Kathmandu University

Department
of
Electrical and Electronics Engineering

Heart Disease Prediction

ETEG 425 Mini Project

By:

Aavash Shrestha (41025)

Date: 10 September 2024

Presentation Outline

- Introduction
- Objectives
- Workflow
- Dataset
- Methodology
- Result and Analysis
- Conclusion

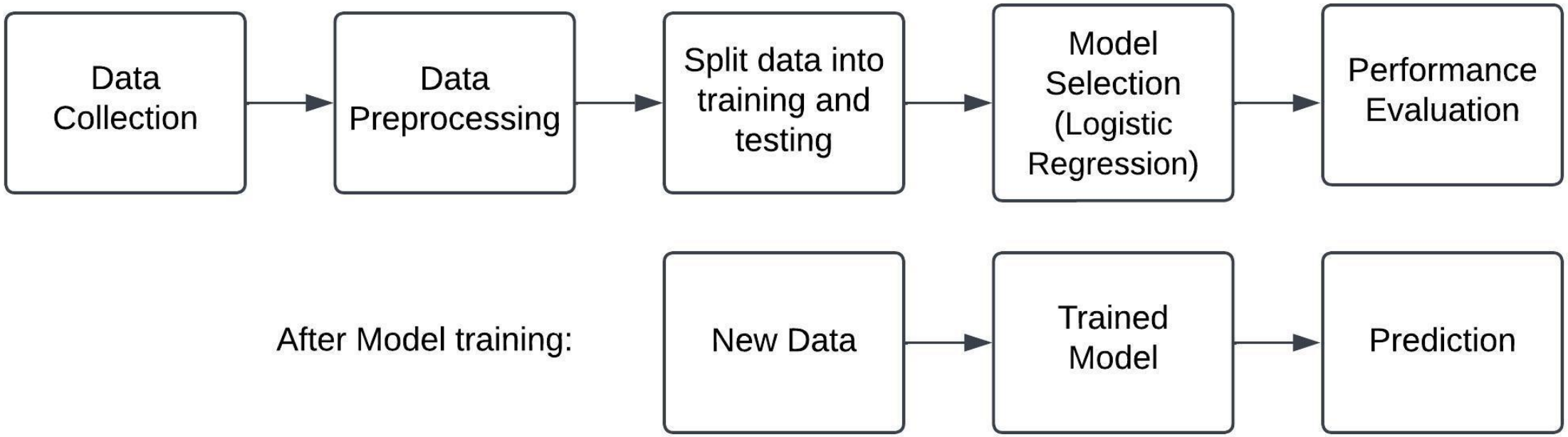
Introduction

- Heart disease is a leading cause of death globally. Accurate prediction can significantly reduce risks by enabling early diagnosis.
- Traditional methods for diagnosing heart disease often involve expensive and invasive procedures. However, machine learning (ML) can provide an opportunity to analyze vast amounts of medical data to predict heart disease more efficiently and accurately.
- By leveraging clinical data, such as patient demographics, medical history, and diagnostic test results, ML algorithms can help predict the likelihood of heart disease

Objective

- To develop a predictive model using logistic regression to determine whether a patient is likely to have heart disease based on input features such as age, cholesterol levels, blood pressure, and other relevant health metrics. Additionally, the project evaluates the impact of different learning rates on model performance to optimize its training process.

Workflow


Data Collection
Data Preprocessing
Split data into training and testing
Model Selection (Logistic Regression)
Performance Evaluation
After Model training:
New Data
Trained Model
Prediction

Dataset

- The data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. **It is integer valued 0 = no disease and 1 = disease.** [1]

- **Attributes:** Age, sex, chest pain type (4 values), resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar > 120 mg/dl, resting electrocardiographic results (values 0,1,2), maximum heart rate achieved, exercise induced angina,oldpeak = ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0-3) colored by fluoroscopy, thal: 0 = normal; 1 = fixed defect

Methodology

1. Data Preprocessing:

- **Splitting the dataset** into training and testing sets (80% training, 20% testing).
- **Scaling the features** manually to normalize the data, ensuring numerical stability and better convergence of the model [2].

Equation Used for scaling:

$$X_{mean} = \frac{X - mean(X)}{std(X)}$$

Subtracting the mean (mean(X)) shifts the data so that its average becomes 0 and Dividing by the standard deviation (std(X)) rescales the data so that it has unit variance (a standard deviation of 1).

2. Logistic Regression Implementation:

- The sigmoid function was used to map the model's output to probabilities between 0 and 1. The sigmoid function is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

- Cross-entropy loss was used as the cost function.
- Gradients of the weights and bias were calculated to update the model parameters using gradient descent.

$$Loss = -\frac{1}{m}\sum_{i=1}^{m}(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i))$$

Where:

- M = number of training samples; y_i = true labeled (0 or 1) for the i^{th} sample
- $\hat{y}_i = sigmoid(z_i) = predicted\ probability\ for\ the\ i^{th}\ sample$; z_i = linear combination of weights, input features and bias

Numerical Instabilities in Loss Calculation:

- The loss calculation resulted in inf due to extreme values passed to the logarithmic function.
- Solution: Applied clipping to the sigmoid outputs to ensure numerical stability:

$$y_{predicted} = max(\epsilon, min(1 - \epsilon, y_{predicted}))$$

where $\epsilon = 1 \times 10^{-10}$ (a small positive constant)

- Clipping ensures that: $\epsilon \leq y_{predicted} \leq 1 - \epsilon$

Methodology

3. **Hyperparameters:**

- Learning Rate: 0.1, 0.01, 0.01
- Epochs: 10,000 iterations

4. **Performance Evaluation:**

- Training and test accuracy were calculated to evaluate the model's effectiveness.
- The training loss and accuracy were tracked over epochs to visualize model convergence.

5. **Visualization:**

- Graphs of training loss and training accuracy over epochs were plotted to ensure proper learning behavior

Result and Analysis

1. **Scaling:** The manual scaling of data ensured that all features were normalized to have a mean of 0 and a standard deviation of 1, avoiding issues with gradient calculations during training.
2. **Model Training:**

- The model's weights and bias were optimized using gradient descent.
- The training loss consistently decreased over epochs, indicating proper learning.
- The training accuracy steadily increased, suggesting that the model captured patterns in the training data.

3. **Model Performance:** (for different values of learning rate)

- Training Accuracy: ~80-85%
- Test Accuracy: ~60-65%

4. **Prediction:** For a new patient input, the model predicted whether the individual had heart disease.

Varying Learning Rate

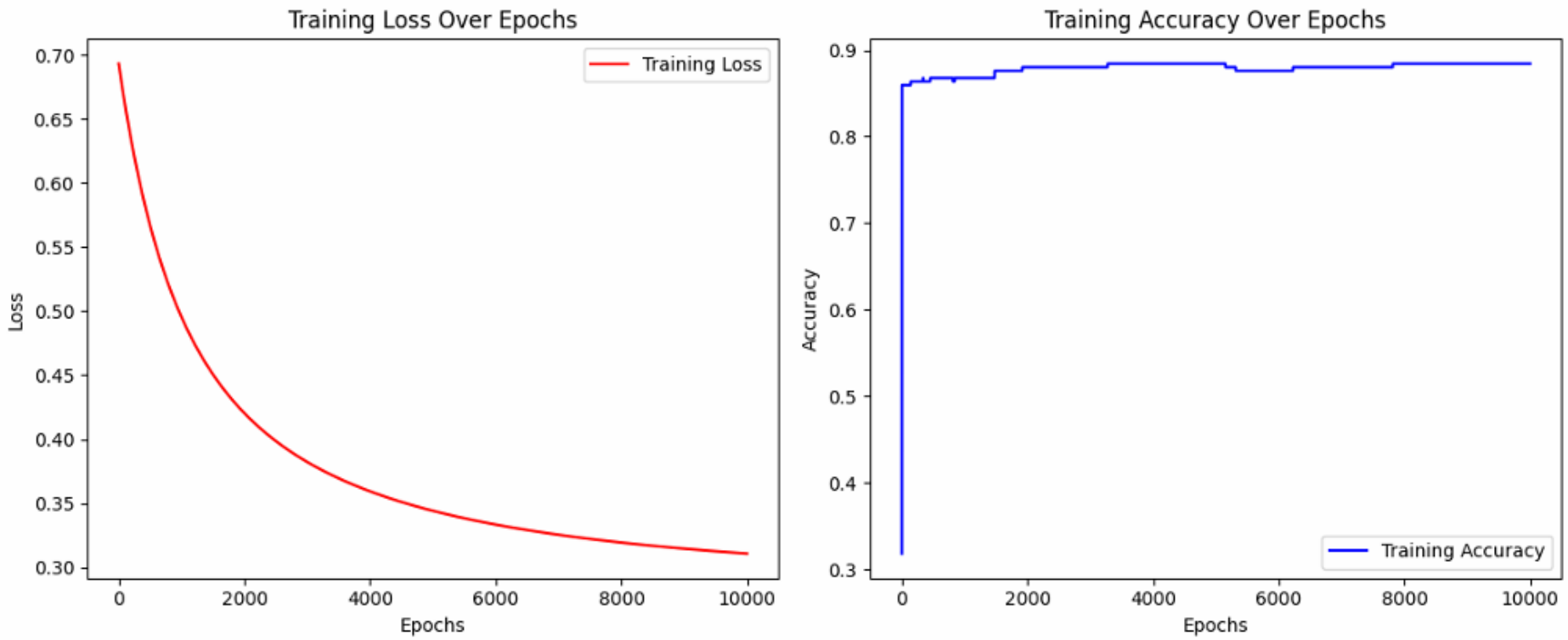


Figure 1 . Training loss and accuracy (learning rate=0.001)

- The loss decreases very slowly and steadily over the epochs
- Accuracy increases steadily but much more slowly than with the higher learning rates.

Varying Learning Rate

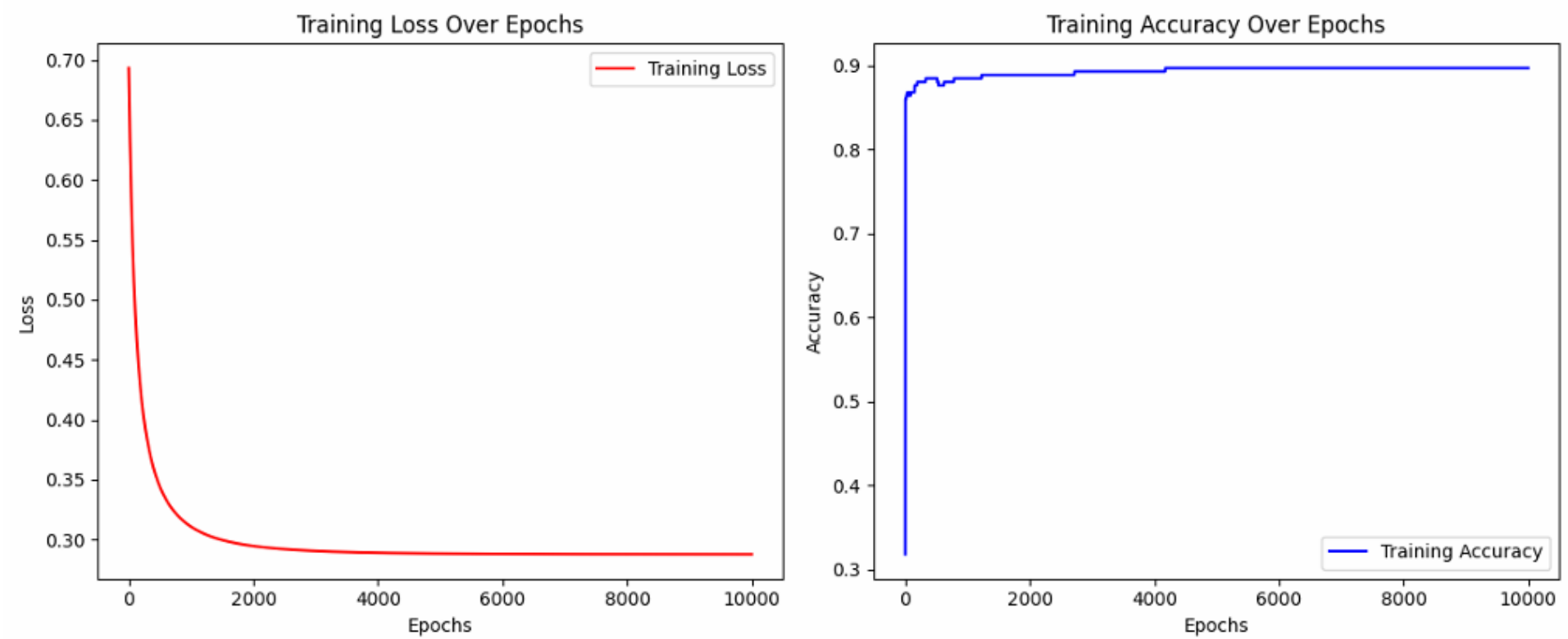


Figure 2 . Training loss and accuracy (learning rate=0.01)

- The loss decreases faster than 0.001 and stabilizes around 0.3 within 10,000 epochs.
- Accuracy improves faster than 0.001 and stabilizes earlier, closer to 90%.

Varying Learning Rate

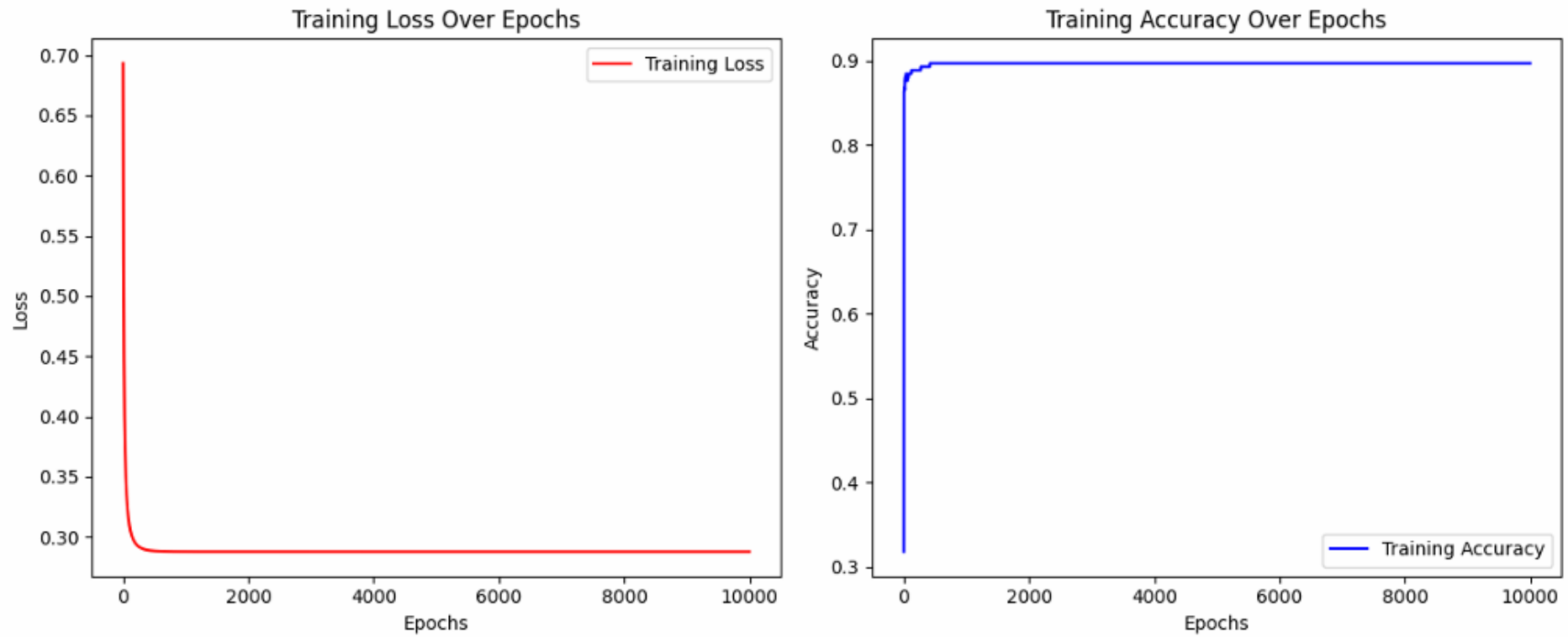


Figure 3 . Training loss and accuracy (learning rate=0.1)

- The loss decreases very rapidly and stabilizes much earlier compared to both 0.01 and 0.001.
- Accuracy reaches 90% the fastest, stabilizing within the first few thousand epochs.

Varying the number of epochs

For different numbers of epochs, the graphs didn’t show any major changes for the same value of learning rate which indicates that the model reaches its optimal performance well before the maximum number of epochs is reached. Some potential reasons for this behavior are:

1) Model Converges Early

- The learning process stabilizes early, and the model's weights reach their optimal values before the later epochs are utilized.
- For example, if the model converges within the first 2,000 epochs, extending the training to 10,000 epochs won’t make a difference in the graph.

Varying the number of epochs

2. Dataset Simplicity

- If the dataset is relatively simple or small, the model might learn everything it needs to early in training.
- In such cases, extending the number of epochs would not affect the performance graphs because the model already achieves optimal performance.

Conclusion

- The logistic regression model provided a reliable and interpretable approach for predicting heart disease. While the model achieved reasonable accuracy, there is potential for further improvement using advanced techniques or additional features. Nevertheless, this project demonstrates the feasibility of using machine learning for healthcare applications.

Refernces

[1] Dua, D., & Graff, C. (2019). UCI Machine Learning Repository: Heart Disease Dataset. University of California, Irvine, School of Information and Computer Sciences. Available at: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[2] Hastie, T., Tibshirani, R., & Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2009.

[3] Goodfellow, I., Bengio, Y., & Courville, A., *Deep Learning*. MIT Press, 2016.

[4] American Heart Association, Understanding Heart Disease Risk Factors. Available at: https://www.heart.org/, 2021.

Thank You for Listening!

For further information:

Aavash Shrestha (41025)
2020

aavashsth@gmail.com