Creating a Data Pipeline to transform, process and analyze Wisconsin Breast Cancer Data

University of Wisconsin-Milwaukee

Venkat Hemant Akurati

Running Head: Creating a Data Pipeline to transform, process and analyze Wisconsin Breast Cancer Data

This Capstone Project was approved by

Advisor: _____ Date: _____
Dr. Jake Luo

Advisor: _____ Date: _____
Dr. Min Wu

Table of Contents

Creating a Data Pipeline to transform, process and analyze Wisconsin Breast Cancer Data

Abstract

Background: I present a method to create a Data pipeline utilizing Wisconsin Diagnostic Breast Cancer (WDBC) for predicting potential malignancy in individuals based on their clinical FNA results. The presented methodology may be incorporated in a variety of applications such as risk management, tailored health communication and decision support systems in healthcare.

Methods: I employed the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which is available through UCI repository, to train various machine learning models for breast cancer prediction. Since the WDBC data is categorical in nature, I employed an ensemble learning approach of data cleaning and data pre-processing before feeding the data into machine learning algorithms. I compared the performance of Random forest, K- Nearest Neighbor, and Support vector machine to predict the risk of breast cancer.

Results: I trained 3 different Machine Learning models on the dataset and after comparing the results I observed that Random Forest was able to outperform the other models most of the time to predict breast cancer among the patients with a higher accuracy of 100%.

Keywords: Breast Cancer, Machine Learning Model, Random Forest

Review of Literature

Disease prediction can be applied to different domains such as risk management, tailored health communication and decision support systems. Risk management plays an important role in health insurance companies, mainly in the underwriting process. Health insurers use a process called underwriting in order to classify the applicant as standard or substandard, based on which they compute the policy rate and the premiums individuals have to pay. Currently, in order to classify the applicants, insurers require every applicant to complete a questionnaire, report current medical status and sometimes medical records, or clinical laboratory results, such as blood test, etc. By incorporating machine learning techniques, insurers can make evidence-based decisions and can optimize, validate and refine the rules that govern their business.

Another domain where disease prediction can be applied is tailored health communication. Disease risk prediction along with tailored health communication can lead to an effective channel for delivering disease specific information for people who will be likely to need it. In addition to population level clinical knowledge, de-identified public datasets represent an important resource for the clinical data mining researchers. While full featured clinical records are hard to access due to privacy issues, de-identified large national public dataset are readily available. Although these public datasets don't have all the variables of the original medical records, they still maintain some of their main characteristics such as data imbalance and the use of controlled terminologies (ICD-9 codes).

Introduction

Breast Cancer – Definition, Symptoms and Statistics

Breast cancer in women has increased significantly recent years. Breast cancer can occur in both men and women, but it's far more common in women. After skin cancer, breast cancer is the most common cancer diagnosed in women in the United States. According to WHO reports, nearly 2.1 million women were impacted by breast cancer. In the year 2018, it is estimated that 627,000 women died of breast cancer. The survival rate has been varied greatly worldwide.

Breast cancer is cancer that forms in the cells of the breasts. Symptoms of breast cancer include a lump in the breast, bloody discharge from the nipple, and changes in the shape or texture of the nipple or breast. Treatment depends on the stage of cancer. It may consist of chemotherapy, radiation, and surgery. It is estimated that 42,690 deaths (42,170 women and 520 men) from breast cancer will occur this year.

Table 1. Estimated New DCIS and Invasive Breast Cancer Cases and Deaths among Women by Age, US, 2019

	DCIS cases		Invasive cases		Deaths	
Age	Number	%	Number	%	Number	%
<40	1,180	2%	11,870	4%	1,070	3%
40-49	8,130	17%	37,150	14%	3,250	8%
50-59	12,730	26%	61,560	23%	7,460	18%
60-69	14,460	30%	74,820	28%	9,920	24%
70-79	8,770	18%	52,810	20%	8,910	21%
80+	2,830	6%	30,390	11%	11,150	27%
All ages	**48,100**		**268,600**		**41,760**	

Estimates are rounded to the nearest 10. Percentages may not sum to 100 due to rounding.

Table 2. Age-specific Ten-year Probability of Breast Cancer Diagnosis or Death for US Women

Current age	Diagnosed with invasive breast cancer	Dying from breast cancer
20	0.1% (1 in 1,479)	<0.1% (1 in 18,503)
30	0.5% (1 in 209)	<0.1% (1 in 2,016)
40	1.5% (1 in 65)	0.2% (1 in 645)
50	2.4% (1 in 42)	0.3% (1 in 310)
60	3.5% (1 in 28)	0.5% (1 in 193)
70	4.1% (1 in 25)	0.8% (1 in 132)
80	3.0% (1 in 33)	1.0% (1 in 101)
Lifetime risk	**12.8% (1 in 8)**	**2.6% (1 in 39)**

Note: Probability is among those who have not been previously diagnosed with cancer. Percentages and "1 in" numbers may not be numerically equivalent due to rounding.

Scope of the Project

Several machine learning techniques were applied to healthcare data sets for the prediction of future health care utilization such as predicting individual expenditures and disease risks for patients. The idea behind this project is to leverage Data Science technologies and Machine Learning algorithms to a large national patient sample database to build predictive models which

can be accurately use in the classification of breast cancer metastasis and predict the distinguishing between benign and malignant tumors. These techniques help in avoiding the possible medical errors by the healthcare providers during the time of diagnosis, and most importantly, it helps in saving time for the healthcare providers. The proposed machine learning technique helps in distinguishing between benign and malignant tumors.

Data Description

The UCI Machine Learning Repository is a database developed and maintained by Center for machine learning and intelligent systems, UC Irvine on 1987 by David Aha. This repository consists of huge collection of databases, domain theories, and data generators. These databases are generally used by the data science professionals for analysis of their ML algorithms.

Wisconsin Diagnostic Breast Cancer (WDBC) dataset collected in 8 groups from the year 1989 to 1991 is used in this project. The data set contains about 669 delimited values of patient characteristics and 32 attributes. These features are computed from Fine needle aspirate (FNA) of breast mass which describe the cell nuclei present in the image. Attributes include patient_ID and diagnosis which is divided into 2 classes “Benign” or “Malignant.” For each cell nucleus 10 features are computed based on the FNA results of the lump. Those include radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. There are no missing attribute values. Class distribution of benign and malignant are 458 and 241 respectively. Each attributes has been given a value in the range of 1-10 based on FNA results.

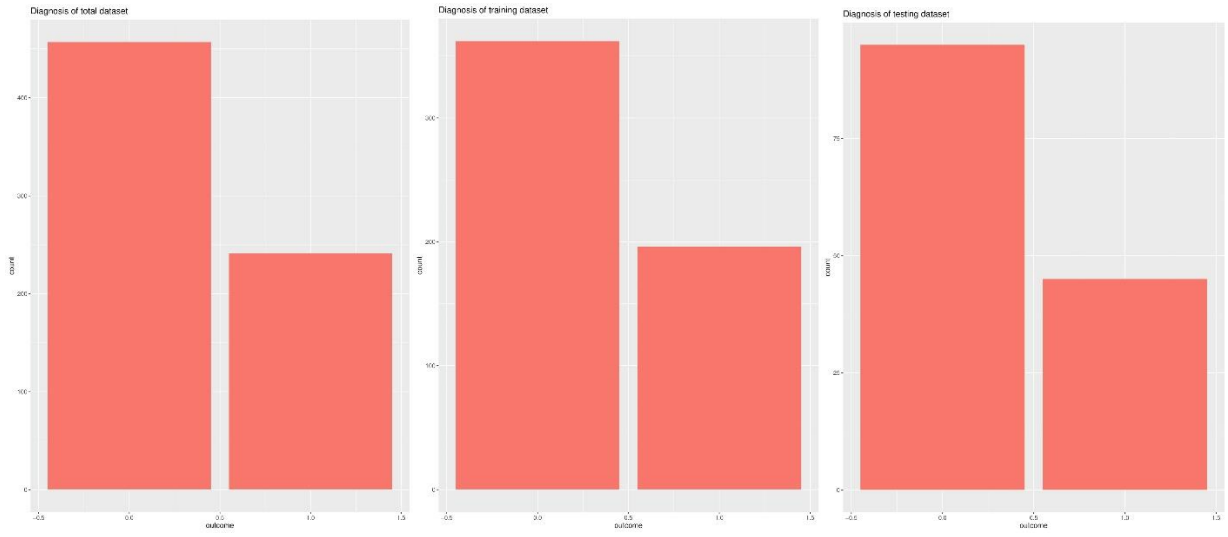
Attribute Information

Sample code number	**Id number**
Clump Thickness	1 – 10

Uniformity of Cell Size	1 – 10
Uniformity of Cell Shape	1 – 10
Marginal Adhesion	1 – 10
Single Epithelial Cell Size	1 – 10
Bare Nuclei	1 – 10
Bland Chromatin	1 – 10
Normal Nucleoli	1 – 10
Mitoses	1 – 10
Class	2 for benign, 4 for malignant

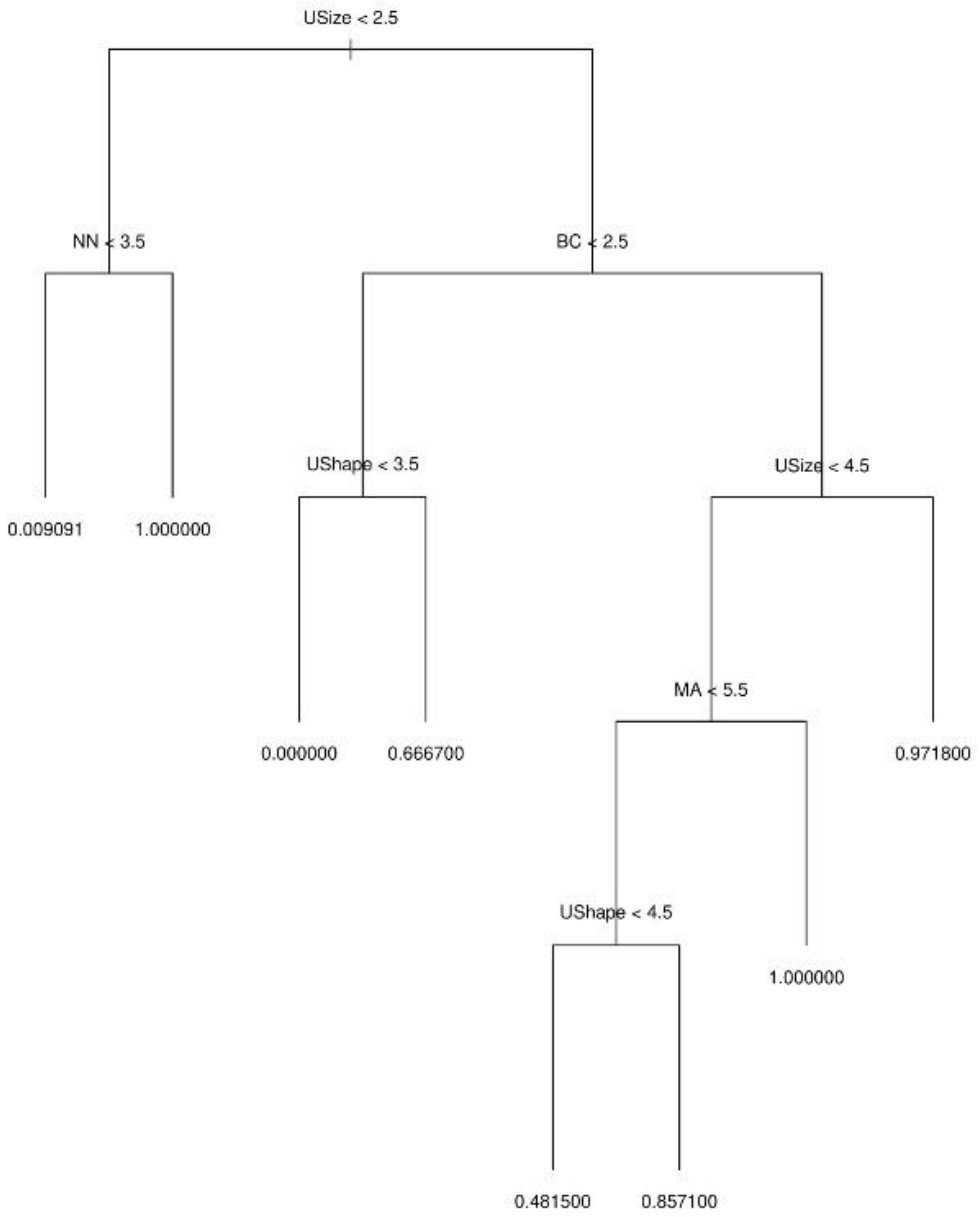
Data Science and ML Modeling

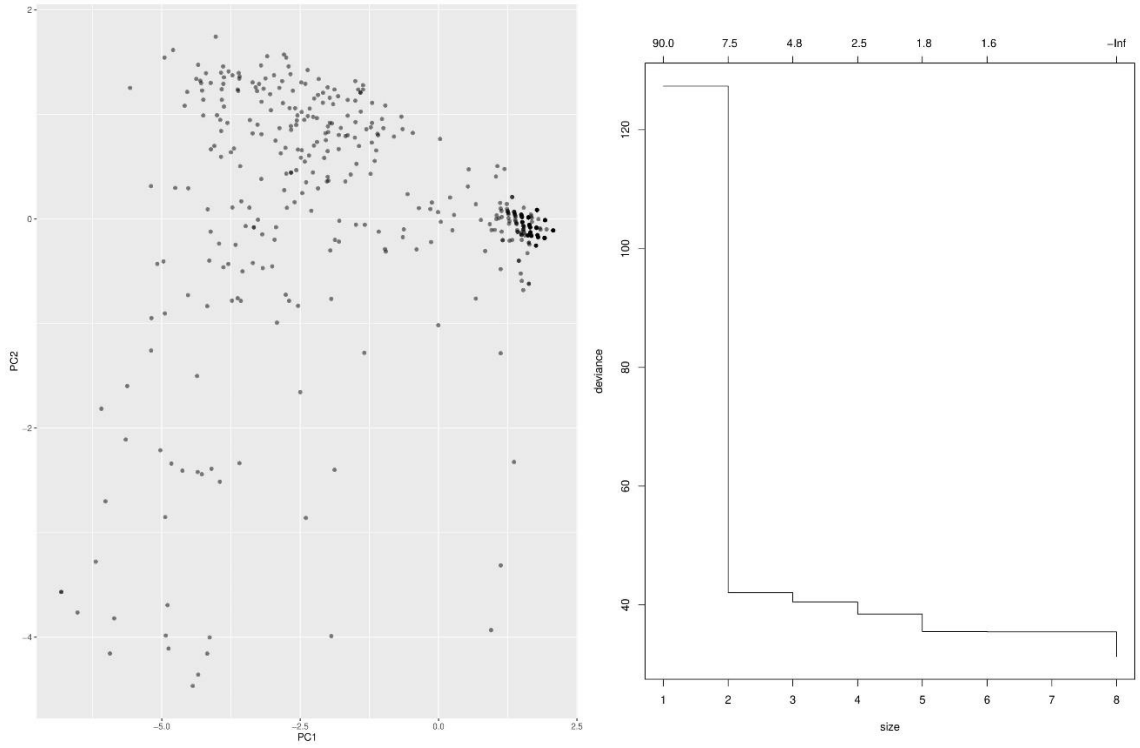
The objective of study is to find an accurate model to predict the incidence of breast cancer based on patients' clinical records. I have employed R programing for the current project to train and test the model. Dataset has been extracted from the source and validated for any missing values. Necessary data cleaning and transformation has been done.



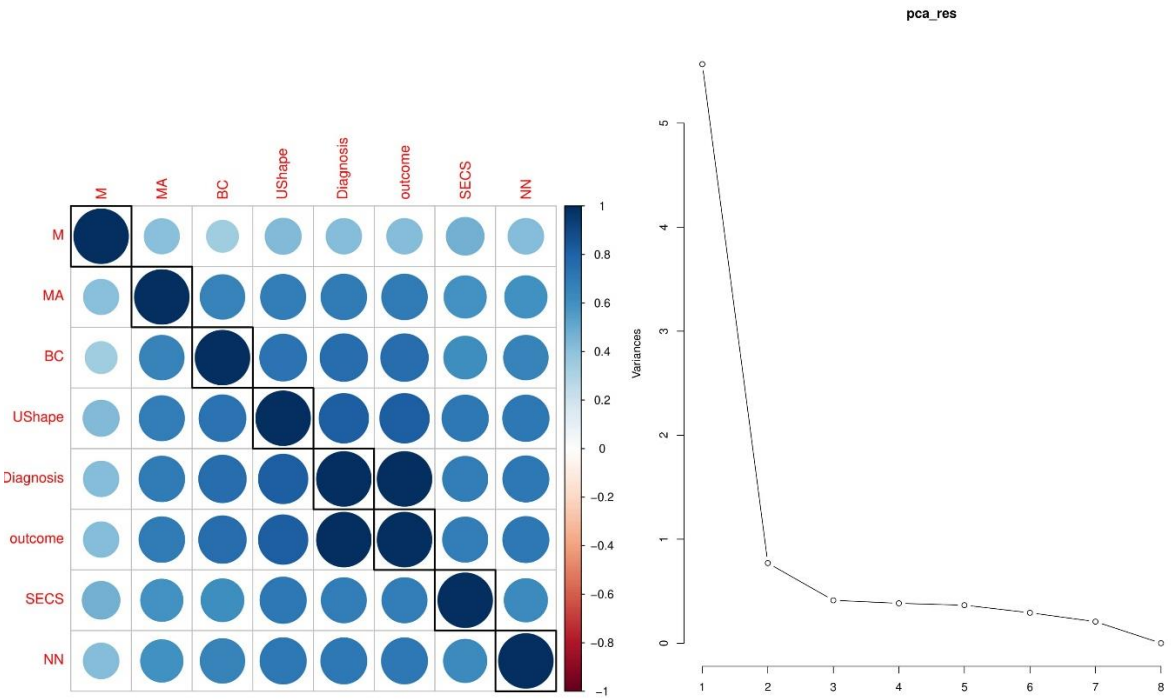
Any missing values are given as NULL. Any co-relations between the features has been thoroughly checked. A new column is added for the disease status in other terms is “0” for Benign and “1” for Malignant. Dataset is loaded and appropriate headers are assigned for the loaded data. Features that are necessary for the analysis has been included and redundant features has been discarded

such as patient personal information including address, phone number, etc. Data is split into training data and testing data in 70:30 ratio and model fitting is performed using generalize linear model (GLM).





Necessary Libraries have been imported such as lattice, plotly, tree, tidyverse, performanceanalysis, etc. Later multiple machine learning algorithms has been used on random samples by setting the seed to random numbers. The accuracy of each model has been calculated and graphs are plotted using ggplot2 package for the most accurate model. Validation is performed using different validation techniques like k-fold cross validation and Leave one out (LOO) validation.



Results

1. Results obtained after random sampling and validation on multiple runs with varying train-test splits.

Data File	**Random Forrest**	**K- Nearest Neighbor**	**SVM**
Wisconsin Diagnostic Breast Cancer	100%	96.4%	96.4%

2. Sample Confusion Matrices

   a. Confusion Matrix for Random Forest

	Normal (Predicted)	Abnormal (Predicted)
Normal (Actual)	95 (TP)	0 (FN)
Abnormal (Actual)	0 (FP)	45 (TN)

Precision = TP/ (TP + FP) = 1

Recall = TP/ (TP + FN) = 1

   b. Confusion Matrix for K- Nearest Neighbor

	Normal (Predicted)	Abnormal (Predicted)
Normal (Actual)	92 (TP)	3 (FN)
Abnormal (Actual)	2 (FP)	43 (TN)

Precision = TP/ (TP + FP) = 0.97

Recall = TP/ (TP + FN) = 0.96

c. Confusion Matrix for SVM

	Normal (Predicted)	Abnormal (Predicted)
Normal (Actual)	93 (TP)	2 (FN)
Abnormal (Actual)	3 (FP)	42 (TN)

Precision = TP/ (TP + FP) = 0.96

Recall = TP/ (TP + FN) = 0.97

Conclusion

In conclusion, the ML methods Random forest, K- Nearest neighbor, and Support vector machine all have performed well in creating and testing a model. However, random forest has given results with 100% accuracy. This model can be used in future predictions to rule out possible diagnosis results. This method developed to validate the important variables that influence the breast cancer survival rate and this type of study also helps to build predictive applications for survival analysis for other disease conditions which helps the healthcare professionals and also save them a lot of time.

REFERENCES

➢ William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

➢ K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).

➢ Ince, M.C. and Karabatak, M. (2009) An expert system for detection of breast cancer based on association rules and neural network. Expert Systems with Applications, 36, 3465-3469. doi:10.1016/j.eswa.2008.02.064

➢ Lieber, J., and Bresson, B., Case-based reasoning for breast cancer treatment decision helping. Proceedings of the 5th European Workshop on Case-Based Reasoning, pp. 173–185, (2000).

➢ Liaw, A. and Watthew, M. (2002) Classification and regression by random forest. R News, 3, 18-22.