PROJECT INTIALIZATION DOCUMENT

1. DETAILS

Team Details: Team consist of three people, and they are Sathvik N G (sathvikng56@gmail.com), Sanketh L Reddy (sankethl2003@gmail.com) ,Thomas S(thomas.sk2504@gmail.com)

Project Name: Life Expectancy Prediction

2. DEFINING PROJECT AND ITS SCOPE

Understanding of the project: The goal of this project is to predict the life expectancy of a person based on various factors. The factors include the year in which the data was recorded, the status of the country, the mortality rate of adults, the number of infant deaths, alcohol consumption, expenditure percentage, hepatitis B vaccination coverage, measles cases, body mass index (BMI), polio vaccination coverage, total expenditure, diphtheria vaccination coverage, HIV/AIDS prevalence, gross domestic product (GDP), population, thinness rates in children aged 1-19 years and 5-9 years, income composition of resources, and schooling.

To achieve this goal, we will create a basic linear regression model to find the linear relationship between life expectancy and the various factors mentioned above. We will also perform exploratory data analysis (EDA) to understand the data and discover patterns, relationships, and anomalies. EDA will involve checking the quality of the data, computing descriptive statistics of the dataset, visualizing the distribution of each variable, analysing relationships between variables and the target variable, exploring categorical variables, identifying missing values, detecting outliers, performing feature engineering, data normalization, and feature selection.

Once the model is built, it will be connected to a centralized database server, and users will be able to upload their data on a web platform. The model will then predict the life expectancy of the person and store the result in a database. Additionally, users will receive their predicted life expectancy in both JSON format and PDF.

Overall, this project aims to automate the process of predicting life expectancy using a machine learning model and make it easier for individuals to understand their life expectancy based on various factors.

Reason for choosing this project: We chose this problem for the reason of its social significance and potential impact on people's lives. Predicting life expectancy is an important problem in public health and can aid policymakers and healthcare providers in making informed decisions. We were also interested in applying machine learning techniques to a real-world problem and gaining experience in data analysis and modelling. Additionally, we felt that this project provided an opportunity to learn and apply statistical and machine learning concepts such as linear regression, feature selection, and model evaluation. As a team, we were excited to collaborate and work on this project and to share our knowledge and insights

Most challenging aspect of the project statement: The most challenging aspect of this project is to identify the relevant variables that have a significant impact on life expectancy and to develop an accurate predictive model. It requires a deep understanding of the underlying concepts and a thorough analysis of the dataset. Another challenge is to deal with missing data and outliers in the dataset, which can affect the performance of the model. Furthermore, selecting the appropriate machine learning algorithm, optimizing its hyperparameters, and evaluating the model's performance is a complex task that requires expertise and experience. Lastly, ensuring that the model is robust and scalable and can be deployed in a real-world scenario is another significant challenge. Overall, this project requires a strong foundation in statistics, data analysis, and machine learning, and it demands a rigorous and systematic approach to problem-solving.

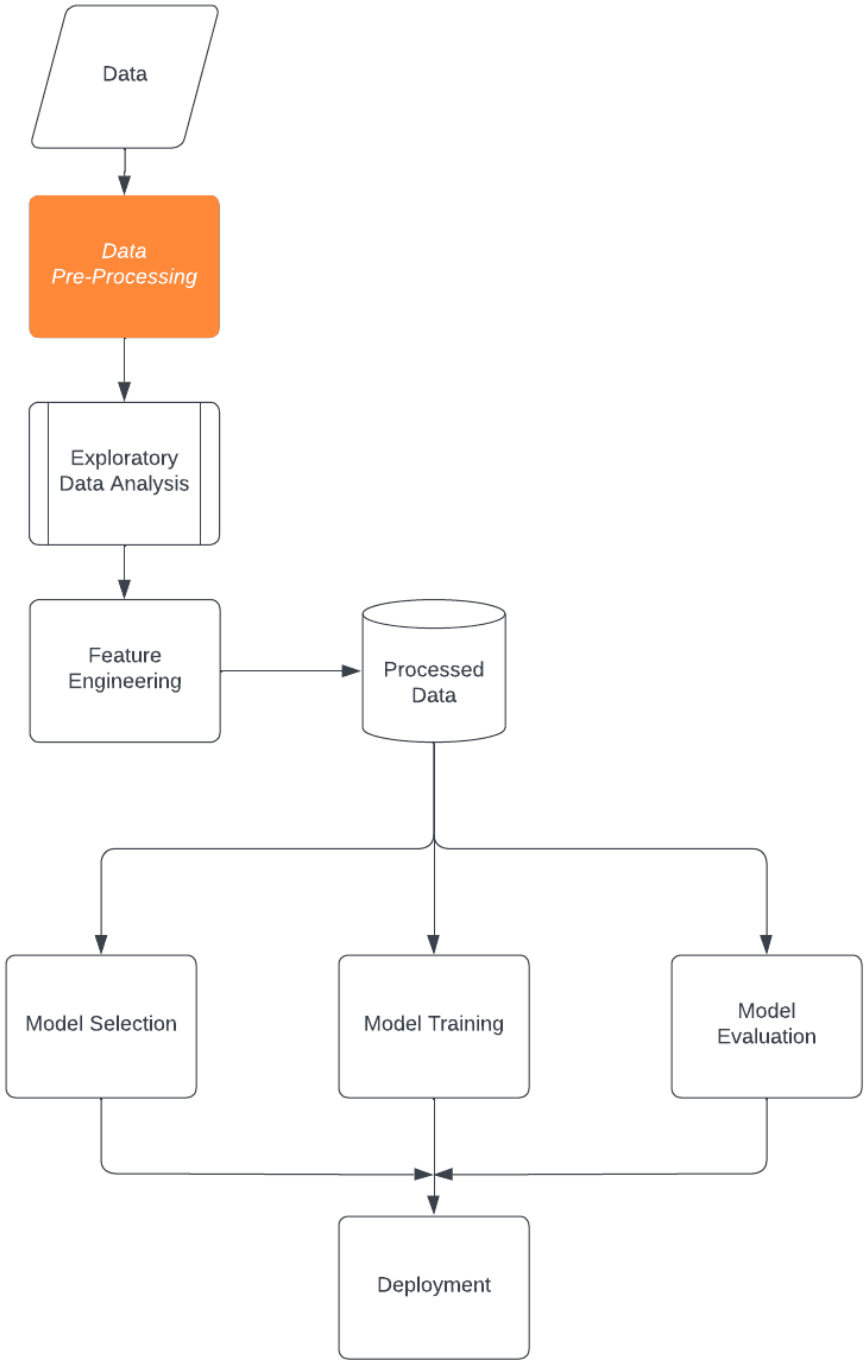
3. APPROACH OF PROBLEM CHOSEN

Approach: Below is the approach that we are going to follow.

- We will collect data on various factors that can affect life expectancy such as age, gender, lifestyle, health conditions, environmental factors, etc. We will use publicly available datasets from reliable sources like WHO, CDC, etc.
- Once we have the data, we will preprocess it by removing any missing or duplicate values, correcting errors, and converting the data into a suitable format for analysis.
- We will perform exploratory data analysis (EDA) on the preprocessed data to gain insights and identify any patterns or correlations between the variables. We will use various visualization techniques like scatter plots, histograms, heatmaps, etc. to analyze the data.
- Based on the results of EDA, we will select the most relevant features that are highly correlated with life expectancy. We will use techniques like correlation matrix, feature importance ranking, etc. to identify the important features.
- We will evaluate various machine learning algorithms like Linear Regression, Random Forest, etc. to identify the best model for our problem. We will use cross-validation techniques like k-fold cross-validation to evaluate the performance of each model.
- Once we have selected the best model, we will train it on the preprocessed data. We will use techniques like regularization, hyperparameter tuning, etc. to improve the performance of the model.
- We will evaluate the performance of the trained model using metrics like mean squared error, R-squared score, etc. We will use test data to evaluate the model and ensure that it generalizes well.
- Once we have a trained and evaluated model, we will deploy it in a production environment. We can create a web-based interface where users can input their details and get their predicted life expectancy. We will use cloud-based services like AWS, Google Cloud, etc. to deploy the model.

Throughout the project, we will be using Python and various libraries like NumPy, Pandas, Matplotlib, Scikit-learn, etc. to implement the machine learning algorithms and perform data analysis. We will also be using Jupyter Notebook to document our progress and share our work with the team.

Diagram/Flowchart (if possible):



Platform/Coding Language/Frameworks (if using): Jupyter Notebook, Python, Pandas, NumPy, Scikit-learn, Matplotlib, Flask

Database/Cloud/Hosting (if using):

External tools (if using):

4. TEAMS ABILITY TO IMPLEMENT WINNING SOLUTION

Background of team members/individual: Sanketh L Reddy ,Sathvik N G and Thomas S are currently pursuing Computer Science and Engineering degree in Sir M.VIT College in Banglore ,Karnataka.All the 3 of us wanted to a internship in our 2nd year which can teach us how to collectively work as team and also to make our CV Better.

Major Expertise of team members/individual: Sathvik N G has the knowledge of Data structures ,JAVA ,Python and C++ programming Languages and he had also developed few web applications in the past.He also has intrest interst in Machine Learning and Artificial Intelligence.Sanketh L Reddy on the other hand have studied JAVA ,C,C++ ,Python,HTML and Data structures and he is also intrested in Machine Learning .Thomas S has the knowledge of DSA,JAVA,Python,C++ and he is intrested in learning about Machine learning and Artificial Intelligence

Roles and responsibilities of team members/individual: Sathvik N G will be collecting data on various factors that can affect life expectancy ,preprocessing the collected data, perform feature engineering for training the model and also deploy the model .Sanketh L Reddy will perform exploratory data analysis (EDA) on the preprocessed data to gain insights and identify any patterns or correlations between the variables. Thomas S will select a suitable machine learning algorithm and will train it on the preprocessed data.All the 3 of us will collectively to evaluate the model and ensure that it generalizes well.

Previous projects undertaken: We have not undertaken any projects in our past. Although this is our first project we are very excited to put in all our efforts to make this project work.

Team strengths: We all have a common interest in Machine Learning. Although we have not worked previously on any group projects but we are extremely motivated to do our first project together as a team. The most important thing is the understanding between our teammates which is extremely important to go ahead and reach the finale. We all have that competitive spirit in us. Our coding fundamentals are clear.

Team/Individual achievements: One of our Team members have won first prize in College level hackathon.

Personal motivation: As our approach is to use machine learning and data science, applying mathematics like statistics and linear algebra to data for visualizing and predicting analysis is very intriguing and above all applying it to a challenging real-life problem like ranking the resume and extracting the information, and giving it a label is exciting