**First Dataset1**

<https://www.kaggle.com/mazharkarimi/heart-disease-and-stroke-prevention>

Context

This is one of the dataset provided by the National Cardiovascular Disease Surveillance System.

The system is designed to integrate multiple indicators from many data sources to provide a comprehensive picture of the public health burden of CVDs and associated risk factors in the United States.

Content

The data are organized by location (national, regional, state, and selected sites) and indicator, and they include CVDs (e.g., heart failure) and risk factors (e.g., hypertension). The data can be plotted as trends and stratified by age group, sex, and race/ethnicity.

2011 to present. BRFSS is a continuous, state-based surveillance system that collects information about modifiable risk factors for chronic diseases and other leading causes of death.

Acknowledgements

Indicators from this data source have been computed by personnel in CDC's Division for Heart Disease and Stroke Prevention (DHDSP).

**Second Dataset2**

<https://www.kaggle.com/andrewmvd/heart-failure-clinical-data>

Dataset from Davide Chicco, Giuseppe Jurman: â€œMachine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020)

| **Feature** | **Explanation** | **Measurement** | **Range** |
| --- | --- | --- | --- |
| Age | Age of the patient | Years | [40,…, 95] |
| Anaemia | Decrease of red blood cells or hemoglobin | Boolean | 0, 1 |
| High blood pressure | If a patient has hypertension | Boolean | 0, 1 |
| Creatinine phosphokinase (CPK) | Level of the CPK enzyme in the blood | mcg/L | [23,…, 7861] |
| Diabetes | If the patient has diabetes | Boolean | 0, 1 |
| Ejection fraction | Percentage of blood leaving | Percentage | [14,…, 80] |
| Sex | Woman or man | Binary | 0, 1 |
| Platelets | Platelets in the blood | kiloplatelets/mL | [25.01,…, 850.00] |
| Serum creatinine | Level of creatinine in the blood | mg/dL | [0.50,…, 9.40] |
| Serum sodium | Level of sodium in the blood | mEq/L | [114,…, 148] |
| Smoking | If the patient smokes | Boolean | 0, 1 |
| Time | Follow-up period | Days | [4,…,285] |
| (target) death event | If the patient died during the follow-up period | Boolean | 0, 1 |

Boolean features  
        \* Sex - Gender of patient Male = 1, Female =0  
        \* Diabetes - 0 = No, 1 = Yes  
        \* Anaemia - 0 = No, 1 = Yes  
        \* High\_blood\_pressure - 0 = No, 1 = Yes  
        \* Smoking - 0 = No, 1 = Yes  
        \* DEATH\_EVENT - 0 = No, 1 = Yes

**Third Dataset3**

https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression

**LOGISTIC REGRESSION - HEART DISEASE PREDICTION**

Introduction  
World Health Organization has estimated 12 million deaths occur worldwide, every year due to Heart diseases. Half the deaths in the United States and other developed countries are due to cardio vascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications. This research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression  
Data Preparation

Source  
The dataset is publically available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD).The dataset provides the patients’ information. It includes over 4,000 records and 15 attributes.  
Variables  
Each attribute is a potential risk factor. There are both demographic, behavioral and medical risk factors.

Demographic:  
• Sex: male or female(Nominal)  
• Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)  
Behavioral  
• Current Smoker: whether or not the patient is a current smoker (Nominal)  
• Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)  
Medical( history)  
• BP Meds: whether or not the patient was on blood pressure medication (Nominal)  
• Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)  
• Prevalent Hyp: whether or not the patient was hypertensive (Nominal)  
• Diabetes: whether or not the patient had diabetes (Nominal)  
Medical(current)  
• Tot Chol: total cholesterol level (Continuous)  
• Sys BP: systolic blood pressure (Continuous)  
• Dia BP: diastolic blood pressure (Continuous)  
• BMI: Body Mass Index (Continuous)  
• Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)  
• Glucose: glucose level (Continuous)  
Predict variable (desired target)  
• 10 year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)  
Logistic Regression  
Logistic regression is a type of regression analysis in statistics used for prediction of outcome of a categorical dependent variable from a set of predictor or independent variables. In logistic regression the dependent variable is always binary. Logistic regression is mainly used to for prediction and also calculating the probability of success.  
The results above show some of the attributes with P value higher than the preferred alpha(5%) and thereby showing low statistically significant relationship with the probability of heart disease. Backward elimination approach is used here to remove those attributes with highest P-value one at a time followed by running the regression repeatedly until all attributes have P Values less than 0.05.  
Feature Selection: Backward elimination (P-value approach)  
Logistic regression equation  
P=eβ0+β1X1/1+eβ0+β1X1P=eβ0+β1X1/1+eβ0+β1X1  
When all features plugged in:  
logit(p)=log(p/(1−p))=β0+β1∗Sexmale+β2∗age+β3∗cigsPerDay+β4∗totChol+β5∗sysBP+β6∗glucoselogit(p)=log(p/(1−p))=β0+β1∗Sexmale+β2∗age+β3∗cigsPerDay+β4∗totChol+β5∗sysBP+β6∗glucose

Interpreting the results: Odds Ratio, Confidence Intervals and P-values  
• This fitted model shows that, holding all other features constant, the odds of getting diagnosed with heart disease for males (sex*male = 1)over that of females (sex*male = 0) is exp(0.5815) = 1.788687. In terms of percent change, we can say that the odds for males are 78.8% higher than the odds for females.  
• The coefficient for age says that, holding all others constant, we will see 7% increase in the odds of getting diagnosed with CDH for a one year increase in age since exp(0.0655) = 1.067644.  
• Similarly , with every extra cigarette one smokes thers is a 2% increase in the odds of CDH.  
• For Total cholesterol level and glucose level there is no significant change.

• There is a 1.7% increase in odds for every unit increase in systolic Blood Pressure.

Model Evaluation - Statistics  
From the above statistics it is clear that the model is highly specific than sensitive. The negative values are predicted more accurately than the positives.  
Predicted probabilities of 0 (No Coronary Heart Disease) and 1 ( Coronary Heart Disease: Yes) for the test data with a default classification threshold of 0.5  
lower the threshold  
Since the model is predicting Heart disease too many type II errors is not advisable. A False Negative ( ignoring the probability of disease when there actually is one) is more dangerous than a False Positive in this case. Hence in order to increase the sensitivity, threshold can be lowered.

Conclusions  
• All attributes selected after the elimination process show P-values lower than 5% and thereby suggesting significant role in the Heart disease prediction.

• Men seem to be more susceptible to heart disease than women. Increase in age, number of cigarettes smoked per day and systolic Blood Pressure also show increasing odds of having heart disease

• Total cholesterol shows no significant change in the odds of CHD. This could be due to the presence of 'good cholesterol(HDL) in the total cholesterol reading. Glucose too causes a very negligible change in odds (0.2%)

• The model predicted with 0.88 accuracy. The model is more specific than sensitive. Overall model could be improved with more data

Appendix  
<http://www.who.int/mediacentre/factsheets/fs317/en/>  
Data Source References  
<https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset/data>

**Fourth Dataset4**

https://www.kaggle.com/volodymyrgavrysh/heart-disease

### Attribute Information:

1. Age: Age
2. Sex: Sex (1 = male; 0 = female)

a. ChestPain: Chest pain (typical, asymptotic, nonanginal, nontypical)

1. RestBP: Resting blood pressure
   * 1. Chol: Serum cholestoral in mg/dl
     2. Fbs: Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
     3. RestECG: Resting electrocardiographic results
     4. MaxHR: Maximum heart rate achieved
     5. ExAng: Exercise induced angina (1 = yes; 0 = no)
     6. Oldpeak: ST depression induced by exercise relative to rest
     7. Slope: Slope of the peak exercise ST segment
     8. Ca: Number of major vessels colored by flourosopy (0 - 3)
     9. Thal: (3 = normal; 6 = fixed defect; 7 = reversable defect)
     10. target: AHD - Diagnosis of heart disease (1 = yes; 0 = no)

Source: <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>

Creators:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Donor: David W. Aha (aha '@' ics.uci.edu) (714) 856-8779

Data Set Information:

This database contains attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. T

he "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence from absence (value 0).

The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

**Fifth Dataset5**

[*https://www.kaggle.com/sulianova/cardiovascular-disease-dataset?select=cardio\_train.csv*](https://www.kaggle.com/sulianova/cardiovascular-disease-dataset?select=cardio_train.csv)

There are 3 types of input features:

* Objective: factual information;
* Examination: results of medical examination;
* Subjective: information given by the patient.

Features:

1. Age | Objective Feature | age | int (days)
2. Height | Objective Feature | height | int (cm) |
3. Weight | Objective Feature | weight | float (kg) |
4. Gender | Objective Feature | gender | categorical code |
5. Systolic blood pressure | Examination Feature | ap\_hi | int |
6. Diastolic blood pressure | Examination Feature | ap\_lo | int |
7. Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
8. Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
9. Smoking | Subjective Feature | smoke | binary |
10. Alcohol intake | Subjective Feature | alco | binary |
11. Physical activity | Subjective Feature | active | binary |
12. Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

All of the dataset values were collected at the moment of medical examination.