Final Year Project Proposal

Heart Diseases Prediction

Haris Bin Irfan **(37888)**

Syed Usama Ahmed Hashmi **(38633)**

Yasir Ul haq **(37728)**

Dr M.Raza

Dr Mansoor Ebrahim

List of Abbreviations and Acronyms

Small and mid-size enterprises (SMEs)
Light Gradient Boosting Machine (LightGBM)
eXtreme Gradient Boosting (XGboost)
Neural Networks (NN)
Artificial Neural Networks (ANN)
Decision Tress (DT)
Naïve Bayes (NB)
Hadoop Distributed File System (HDFS)
Support Vector Machine (SVM)
Data Mining (DM)
Machine learning (ML)
Random forest (RF)

Table of Contents

Final Year Project Proposal

Section – 1

1.1 Project Identification

Project Title: Heart Diseases Prediction				
Group Leader (GL):				
1.	Name:	**Haris Bin Irfan**		
	Reg #:	37888		
	CGPA:	2.97		
	Mobile # :	03042723854	Email:	haris.bin.irfan2012@outlook.com
	Signature:			
Group Members (GM's):				
2.	Name:	**Yasir Ul Haq**		
	Reg #:	37728		
	CGPA:	2.68		
	Mobile # :	03350237955	Email:	yasirkhanchandia8@gmail.com
	Signature:			
3.	Name:	**Syed Usama Ahmed Hashmi**		
	Reg #:	38633		
	CGPA:	2.35		
	Mobile # :	03423579322	Email:	uxamasyed60@gmail.com
	Signature:			

What technology is core to your product? *(Please mark ☑ where applicable)*

[ ] 3D/4D Printing
[ ] Augmented Reality / Virtual Reality
[ ] Big Data, Artificial Intelligence
[ ] Blockchain
[ ] Cloud
[ ] Neurotech
[ ] Robotics
[ ] Shared economy
[ ] The Internet of Things
[ ] Wearables, Implantables
[✓] Others (specify): Machine Learning

What is the target market(s) for the products? *(Please mark ☑ where applicable)*

[ ] Automotive, aviation, marine
[ ] Business, marketing, finance
[ ] Defence, security, safety
[ ] Education and training
[ ] Environment, water management
[ ] Entertainment, tourism, sport/recreation
[ ] Food, livestock, agribusiness
[✓] Healthcare
[ ] Infrastructure, housing & transport
[ ] Mining equipment technology & services
[ ] Oil, gas, energy
[ ] Textiles, clothing, footwear
[ ] Others (specify):

Other Organizations Involved in the Project: *(Please identify all affiliated organizations collaborating in the project, and describe their role/contribution to the project.)*

Academic Organizations:

#	*Organization Name*	*Role / Contribution*
1.	IQRA University	Guidance/Platform provide

Industrial Organizations:

#	*Organization Name*	*Role / Contribution*
1.	n/a	n/a

Funding Organizations:

#	*Organization Name*	*Role / Contribution*
1.	n/a	n/a

Key Words: *(Please provide a maximum of 5 key words that describe the project)*

Heart disease, Machine learning, Voting classifier, Prediction model, Decision Support System, classification techniques

Research and Development Theme: *(please identify the Research Theme.)*

Predicting the presence of heart disease by using Data Mining Techniques

Project Status: (Please mark ☑)

[✓] New [ ] Modification to previous Project

[ ] Extension of existing project

Project Duration: 6 Months

Proposed Budget: Rs.1,309,606 PKR

The Problem:

The heart is a very critical part of the human body. It pumps blood into the whole body If the flow of blood in the body becomes insufficient, organs like the brain suffer, and if the heart stops working completely, death occurs within minutes.
Some of the heart disease risk factors are:

1. Smoking: Smokers are twice as likely to have a heart attack as non smokers.
2. Cholesterol: A low cholesterol diet and Tran's saturated fat can help lower cholesterol levels and reduce heart disease risk.
3. Blood pressure: High BP leads to heart Attack.
4. Diabetes: If not controlled, diabetes can result in severe heart damage including heart attack and death.
5. Sedentary life style: Simple leisure activities, such as gardening and walking, can reduce our risk of heart disease.
6. Eating Habits: A healthy diet of the soul, low in salt, calories, saturated fat, trans fat, cholesterol and refined sugars can decrease our risk of heart disease.
7. Stress: Poorly controlled stress can result in heart attacks and strokes.

There are many methods related to prediction of disease. Yet heart-related disease in particular has been analyzed and the level of risk is produced. But there are usually no such tools that are used for specific disease prediction.
The main objective is to predict the Boolean class heart disease prediction, which represents whether a patient has heart disease or not:
False does not represent heart disease.
True represents heart disease present

Following are some of the well-known (identify the best known if possible) existing solutions to this problem. Their known strengths and weaknesses are also provided.

(Maximum 200 words.)

The pre-existing system operates on deep learning as well as data mining collections. By applying the powerful prediction algorithm, the existing system modules produce a comprehensive report. The main aim of the existing system was to compare and check the pre-patient with disease outputs and new patient disease and then identify future possibilities of cardiac disease for a specific patient. By implementing following classification models:

- Naive Bayes
- Logistic Regression
- Decision Tree
- Random Forest
- LightGBM
- XGboost.

But the accuracy is much less by using all the existing systems.

Drawbacks of the existing system

1.) It's very difficult to maintain the system.
2.) There is a high chance of obtaining inaccurate results.
3.) GUI is much less user-friendly.
4.) It takes longer to process the activities.

Our solution will address the following weaknesses of above mentioned solutions.

(Maximum 200 words.)

The proposed system has details that are categorized according to features in it if patients have heart disease or not. This proposed system will attempt to use this dataset to create a model that tries to predict whether or not a person has this disease.In this proposed system, we use 4 Classification algorithums BaggingClassifier(GaussianNB), MLPClassifier, SVC and AdaBoostClassifier(RandomForestClassifier). Calculating the score using the sklearn library. Implements VotingClassifier to get best accuracy results. Using the Comparing Models and Confusion Matrix to finally analyze the results. It should be grouped into separate structured data depending on the features of the patient's heart from the dataset we have. First, we have to import the dataset. Read the dataset, the data should contain different variables like age, gender, sex, cp(chest pain),slope, target etc. The data should be explored so that the information is verified. The records are divided into two datasets: dataset training 80% and dataset testing 20%. To avoid bias, the records are selected randomly for each set. Using VotingClassifier, we need to create a model that predicts the disease of the patient.

We will use the following techniques to achieve improvements mentioned above.

(Maximum 200 words.)

We will be using below techniques and algorithms to achieve our desired tasks:

- **Supervised Machine learning :** Supervised learning is the task of learning a function that maps an input to an output based on examples of pairs of input-output. It implies a feature consisting of a collection of training examples from marked training data.

- **MLPClassifier :** A multilayer perceptron (MLP) is a feedforward artificial neural network generating a set of outputs from a set of inputs. An MLP is characterized by multiple layers of input nodes connected as a guided graph between input and output layers.

- **SVM :** A Support Vector Machine (SVM) is formally defined by a separate hyperplane as a discriminatory classifier. In other words, given the labeled training data (supervised learning), an optimal hyperplane is produced by the algorithm that categorizes new examples.

- **GaussianNB :** Gaussian Bayes of Naive. A special type of NB algorithm is a Gaussian Naive Bayes algorithm. When the apps have constant values, it is explicitly used. It is also assumed that all features follow the normal distribution of a gaussian distribution.

- **AdaBoostClassifier(RandomForestClassifier) :** AdaBoost, short for Adaptive Boosting, is a meta-algorithm for machine learning invented by Yoav Freund and Robert Schapire who received the Gödel Prize for their research in 2003. It can be used to improve performance in conjunction with many other learning algorithms.

- **EnsembleVoteClassifier:** Voting is one of the easiest ways to combine the predictions of several algorithms in machine learning. Voting classifier is not an actual classifier but a wrapper for a set of different classifiers that are trained and priced in parallel to exploit that algorithm's different peculiarities.

Synopsis:

*(A brief description of the idea, in non-technical language, explaining product benefit, target market, basic technology, commercial partners, investors, and potential customers. **Maximum 200 words.**)*

It may have happened so many times that you or someone else seek support from doctors without delay but for some reason they are not available. The application for Heart Disease Prediction is an end-user support and online consultation project. Here, through an intuitive online system, we propose a web application that allows people to get immediate feedback on their heart disease. Different details and the heart disease associated with these details are incorporated into the application. The software helps users to share their heart-related symptoms and then process user-specific details with an intelligent machine learning model to test for detection of heart disease. Here we use some intelligent machine learning algorithms to guess the most accurate result that can be linked to the details of the patient.

Section – 2

2.1 Background

Scope of the Project:

The aim of the project here is that the incorporation of clinical decision support with computer-based patient records will eliminate medical errors, increase patient safety, decrease unintended discrepancies in training, and improve patient results.This proposal is encouraging because methods of data modeling and evaluation,e.g data mining, have the potential to generate an intelligence-rich enviroment that can help improve the quality of medical decisions dramatically.Prediction of heart disease also helps to reduce the cost of treatment and also increase visualization and analysis services.With deep knowledge in this field and accurate data, large companies are investing heavily in this type of activity to help focus on future incidents and risks involved.This work brings together all historical and current data available as a framework for establishing reasonable expectations of the future.

Literature Review: *(Detailed summary of what all has been done internationally in the proposed area quoting references and bibliography. **Maximum 1500 words.**)*

There have been numerous studies that focus on heart disease diagnosis. Various data mining techniques have been applied for diagnosis and different probabilities have been achieved for different methods.

In this chapter, numerous DM techniques implemented within recent years have been examined and updated for prediction of heart disease. One of the greatest strengths that can be extended to various health science issues is the wide range of methodologies and techniques in DM [2]. Scientists have used specific DM approaches such as association rule mining, clustering, classification to improve disease diagnosis with good accuracy and low risk of errors. Existing literature indicates that DM plays an effective role in the predictive mode of heart disease over clustering, association rule and regression through classification. The study also highlights some research using only one DM method for the diagnosis of heart disease, while many of the other studies have used ensemble / hybrid DM methods in the search for better model accuracy and reliability.

Lee et al suggested a numerical and classification-based approach for the creation of a multi-parameter linear and nonlinear heart rate variability (HRV) relationship [3]. Experimental work was carried out using linear and nonlinear HRV parameters while applying Naïve Bayesian, association rules and SVM classifiers. SVM had better accuracy than other classifiers. Tan and Teoh proposed a hybrid approach approach based on a wrapper approach based on classification and GAs. The SVM classified the patterns into desired classes based on the sub-set attribute identified by GA. The data set was used for the UCI Machine Learning Repository and the study showed the GA – SVM hybrid approach's effectiveness. In the multi-class setting, an average accuracy of 84.07 percent improved the efficiency of the GA – SVM hybrid model.

(Polaraju, Durga Prasad, & Tech Scholar, 2017)Proposed Heart Disease Prediction using Multiple Regression Model and it shows that Multiple Linear Regression is suitable for predicting possibility of heart disease. Proposed Heart Disease Prediction using Multiple Regression Model and it shows that Multiple Linear Regression is suitable for predicting possibility of heart disease. The work is carried out using training data set consisting of 3000

instances with 13 different attributes previously mentioned. The data set is divided into two parts, 80% of the data used for training and 20% used for testing [4].

(Deepika & Seema, 2017) ☐Focuses on techniques capable of predicting chronic disease by using Naïve Bayes, Decision Tree, Support Vector Machine (SVM) and Artificial Neural Network (ANN) data from historical health records. In order to measure better performance at an accurate rate, a comparative study is performed on classifiers. SVM gives the highest accuracy rate from this experiment, while Naïve Bayes gives the highest accuracy for diabetes.

(Sai & Reddy, 2017) ☐ Proposed prediction of heart disease using data mining ANN algorithm. There was a need to develop a new method that can predict heart disease due to increased costs of treatment for heart disease. Prediction model is used after evaluation to predict the patient's condition based on different parameters such as heart beat rate, blood pressure, cholesterol, etc. The system's reliability is demonstrated in java [5].

(Sultana, Haider, & Uddin, 2017) Proposed cardiovascular disease study. This paper proposed techniques for data mining to predict the disease. It is intended to provide the overview of current techniques to extract information from the database and will be useful for practitioners of health care. Based on the time taken to build the decision tree for the system, performance can be obtained. The primary goal is to predict the disease with fewer attributes [6].

(A & Naik, 2016) ☐ It is recommended that a predictive system be developed to diagnose heart disease from the medical data set of the patient. 13 The system was considered to be based on risk factors of input attributes. Data cleaning and data integration were performed after analyzing the data from the dataset. To predict heart disease, he used k-means and naive Bayes. This paper is designed to build the system using the diagnostic historical heart database. For the construction of the system, 13 attributes were considered. Data mining techniques such as clustering, classification methods can be used to retrieve information from the database. The Cleveland Heart Database used 13 attributes with a total of 300 files. This model is intended to predict whether or not the patient has heart disease based on 13 attribute values [7].

(Soni, Ansari, & Sharma, 2011) ☐Proposed use of non-linear classification algorithm to predict heart disease. It is suggested to use bigdata tools like Hadoop Distributed File System (HDFS), Map reduction along with SVM for heart disease prediction with optimized set of attributes. This work carried out an investigation into the use of various techniques of data mining to predict heart disease. It suggests using HDFS to store large data in different nodes and use SVM simultaneously to execute the prediction algorithm in more than one node. SVM is used in parallel fashion, resulting in higher computational time than sequential SVM [8].

(Yan and Zheng Proposed) an actual GA-based diagnostic system for heart disease while applying the sub-setting of critical clinical features. The prediction system was designed to identify five major heart diseases, using 352 heart disease data instances with their corresponding diagnostic weights to support or deny each heart disease diagnosis. This provided the functional prediction framework for heart disease with reasonably high accuracy. Based on the machine learning literature and DM, Austin et al developed alternative classification schemes that include bootstrap aggregation bagging, RFs and boosting and SVMs. Research has shown that current DM and ensemble methods offer high precision and performance advantages [9].

Current State of the Art: *(Please describe the current state of the art specific to this research topic. **Maximum 500 words.**)*

It is a huge task, always requiring the involvement of good SMEs from the analyzed field, and allowing to answer questions of what and how, but not when. Entire exercise is based on the fact that all problems faced by man-made systems are the result of past decisions, and how we will solve them now will also determine future problems. There is also typically a limit in the knowledge / bias of small and medium-sized companies based on current technology, if truly unexpected new arrival occurs, it can have a huge impact on the entire picture. But even if the answers vary from major problems, or if the analysis is diligent enough, challenges will remain.
Purusothaman Get al (2015) proposed a hybrid model (accuracy 96%) using Decision Tree (accuracy 76%), Associative Rules (accuracy 55%), K-NN (accuracy 58%), Artificial Neural Networks (accuracy 85%), SVM (accuracy 86%) and Naive Bayes (accuracy 69%),
Srinivas K et al (2010) used Cleveland hospital dataset with 15 attributes and trained multiple models, DT (C4.5) (accuracy 82.5%), Naïve Bayes (accuracy 82%), SVM (accuracy 82.5%), NN (MLP) (accuracy 89.75%), using that dataset and came to a conclusion and proposed a NN (MLP) model with the heights accuracy of 89.75%.
John Peter T et al (2012) used Cleveland hospital dataset with 13 attributes and trained multiple models, Naïve Bayes (accuracy 83.70%), Decision Tress (accuracy 76.66%), K-NN (accuracy 75.18%), NN (accuracy 78.485%), using that dataset and came to a conclusion and proposed a Naïve Bayes model with the heights accuracy of 83.70%.

Challenges: *(Please describe the challenges, specific to this research topic, currently being faced internationally. **Maximum 500 words.**)*

Medical diagnosis is considered to be an important yet complex task that needs to be accurately and efficiently carried out. It would be very useful to automate the same. Clinical decisions are often made based on the expertise and experience of the practitioner rather than the knowledge-rich information that is hidden in the server. This practice leads to unacceptable biases errors and high medical costs that impact patient's quality of service. Data mining has the potential to create a knowledge-rich environment that can significantly enhance the quality of clinical decision-making. Choosing the best algorithm for the purpose of training. Implementing the high Accuracy rate algorithm in system to generate accurate heart disease predictions.

Motivation and Need: *(Please describe the motivation and need for this work. **Maximum 500 words.**)*

Healthcare organizations face a major challenge in providing quality facilities at affordable rates. Quality service includes correct diagnosis of patients and effective treatments. Bad medical decisions can lead to catastrophic and therefore unacceptable consequences. Hospitals also need to reduce the expense of clinical testing. By using effective computer-based data and/or decision support systems, they can achieve these outcomes. Many hospitals today use some kind of hospital information systems to monitor their health care or patient information. Usually, these systems create huge amounts of data in the form of numbers, text, graphs and photographs. These data are, sadly, rarely used to assist medical decision-making. Such data contain a variety of hidden information that is largely unused. This raises a critical question: "How will we have a tendency to turn data into helpful knowledge that helps health care professionals to create smart medical decisions?" This is the main driving force behind this study.

2.2 Outcomes and Benefits

Expected Outcomes: *(Provide a list of proposed project outputs including publications, databases etc.)*

- Full Project Proposal Report with all the details of the project
- Complete Web Application with all feature mentioned above
- Server for Database
- Backup Database

Key Benefits and Beneficiaries: *(Please identify clearly the benefits and potential customers/beneficiaries of the project.)*

Predictive modeling is useful as it offers accurate insight into any problem and allows users to create outcomes. In order to maintain a competitive advantage, insight into future events and results that challenge key assumptions is crucial.

Technology Transfer/Diffusion Approach: *(Please describe how the outputs of the project will be transferred to the beneficiaries/customers. **Maximum 500 words.**)*

An advertisement of the product will be placed on some social and high traffic websites through which user would be able to get more traffic on our heart disease prediction website
The banner is placed in various universities as well.

2.3 Objectives

*(Please describe the measurable objectives of the project and define the expected results. Use results-oriented wording with verbs such as 'to develop..', 'to implement..', 'to research..', 'to determine..', 'to identify..' **The objectives should not be statements and should actually specify in simple words what the project team intends to achieve (something concrete and measurable/ deliverable). Fill only those objectives that are applicable to the proposed project**.)*

Research Objectives: *(if any)*

- Prediction system for heart disease is aimed at exploiting data mining techniques on medical data set to assist in predicting heart disease. Provides new approach to the data's hidden patterns.

Academic Objectives: *(if any)*

- Understanding of different Data Mining Techniques in building new model with high accuracy result.

2.4 Research Approach

Development / Research Methodology:

(Please describe the technical details and justification of your development and research plan. The block diagrams, system flow charts, high level algorithm details etc. have to be provided in this section. ***Maximum 3000 words.****)*

Methodology is a system that includes steps to convert raw data into recognized data patterns to extract user knowledge. The approach suggested involves measures, referred to as the pre-processing stage where the information are thoroughly analyzed. It will deal with missing values, balance information and normalize attributes depending on the algorithms used. Using classification models and EnsembleVoteClassifier, predictive analysis of the data is done after pre-processing of data. Eventually, prescriptive modeling is performed, where different performance metrics are used to test the predictive model in terms of performance and accuracy.

Data set for experiment:

The data set for this research has been taken from the UCI data repository. Used data is freely available from the UCI Machine Learning Repository [10]. The Cleveland data were collected from the above-mentioned DM warehouse. This database includes 76 attributes and 14 attributes have been picked after neglecting redundant and obsolete attributes.

The list of 14 attributes and their brief description are shown in above table. In particular, several researchers used the Cleveland datasets and found that they were appropriate for the creation of a mining model due to lower missing values and outliers. Before they were submitted to the proposed algorithm for training and testing, the data were cleaned and preprocessed. Therefore, 303 are the appropriate instances for the development of supervised machine-learning model building. Attribute selection technique was used for further data reduction to make patterns easier and more comprehensible, but negligible effects were found on model performance observations undertaken in this study. All 13 attributes are considered in order to develop a model for the classifier and to achieve a predictive outcome for heart disease. GaussianNB, MLPClassifier, SVM, RandomForestClassifier algorithm are the classification techniques used in this research. The EnsembleVoteClassifier was used to evaluate the algorithms involved in the classification. The model was built using the Google Colab tool. In these experiments, 5-fold cross-validations were used to divide the data set into training and test sets that meets the model training and testing purpose requirement. As a result, the accuracy rate of this study was over 90%.

S no	Input variables	Description	Options
1	Age	Age in years	Continuous value
2	Sex	1 = male, 0= female	Male, female
3	Cp	Chest pain type	Chest pain type. Values from 1 to 4. 1: typical angina. 2: atypical angina. 3: non-anginal pain. 4: asymptomatic.
4	Trestbps (blood pressure)	Resting blood pressure in mmHg	Continuous value in mmHg
5	Chol (cholesterol)	Serum cholesterol in mm/dL	Continuous value in mm/dL
6	Fbs (fasting blood sugar)	Fasting blood sugar in mg/dL	Fasting blood sugar attributes value "1" for greater than 120 mg/dL, else the attribute value is 0 (false). Value 1 = true. Value 0 = false.
7	Restecg (ECG)	Electrocardiographic results (ECG result)	Resting electrocardiographic results value ranging from 0 to 2. 0: normal. 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV). 2: showing probable or definite left ventricular hypertrophy
8	Thalach (heart rate)	Maximum heart rate	Continuous value categorized into normal and abnormal
9	Exang	Exercise Induced angina	Exercise induced angina. Values from 0 to 1. Value 1 = yes. Value 0 = no.
10	Old peak	ST depression induced by exercise relative to rest	Continuous values
11	Slope	Slope of the peak exercise ST segment	Measure of slope for peak exercise. Values can be 1, 2, or 3. Value 1: up sloping. Value 2: flat. Value 3: down sloping.
12	Ca	Number of major vessels colored by fluoroscopy	Number of major vessels from 0 to 3
13	Thal	Heart rate of patient	Represents heart rate of the patient. It can take values 3, 6, or 7. Value 3 = normal. Value 6 = fixed defect. Value 7 = reversible defect.
14	Class	Class labels (predicted outcome)	Contains a numeric value between 0 and 1. Each value represents heart disease or absence of disease. Value 0: absence of heart disease. Value 1: presence of heart disease.

Classifiers used for experiment

GaussianNB:
A special type of NB algorithm is a Gaussian Naive Bayes algorithm. It is specifically used when there are continuous values for the features. It is also assumed that all features follow the normal distribution of a gaussian distribution.

Bayes' Theorem
Theorem of Bayes considers the probability of an event occurring given the likelihood of another event already occurring. Mathematically, the theorem of Bayes is stated as the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where are events A and B and events P(B)? 0.

- Basically, given that event B is true, we are trying to find the probability of event A. Event B is also known as proof.
- P(A) is the priori of A (previous probability, i.e. probability of occurrence before proof is seen). The proof is an undefined instance's attribute value (here it is case B).
- P(A) is a posteriori probability of B, i.e. probability of occurrence after proof.

Now, with regard to our dataset, Bayes ' theorem can be applied as follows:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

Where y is a variable of class and X is a vector of dependence (size n) where:

$$X = (x_1, x_2, x_3, ..., x_n)$$

MLPClassifier:
A multilayer perceptron (MLP) is an artificial neural feedforward network producing a collection of outputs from a set of inputs. Multiple layers of input nodes connected as a direct graph between the input and output layers characterize an MLP. MLP uses network learning backpropogation. MLP is a form of deep learning

SVM:
SVM is a rough realization of structure-based risk minimization and linear / non-linear data categorization. SVMs optimize the dividing hyperplane gap. Each training sample sub-set (support vectors) defines boundary decision functions. SVM includes three steps, the creation of the support vector, the formation of the maximum distance between the found points and the boundary of the perpendicular decision.
For many practical problems, the maximum margin linear classifier is an inapplicable approach. For practical applications, where the non-linear separable data set is to be separated by hyperplane, the non-linear data is to be mapped via kernel functions to another feature space. The data points are separated by hyperplane in the original input space, and the functions of the kernel map non-linear training samples to high-dimensional space. A search for the best hyperplane is then performed to split the transformed data into two different classes. For classification purposes, the margin of the hyperplane is maximized thus reducing classification errors. For multi-dimensional hyperplane, the algorithm predicts the risk of heart disease and optimally categorizes the data into different labels by generating the margin between data clusters.

RandomForestClassifier:
RandomForestClassifier, one of the most accurate algorithms for machine learning, is a decision-based ensemble classifier approach that contains a flowchart like a tree structure. RF is a combination of { h(x, n) } tree-structured classifiers where random trees for data classification are distributed for "x" data input and "n.". The decision tree in a random tree-structured forest casts a vote indicating the knowledge group decision. RF uses the Gini index to assess that tree's final category. This algorithm selects optimal attributes for each tree from the total number of input attributes "M." Using this selected attribute to develop a decision tree template, the best possible split is generated using the Gini index. For each of the branches, this is an iterative process until the terminating nodes are too small to further break. For data set x classes with "n", You can define Gini-index, Gini(x) by:

$$Gini(X) = \sum_{i=1}^{n} (Rj)^2$$

Where "Rj" in the data set "X" is the relative frequency of class j The split with the lowest Gini index is picked at the split value in RF.

AdaBoostClassifier:
AdaBoost, short for Adaptive Boosting, is a meta-algorithm for machine learning developed by Yoav Freund and Robert Schapire, who received the Gödel Prize for their research in 2003. It can be used to improve performance in conjunction with many other forms of learning algorithms.

EnsembleVoteClassifier:
Ensemble method is a well-proven technique used in research to achieve highly accurate data classification by hybridizing multiple classifiers in order to achieve more reliable and accurate prediction results. The enhanced quality of prediction is a well-known in-built ensemble methodology feature. The EnsembleVoteClassifier is a meta-classifier for combining, for majority or plurality voting, similar or conceptually different machine learning classifiers. Through hard voting, we predict the label of the final class as the class label that the classification models predicted most frequently. Hard voting is where a model to make the final prediction by a simple majority vote for accuracy is chosen from an ensemble. Soft Voting can only be achieved if all the classifiers can measure the outcomes probabilities.

Key Milestones and Deliverables:

*(Please list and describe the principal milestones and associated deliverables of the project. The timing of milestones is also to be shown in the Gantt chart in Annexure-A. **Quarterly deliverables are preferred**.)*

S. No.	Elapsed time since start of the project	Milestone	Deliverable
1.	Week 01 – 02	Project Idea	Report on approved idea.
2.	Week 03 – 06	System Requirements Gathering.	Functional/nonfunctional
3.	Week 07 – 10	Designing System Diagrams	Diagram (UML, ERD)
4.	Week 10 – 13	Logo Design, Product name and prototypes.	Logo and Prototype
5.	Week 13 – 15	Complete Report	Final Report and Poster
6.	Week 16 – 20	Implementation	-
7.	Week 20 – 22	Implementation	-
8.	Week 22 – 24	Implementation	-
9.	Week 24 – 25	Testing	Complete system with report (user and technical)
10.	Week 26 – 27	Delivery/Training	Delivery/Training of project to end user to handle the system

(Please add more rows if required.)

2.5 Risk Analysis

(Please list the risks that may cause delays in, or prevent implementation of, the project. For each risk estimate the likelihood, likely impact/consequences on the project and steps to minimize/avoid the risk.)

Risk	Likelihood (Low, Med, High)	Impact	Mitigation
Project completion delays	Med	Serious	Paying a lot of attention to project planning and Constantly track and measure the progress.
The budget is not enough / exceeds	Med	High	Place a condition in the contract if more expenses are incurred, it must be covered by the funded party to avoid the risk.
Lack of team members cooperation	Med	High	Tools for project management will be used to support individual activities.
Technology	High	Critical	Best performance servers will be used for hosting of the website.

Section – 3

3.1 Resources & Other Requirements

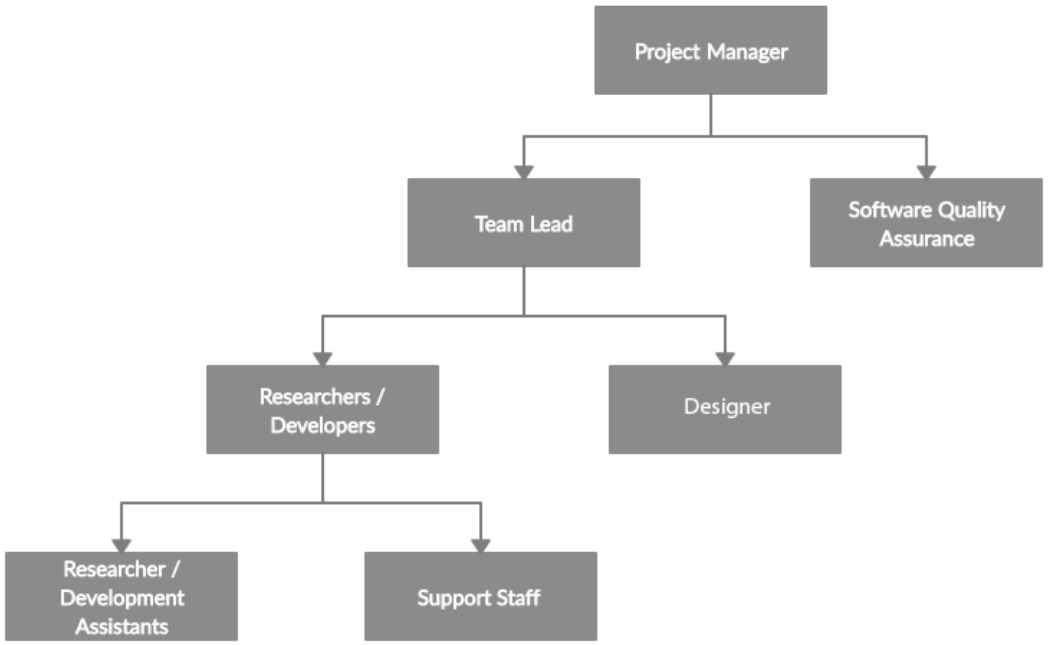
Project Team:

The numbers in the table below must tally with the HR Cost sheet in the Budget file.

Title / Position	*Number*
Project Manager	1
Team Leads	1
Researchers / Developers	2
Researcher / Development Assistants	1
Designers	1
Support Staff	1
Software Quality Assurance	1
Contract Staff (please specify)	
Others (please specify)	
Add more rows if required	

Team Structure:

(Please define the team structure (organogram) and role/key responsibilities of each member. If in collaboration with another partner, the division of manpower at various locations of partners be provided.)



Title/Position (of each member)	Role/Key Responsibilities	Minimum Qualification Required	Expertise / Background Required	Minimum Experience Required (years)
Project Manager	Overall responsibility for a successful planning, manage all processes in the project	Masters	Leadership, Negotiation, Scheduling, Risk Management, business skills.	4
Team Leads	Team leader will lead a team within the project	Master	Strong Organization Skills, Confident in the Team, Respectful to Others, Fair and Kind	3
Researchers / Developers	Research about the project and new innovation and ideas/ Develop back end, algorithm	MS(SE)	Attention to detail ,Technical skills, Critical thinking, Communication, Planning and scheduling / Web Development	3
Researcher / Development Assistants	Analyze and evaluate clinical data gathered during research.	BS(SE)	Technical skills, Critical thinking, Communication, Planning and scheduling / Web Development	3
Designers	Develop IU graphic Design	BS(CS)/ BS(SE)	Web Designing /Graphic Designing	3
Support Staff	Supports the crucial front line, those agents who are directly interacting with customers, fielding their inquiries and solving problems so they remain your customers.	BS(CS)/ BS(SE)	Time management skills, Attentiveness, Knowledge of the Product.	2
Software Quality Assurance	For performing software life cycle, and for testing overall system	BS(SE)	SDLC process, Quality Engineering	4

Remarks:

1. Name & Signature of Supervisor: _____

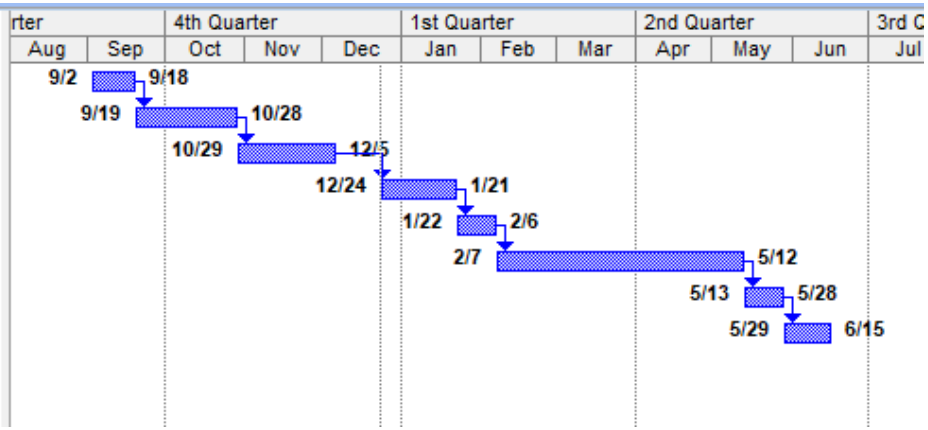
2. Name & Signature of Co-Supervisor: _____

3. Name & Signatures of FYP Coordinator: _____

Annexure–A: Project Schedule / Milestone Chart

(Project schedule using MS-Project (or similar tools) with all tasks, deliverables, milestones, cost estimates, payment schedules clearly indicated are preferred.)

	Task Name	Duration	Start	Finish	Predecessors
	Project Idea	13 days	Mon 9/2/19	Wed 9/18/19	
	System Requirements Gathering	28 days	Thu 9/19/19	Mon 10/28/19	1
	Designing System Diagrams	28 days	Tue 10/29/19	Thu 12/5/19	2
	Logo Design, Product name and prototypes.	21 days	Tue 12/24/19	Tue 1/21/20	3
	Complete Report	12 days	Wed 1/22/20	Thu 2/6/20	4
	Implementation	68 days	Fri 2/7/20	Tue 5/12/20	5
	Testing	12 days	Wed 5/13/20	Thu 5/28/20	6
	Delivery/Training	12 days	Fri 5/29/20	Mon 6/15/20	7



Annexure–B: Proposed Budget

Please use the embedded Excel Worksheet for providing budget details.

Double click the icon to open the worksheet.

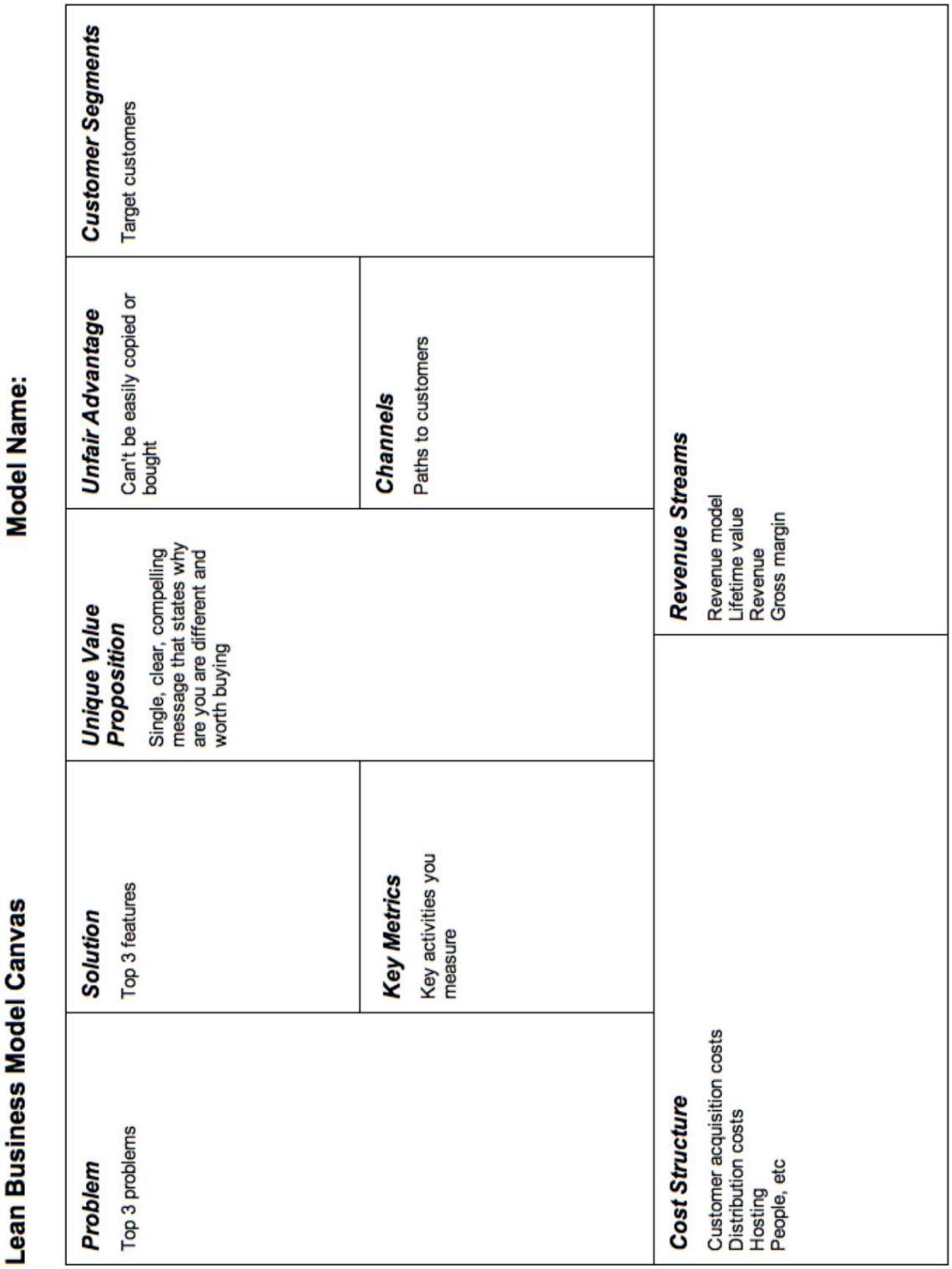
(Do not enter the summary amounts (Sr.# 1-6) as they are automatically updated S

Sr.	Description	Amount (Rs)
	Heads of Expenditure	
1	Technical HR Deployment Cost	Rs.60,000
2	Support Staff	Rs.150,000
3	Equipments	Rs.452,606
4	Traveling	Rs.90,000
5	Boarding & Lodging	Rs.180,000
6	Miscellaneous	Rs.102,000
	Sub Total:	**Rs.1,034,606**
7	**Audit Charges**	Rs.25,000
8	**Contingency**	Rs.50,000
9	**Institutional/Organizational Overheads**	Rs.200,000
	Total Budget:	**Rs.1,309,606**

Funding Sources: *(Please indicate funding sources for the project)*

Sr.	Funding Source	Amount (Rs)
1	National ICT R&D Fund	Rs.1,309,606
2	Internal Funds	
3	Other Sources (specify)	
4	(Add more entries, if required)	
	Total:	**Rs.1,309,606**

Annexure–C: Business Canvas Model



Bibliography

[1] L. NM, "Data Mining for Cancer Management in Egypt Case Study," 2007. [Online].

[2] L. HG, N. KY and R. KH, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV. In: Washio T, Zhou ZH, Huang JZ, et al, editors. Emerging Technologies in Knowledge Discovery and Data Mining. PAKDD 2007, Lecture Notes in Computer Scien," 2007. [Online].

[3] C. Blake and . C. Mertz, "UCI Machine LearningDatabases," 2004. [Online]. Available: http://mlearn.ics.uci.edu/databases/heartdisease/.

[4] K. Polaraju, D. Durga Prasad and . M. Tech Scholar, "Prediction of Heart Disease using Multiple Linear Regression Model.," 2017. [Online]. Available: International Journal of Engineering Development and Research, 5(4), 2321–9939. Retrieved from www.ijedr.org.

[5] P. P. &. R. C. Sai, "HEART DISEASE PREDICTION USING ANN ALGORITHM IN DATA MINING.," 2017. [Online]. Available: International Journal of Computer Science & Mobile Computing, 6(4), 168–172. Retrieved from www.ijcsmc.com.

[6] M. H. A. &. U. Sultana, "Analysis of data mining techniques for heart isease prediction. In 2016 3rd International Conference n Electrical Engineering and Information and Communication Technology, iCEEiCT 2016 (pp. 1–5).," 2017. [Online]. Available: https://doi.org/10.1109/CEEICT.2016.7873142.

[7] A. S. &. N. C. A, "Different Data Mining Approaches for Predicting Heart Disease, 277–281.," 2016. [Online]. Available: https://doi.org/10.15680/IJIRSET.2016.0505545.

[8] J. A. U. &. S. D. Soni, "Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers. Heart Disease," 2011. [Online]. Available: Analysis of data mining techniques for heart isease prediction. In 2016 3rd International Conference n Electrical Engineering and Information and Communication Technology, iCEEiCT 2016 (pp. 1–5). https://doi.org/10.1109/CEEICT.2016.7873142.

[9] Z. J. Yan H, "Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm.," 2008. [Online]. Available: J Appl Soft Comput..

[10] U. M. L. Repository., Accessed November 01, 2018. [Online]. Available: Available from: https://archive.ics.uci.edu/ml/index.php. .