**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Rohit Kumar Sharma [rohitraw6@gmail.com](mailto:rohitraw6@gmail.com) |
| **Please paste the GitHub Repo link.** |
| GitHub Link: - https://github.com/Rohitkumar1011/Cardiovascular-Risk-Prediction |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| **Summary**  Heart disease is the leading cause of morbidity and mortality worldwide, killing more people each year than any other cause. For example, an estimated 17.9 million people died from heart disease in 2016, accounting for 31% of all global deaths. More than three-quarters of these fatalities occurred in low- and middle-income countries.  Most heart diseases are highly preventable, and simple lifestyle changes (such as quitting smoking, eating healthily, losing weight, and exercising) combined with early treatment greatly improve prognoses. However, identifying high-risk patients is difficult due to the multi-factorial nature of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, and so on. Because of these constraints, scientists have turned to modern approaches for disease prediction such as Data Mining and Machine Learning.  Machine learning (ML) is useful in supporting decision making and risk assessment from the huge amount of data produced by the healthcare industry on heart disease due to its supremacy in pattern identification and categorization.  I have used data from an ongoing cardiovascular study of Framingham, Massachusetts, inhabitants. The categorization purpose is to determine whether the patient will develop coronary heart disease in the next ten years (CHD). The dataset contains information on the patients. It has about 4,000 records with 15 qualities. Variables Each characteristic is a possible risk factor. There are risk variables that are demographic, behavioural, and medical in nature.  I have handled missing values at very start because it can cause error, then I have used Boruta algorithm to select top features. In this analysis I have used four algorithms namely: - Logistic Regression, Decision Tree, K nearest neighbour, Support vector machine.  Dataset which I have used is imbalanced, so with help of oversampling technique smote I have balanced it. Feature scaling is also done which rescale data distribution to mean of zero and standard deviation of 1. Rescaling increases the speed of algorithms and reduce computation time.  This model should highly accurate because it is related to healthcare industry that is why I have used four algorithm and compared them precisely. Support vector machine is showing very good accuracy according to various evaluation metrics this is the reason SVM is our final model. |