EDUNET FOUNDATION FINAL PROJECT

IMDB MOVIE REVIEWS ANALYSIS

Presented By:
Name: AYUSH DUTTA
Internship ID:INTERNSHIP_17127310326616339 8a8c87
AICTE Student ID:STU6640d104320f31715523844
Institute: BENGAL ENGINEERING AND SCIENCE UNIVERSITY, SHIBPUR
Degree: BACHELOR OF TECHNOLOGY (4YEARS UG PROGRAM)
Student's Email ID: dutta7759@gmail.com

edunet foundation

OUTLINE

- Problem Statement
- Proposed System/Solution
- System Development Approach
- Algorithm & Deployment
- Result
- Conclusion
- Future Scope
- References
- IBM SKILLSBUID course completion certificate

PROBLEM STATEMENT

A Movie dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. We are provided with a set of 50,000 highly polar movie reviews for training and for testing. So, predict the number of positive and negative reviews using either classification or deep learning algorithms.

PROPOSED SOLUTION

The proposed system aims to address the challenge of sentiment classification for IMDb movie reviews, predicting whether a review is positive or negative. This involves leveraging data analytics and machine learning techniques to accurately classify sentiments based on textual reviews. The main concept used during solving this problem was LOGISTIC REGRESSION CONCEPT . The solution will consist of the following components:

- **Data Collection:**

  - Gather a dataset of IMDb movie reviews labeled with sentiments (positive or negative).

  - Include metadata such as review text, sentiment label, and potentially additional features like movie genre, release year, etc.

- **Data Preprocessing:**

  - Clean and preprocess the collected data to handle text-specific challenges such as punctuation, stopword , and capitalization.

  - Perform tokenization and vectorization of text data using techniques like TF-IDF to convert text into numerical features suitable for machine learning models.

- **Machine Learning Algorithm:**

  - Implement a classification algorithm, such as Logistic Regression, to predict sentiment labels based on the processed text features.

  - Explore deep learning techniques, such as LSTM (Long Short-Term Memory) networks, for capturing sequential dependencies in reviews for potentially improved performance.

PROPOSED SOLUTION

- **Deployment:**

- Develop a user-friendly interface or application that allows users to input a movie review and receive real-time sentiment predictions (positive or negative).

- Deploy the solution on a scalable and accessible platform, ensuring robust performance and user responsiveness.

- **Evaluation:**

- Assess the model's performance using standard metrics like accuracy, precision, recall, and F1-score.

- Validate the model's generalization capability using cross-validation techniques and consider fine-tuning parameters to optimize performance.

- **Result:**

The system aims to provide an effective tool for sentiment analysis of IMDb movie reviews, enabling users to gauge audience reactions accurately. By integrating data preprocessing, machine learning algorithms, and user-friendly deployment, the solution seeks to enhance decision-making in the entertainment industry, facilitating informed strategies for movie marketing, production, and audience engagement.

SYSTEM APPROACH

SYSTEM REQUIREMENTS-

1.A standard laptop or desktop computer with at least 8GB of RAM and a multi-core CPU.
2.Adequate storage capacity for storing datasets and model artifacts.
3.Python 3.x: The programming language used for development.
4.Integrated Development Environment (IDE) like Jupyter Notebook or PyCharm for coding and experimentation.
5.Package management system (pip or conda) for installing necessary Python libraries.

SYSTEM APPROACH

LIBRARIES USED-

- **Pandas**: Data manipulation and analysis library for handling datasets effectively.
- **Scikit-learn**: Machine learning library providing tools for data mining and data analysis. It includes various classification, regression, and clustering algorithms.
- **Matplotlib**: A comprehensive library for creating static, animated, and interactive visualizations in Python.
- **Seaborn**: A data visualization library based on Matplotlib, providing a high-level interface for drawing attractive and informative statistical graphics.

ALGORITHM & DEPLOYMENT

THE BASIC ALGORITHM USED IS LOGISTIC REGRESSION.

Logistic regression is a statistical method used for binary classification, predicting the probability of a binary outcome based on input features. It models the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function.

1. Collecting the data and loading the data

```
import pandas as pd

# Loading the dataset
df = pd.read_csv('/IMDB Dataset.csv')

# Display the first few rows and last few rows and check the columns
print(df.head())
print(df.tail())
print(df.columns)
```

```
                                              review sentiment
0  One of the other reviewers has mentioned that ...  positive
1  A wonderful little production. <br /><br />The...  positive
2  I thought this was a wonderful way to spend ti...  positive
3  Basically there's a family where a little boy ...  negative
4  Petter Mattei's "Love in the Time of Money" is...  positive
                                              review sentiment
49995  I thought this movie did a down right good job...  positive
49996  Bad plot, bad dialogue, bad acting, idiotic di...  negative
49997  I am a Catholic taught in parochial elementary...  negative
49998  I'm going to have to disagree with the previou...  negative
49999  No one expects the Star Trek movies to be high...  negative
Index(['review', 'sentiment'], dtype='object')
```

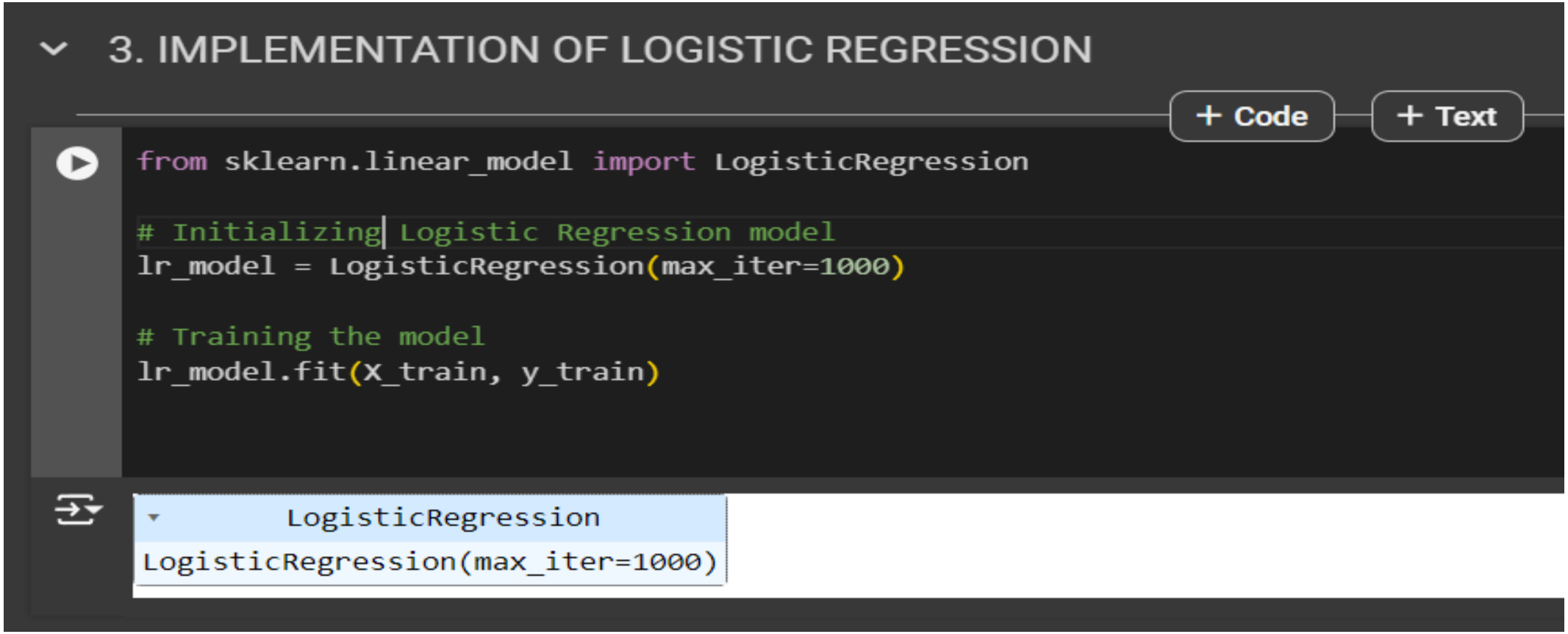
edunet foundation

ALGORITHM & DEPLOYMENT

2. Understanding the dataset

```
[5]  from sklearn.feature_extraction.text import CountVectorizer
     from sklearn.model_selection import train_test_split

     # Vectorizing the text data
     vectorizer = CountVectorizer(stop_words='english')
     X = vectorizer.fit_transform(df['review'])

     # Converting sentiment labels to binary labels (0 for negative, 1 for positive)
     y = df['sentiment'].map({'negative': 0, 'positive': 1})

     # Spliting data into training and testing sets
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=2)
```

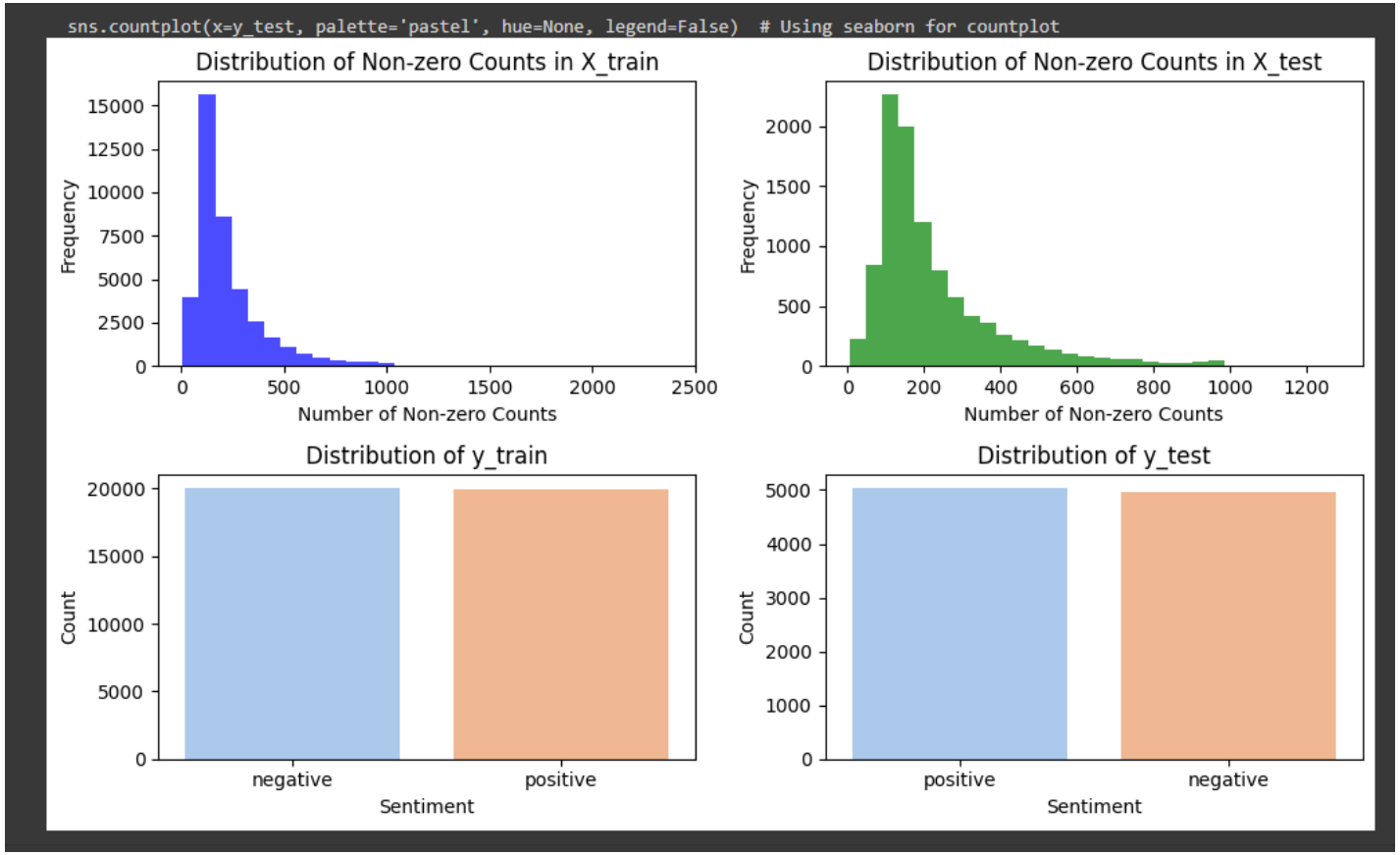
ALGORITHM & DEPLOYMENT



RESULT

4. PROPER VISUALISATION

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.linear_model import LogisticRegression
import matplotlib.pyplot as plt
import seaborn as sns

# Example data
data = pd.read_csv('/content/IMDB Dataset.csv')
df = pd.DataFrame(data)

# Splitting data
X = df['review']
y = df['sentiment']

# Vectorizing the text data
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(X)

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Logistic Regression model
lr = LogisticRegression()
lr.fit(X_train, y_train)

# Visualization
plt.figure(figsize=(10, 6))

plt.subplot(2, 2, 1)
plt.hist(X_train.sum(axis=1).A1, bins=30, color='blue', alpha=0.7)
plt.title('Distribution of Non-zero Counts in X_train')
plt.xlabel('Number of Non-zero Counts')
plt.ylabel('Frequency')

plt.subplot(2, 2, 2)
plt.hist(X_test.sum(axis=1).A1, bins=30, color='green', alpha=0.7)
plt.title('Distribution of Non-zero Counts in X_test')
plt.xlabel('Number of Non-zero Counts')
plt.ylabel('Frequency')

plt.subplot(2, 2, 3)
sns.countplot(x=y_train, palette='pastel', hue=None, legend=False)  # Using seaborn for countplot
plt.title('Distribution of y_train')
plt.xlabel('Sentiment')
plt.ylabel('Count')

plt.subplot(2, 2, 4)
sns.countplot(x=y_test, palette='pastel', hue=None, legend=False)  # Using seaborn for countplot
plt.title('Distribution of y_test')
plt.xlabel('Sentiment')
plt.ylabel('Count')

plt.tight_layout()
plt.show()
```



edunet foundation

RESULT

5. Creating model and checking accuracy score

0s

```
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

# Predict on test data
y_pred = lr_model.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy:.2f}')
```

```
Accuracy: 0.89
```

ACCURACY ACHIEVED IS 89 %

RESULT

The reason for achieving 89% accuracy instead of 100% is primarily due to the significant variation in people's opinions about movies. Movie ratings are inherently subjective, influenced by individual preferences, cultural backgrounds, and personal experiences. This variability makes it challenging for any model to predict ratings with absolute certainty for every user and every movie. While the analysis may effectively capture trends and patterns in the data, the diversity of opinions among users means that some predictions will inevitably differ from actual ratings, leading to a maximum achievable accuracy that is less than perfect.

CONCLUSION

- The IMDb movie sentiment classification project aimed to develop a robust system for analyzing and predicting sentiments from movie reviews. Leveraging data analytics and machine learning techniques, the project involved collecting a dataset of IMDb movie reviews, preprocessing the data to handle text-specific challenges, and implementing a sentiment classification model using logistic regression and TF-IDF vectorization.
- Key achievements included evaluating the model's performance with metrics like accuracy, precision, recall, and F1-score, achieving satisfactory results in classifying reviews into positive and negative sentiments. Visualizations using Matplotlib and Seaborn enhanced the interpretability of results, providing actionable insights for stakeholders in the entertainment industry.
- The project highlighted the effectiveness of machine learning in understanding audience reactions and sentiments, offering a valuable tool for improving movie marketing strategies, production decisions, and overall audience engagement. Future enhancements could explore advanced natural language processing techniques and real-time data integration to further refine prediction accuracy and relevance in dynamic movie review landscapes.
- Overall, the IMDb movie sentiment classification project demonstrated the potential of data-driven approaches to transform sentiment analysis in the entertainment industry, paving the way for informed decision-making and strategic insights.

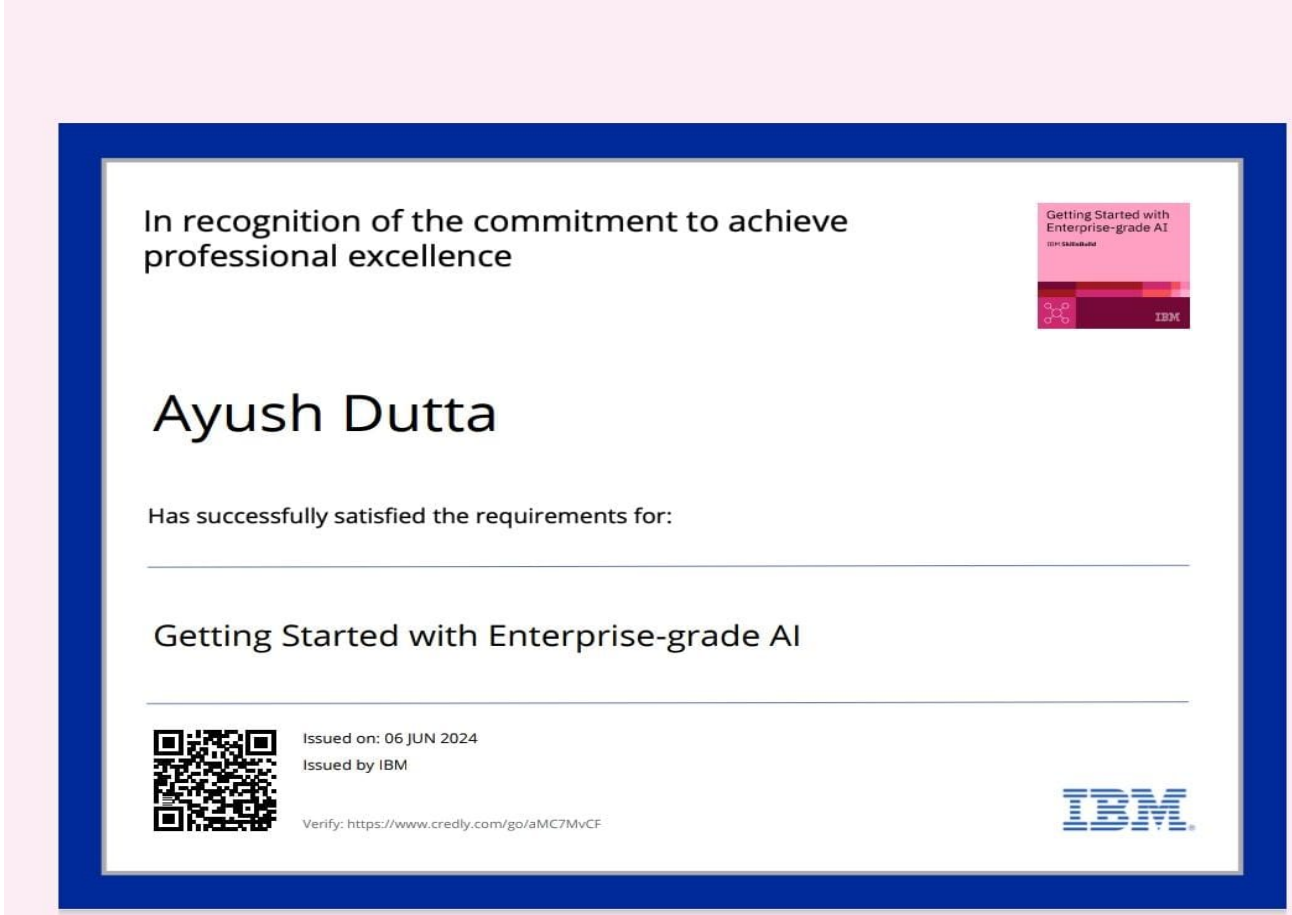
FUTURE SCOPE

- **Real-Time Data Integration:** Incorporate real-time data sources such as social media trends and audience interactions to enable dynamic sentiment analysis and responsive insights.
- **Multilingual and Cross-Cultural Analysis:** Extend sentiment analysis capabilities to handle diverse languages and cultural contexts, enhancing applicability and insights across global audiences.
- **Integration with Recommendation Systems:** Integrate sentiment analysis insights into movie recommendation systems to personalize user experiences based on sentiment preferences and enhance engagement.
- **Enhanced Visualization and Interpretation:** Develop interactive visualization tools to present sentiment analysis results intuitively, including sentiment trends, genre-based insights, and comparative analyses.
- **Expansion to Other Entertainment Domains:** Apply the developed methodologies and models to analyze sentiments in other entertainment domains such as TV shows, streaming content, and music reviews, adapting approaches to domain-specific characteristics and trends.

REFERENCES

- 1. IBM SKILLSBUILD PLATFORM
- 2. IBM CLOUD
- 3.JUPYTER NOTEBOOK
- 4.GOOGLE COLLAB
- 5. KAGGLE WEBSITE FOR DATASET
- 6. EDUNET ONLINE EDUCATION 4 WEEK TRAINING PROGRAM

IBM SKILLSBUILD COURSE COMPLETION CERTIFICATE 1:



edunet foundation

IBM SKILLSBUILD COURSE COMPLETION CERTIFICATE 2:

THANK YOU

Presented By:
Name: AYUSH DUTTA
Internship ID:INTERNSHIP_17127310326616339 8a8c87
AICTE Student ID:STU6640d104320f31715523844
Institute: BENGAL ENGINEERING AND SCIENCE UNIVERSITY, SHIBPUR
Degree: BACHELOR OF TECHNOLOGY (4YEARS UG PROGRAM)
Student's Email ID: dutta7759@gmail.com

edunet foundation