor (ADSF) or Resistin levels
* Monocyte Chemoattractant Protein- 1 (MCP-1) levels

These measurements were taken from 64 breast cancer patients and 52 healthy individuals. If the correct diagnosis, of breast cancer, can be done with these parameters early diagnosis will increase and by consequence the survival rate of patients.

The following represents a brief documentation of each variable, their interactions and role in breast cancer:

### Age

Age affects the way cells interact and behave. The older an individual is the more likely abnormal cells can be generated. The more these abnormal cells occur the higher the probabilities are for men and women to develop breast cancer. The rate of breast cancer diagnosis for women start increasing after 40 years of age and 68 years of age for men.

### BMI

Body Mass Index (BMI) is a measure created to define body weight in four different status underweight, normal, overweight and obese. The magnitudes of BMI and its corresponding classification can be found in Table1.

|  |  |
| --- | --- |
| **BMI** | **Body Weight Status** |
| 18.4 and lower | Underweight |
| 18.5 to 24.9 | Normal |
| 25.0 to 29.9 | Overweight |
| 30.0 and greater | Obese |

Table 1 Body Mass Index Parameters

Many studies have linked BMI to breast cancer. These studies have concluded that BMI affects women chances to develop breast cancer differently before and after menopause. If a woman has not yet reach menopause, and has a BMI range of 25 and greater she will have 20 to 40 percent lower risk of breast cancer compared to other women with lower BMI index. This risk shifts after menopause. A woman who lands on the “obese” or “overweight” category after menopause increases her risk by 30 to 60 percent. Researchers have link body weight status to estrogen production. Women with higher estrogen levels has an increased risk of developing breast cancer compared to women with lower estrogen levels.

The reason for this risk shift is due to where estrogen is produced before menopause and after menopause. Before menopause estrogen is mainly produced in the ovaries when a woman hits menopause fatty tissue is the main source of estrogen. Fatty tissue contains an enzyme called aromatase (mainly in fatty tissue) which converts the hormone androgen to estrogen. The more “fatty tissue” a woman has, the higher the production of estrogen in her body. Which, is positively correlated to a higher risk of breast cancer.

### Glucose

Blood sugar or glucose is the main source of energy for humans. The human body creates glucose out of the nutrients received during the day from proteins, fats and carbohydrates. Many research papers link high-normal glucose levels to increased breast cancer risk (David Spero, 2018).

Maintaining healthy glucose levels not only reduce the risk of breast cancer but also the appearance of many other health problems. Glucose levels are a consequence of an individual’s diet. The more a person ingests sugars and starches the higher their blood sugar. Glucose levels are also affected by age. The older a person gets the more glucose the liver might produce. The excess of glucose is known as “gluconeogenesis” which indicates an excess production of glucose regardless of the person’s diet. Another important factor link to the rise pf glucose level in the body is BMI. The “heavier” a person is the greater is their BMI and therefor, their glucose levels. Table2 showcases the recommended levels of glucose according to the individuals ingest of food.

|  |  |
| --- | --- |
| **Blood Sugar** | |
| **Fasting** | |
| Individual without diabetes | 70 – 99 mg/dl |
| ADA Recommendation for people with diabetes | 80 – 130 mg/dl |
| **Two hours after a meal** | |
| Individual without diabetes | Less than 140 mg/dl |
| ADA Recommendation for people with diabetes | Less than 180 mg/dl |

Table 2 Ideal Blood Sugar Levels

### Insulin

Insulin is a hormonal stimulator for cellular proliferation (MedicineNet, 2012) and enables the utilization of glucose. Insulin turns blood sugar into energy (David McCulloch and Kaiser Permanente, 2014). This hormone also promotes the storage of glucose into the muscles, fatty tissue and liver as a reserve for when this energy is needed. People who are heavier tend to have higher levels of insulin.

Many studies have link postmenopausal women with higher levels of insulin while, the opposite is true for premenopausal women. These studies have also showcased a positive linear correlation between higher levels of insulin and a worse prognosis regarding breast cancer.

### HOMA

Homeostatic model assessment (HOMA) is a method for assessing beta-cell function and insulin resistance (Tara M. Wallace, 2004). The HOMA model provides us with an estimate of insulin sensitivity from fasting and glucose levels in the body.

Studies have link insulin resistance with women suffering from breast cancer. Body tissues tend to become insulin resistant when a person is obese or overweight. Insulin resistance can become so high that a patient or individual can suffer from hyperinsulinemia, high activation of insulin pathways. This symptom can aid in the proliferation of cancer cells indicating that women with insulin resistance are at a higher risk of developing breast cancer.

### Leptin

Leptin also known as the “obesity, fat or starvation hormone” is a protein created in the fat cells. This protein signals to the brain the stored energy condition in your body. Every individual has a different leptin threshold desired by the individual’s body. If a person starts dieting the brain notices a lack of leptin and signals to the body a starvation warning. This warning tells the body to fell more hunger, store more nutrients and fat to bring the leptin levels back to normal. So far, leptin is widely known as a weight regulating protein but new evidence shows leptin involvement in tumorigenesis specially in the development of breast cancer (Surmacz, 2007).

Leptin’s interaction with breast cancer was studied by the utilization of breast cancer models. In this studies Leptin showcased the following behavior:

* Leptin aid with the proliferation of cancerous cells.
* Cross-signaling with breast cancer cells.
* Modulation of estrogen receptor (elevates estrogen levels).

Several other studies confirm the above behavior of leptin. Even though studies have confirmed leptin is an active factor in breast cancer researchers are still unable to track the origin of leptin acting with breast cancer cells. The most important conclusion of these studies is, breast carcinogenesis can induce an overabundance of leptin which in consequence causes resistance to breast cancer treatments.

### Adiponectin

Adiponectin is a protein hormone exclusively produced and secreted by fat cells which aids in the regulation of lipids and glucose. This protein heavily influences the body’s response to insulin (MedicineNet, n.d.). One of the main proteins produced by adiponectin is leptin. Different research papers have concluded that lower levels of adiponectin are correlated to obesity and influences insulin resistance. So far, studies are inconclusive about adiponectin role in breast cancer but, there is a popular speculation that higher levels of adiponectin may decrease the risk of breast cancer.

### Resistin

Resistin also known as adipocyte-specific secretory factor (ADSF) is a protein exclusively secreted by adipose tissue. This protein is actively involved in the regulation of adipogenesis and insulin resistance. Studies revolving the role of ADSF in breast cancer are still inconclusive but one fact has been proven is several studies. The high levels of resistin are positively correlated to increased breast cancer risk.

### MCP-1

Monocyte Chemoattractant Protein- 1 (MCP-1) is one of the many chemokine receptors which oversee regulating cell traffic throughout the body. This receptor has been strongly linked to triple breast cancer, the most aggressive type of breast cancer. Studies have shown that MCP-1 regulate cancerous cell invasion to other organs. Studies suggest the regulation of MCP-1 mediated pathways to target and reduce cancerous cells. The greatest source of MCP-1 is adipose tissue being severely correlated to a high BMI and insulin resistance.

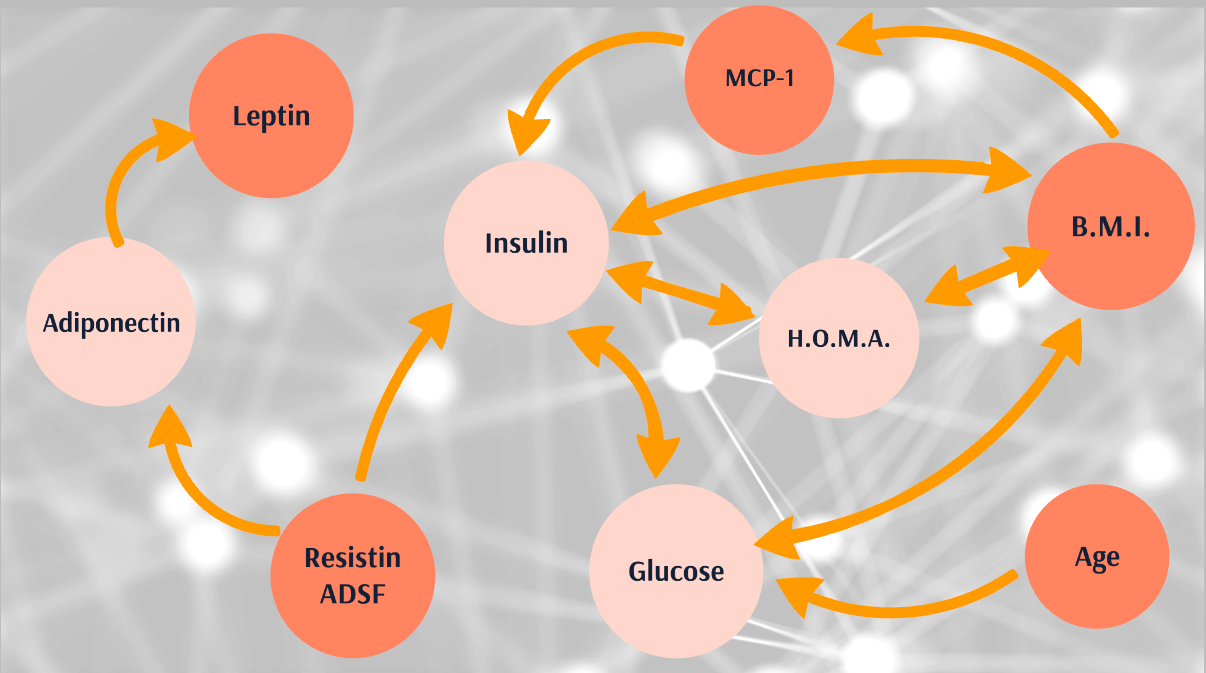


Figure 1 Relationships among variables

Figure1 showcases the different interactions among variables.

## Hypothesis

Individuals with higher BMI and glucose index will fall under the “patient” classification without the need to further normalize the dataset.

## Statistical Analysis

Given that the dataset is very small the following classifications will be implemented:

* K- Nearest Neighbors
* Logistic Regression
* Random Forest

These classification methods will judge based on the accuracy that they provide.

# Data Exploration

## 

## Data Distribution

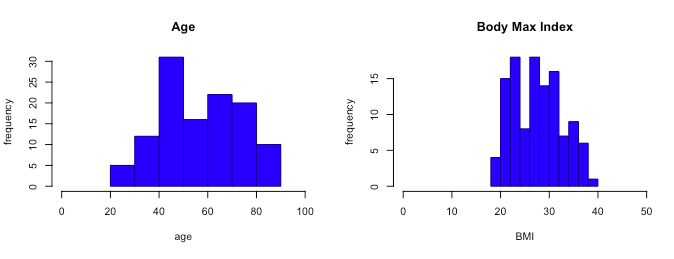


Figure 2 Age and BMI distributions

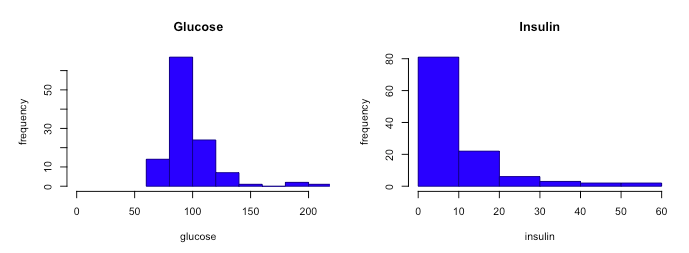


Figure 3 Glucose and insulin distributions.

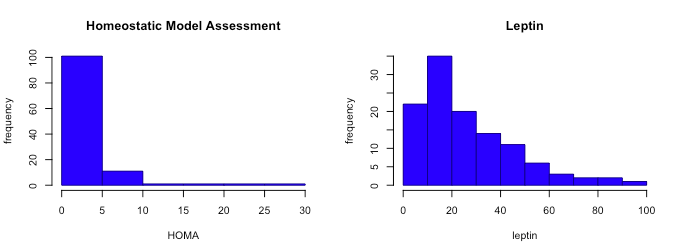


Figure 4 HOMA and Leptin distributions

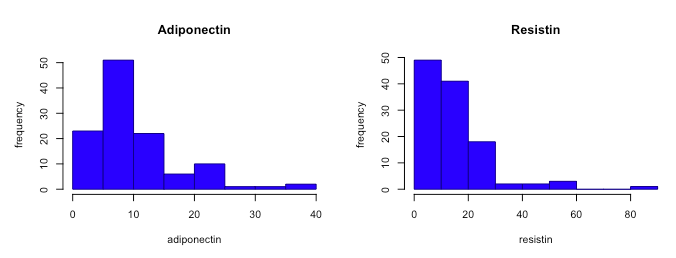


Figure 5 Adiponectin and resistin distributions.

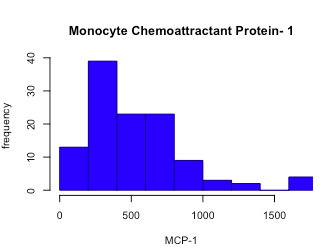


Figure 6 MCP-1 distribution.

Given the data distributions we can easily see the following:

1. Normalization cannot be assumed for any variable
2. The dataset has outliers that will make the classification algorithms bias.
3. There might be sufficient evidence to prove the correlation among the variables that regulate the production of cells such as glucose, insulin and adiponectin.

## Correlation

When computing the correlation matrix, the highest correlation expectations are among the variables glucose, BMI, HOME, MCP-1, ad insulin. The expectations for the dataset are met with the following values in descending order:

1. HOMA and Insulin correlation value = 1
2. HOMA vs Glucose correlation value = (0.6, 0.8)
3. BMI vs Leptin correlation value = (0.4, 0.6)
4. Glucose vs Insulin correlation value = (0.4, 0.6)
5. Glucose vs Classification & Resistin vs MCP-1 correlation value (0.2, 0.4)

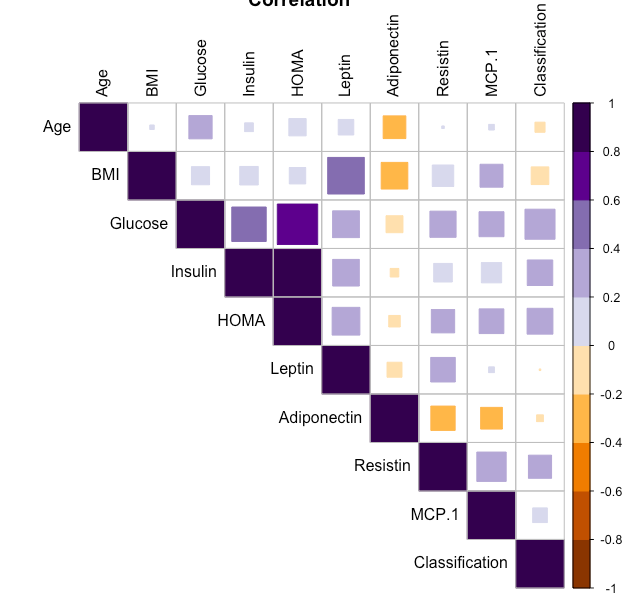


Figure 7 Correlation plot for all variables.

## Balancing Dataset

The dataset provided shows clear signs of imbalance therefore, adjustments must be made to increase the accuracy while performing classification. Machine learning classifiers are affected by the proportions of the different classes. Most of these algorithms tend to favor the most predominant class which can lead to misleading accuracies or classification of a new observation. Given that breast cancer is a serious medical condition avoiding misleading accuracies and classifications is a must to ensure the correct diagnosis of future patients.

There are many ways to deal with imbalance data of which the following have been chosen for this task; oversampling (blnc\_over), undersampling (blnc\_under), a mixture within over and undersampling (blnc\_both) and synthetic data (data,rose). “All these methods modify the class proportion and the size of the original dataset” (Team, 2016).

## Normalizing the datasets

When performing the planned algorithms, the accuracy was very low indicating that normalization is an issue. Therefore, all datasets were normalized by the formula below and re-tested in all the algorithms.



Figure 8 Normalization Function

After normalization, there was an accuracy improvement but it was not significant enough to establish a successful algorithm.

# Statistical Analysis

For the statistical analysis 3 main algorithms were used for each of the balance datasets. All 3 algorithms need improvement; accuracy was based on the confusion matrix and the Receiver Operating Characteristic (ROC). The confusion matrix allows us to determine a ratio or percentage of the correctly classified observations versus the incorrectly classified. The addition of the ROC curve is used to evaluate and compare the models against each other. Roc curves allow us to see how a predictive model distinguishes between true positives and negatives by, utilizing sensitivity and specificity.

## K-Nearest Neighbor

The reasoning behind the utilization of K nearest neighbor algorithm is based on the hypothesis that observations containing large BMI will be placed relatively close together. Given that KNN classifies observations based on the distance between them the expectation was to have a higher true positive classification number than other algorithms.

During the trials this model seem to work better for the oversampling normalized dataset. The overall classification has room for improvement.

## Logistic Regression

Logistic regression model was chosen do to its popularity; for the classification of binary data. This model performed better than the others but still showcase large amounts of miss-classification. The model showcased a more uniform performance over all balanced datasets. The dataset that was better predicted by this model was the synthetic data created with the ROSE package.

## Random Forest for Classification

Random forest is a very popular classification model known for its learning capabilities by adjusting the weights and nodes to better the algorithm. The main goal for the utilization of this algorithm was to encounter a learning path that will deliver a minimum of 95% confidence interval. Given that the data set is very small and unbalance the random forest didn’t perform as expected, regardless of the balancing and normalization of the given dataset. If more synthetic data is generated or more real observations are added to this model the overall performance will increase exponentially.

This algorithm performs mostly uniformly among all balance dataset with a clear winner being the synthetic balanced normalize data since it contained the most observations.

## ROC Curves

The following graphs showcase the ROC curve for all algorithms that were utilized. The datasets utilized for each run are colored coded on the following way:

|  |  |
| --- | --- |
| **Type of balanced data** | **Color** |
| Oversampling | Blue |
| Undersampling | Purple |
| Over and Under Sampling | Orange |
| Synthetic Data | Green |

Table 3 ROC curve color-code

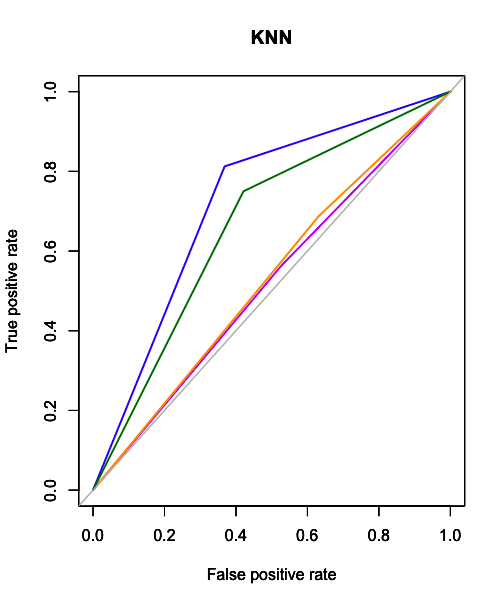


Figure 9 ROC curve for KNN

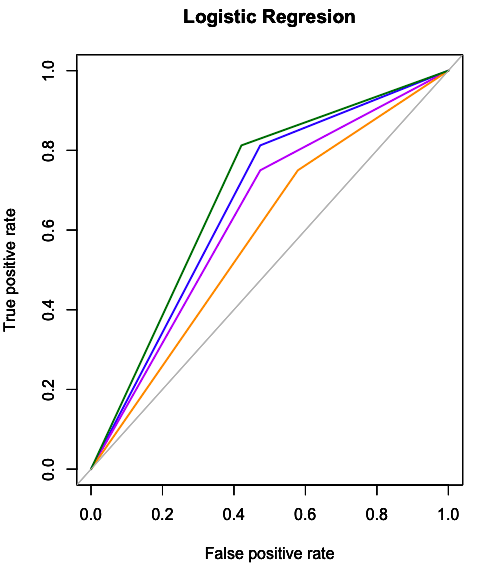


Figure 10 ROC curve for Logistic Regression

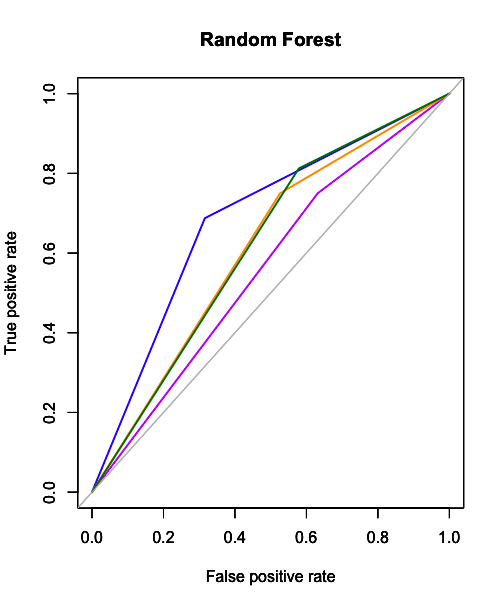


Figure 11 ROC curve for Random Forest

# Conclusions

### Hypothesis Conclusion

The hypothesis has been rejected since normalization is a key component for accurate classification. Regardless of normalization being done in the dataset classification algorithms were still heavily influenced by the extremas of the dataset. It is necessary to find a new method for normalization that will more accurately normalize the dataset.

## Ongoing Work

In the ongoing work for this project is as follows:

* Fit in a median value for each variable to achieve a better normality distribution for all variables. This will give us a more accurate normalization.
* Include a 95% confidence interval to judge model’s accuracy. Given that breast cancer is a very predominant and aggressive disease the correct diagnosis by these parameters is a must to lower the casualty rate.
* Combine with present dataset with a larger one to have a balance data without the need of over/under sampling or generating new observations. By increasing our dataset with recorded observations instead of synthetic the algorithm will have a more realistic dataset sample to learn from.
* Create a visualization in which showcases the algorithms classifications as well as the correct classification of the observations.

# 

# References

[Patricio, 2018] Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer, 18(1).

(n.d.).

(2018, 03 06). Retrieved from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra#

American Cancer Society. (2017, 08 1). *American Cancer Society*. Retrieved from What is Breast Cancer?: https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html

American Cancer Society. (2017, 07 01). *American Cancer Society*. Retrieved from How Common is Breast Cancer?: https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html

BreasrCancer.org. (2015, 11 19). *Higher Insulin Levels Linked to Worse Prognosis of Metastatic Breast Cancer*. Retrieved from BreasrCancer.org: https://www.breastcancer.org/research-news/insulin-levels-linked-to-mets-prognosis

Breast Cancer Now's. (2017, 11). *Reduce your risk of developing breast cancer after the menopause by maintaining a healthy weight throughout your life*. Retrieved from Weight and breast cancer risk: http://breastcancernow.org/about-breast-cancer/want-to-know-about-breast-cancer/what-can-cause-breast-cancer/reduce-risk-of-breast-cancer/weight-and-breast-cancer-risk

Cancer.Net. (2018, 01). *Breast Cancer: Statistics*. Retrieved from Cancer.Net: https://www.cancer.net/cancer-types/breast-cancer/statistics

David McCulloch and Kaiser Permanente. (2014, 03 01). *Kaiser Permanente*. Retrieved from How Insulin Works: https://wa.kaiserpermanente.org/healthAndWellness/index.jhtml?item=%2Fcommon%2FhealthAndWellness%2Fconditions%2Fdiabetes%2FinsulinProcess.html

David Spero, B. R. (2018, 03 27). *What Is a Normal Blood Sugar Level?* Retrieved from Diabetes Self-Management: https://www.diabetesselfmanagement.com/blog/what-is-a-normal-blood-sugar-level/

Jee-Hyun Kang, B.-Y. Y.-S. (2007, 02 28). *Relationship of Serum Adiponectin and Resistin Levels with Breast Cancer Risk*. Retrieved from US National Library of Medicine National Institutes of Health Search databaseSearch term Search : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2693546/

Kee-Hong Kim, L. Z. (2004, 04 27). *Dominant inhibitory adipocyte-specific secretory factor (ADSF)/resistin enhances adipogenesis and improves insulin sensitivity*. Retrieved from US National Library of Medicine National Institutes of Health Search databaseSearch term Search : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC404122/

Li-Yuan Liu, M. W.-B.-X.-Z.-G. (2013, 08 22). *The Role of Adiponectin in Breast Cancer: A Meta-Analysis*. Retrieved from US National Library of Medicine National Institutes of Health: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3749999/

Lunardo, N. (n.d.). *accuracy,meas*. Retrieved from RDocumentation: https://www.rdocumentation.org/packages/ROSE/versions/0.0-3/topics/accuracy.meas

MedicineNet. (2012, 03 19). *Medical Definition of Glucose*. Retrieved from MedicineNet.com: https://www.medicinenet.com/script/main/art.asp?articlekey=3608

MedicineNet. (n.d.). *Medical Definition of Adiponectin*. Retrieved from MedicineNet: https://www.medicinenet.com/script/main/art.asp?articlekey=17982

National Breast Cancer Foundation, INC. (n.d.). *Male Breast Cancer*. Retrieved from National Breast Cancer Foundation, INC.: http://www.nationalbreastcancer.org/male-breast-cancer

Santacruz, A. (2016, 09 20). *amsantac.co Smart Solutions for Geospatial Data Processing*. Retrieved from Why it is important to work with a balance classification datasset: http://amsantac.co/blog/en/2016/09/20/balanced-image-classification-r.html

Sara Suchy, T. G. (2012, 10 15). *A Brief History of Breast Cancer*. Retrieved from healthcentral live bold live now: https://www.healthcentral.com/slideshow/a-brief-history-of-breast-cancer#slide=5

Satish L. Deshmane, S. K. (2009, 06 29). *Monocyte Chemoattractant Protein-1 (MCP-1): An Overview*. Retrieved from US National Library of Medicine National Institutes of Health: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2755091/

Sbarro Health Research Organization (SHRO). (2013, 04 04). *Assesing insulin resistance can inform about breast cancer risk*. Retrieved from Science News: https://www.sciencedaily.com/releases/2013/04/130404135200.htm

Surmacz, E. (2007, 01 26). *Obesity hormone leptin: a new target in breast cancer?* Retrieved from Breast Cancer Research: https://breast-cancer-research.biomedcentral.com/articles/10.1186/bcr1638

Tape, T. G. (n.d.). *The Area Under an ROC Curve*. Retrieved from Interpreting Diagnostic Tests: http://gim.unmc.edu/dxtests/roc3.htm

Tara M. Wallace, M. J. (2004, 02 25). *Use and Abuse of HOMA Modeling*. Retrieved from Diabetes Care: http://care.diabetesjournals.org/content/27/6/1487

Team, A. V. (2016, 03 28). *Practical Guide to deal with Imbalanced Classification Problems in R*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/

*The Facts on Leptin: FAQ*. (n.d.). Retrieved from Katherine Kam: https://www.webmd.com/diet/obesity/features/the-facts-on-leptin-faq#1

The Susan G. Komen Breast Cancer Foundation INC. (2017, 10 18). *The who, what, where, when and sometimes, why.* Retrieved from Susan G. Komen: https://ww5.komen.org/AboutBreastCancer/RiskFactors/FactorsAffectingBreastCancerRisk/FactorsAffectingBreastCancerRisk.html

Trichopoulos, C. M. (2004, 03 01). *Adiponectin and Breast Cancer Risk*. Retrieved from The Journal of Clinical Endocrinology & Metabolism: https://academic.oup.com/jcem/article/89/3/1102/2844088

Vadgama, P. D. (2018, 08 28). *MCP-1 is overexpressed in triple-negative breast cancers and drives cancer invasiveness and metastasis*. Retrieved from SpringerLink: https://link.springer.com/article/10.1007%2Fs10549-018-4760-8

William Faloon, L. H. (2013, 02). *Life Extension Magazine*. Retrieved from Elevated Glucose Increases Incidence of Breast Cancer and Brain Shrinkage: http://www.lifeextension.com/Magazine/2013/2/elevated-glucose-increases-incidence-of-breast-cancer-and-brain-shrinkage/Page-01