

**CIS 666 – DATA MINING**

**Course project proposal on Prediction of Cardiovascular Disease Risk**

Submitted by

Pooja Reddy Arava - 2871945

Dharmatejaswi Gogineni - 2889146

1. **INTRODUCTION :**

Our evaluation has the data set, which focuses on the risk prediction of cardiovascular disorders, was taken from Kaggle. Our goal in using this information is to forecast each person's risk of cardiovascular illnesses based on their specific data.

1. **Data Sets:**

There are 308,855 rows and 19 columns in this dataset, which was obtained via Kaggle. General Health, Checkup, Exercise, Heart Disease, Skin Cancer, Other Cancer, Depression, Diabetes, Arthritis, Sex, Age Category, Height (cm), Weight (kg), BMI, Smoking History, Alcohol Addiction, Fruit Intake, Green Vegetable Intake, and Fried Potato Intake are among the variables included in the dataset.

The target factors of interest for this predictive study are diabetes, gender, and overall health. These factors were chosen because they were thought to be important in determining the risk of cardiovascular disease. The "Cardiovascular Diseases Risk Prediction Dataset" dataset was gathered from the Kaggle platform <https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset/data>.

**3. Issues with Research :**

1. This study seeks to provide insights into the best models for predicting cardiovascular disease (CVD) risk using individual lifestyle characteristics by comparing the efficacy of several machine learning algorithms.

2. The study shows that hyperparameter adjustment is important for improving the Logistic Regression model's efficacy. This demonstrates how important it is to choose the best model setups.

3. The primary objective of the study is to determine important personal characteristics that might provide medical practitioners with important information about the particular factors that contribute to the risk of cardiovascular disease (CVD). This enables them to develop individualized prevention measures and focused intervention techniques.

1. **Potential Solutions :**
2. **Comparing Machine Learning Algorithms for CVD Risk Prediction:**

First, we clean and organize the data to handle any missing values or unusual data points. Then, we look into various lifestyle factors that could affect cardiovascular disease (CVD) risk and select the most relevant ones. After that, we test out different machine learning algorithms like logistic regression, decision trees, random forest, and support vector machines to see which one performs best in predicting CVD risk based on these factors. We use techniques like N-fold cross-validation to make sure our models are reliable and evaluate them using metrics like accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC-ROC). Finally, we analyze the results to see which model gives us the most accurate predictions and better insights into CVD risk based on lifestyle factors.

**B . Hyper parameter Tuning for Logistic Regression:**

For logistic regression, we fine-tune the model's parameters using methods like grid search or random search to find the best combination. We use cross-validation to ensure that the parameters we choose work well with new data and prevent our model from becoming too specialized to the training data. Then, we evaluate the performance of our tuned logistic regression model using appropriate metrics, making sure it performs well across different scenarios. Additionally, we conduct sensitivity analysis to understand how changes in parameters affect the model's performance.

**C . Identifying Significant Personal Attributes for CVD Risk Prediction:**

Here, we focus on understanding which personal attributes have the most impact on predicting CVD risk. We analyze the importance of each attribute using methods like examining coefficient magnitudes in logistic regression or feature importance scores in decision trees/random forest.

Furthermore, we conduct statistical tests like t-tests or ANOVA to determine the significance of these attributes. We then interpret the results and visually represent the relationships between personal attributes and CVD risk using charts or plots. Finally, we provide healthcare professionals with actionable insights based on our findings, helping them tailor intervention strategies and preventive measures for individuals based on their specific risk factors.

1. **Evaluations :**

In our project, we will employ N-fold cross-validation to evaluate the performance of our predictive models for predicting cardiovascular disease risk. By utilizing this technique, we aim to obtain reliable estimates of model performance and ensure the robustness of our findings. Additionally, it will enable us to compare the effectiveness of different models and make informed decisions about model selection and hyper parameter tuning.

1. **Expected Outcomes :**

In this project, we've conducted a comprehensive analysis aimed at predicting the risk of cardiovascular disease (CVD) by considering various lifestyle factors of individuals. We began by thoroughly preparing our dataset for analysis, which involved cleaning up any missing values, handling outliers, and making sure our features were scaled appropriately. Next, we explored the effectiveness of different machine learning algorithms such as logistic regression, decision trees, random forest, and support vector machines in predicting CVD risk. By comparing these algorithms, we gained insights into which ones performed better in this task. We also paid close attention to the importance of fine-tuning the parameters of our logistic regression model to enhance its effectiveness. Through careful evaluation using techniques like N-fold cross-validation and relevant evaluation metrics, we were able to assess the performance of our models accurately. Moreover, our analysis didn't stop at just predicting risk. We delved deeper into understanding the significant personal attributes that contribute to CVD risk. By leveraging techniques like feature importance analysis and statistical tests, we pinpointed which attributes had the most significant impact on predicting CVD risk. Ultimately, our project aims to provide healthcare professionals with actionable insights derived from our analysis. By understanding the key factors contributing to CVD risk, they can tailor intervention strategies and personalized preventive measures for individuals, potentially helping to reduce the prevalence of cardiovascular diseases.