**Semester VI**
**Group 6**

# CARDIAB

## Cardiovascular Diseases and Diabetes Prediction Using ML

**Group Members:**

1. Shefali Jindal [A07]

2. Tithi Agrawal [A14]

3. Abbas Husain [A32]

4. Atharva Nimbalwar [A41]

5. Shubham Saboo [A76]

**Project Guide:**

**Prof. Heena Agrawal**

# Problem

- Covid -19
- Deaths related to CVD alone
- Deaths attributable to Diabetes alone
- Overburdened Healthcare system

**Hospitalizations were 6 times higher and deaths 12 times higher for COVID-19 patients with reported underlying conditions***

MOST FREQUENTLY REPORTED UNDERLYING CONDITIONS

CARDIOVASCULAR DISEASE

DIABETES

CHRONIC LUNG DISEASE

*compared to those with no reported underlying health conditions

CDC.GOV

bit.ly/MMWR61520

MMWR

# Solution/Application

- Web Application which uses machine learning models to predict the chances of CVD and Diabetes.
- Tips and precautions for maintaining a healthy life
- Generating a PDF report of the result
- Emailing the report to the patient
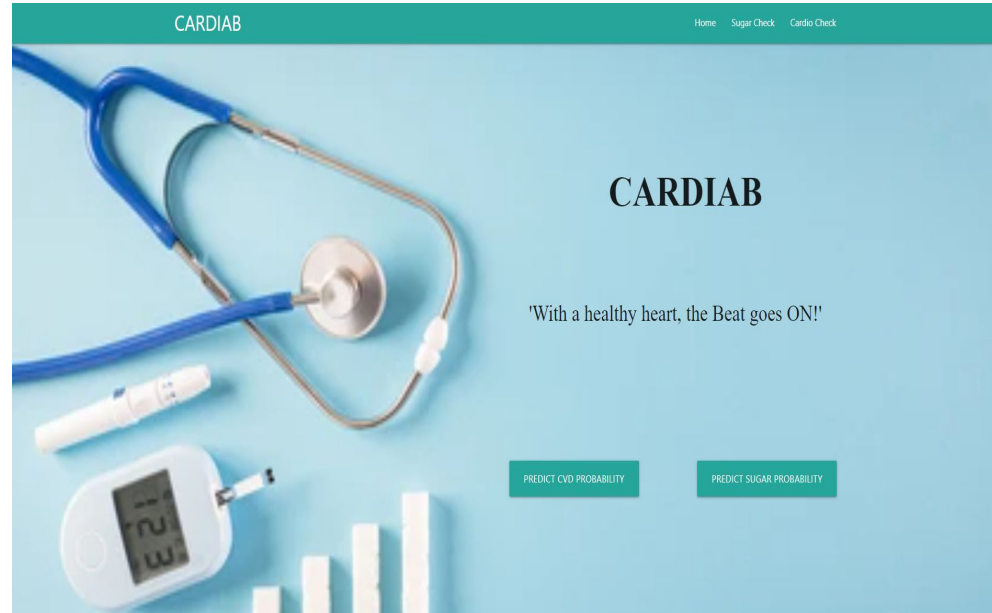
# Existing Work

The Diabetes Model

1.      From (Mahesh Barale et al 2016)

·       They removed the cases with more than one value missing and then finally input this data on 534 out of 768 values which is very less for model training. Also, about 30% of the data is removed.

2.      From (Naveen Kishore et al 2020)

·       This research paper has illustrated the training of model using 4 algorithms with the following accuracies: SVM 73.43%, Random Forest 74.4%, KNN 71.3% and Logistic Regression 72.39%.

# Existing Work

The CVD Model

1.      From (V. V. Ramalingam et al 2018)

·       This research paper highlights the importance of feature selection and feature engineering but it uses a dataset (the Cleveland dataset) which consists of less than 400 records.

2.      From (Adil Hussain Seh et al 2019)

·       In this research paper the authors have removed the outliers using known values of those attributes and highlighted the difference in the results. Their model seems to be over-fitted.
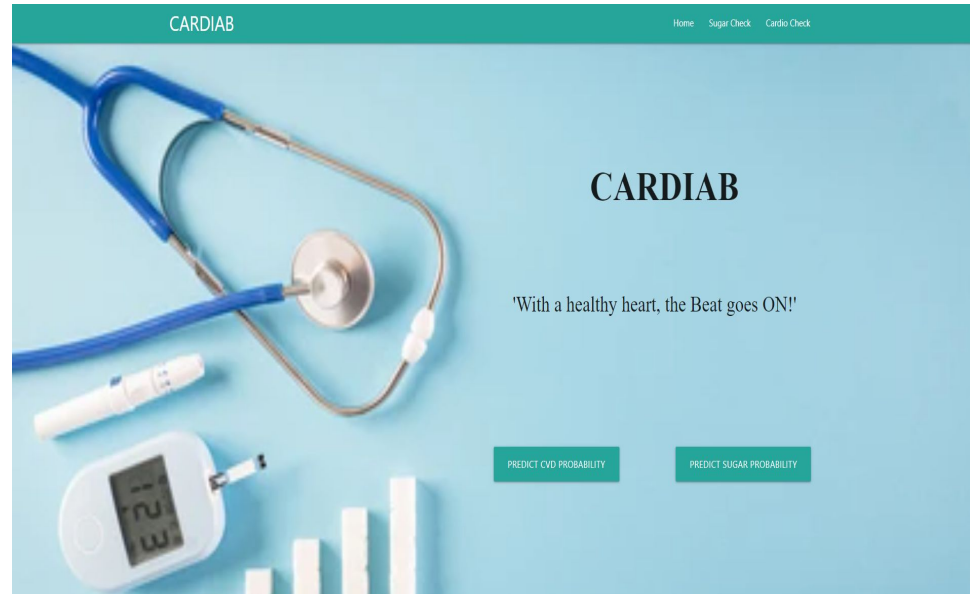
# Technologies and Tools

1. Hypertext Markup Language [HTML]
2. Cascading Style Sheet [CSS]
3. Visual Studio Code
4. Flask
5. Materializecss.com
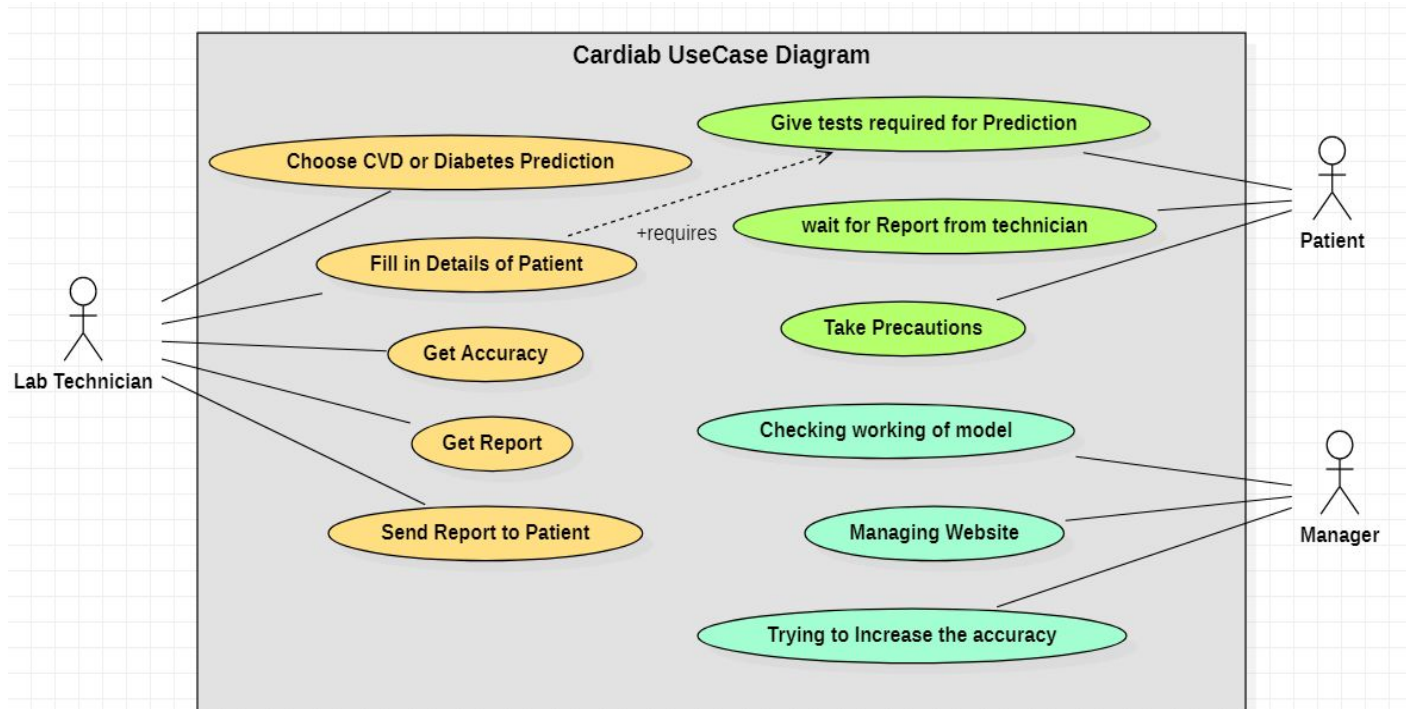6. Heroku

# Application Link

## https://cvd-dia-predict.herokuapp.com/

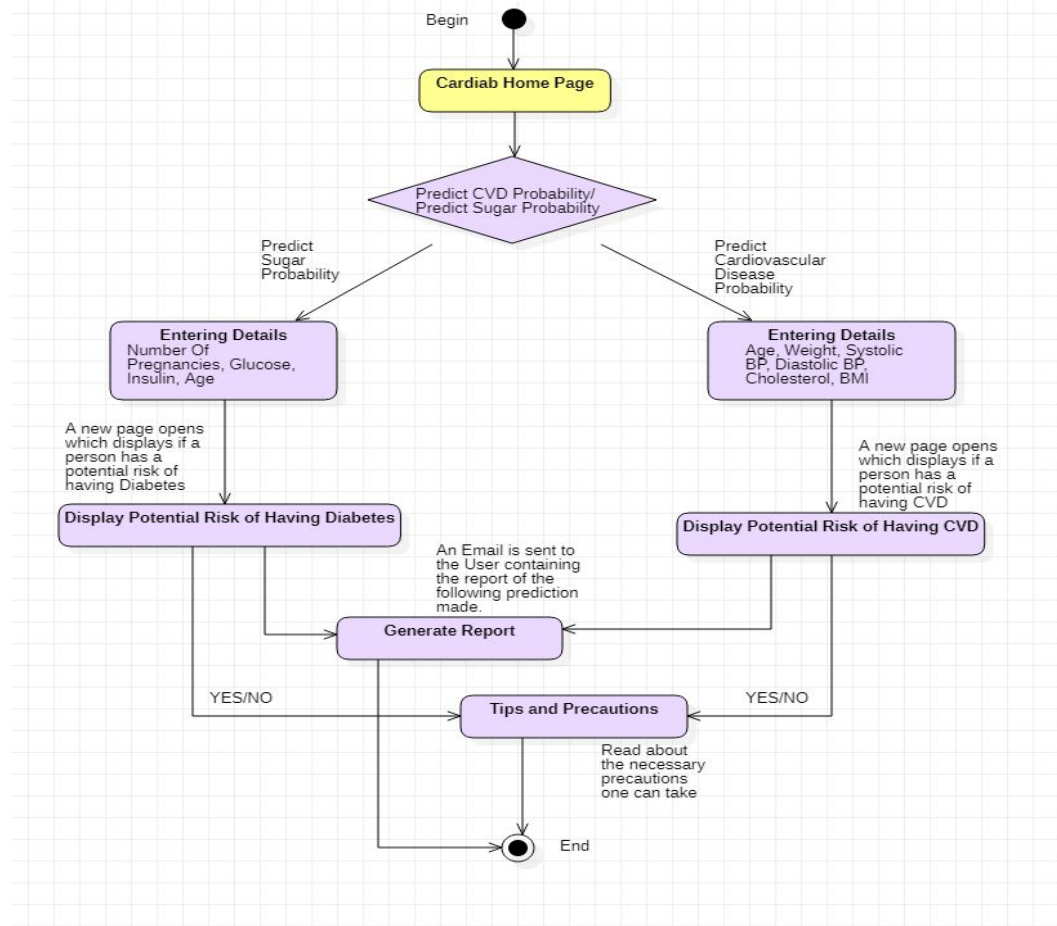# UseCase Diagram

Use case diagram depicts all the actors involved in this project and their functionalities throughout the project.
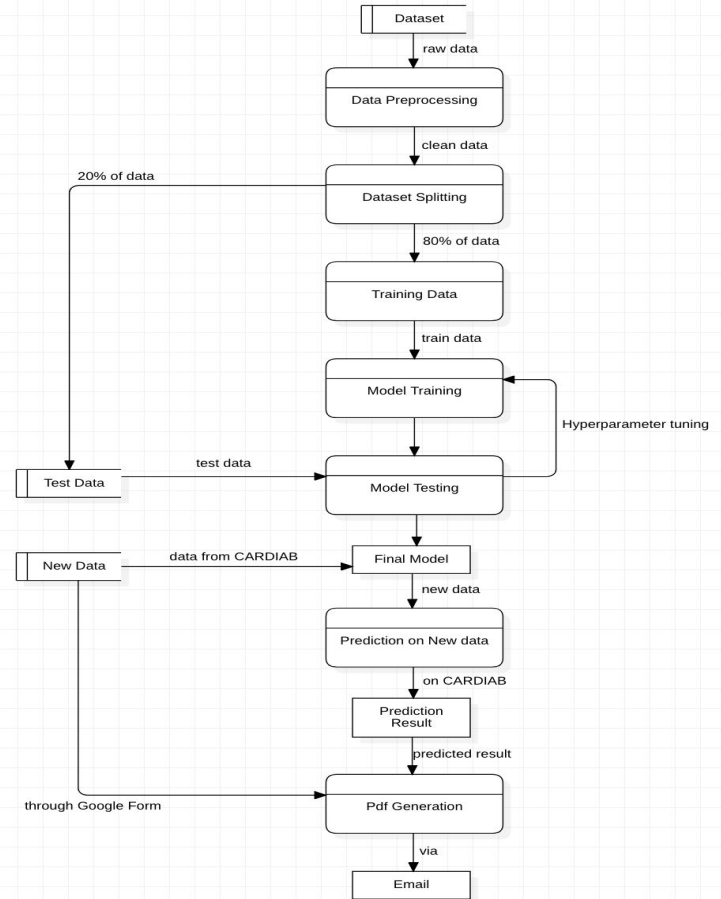


Cardiab UseCase Diagram

# Statechart Diagram (Front-end)

This depicts different states and levels of the UI of our web application.
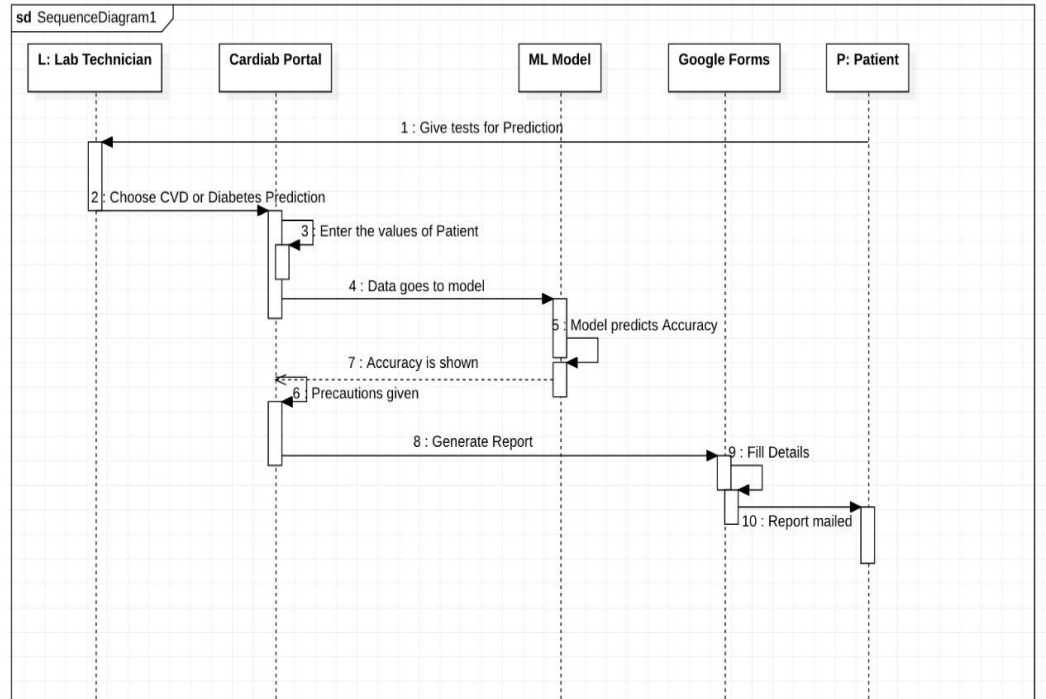
# 2-Level Data Flow Diagram

DFD represents communication, manipulation and distribution of data between the system.

# SEQUENCE DIAGRAM

This depicts the actual flow of events throughout the process between the participants.

# Accuracy of algorithms

**Table 1: Accuracies of the two models on different algorithms (before hyperparameter tuning):**

| Algorithm | CVD Model | Diabetes Model |
|---|---|---|
| KNN | 73.16% | 75% |
| SVM | 73.15% | 78.94% |
| Logistic Regression | 72.59% | 77.63% |
| Random Forest | 69.47% | 75.65% |

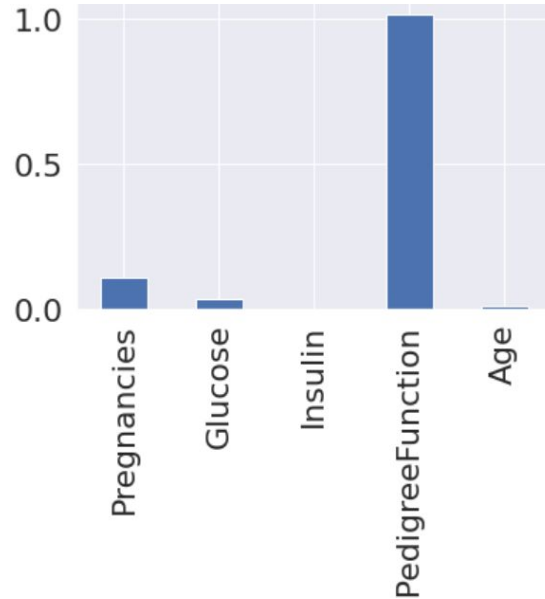**Table 2: Accuracies of the two models on best algorithms (after hyperparameter tuning):**

| Models | Algorithm | Accuracy |
|---|---|---|
| CVD Model | KNN | 73.25% |
| Diabetes Model | Logistic Regression | 78.94% |

**1** Data Cleaning and Preprocessing
1. Identifying missing values and replacing it with mean or median values.
2. Removing Outliers

**2** Data Visualization & Feature Selection
Analyzing correlation matrix. and removing the least significant attributes.

**3** Splitting the Data and training the model
Using 4 different ML algorithms: KNN, SVM, RFC ,Logistic regression

**4** Finding out the best Model & Tuning
Selection Of best model Hyperparameter tuning

**5** Result Evaluation
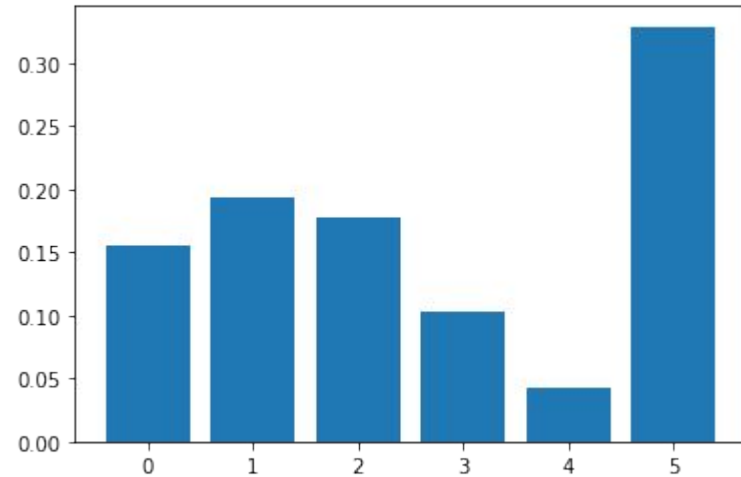F1 score, precision, recall, accuracy

# Feature Importance



**Diabetes Model**

**CVD Model**

0:Age , 1:Weight , 2:ap_hi , 3:ap_lo
4:Cholesterol , 5:BMI

# Future Scope

- Increasing the accuracy of the system and decreasing the number of false negatives. This will make the ML model more reliable.
- Our application can be extended to apply this method, for other diseases like chronic lung diseases, etc.
- Find a mechanism to continuously update the training data with new data.
- Incorporating more modules like ECG values for prediction of cardiovascular diseases if suitable datasets are available.

# References

[1] Barale, Mahesh & Shirke, Digambar. (2016). Cascaded Modeling for PIMA Indian Diabetes Data. International Journal of Computer Applications. 139. 1-4. 10.5120/ijca2016909426.

[2] Naveen Kishore G, V.Rajesh, A.Vamsi Akki Reddy, K.Sumedh, T.Rajesh Sai Reddy. (2020) Prediction Of Diabetes Using Machine Learning Classification Algorithms. ISSN 2277-8616

[3] Subbalakshmi, G. & Ramesh, K. & Rao, M.. (2011). Decision Support in Heart Disease Prediction System using Naive Bayes. Ind. J. Comput. Sci. Eng. (IJCSE). 2. 170-176.

[4] Seh, Adil Hussain. (2019). A Review on Heart Disease Prediction Using Machine Learning Techniques.

[5] Ramalingam, V V & Dandapath, Ayantan & Raja, M. (2018). Heart disease prediction using machine learning techniques: A survey. International Journal of Engineering & Technology. 7. 684. 10.14419/ijet.v7i2.8.10557.

[6] Yu, W., Liu, T., Valdez, R., Gwinn, M., Khoury, M.J., 2010. Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making 10. doi:10.1186/1472-6947-10-16 | arXiv:arXiv:1011.1669v3

[7] Sellappan Palaniappan, Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, 978-1-4244-1968- 5/08/$25.00 ©2008 IEEE.

[8] Andreas G. K. Janecek ,WilfriedN.Gansterer and Michael A.Demel,‖On the Relationship Between Feature Selection and Classification Accuracy‖,JMLR: Workshop and Conference Proceedings 4: 90-105

[9] Marvin L. Brown and John F. Kros, Data Mining and the Im-pact of Missing Data, Industrial Management & Data Systems,Volume 103, ISS: 611–621, (2003).

[10] Bamnote, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763–770 | doi:10.1007/978-3-319-11933-5.