VIEHGROUP

PROJECT INTIALIZATION DOCUMENT

1. DETAILS

Team Details: Team consist of only me (sarthakwakchaure88@gmail.com)

Project Name: Breast Cancer Prediction

2. DEFINING PROJECT AND ITS SCOPE

Understanding of the project: Breast Cancer is one of the leading cancer developed in many countries including India. Though the endurance rate is high-with early diagnosis 97% women can survive for more than 5 years. Statistically, the death toll due to this disease has increased drastically in last few decades. The main issue pertaining to its cure is early recognition. Hence, apart from medicinal solutions some Data Science solution needs to be integrated for resolving the death causing issue. This analysis aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the breast cancer is benign or malignant. To achieve this I have used machine learning classification method (Logistic Regression) to fit a function that can predict the discrete class of new input.

Reason for choosing this project: I choose this problem for the reason of technology and interest. I also quite interested in the concept of machine learning and classification techniques for prediction. I have done some mini projects on machine learning, and it would be interesting to use it in real-life applications. Also, I have good grasp of python programming language technologies which I am going to use in our project.

Most challenging aspect of the project statement: : Solving any problem using machine learning is quite challenging. As given dataset contains so many features, so selection of features from the given dataset is also challenging task because it can affect the output of project. The decision of selecting the most suitable algorithm based on the most effective output is challenging.

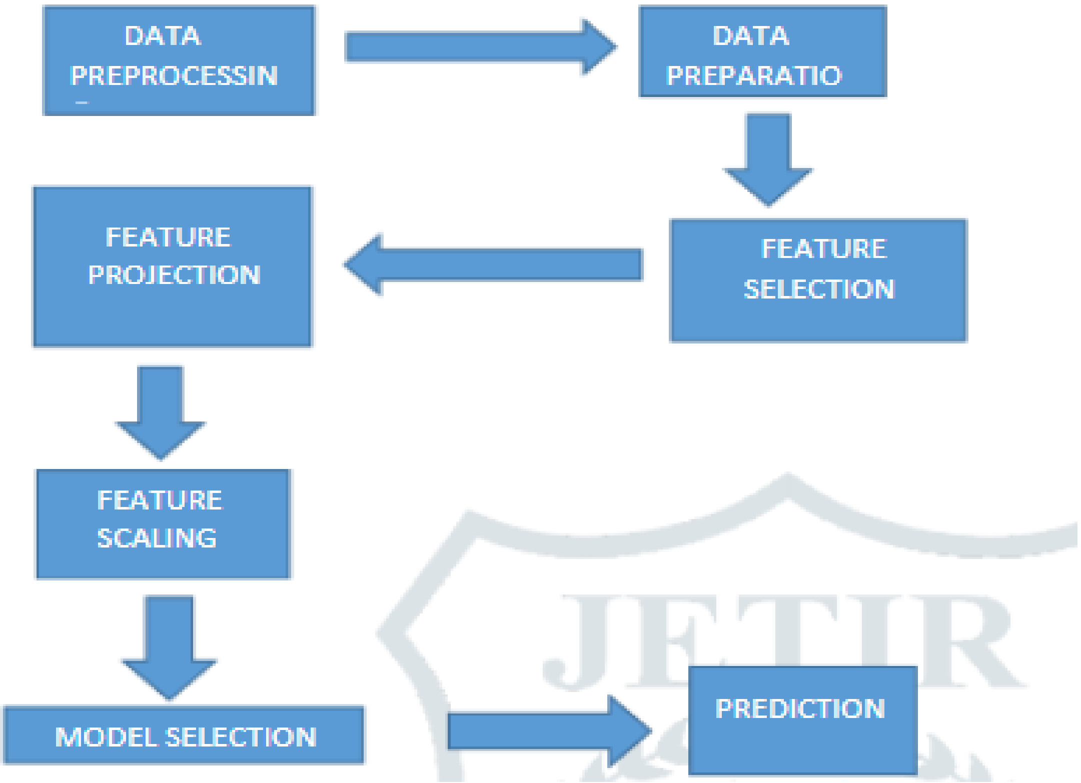
3. APPROACH OF PROBLEM CHOSEN

Approach: Below is the approach that I am going to follow.

1) Initially I will collect dataset and will perform data pre-processing on it. It includes steps like Data Cleaning, Data Integration, Data Transformation, Data Reduction, Data Discretization.
2) After that, I will perform EDA on the dataset to understand the relationship between the parameters. Here I will use visualization tools like Matplotlib to visualize the data for better understanding.

3) The above step will help us to identify the outliers or the most important features for the cancer prediction.
4) I will be using Python libraries like NumPy, Sci-kit learn, TensorFlow, matplotlib for machine learning using Jupyter notebook.
5) Then I will split the dataset randomly into training and testing data.
6) After that I will apply Logistic Regression algorithm on training data to train model.
7) Then model testing and evaluation will be done using classification metrics.
8) Then the model will be deployed and predicted cancer type will be displayed.

Diagram/Flowchart:



Platform/Coding Language/Frameworks (if using): Jupyter Notebook, Google Colab, Visual studio Code, Pytho, HTML, CSS, JavaScript, Pandas, NumPy, Ski-kit learn, Flask/Django.

4. TEAMS ABILITY TO IMPLEMENT WINNING SOLUTION

Background of team members/individual: I am Sarthak Wakchaure pursuing B.Tech in Computer Engineering from MIT Academy of Engineering college, Pune. My areas of interest are ML, Web development and Cloud Computing.

Major Expertise of team members/individual: I am Sarthak Wakchaure having a knowledge of DSA and been doing competitive programming since a long time. I also have an interest in Machine Learning and Web development.

Roles and responsibilities of team members/individual: As I am along in project so all the responsibilities like dataset collection, feature selection, training of model, testing, evaluation and deployment will be taken by me.

Previous projects undertaken: Below are the projects undertaken by me.

1) Student management system using python (Application)
2) Text to speech converter (web application)
3) Blood management system (Website)

Team/Individual strengths: I am having an interest in Machine Learning and Data Science. I also have working experience as individuals as well as group projects. My coding fundamentals and DSA are clear.

Team/Individual achievements: Took participation in Datathon arranged by Bajaj company.

Personal motivation: As my approach is to use machine learning, 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 taking input from user and predicting type of cancer is interesting.