VISVESVARAYA TECHNOLOGICAL UNIVERSITY
BELGAUM 59014

Big Data Analytics Project Report on

"Predicting Arrhythmia Using Cardiac History"

By

Deepthi Bhat (1BM16CS003)

Aditi Awasthi (1BM16CS008)

Medhini Oak (1BM16CS047)

Under the Guidance of

Mrs. K. Panimozhi

Assistant Professor, Department of CSE

BMS College of Engineering

BDA Project Development carried out at

Department of Computer Science and Engineering

BMS College of Engineering

(Autonomous college under VTU)

P.O. Box No.: 1908, Bull Temple Road, Bangalore-560 019

2019-2020

BMS COLLEGE OF ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Big Data And Analytics project titled "**Predicting Arrhythmia Using Cardiac History"** has been carried out by Deepthi Bhat (1BM16CS003), Aditi Awasthi (1BM16CS008) and Medhini Oak (1BM16CS047) during the academic year 2019-2020.

Signature of the guide
Mrs. K. Panimozhi
Assistant Professor
Department of Computer Science and Engineering
BMS College of Engineering, Bangalore

BMS COLLEGE OF ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

We, Deepthi Bhat (1BM16CS003), Aditi Awasthi (1BM16CS008) and Medhini Oak (1BM16CS047), students of 7th Semester, B.E, Department of Computer Science and Engineering, BMS College of Engineering, Bangalore, hereby declare that, this Big Data And Analytics project development work entitled "**Predicting Arrhythmia Using Cardiac History**" has been carried out by us under the guidance of Mrs. K. Panimozhi, Assistant Professor, Department of CSE, BMS College of Engineering, Bangalore during the academic semester Aug - Dec 2019.
We also declare that to the best of our knowledge and belief, the development reported here is not from part of any other report by any other students.

Signature

Deepthi Bhat (1BM16CS003)

Aditi Awasthi (1BM16CS008)

Medhini Oak (1BM16CS047)

1. Abstract

Cardiac arrhythmia (abnormal heart rate or rhythm) is a very common affliction with more than 10 million cases per year. Some cases of arrhythmia can be critical and only with quick response can the patient reduce risks of complications. The diagnosis of arrhythmia involves handling of huge amount of ECG data which poses the risk of human error in the interpretation of data. If a health care provider failed to diagnose a severe case of arrhythmia, they can even be held liable for medical malpractice.

2. Introduction

This project aims at computer assisted analysis of the ECG data of a given subject and arrhythmia detection and classification and can thus, play a huge role as a decision support system to the doctors. It tackles this goal in the form of a supervised learning problem.
Different classes of cardiac arrhythmia include:

- Normal
- Ischemic changes (Coronary Artery)
- Old Anterior Myocardial Infarction
- Old Inferior Myocardial Infarction
- Sinus tachycardia (fast heartbeat >100 beats/min)
- Sinus bradycardia (slow heartbeat < 60 beats/min)
- Ventricular Premature Contraction (PVC)
- Supraventricular Premature Contraction
- Left bundle branch block
- Right bundle branch block
- Left ventricle hypertrophy
- Atrial Fibrillation or Flutter (irregular heartbeat)
- Others

3. Software Requirements

- Python 3.7.3

  Libraries used:

  - Scikit-learn
  - Tkinter, EasyGUI
  - Pickle
  - Matplotlib
  - Pymongo
  - NumPy
- Mongo 3.6.8

4. Main aspects of the project

1. No-SQL Database

Library used: pymongo

MongoDB is a DBMS that uses a document-oriented database model. Instead of using tables and rows as in relational databases, the MongoDB architecture is made up of collections and documents. The database will store the cardiac data of earlier subjects as well as the target outcomes for each subject which help us train the model which will later be used for prediction.
The main reason for the use of MongoDB is the fact that it allows us to pickle (conversion of python object to serialized binary form) our trained model and store it for future use without needing to retrain the model each time it is needed.

2. Classification Algorithm

Library used: Scikit-learn

In order to classify the subjects into various types of arrhythmia, we need a classification algorithm e.g. logistic regression, Naive Bayes classifier, KNNs, Random Forests, Fisher's linear discriminant, etc. For this project, we have employed the logistic regression model.

3. Data Visualization

Library used: Matplotlib

By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.
Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, Qt, or GTK+.

5. Design

The main modules of our project include:

1. Basic CRUD Operations on Data

Efficient handling of big data by making use of MongoDB for storage and performing create, read, update and delete operations on it. The dataset obtained from the UCI repository: https://archive.ics.uci.edu/ml/datasets/Arrhythmia

Each row represents the medical record of a different patient. There are 175 attributes like age, weight and patient's ECG related data. General attributes like age and weight have discrete integral values while other ECG features like QRS duration have real values. The variable Class is our target variable.

In case needed, aggregations can be performed on the training data set, including finding the minimum or maximum of certain fields in the dataset and even average, since the data is entirely numerical in nature

2. Training a logistic regression model using the historical data

Logistic regression is a statistical model that uses a logistic function to model a dependent variable. Since the logistic regression is used for binary classification of datasets with dependent features, in order to apply logistic regression to our multi-class dataset, we firstly classified our instances into two major classes, class 1 and class NOT-1. Then further classification into 13 classes is done.

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}(1-y^{(i)})\log(1-h_\theta(x^{(i)})) + y^{(i)}\log h_\theta(x^{(i)})\right]$$
$$= -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=0}^{1} 1\left\{y^{(i)}=j\right\}\log p(y^{(i)}=j|x^{(i)};\theta)\right]$$

3. Classifying new subjects

A .csv file containing the cardiac information of the new subject under consideration is fed into our trained model and a prediction is made regarding which one of the 13 possible classes this particular subject is most likely to lie in. The ECG information consists of features like Age, Sex, Height, Weight, QRS duration, P-R interval, Q-T interval, T interval, P interval, Heart rate, etc.

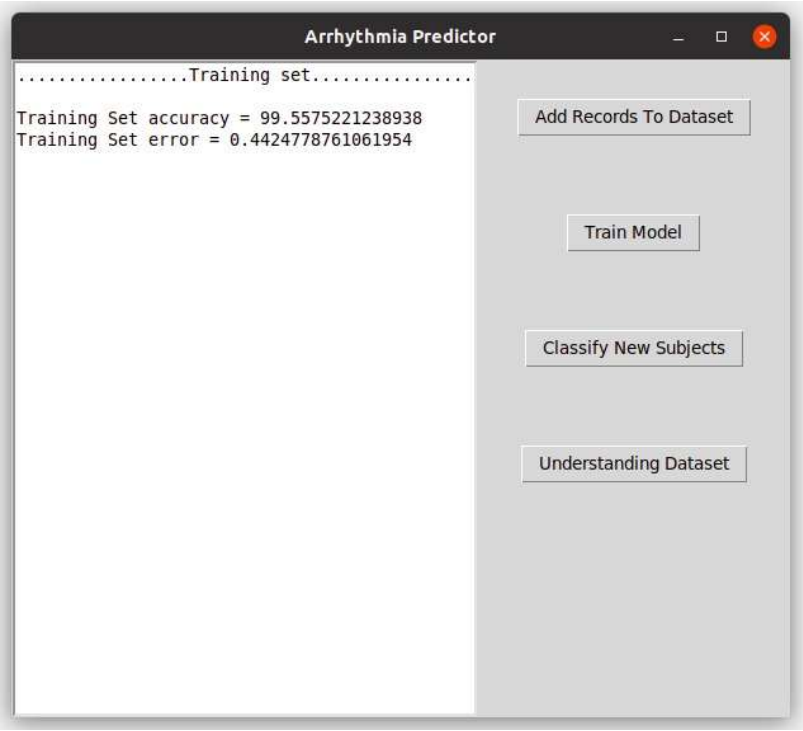
It is to be noted that the results produced are not absolutely concrete and need to be verified by a practicing cardiologist.

4. Observing trends across dataset

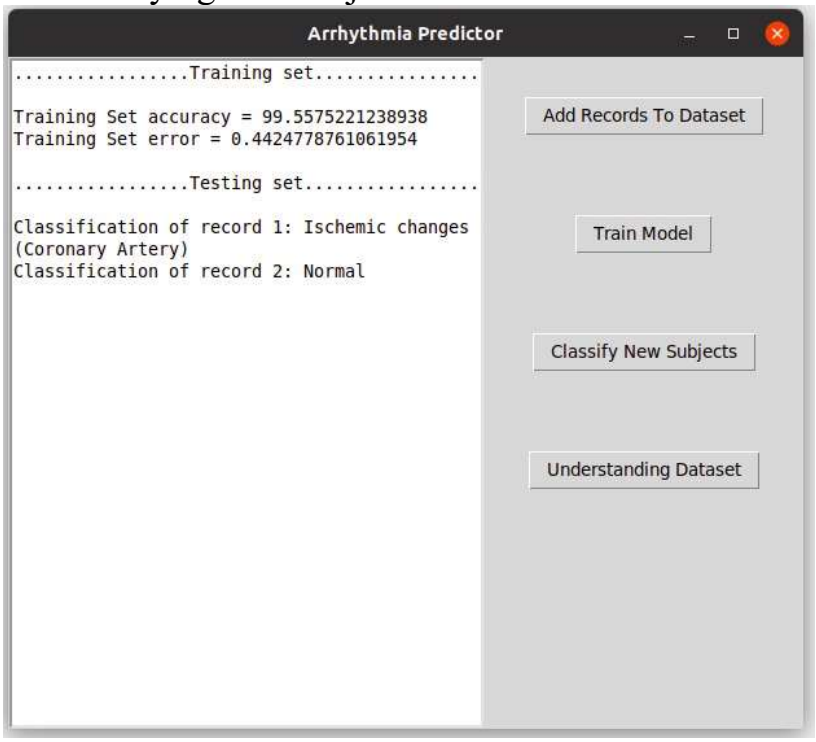
Visualization of salient features of the dataset is available in the form of bar graphs, showing the frequency of occurrence of various classes of arrhythmia across various age and weight groups as well as the general trend of average heart rate observed in the respective cases. This will help give the user a broad picture about the groups that are specially affected by that class of arrhythmia.

6. Screenshots

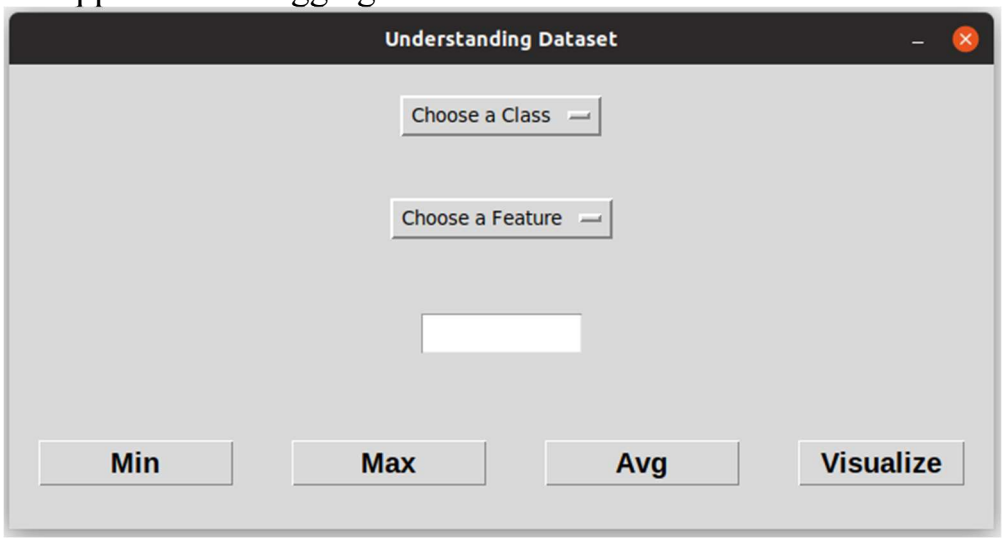
1. Once the model is trained

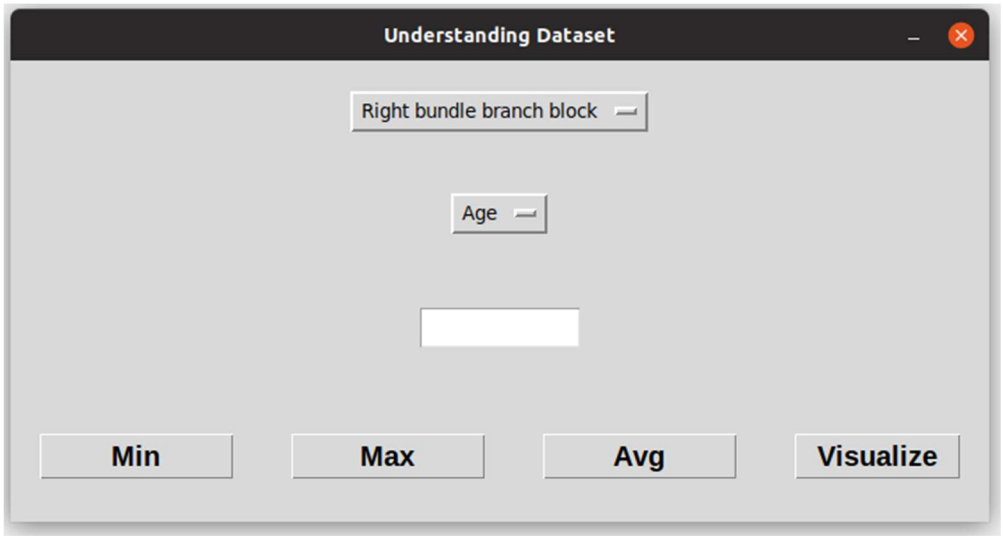


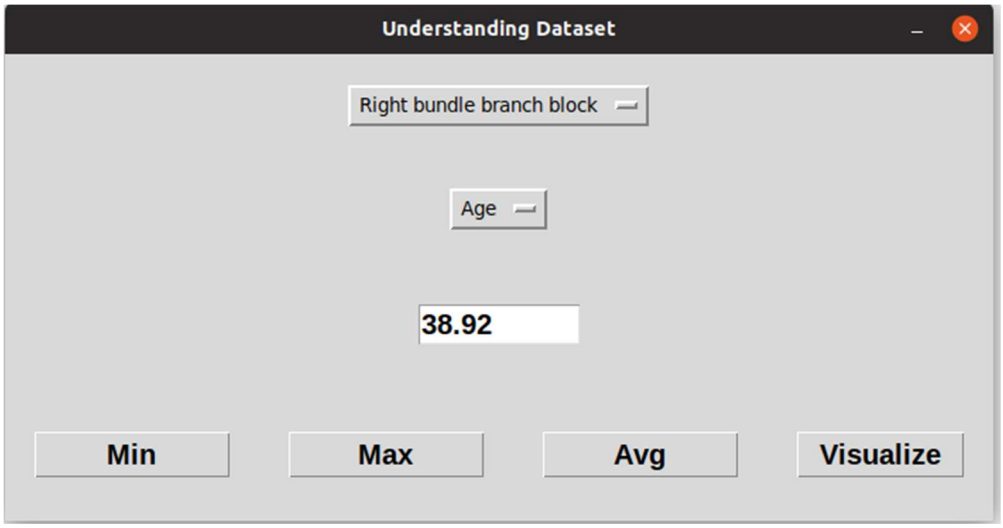
2. Classifying new subject



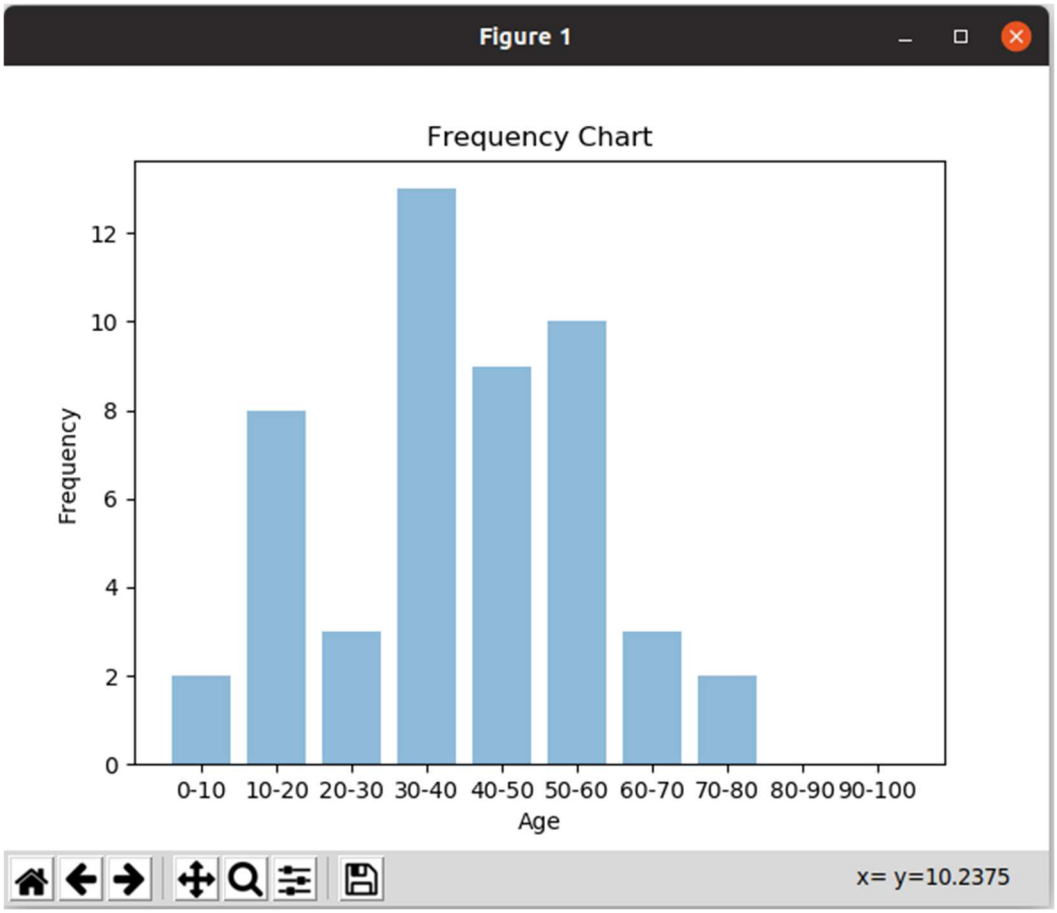
3. Application of aggregation functions on dataset







4. Visualization of different features across classes of arrhythmia



7. Results

The model is trained with an accuracy of 99.6% and is accurately able to classify a subject's ECG data. The total number of deaths due to cardiovascular diseases read 17.3 million a year according to the WHO causes of death. Thus, prediction of cardiac arrhythmia in real life is of great significance. The technique illustrated in our project can be deployed in hospitals where a large dataset is available and can help doctors in making more precise decisions and to cut down the number of casualties due to heart diseases in the future.

This project showcases the power of big data in the field of health and medicine. If used extensively and updated rigorously by medical professionals around the world, this repository can be a rich storehouse for historical cardiac data, with the model becoming better as time goes on and acting as a lifesaver to millions of people.