# **SVKM’s NMIMS**

# **School of Technology Management & Engineering, Navi Mumbai**

# 

# **Topic:** DMA project

# **B. TECH - SEM. 6th**

# **(Computer Science & Business System)**

Avantika Jalote A061

Karim Khan A069

Kevin Mathew A068

Mansi Nandkar A058

# AIM:

The main aim of the project is to predict heart attacks. It makes use of the attributes and their values to predict the possibility of a heart attack.

# PROCEDURE:

1. **Data collection:**
   1. The data was collected from <https://archive.ics.uci.edu/ml/datasets/heart+disease>
   2. The dataset here had data from 4 clinics. The data was present in .data files. The unprocessed data was present with 75+1 attributes.. We tried to filter the data but ran into problems.
   3. We collected the processed files with 14 attributes each and then moved to the next step of data integration.
   4. The attributes included:
      1. **Age**: Age of the patient
      2. **Sex**: Sex of the patient
      3. **cp**: Chest Pain type chest pain type
         1. Value 1: typical angina
         2. Value 2: atypical angina
         3. Value 3: non-anginal pain
         4. Value 4: asymptomatic
      4. **trtbps**: resting blood pressure (in mm Hg)
      5. **chol**: cholesterol in mg/dl fetched via BMI sensor
      6. **fbs**:(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
      7. **rest\_ecg**: resting electrocardiographic results
         1. Value 0: normal
         2. Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
         3. Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
      8. **thalachh**: maximum heart rate achieved
      9. **exng**: exercise induced angina (1 = yes; 0 = no)
      10. **oldpeak**: Previous peak
      11. **slp**: Slope
      12. **caa**: number of major vessels (0-3)
      13. **thal**: Thal rate
          1. 3 = normal
          2. 6 = fixed defect
          3. 7 = reversible defect
      14. **Output**:
          1. 0 = less chance of heart attack
          2. 1 = more chance of heart attack
2. **Data integration:**
   1. We merged the processed data for all the 4 clinics
   2. We also merged the oxygen saturation statistics with the main data.
3. **Data cleaning :**
   1. With the Cleveland data:
      1. No NULL values were found
   2. With merged data
      1. NULL values are found and are replaced with mean values
4. **Data selection**:
   1. Data anFeatures were already selected from a pool of 76 attributes to a mere 14+1 attributes.
5. **Data transformation**:
   1. The data was then normalized using the StandardScalar

# MODELS USED:

We have created a test data set and split it into training and testing data. We have standardized the data using a standard scaler.

By implementing logistic regression , we got an accuracy of:

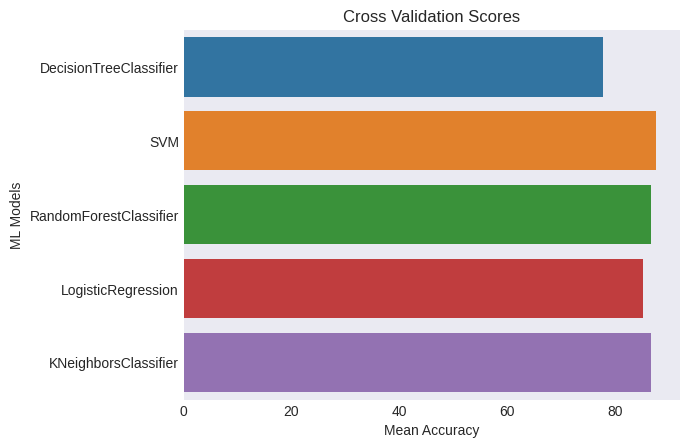
**Training Dataset:88.67%**

**Testing Dataset: 58.0%**

We have tried other models: Logistic, SVM, Random Tree, Decision Tree, and KNN classifier.

By this, we got the result with the following:

| **MODEL** | **ACCURACY** |
| --- | --- |
| DECISION TREE | 77.85714285714286 |
| SVM | 87.66666666666667 |
| RANDOM FOREST | 86.66666666666666 |
| LOGISTIC REGRESSION | 85.19047619047619 |
| KNN CLASSIFIER | 86.73809523809524 |



In conclusion , we got SVM is giving us the highest accuracy of 87.67%

# 

# 

# IPYNB FILE : [Google Colab Notebook](https://colab.research.google.com/drive/1HoDNR7L5L4ACL-R7UNV4pciI9zxtgT0Y?authuser=2#scrollTo=c0JoaCIh-BWJ)

# 

# CONCLUSION:

Amongst all the other models SVM is the most accurate.