University of Wisconsin - Milwaukee
Milwaukee, Wisconsin
Graduate School

# Create a Big Data Pipeline to transform, process and analyze a large National Patient Database

A Project Submitted in Partial Fulfillment of the Requirements of the Master of Science in Computer Science – Professional Track

Jay Tank
May 2020
Student ID – 991343373

This Capstone Project was approved by

Advisor: ______________________________________________ Date: ____________
Dr. Jake Luo

Advisor: ______________________________________________ Date: ____________
Dr. Christine Cheng

# Acknowledgements

I would like to take this time to express my sincere gratitude to my advisors Dr. Jake Luo and Dr. Christine Cheng for providing their invaluable guidance, comments and suggestions throughout the course of this project. I would especially thank Dr. Jake Luo for providing me with NIS data samples, helping me to understand and utilize tools/resources required for the completion of this project. I would also like to thank Dr. Christine Cheng for constantly motivating me to work harder and supporting me during difficult times in the journey of completing my master's degree. This project would have not been possible without their help and insight.

I want to thank my parents for being great role models in work hard and perseverance can accomplish. At last but not the least, I am thankful to all my teachers and friends who have always been supporting and encouraging me throughout the year. The journey in doing this project and completing my degree has been a great learning experience and one I take pride in accomplishing.

# Table of Contents

# Abstract

**Background:** I present a method to create a Big Data pipeline utilizing Healthcare Cost and Utilization Project (HCUP) dataset for predicting disease risk of individuals based on their medical diagnosis history. 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 National Inpatient Sample (NIS) data, which is available through Healthcare Cost and Utilization Project (HCUP), to train various machine learning models for breast cancer prediction. Since the HCUP data is categorical in nature and highly imbalanced, 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 Logistic Regression, Decision Tree and Naive Bayes Classifier 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 Decision tree was able to outperform the other models 97% of the time to predict breast cancer among the patients with a higher accuracy.

**Conclusion:** I was able to overcome the class imbalance problem and achieve promising results. Using the national HCUP data set and leveraging the data science technologies I was successfully able to build a predictive model which gives an average AUC of 85%.

# Review of Literature

The reporting requirements of various US governmental agencies such as Center for Disease Control (CDC), Agency for Health Care Quality (AHRQ) and US Department of Health and Human Services Center for Medicare Services (CMS) have created huge public datasets that, I believe, are not utilized to their full potential. For example, CDC https://www.cdc.gov/ makes available National Health and Nutrition Examination Survey (NHANES) data which can be used to predict diabetes risk. CMS https://www.cms.gov/ uses the Medicare and Medicaid claims to create the minimum dataset (MDS). Herbert and others use MDS data to identify people with diabetes. For this project, I use the National Inpatient Sample (NIS) data created by AHRQ http://www.ahrq.gov Healthcare Utilization Project (HCUP), which captures eight chronic diseases  to predict the risk for breast cancer.

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, deidentified 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, deidentified large national public dataset are readily available [6]. 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 is cancer that forms in the cells of the breasts. After skin cancer, breast cancer is the most common cancer diagnosed in women in the United States. Breast cancer can occur in both men and women, but it's far more common in women. 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.

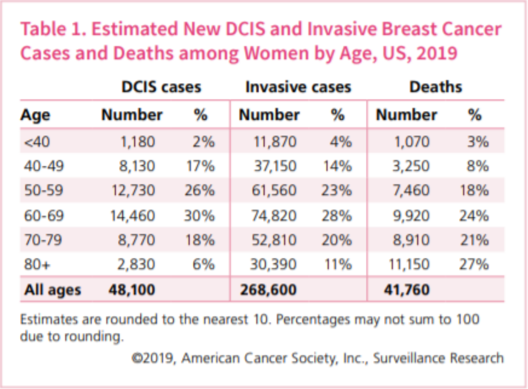


Table 1. Estimated New DCIS and Invasive Breast Cancer Cases and Deaths among Women by Age, US, 2019

| | 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.

©2019, American Cancer Society, Inc., Surveillance Research

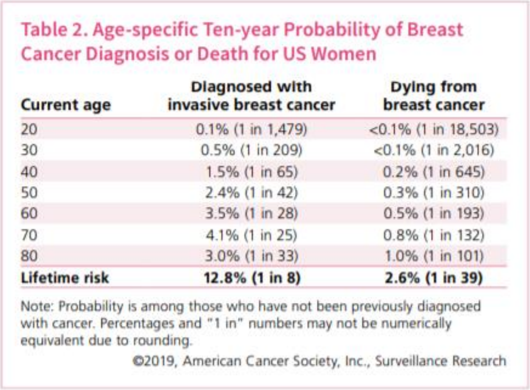


Table 2. Age-specific Ten-year Probability of Breast Cancer Diagnosis or Death for US Women

| 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.

©2019, American Cancer Society, Inc., Surveillance Research

## 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 risk of breast cancer.
After loading dataset into a database like PostgreSQL, I performed data cleaning to remove co-relations between the features and trained different Machine Learning models on the random sub-samples in order to obtain the best working model for the data.

# Data Description

## Data Sources

The Nationwide Inpatient Sample (NIS) is a database of hospital inpatient admissions that dates back to 1988 and is used to identify, track, and analyze national trends in health care utilization, access, charges, quality, and outcomes. The NIS database is developed by the Healthcare Cost and Utilization Project (HCUP) and sponsored by the Agency for Healthcare Research and Quality (AHRQ). The NIS data contains discharge level information on all inpatients from a 20% stratified sample of hospitals across the United States, representing approximately 90% of all hospitals in the country. HCUP data from the year 2012 to 2016 are used in this project. The data set contains about 8 million records of hospital stays, with 126 clinical and nonclinical data elements for each visit. Nonclinical elements include patient demographics, hospital identification, admission date, zip code, calendar year, total charges and length of stay. Clinical elements include procedures, procedure categories, diagnosis codes and diagnosis categories. Every record contains a vector of 15 diagnosis codes. The diagnosis codes are represented using the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). The International Statistical Classification of Disease is designed and published by the World Health Organization (WHO). The ICD-9 codes are alphanumeric codes, 3-5 characters long and used by hospitals, insurance companies and other facilities to describe health conditions of the patient. Every code represents a disease, condition, symptom, or cause of death. There are numerous codes, over 14,000 ICD-9 codes and 3,900 procedures codes. Every ICD-9 code has a corresponding diagnosis category and every category contains a set of ICD-9 codes. Demographics such as age, race and sex are also included in the data set. The data set is highly imbalanced.

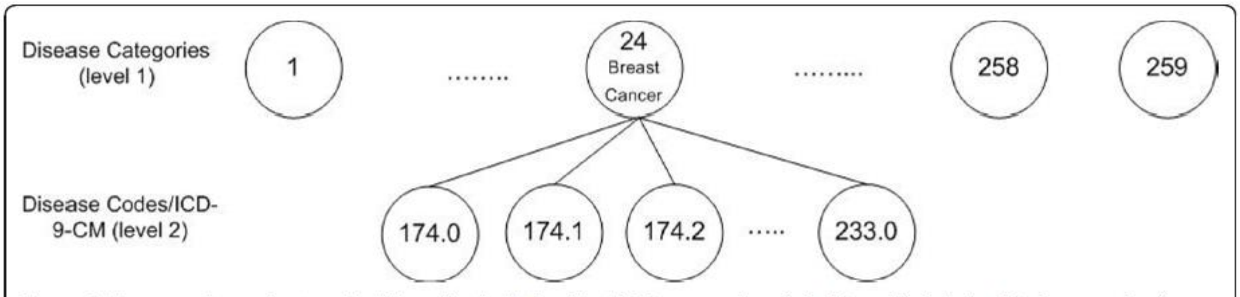


**Figure 1 Disease codes and categories hierarchical relationship**. This is a snap shot of the hierarchical relationship between the diseases and disease categories. For instance, disease category 49 (diabetes) has a children that are represented in disease codes (ICD-9-CM).

PLEASE NOTE: data files after year 2016 contains ICD- 10 codes for disease diagnosis categories.

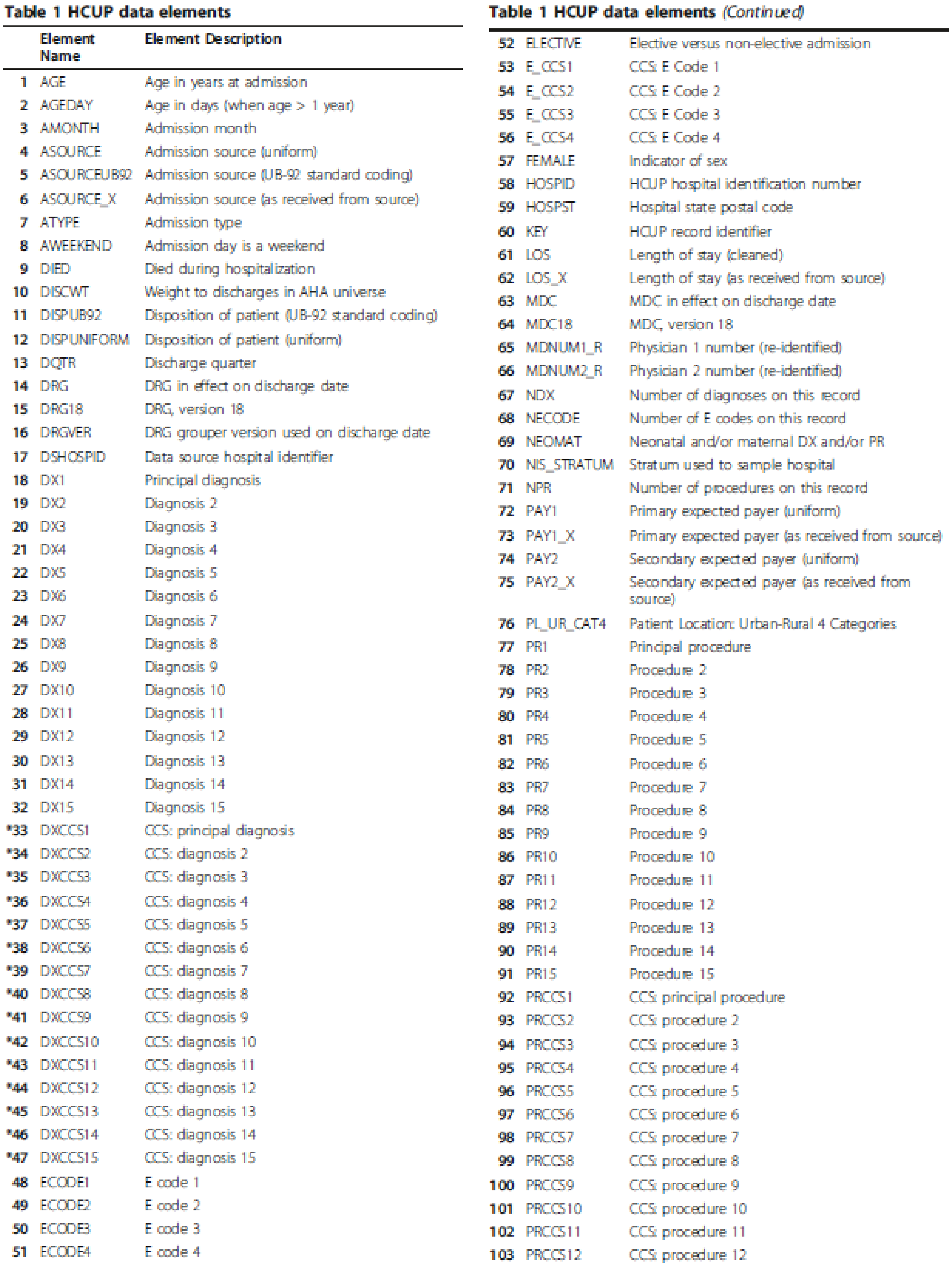


**Table 1 HCUP data elements**

| | Element Name | Element Description |
|---|---|---|
| 1 | AGE | Age in years at admission |
| 2 | AGEDAY | Age in days (when age > 1 year) |
| 3 | AMONTH | Admission month |
| 4 | ASOURCE | Admission source (uniform) |
| 5 | ASOURCEUB92 | Admission source (UB-92 standard coding) |
| 6 | ASOURCE_X | Admission source (as received from source) |
| 7 | ATYPE | Admission type |
| 8 | AWEEKEND | Admission day is a weekend |
| 9 | DIED | Died during hospitalization |
| 10 | DISCWT | Weight to discharges in AHA universe |
| 11 | DISPUB92 | Disposition of patient (UB-92 standard coding) |
| 12 | DISPUNIFORM | Disposition of patient (uniform) |
| 13 | DQTR | Discharge quarter |
| 14 | DRG | DRG in effect on discharge date |
| 15 | DRG18 | DRG, version 18 |
| 16 | DRGVER | DRG grouper version used on discharge date |
| 17 | DSHOSPID | Data source hospital identifier |
| 18 | DX1 | Principal diagnosis |
| 19 | DX2 | Diagnosis 2 |
| 20 | DX3 | Diagnosis 3 |
| 21 | DX4 | Diagnosis 4 |
| 22 | DX5 | Diagnosis 5 |
| 23 | DX6 | Diagnosis 6 |
| 24 | DX7 | Diagnosis 7 |
| 25 | DX8 | Diagnosis 8 |
| 26 | DX9 | Diagnosis 9 |
| 27 | DX10 | Diagnosis 10 |
| 28 | DX11 | Diagnosis 11 |
| 29 | DX12 | Diagnosis 12 |
| 30 | DX13 | Diagnosis 13 |
| 31 | DX14 | Diagnosis 14 |
| 32 | DX15 | Diagnosis 15 |
| *33 | DXCCS1 | CCS: principal diagnosis |
| *34 | DXCCS2 | CCS: diagnosis 2 |
| *35 | DXCCS3 | CCS: diagnosis 3 |
| *36 | DXCCS4 | CCS: diagnosis 4 |
| *37 | DXCCS5 | CCS: diagnosis 5 |
| *38 | DXCCS6 | CCS: diagnosis 6 |
| *39 | DXCCS7 | CCS: diagnosis 7 |
| *40 | DXCCS8 | CCS: diagnosis 8 |
| *41 | DXCCS9 | CCS: diagnosis 9 |
| *42 | DXCCS10 | CCS: diagnosis 10 |
| *43 | DXCCS11 | CCS: diagnosis 11 |
| *44 | DXCCS12 | CCS: diagnosis 12 |
| *45 | DXCCS13 | CCS: diagnosis 13 |
| *46 | DXCCS14 | CCS: diagnosis 14 |
| *47 | DXCCS15 | CCS: diagnosis 15 |
| 48 | ECODE1 | E code 1 |
| 49 | ECODE2 | E code 2 |
| 50 | ECODE3 | E code 3 |
| 51 | ECODE4 | E code 4 |
| 52 | ELECTIVE | Elective versus non-elective admission |
| 53 | E_CCS1 | CCS: E Code 1 |
| 54 | E_CCS2 | CCS: E Code 2 |
| 55 | E_CCS3 | CCS: E Code 3 |
| 56 | E_CCS4 | CCS: E Code 4 |
| 57 | FEMALE | Indicator of sex |
| 58 | HOSPID | HCUP hospital identification number |
| 59 | HOSPST | Hospital state postal code |
| 60 | KEY | HCUP record identifier |
| 61 | LOS | Length of stay (cleaned) |
| 62 | LOS_X | Length of stay (as received from source) |
| 63 | MDC | MDC in effect on discharge date |
| 64 | MDC18 | MDC, version 18 |
| 65 | MDNUM1_R | Physician 1 number (re-identified) |
| 66 | MDNUM2_R | Physician 2 number (re-identified) |
| 67 | NDX | Number of diagnoses on this record |
| 68 | NECODE | Number of E codes on this record |
| 69 | NEOMAT | Neonatal and/or maternal DX and/or PR |
| 70 | NIS_STRATUM | Stratum used to sample hospital |
| 71 | NPR | Number of procedures on this record |
| 72 | PAY1 | Primary expected payer (uniform) |
| 73 | PAY1_X | Primary expected payer (as received from source) |
| 74 | PAY2 | Secondary expected payer (uniform) |
| 75 | PAY2_X | Secondary expected payer (as received from source) |
| 76 | PL_UR_CAT4 | Patient Location: Urban-Rural 4 Categories |
| 77 | PR1 | Principal procedure |
| 78 | PR2 | Procedure 2 |
| 79 | PR3 | Procedure 3 |
| 80 | PR4 | Procedure 4 |
| 81 | PR5 | Procedure 5 |
| 82 | PR6 | Procedure 6 |
| 83 | PR7 | Procedure 7 |
| 84 | PR8 | Procedure 8 |
| 85 | PR9 | Procedure 9 |
| 86 | PR10 | Procedure 10 |
| 87 | PR11 | Procedure 11 |
| 88 | PR12 | Procedure 12 |
| 89 | PR13 | Procedure 13 |
| 90 | PR14 | Procedure 14 |
| 91 | PR15 | Procedure 15 |
| 92 | PRCCS1 | CCS: principal procedure |
| 93 | PRCCS2 | CCS: procedure 2 |
| 94 | PRCCS3 | CCS: procedure 3 |
| 95 | PRCCS4 | CCS: procedure 4 |
| 96 | PRCCS5 | CCS: procedure 5 |
| 97 | PRCCS6 | CCS: procedure 6 |
| 98 | PRCCS7 | CCS: procedure 7 |
| 99 | PRCCS8 | CCS: procedure 8 |
| 100 | PRCCS9 | CCS: procedure 9 |
| 101 | PRCCS10 | CCS: procedure 10 |
| 102 | PRCCS11 | CCS: procedure 11 |
| 103 | PRCCS12 | CCS: procedure 12 |

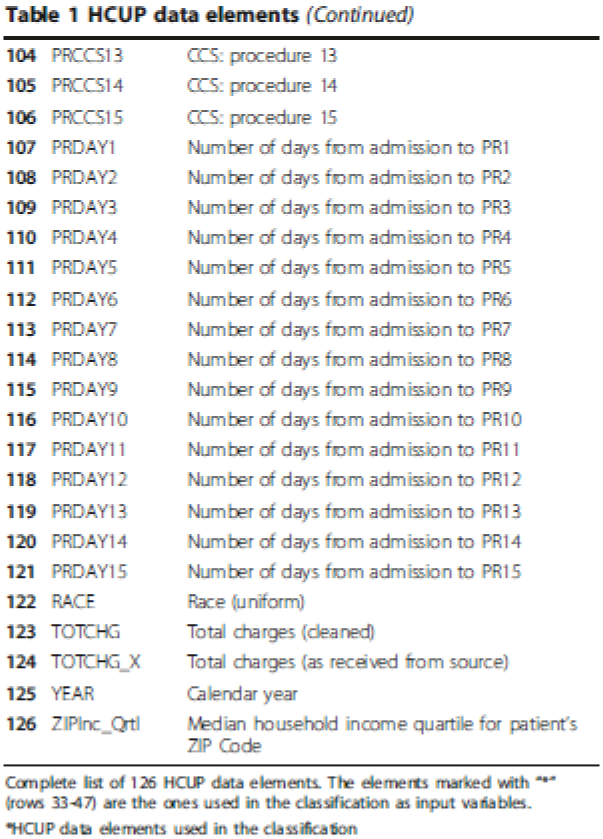
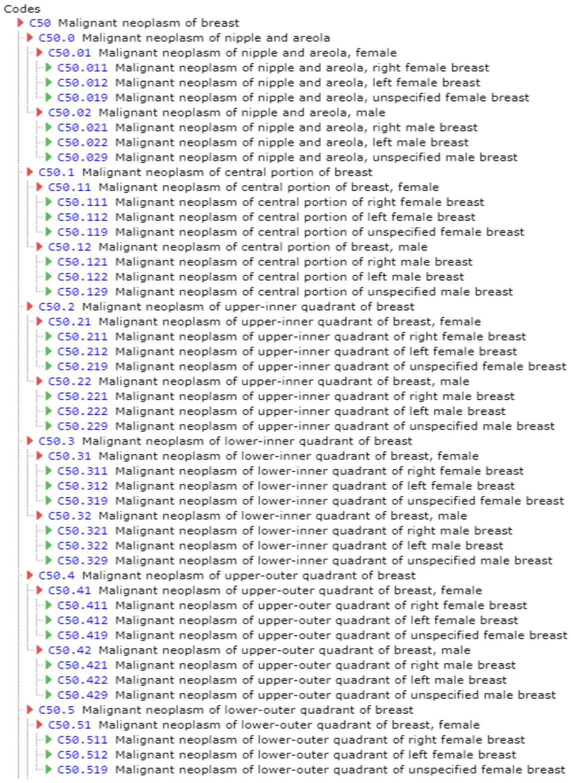


**Table 1 HCUP data elements** *(Continued)*

| | | |
|---|---|---|
| **104** | PRCCS13 | CCS: procedure 13 |
| **105** | PRCCS14 | CCS: procedure 14 |
| **106** | PRCCS15 | CCS: procedure 15 |
| **107** | PRDAY1 | Number of days from admission to PR1 |
| **108** | PRDAY2 | Number of days from admission to PR2 |
| **109** | PRDAY3 | Number of days from admission to PR3 |
| **110** | PRDAY4 | Number of days from admission to PR4 |
| **111** | PRDAY5 | Number of days from admission to PR5 |
| **112** | PRDAY6 | Number of days from admission to PR6 |
| **113** | PRDAY7 | Number of days from admission to PR7 |
| **114** | PRDAY8 | Number of days from admission to PR8 |
| **115** | PRDAY9 | Number of days from admission to PR9 |
| **116** | PRDAY10 | Number of days from admission to PR10 |
| **117** | PRDAY11 | Number of days from admission to PR11 |
| **118** | PRDAY12 | Number of days from admission to PR12 |
| **119** | PRDAY13 | Number of days from admission to PR13 |
| **120** | PRDAY14 | Number of days from admission to PR14 |
| **121** | PRDAY15 | Number of days from admission to PR15 |
| **122** | RACE | Race (uniform) |
| **123** | TOTCHG | Total charges (cleaned) |
| **124** | TOTCHG_X | Total charges (as received from source) |
| **125** | YEAR | Calendar year |
| **126** | ZIPInc_Qrtl | Median household income quartile for patient's ZIP Code |

Complete list of 126 HCUP data elements. The elements marked with "*" (rows 33-47) are the ones used in the classification as input variables.
*HCUP data elements used in the classification



```
Codes
C50 Malignant neoplasm of breast
  C50.0 Malignant neoplasm of nipple and areola
    C50.01 Malignant neoplasm of nipple and areola, female
      C50.011 Malignant neoplasm of nipple and areola, right female breast
      C50.012 Malignant neoplasm of nipple and areola, left female breast
      C50.019 Malignant neoplasm of nipple and areola, unspecified female breast
    C50.02 Malignant neoplasm of nipple and areola, male
      C50.021 Malignant neoplasm of nipple and areola, right male breast
      C50.022 Malignant neoplasm of nipple and areola, left male breast
      C50.029 Malignant neoplasm of nipple and areola, unspecified male breast
  C50.1 Malignant neoplasm of central portion of breast
    C50.11 Malignant neoplasm of central portion of breast, female
      C50.111 Malignant neoplasm of central portion of right female breast
      C50.112 Malignant neoplasm of central portion of left female breast
      C50.119 Malignant neoplasm of central portion of unspecified female breast
    C50.12 Malignant neoplasm of central portion of breast, male
      C50.121 Malignant neoplasm of central portion of right male breast
      C50.122 Malignant neoplasm of central portion of left male breast
      C50.129 Malignant neoplasm of central portion of unspecified male breast
  C50.2 Malignant neoplasm of upper-inner quadrant of breast
    C50.21 Malignant neoplasm of upper-inner quadrant of breast, female
      C50.211 Malignant neoplasm of upper-inner quadrant of right female breast
      C50.212 Malignant neoplasm of upper-inner quadrant of left female breast
      C50.219 Malignant neoplasm of upper-inner quadrant of unspecified female breast
    C50.22 Malignant neoplasm of upper-inner quadrant of breast, male
      C50.221 Malignant neoplasm of upper-inner quadrant of right male breast
      C50.222 Malignant neoplasm of upper-inner quadrant of left male breast
      C50.229 Malignant neoplasm of upper-inner quadrant of unspecified male breast
  C50.3 Malignant neoplasm of lower-inner quadrant of breast
    C50.31 Malignant neoplasm of lower-inner quadrant of breast, female
      C50.311 Malignant neoplasm of lower-inner quadrant of right female breast
      C50.312 Malignant neoplasm of lower-inner quadrant of left female breast
      C50.319 Malignant neoplasm of lower-inner quadrant of unspecified female breast
    C50.32 Malignant neoplasm of lower-inner quadrant of breast, male
      C50.321 Malignant neoplasm of lower-inner quadrant of right male breast
      C50.322 Malignant neoplasm of lower-inner quadrant of left male breast
      C50.329 Malignant neoplasm of lower-inner quadrant of unspecified male breast
  C50.4 Malignant neoplasm of upper-outer quadrant of breast
    C50.41 Malignant neoplasm of upper-outer quadrant of breast, female
      C50.411 Malignant neoplasm of upper-outer quadrant of right female breast
      C50.412 Malignant neoplasm of upper-outer quadrant of left female breast
      C50.419 Malignant neoplasm of upper-outer quadrant of unspecified female breast
    C50.42 Malignant neoplasm of upper-outer quadrant of breast, male
      C50.421 Malignant neoplasm of upper-outer quadrant of right male breast
      C50.422 Malignant neoplasm of upper-outer quadrant of left male breast
      C50.429 Malignant neoplasm of upper-outer quadrant of unspecified male breast
  C50.5 Malignant neoplasm of lower-outer quadrant of breast
    C50.51 Malignant neoplasm of lower-outer quadrant of breast, female
      C50.511 Malignant neoplasm of lower-outer quadrant of right female breast
      C50.512 Malignant neoplasm of lower-outer quadrant of left female breast
      C50.519 Malignant neoplasm of lower-outer quadrant of unspecified female breast
```

## Data Pre-Processing

The data set was provided in a large ASCII files, each containing approximately 7 million records. All the required data files are named as "**Core**" files. Along with the core files, there are "**Hospital Weights File"** and "**Severity Measures File"** which contains and several other features.

The first step was to parse the data set using "**File Specification**" files. These files specify the starting column and the ending column in the ASCII file for each data element (length of data element). Load all the data into PostgreSQL database dynamically. The running time for dynamic loading takes around 48 hours to complete.

The second step is to update all the tables in database by adding extra column called outcome which contains the disease code for breast cancer from every record. If value is not found, then insert null.

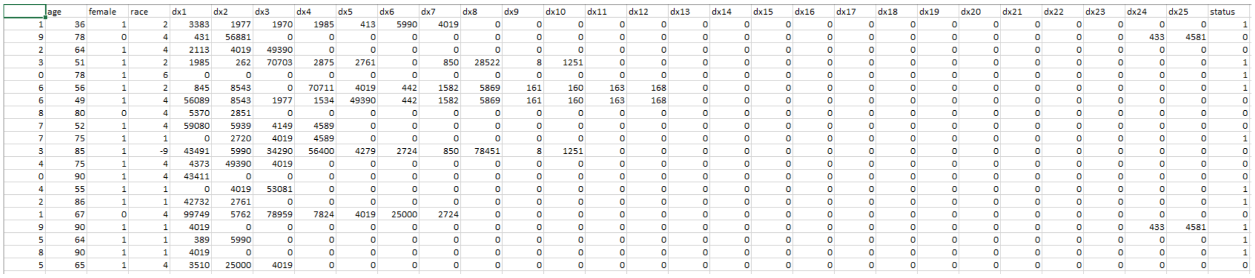
The third step in to randomly select N records and extract a set of relevant features to pass to machine learning algorithm. Every record is a sequence of characters that are not delimited.

# Data Science and Machine Learning Modelling

The final step is to perform any required data cleaning like converting the data types using type casting, performing label encoding for categorical features, fill all the null values, checking for any co-relations between the features before passing the data to machine learning model.

## Feature Selection

For every record, I extracted the age, gender, race and 25 -30 diagnosis categories. I denote the samples that contain a given disease category as "active" and the remaining ones as "inactive". The active and inactive data samples are defined only from the point of view of the disease being classified. We cannot include the features which are likely to have a co-relation with our desired output A snippet of **"pandas dataframe"** is shown below showing active and inactive samples



| | age | female | race | dx1 | dx2 | dx3 | dx4 | dx5 | dx6 | dx7 | dx8 | dx9 | dx10 | dx11 | dx12 | dx13 | dx14 | dx15 | dx16 | dx17 | dx18 | dx19 | dx20 | dx21 | dx22 | dx23 | dx24 | dx25 | status |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 36 | 1 | 2 | 3383 | 1977 | 1970 | 1985 | 413 | 5990 | 4019 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 9 | 78 | 0 | 4 | 431 | 56881 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 433 | 4581 | 0 |
| 2 | 64 | 1 | 4 | 2113 | 4019 | 49390 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 51 | 1 | 2 | 1985 | 262 | 70703 | 2875 | 2761 | 0 | 850 | 28522 | 8 | 1251 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 78 | 1 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 6 | 56 | 1 | 2 | 845 | 8543 | 0 | 70711 | 4019 | 442 | 1582 | 5869 | 161 | 160 | 163 | 168 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 6 | 49 | 1 | 4 | 56089 | 8543 | 1977 | 1534 | 49390 | 442 | 1582 | 5869 | 161 | 160 | 163 | 168 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 80 | 0 | 4 | 5370 | 2851 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 52 | 1 | 4 | 59080 | 5939 | 4149 | 4589 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 75 | 1 | 1 | 0 | 2720 | 4019 | 4589 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 3 | 85 | 1 | -9 | 43491 | 5990 | 34290 | 56400 | 4279 | 2724 | 850 | 78451 | 8 | 1251 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 75 | 1 | 4 | 4373 | 49390 | 4019 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 90 | 1 | 4 | 43411 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 55 | 1 | 1 | 0 | 4019 | 53081 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 86 | 1 | 1 | 42732 | 2761 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 67 | 0 | 4 | 99749 | 5762 | 78959 | 7824 | 4019 | 25000 | 2724 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 90 | 1 | 1 | 4019 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 433 | 4581 | 1 |
| 5 | 64 | 1 | 1 | 389 | 5990 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 8 | 90 | 1 | 1 | 4019 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 5 | 65 | 1 | 4 | 3510 | 25000 | 4019 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

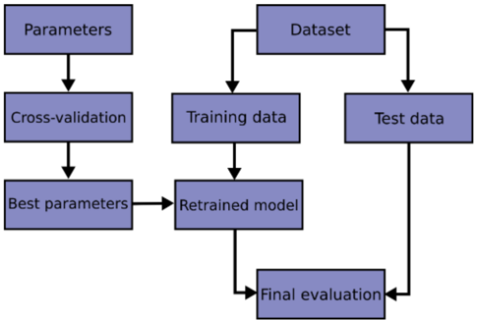
## Learning from Imbalanced Data

A data set is class-imbalanced if one class contains significantly more samples than the other. In such cases, it is challenging to create an appropriate testing and training data sets, given that most classifiers are built with the assumption that the test data is drawn from the same distribution as the training data. Presenting imbalanced data to a classifier will produce undesirable results such as a much lower performance on the testing that on the training data.

## Machine Learning Modelling and Validation

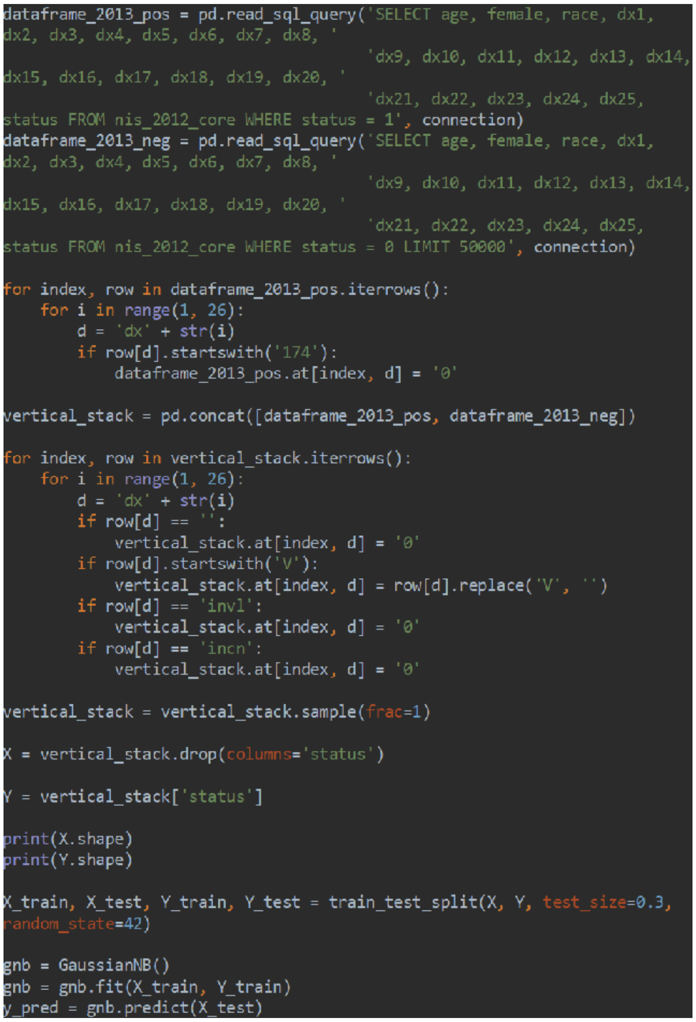
For this project, I initially split my data into testing data and training data in the ratio 70:30 and performed multiple runs with random samples for Logistic regression, Decision Tree and Naïve Bayes models to observe the accuracy and compare the results obtained

For eg: data file **'nis_2014_core'** contains a total of 7071762 records from which 34901 are active samples. For the project requirements, I took all the active samples and 50,000 inactive samples. After shuffling all the records, I split the data into 70% training data and 30% testing data performed multiple runs to record the observations. I validated the observations by splitting the data into 80% training data and 20% testing data. I also performed validation using different validation techniques like k-fold cross validation and Leave one out (LOO) validation.



*A flowchart of typical cross validation workflow in model training*

Below is a code snippet for Naïve Bayes Classifier performed on **'nis_2013_core'** which has 28 features and the data is split into70 : 30 (train : test).



```
dataframe_2013_pos = pd.read_sql_query('SELECT age, female, race, dx1,
dx2, dx3, dx4, dx5, dx6, dx7, dx8, '
                                       'dx9, dx10, dx11, dx12, dx13, dx14,
dx15, dx16, dx17, dx18, dx19, dx20, '
                                       'dx21, dx22, dx23, dx24, dx25,
status FROM nis_2012_core WHERE status = 1', connection)
dataframe_2013_neg = pd.read_sql_query('SELECT age, female, race, dx1,
dx2, dx3, dx4, dx5, dx6, dx7, dx8, '
                                       'dx9, dx10, dx11, dx12, dx13, dx14,
dx15, dx16, dx17, dx18, dx19, dx20, '
                                       'dx21, dx22, dx23, dx24, dx25,
status FROM nis_2012_core WHERE status = 0 LIMIT 50000', connection)

for index, row in dataframe_2013_pos.iterrows():
    for i in range(1, 26):
        d = 'dx' + str(i)
        if row[d].startswith('174'):
            dataframe_2013_pos.at[index, d] = '0'

vertical_stack = pd.concat([dataframe_2013_pos, dataframe_2013_neg])

for index, row in vertical_stack.iterrows():
    for i in range(1, 26):
        d = 'dx' + str(i)
        if row[d] == '':
            vertical_stack.at[index, d] = '0'
        if row[d].startswith('V'):
            vertical_stack.at[index, d] = row[d].replace('V', '')
        if row[d] == 'invl':
            vertical_stack.at[index, d] = '0'
        if row[d] == 'incn':
            vertical_stack.at[index, d] = '0'

vertical_stack = vertical_stack.sample(frac=1)

X = vertical_stack.drop(columns='status')

Y = vertical_stack['status']

print(X.shape)
print(Y.shape)

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.3,
random_state=42)

gnb = GaussianNB()
gnb = gnb.fit(X_train, Y_train)
y_pred = gnb.predict(X_test)
```

# Results

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

| Data File | Logistic Regression | Decision Tree | Naïve Bayes |
|---|---|---|---|
| nis_2012_core | 75.50% | 82.63% | 48.02% |
| nis_2013_core | 78.50% | 78.63% | 46.77% |
| nis_2014_core | 72.83% | 84.16% | 50.42% |
| nis_2015_core | No relevant features | No relevant features | No relevant features |
| nis_2016_core | 73.76% | 87.88% | 55.24% |

2. Sample Confusion Matrices (for data file **'nis_2013_core'**)

   a. Decision Tree Classifier gives us the following confusion matrix

| | Actual Values | |
|---|---|---|
| Predicted Values | TP = 9990 | FP = 4980 |
| | FN = 203 | TN = 10549 |

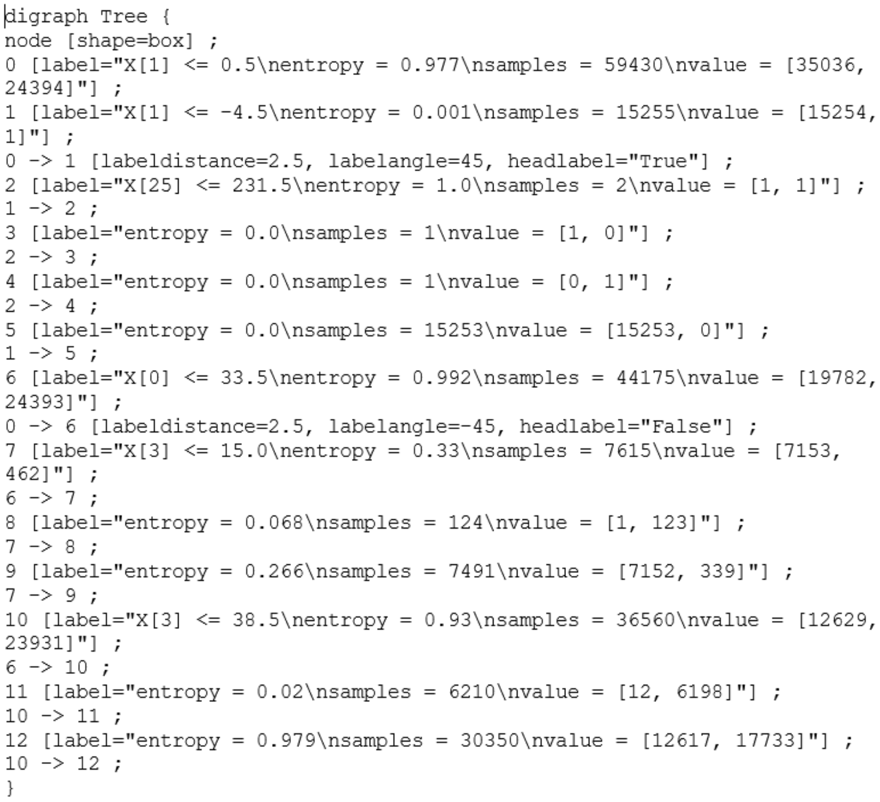
Precision = TP/ (TP + FP) = 0.67
Recall = TP/ (TP + FN) = 0.49

   b. Naïve Bayes gives us following confusion matrix

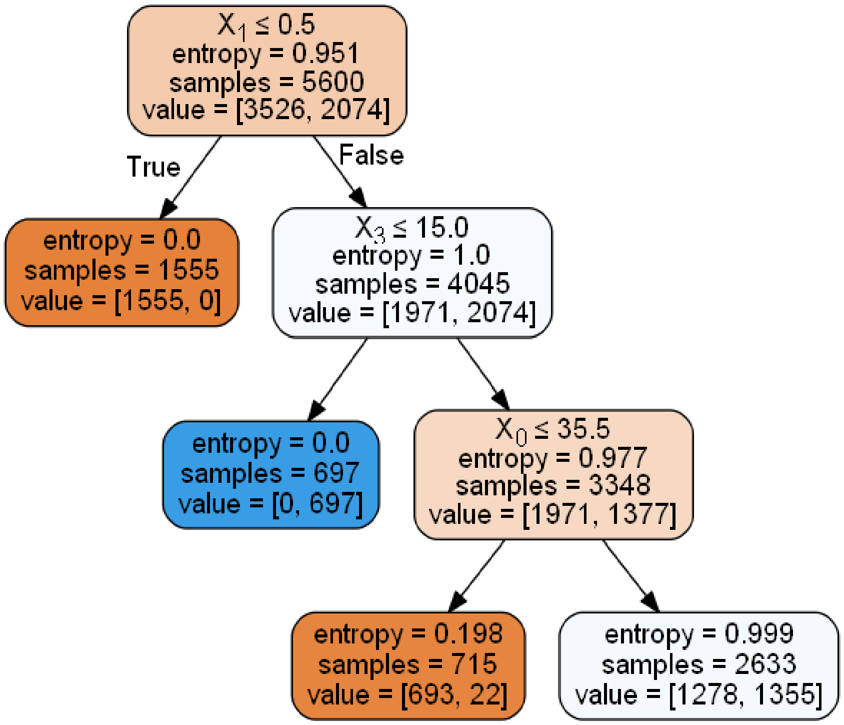
| | Actual Values | |
|---|---|---|
| Predicted Values | TP = 1611 | FP = 28 |
| | FN = 553 | TN = 10060 |

Precision = TP/ (TP + FP) = 0.98
Recall = TP/ (TP + FN) = 0.14

After rigorous training the models with random sampling and validation, I observed that Decision Tree classifier gives the best accuracy as compared to other models. Please find the sample outputs below for Decision Tree Classifier run on **'nis_2014_core'**



```
digraph Tree {
node [shape=box] ;
0 [label="X[1] <= 0.5\nentropy = 0.977\nsamples = 59430\nvalue = [35036,
24394]"] ;
1 [label="X[1] <= -4.5\nentropy = 0.001\nsamples = 15255\nvalue = [15254,
1]"] ;
0 -> 1 [labeldistance=2.5, labelangle=45, headlabel="True"] ;
2 [label="X[25] <= 231.5\nentropy = 1.0\nsamples = 2\nvalue = [1, 1]"] ;
1 -> 2 ;
3 [label="entropy = 0.0\nsamples = 1\nvalue = [1, 0]"] ;
2 -> 3 ;
4 [label="entropy = 0.0\nsamples = 1\nvalue = [0, 1]"] ;
2 -> 4 ;
5 [label="entropy = 0.0\nsamples = 15253\nvalue = [15253, 0]"] ;
1 -> 5 ;
6 [label="X[0] <= 33.5\nentropy = 0.992\nsamples = 44175\nvalue = [19782,
24393]"] ;
0 -> 6 [labeldistance=2.5, labelangle=-45, headlabel="False"] ;
7 [label="X[3] <= 15.0\nentropy = 0.33\nsamples = 7615\nvalue = [7153,
462]"] ;
6 -> 7 ;
8 [label="entropy = 0.068\nsamples = 124\nvalue = [1, 123]"] ;
7 -> 8 ;
9 [label="entropy = 0.266\nsamples = 7491\nvalue = [7152, 339]"] ;
7 -> 9 ;
10 [label="X[3] <= 38.5\nentropy = 0.93\nsamples = 36560\nvalue = [12629,
23931]"] ;
6 -> 10 ;
11 [label="entropy = 0.02\nsamples = 6210\nvalue = [12, 6198]"] ;
10 -> 11 ;
12 [label="entropy = 0.979\nsamples = 30350\nvalue = [12617, 17733]"] ;
10 -> 12 ;
}
```



# Summary and Future Work

In this study I used the NIS dataset (HCUP) created by AHRQ. I was successfully able to load all the required data into database, performed data pre-processing and create different machine learning models that have a good predictive power for the diagnosis of breast cancer. Also, I found out that for NIS data we can use Decision based classifier to make better predications and overall give a higher accuracy as compared to other models.

For this project, I have focused on prediction of breast cancer, but this method can be used to predict the risk for any disease. Various other predictive and statistical analysis can be performed using other data sets (Hospital Weights File and Severity Measures File). NIS have good amount of data but lack of early indication and early diagnosis. So, we can develop a method which will help to overcome this problem that will aid patients for early detection and proper diagnosis to prevent further risk.

# References

1. Agency for Healthcare Research and Quality - https://www.ahrq.gov/
2. NIS File Specifications –
https://www.hcup-us.ahrq.gov/db/nation/nis/nisfilespecs.jsp#2017NIS
3. NIS data elements description - https://www.hcup-us.ahrq.gov/db/nation/nis/nisdde.jsp
4. https://www.cancer.net/cancer-types/breast-cancer/statistics
5. https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470
6. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf
7. https://www.postgresqltutorial.com/
7. https://scikit-learn.org/stable/
8. https://towardsdatascience.com/machine-learning/home