PROJECT PART1

Team Members

SaiSunilKumar Ponduri-700741913

Kadire Sanjay Kumar Reddy-700741058

Jayanth Sri Sai Dulla- 700734068

Jagadeeswar Chimata- 700731649

Life Expectancy (WHO) with several ML techniques

Problem statement:

Taking a look at the life expectancy trend globally has an increasing trend till the year of 2019. The life expectancy between the years 2000-2019 is 66.8 years. If we look at the individual years the life expectancy number in the year of 2000 is 63.7. There is an 8% percent increase in the life expectancy that is from 58.3 to 63.7 years . But HLE(Healthy Life Expectancy) for a few countries is not in an increasing trend so there is a need to understand the factors that affect life expectancy. GHO(Global Health Organisation) under WHO(World Health Organisation) keeps the track of all the records of 193 countries. The challenging task is that these factors for various countries differ one to another. Understanding the data selecting the best features is a cumbersome task.

Literature survey:

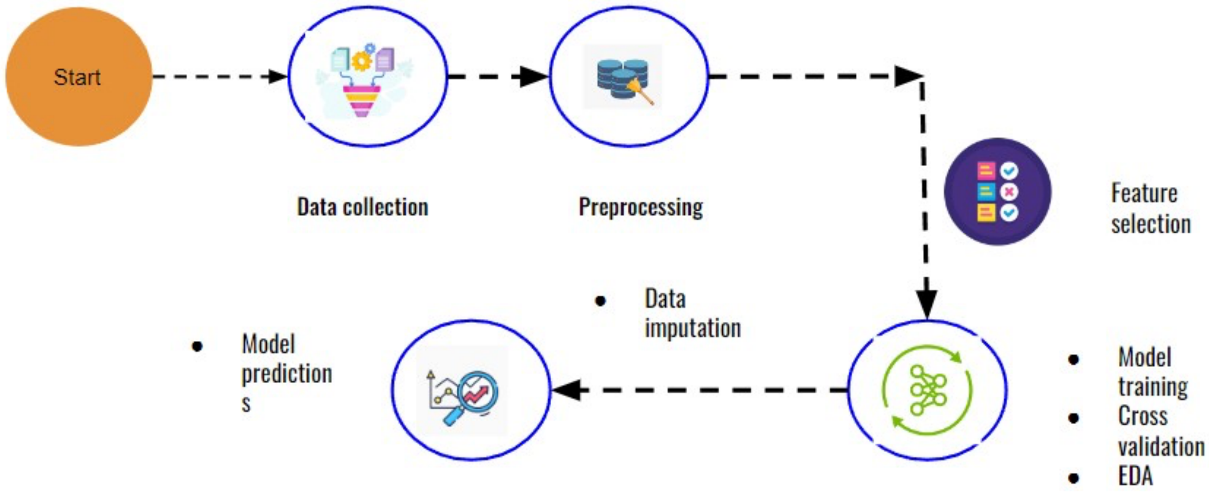
Before conducting the experimental analysis we need to make sure that data is error free and clean. For this task all the preprocessing techniques are applied. In the existing models like linear regression they can not differentiate between important and less important features; this condition leads to overfitting of the model.Linear regression challenges and problems are overcome by Ridge Regression.Ridge regression features can differentiate the important and less important features.Decision Tree algorithm can be used for both regression and classification problems the algorithm splits the data into multiple subsets and trains the data.The advantage of the algorithm is that it can easily comprehend the data can be of any dimensions and produce good results [1]. In this paper we have studied the life expectancy of the HIV/AIDS patients mortality data.To find the life expectancy number we have used Random Forest Regressor, Decision Tree Regressor and linear regressor. Amongst these algorithms Random Forest algorithms outperforms the other algorithms.For the evaluation R squared error, Mean Squared Error, Mean Absolute Error is used. Random Forest regressor achieves 0.99 on training data and 0.95 on testing data [2]. Though the overall life expectancy crossed 70 years in a few countries the life expectancy is very low and needs to be analyzed.In this project we are analyzing the life expectancy data with 18 features GDP, disease, school index.In this paper 3 MLR(Multi Linear Regression) models are implemented and R^2 values are recorded for both developing countries and developed countries [3].

The life expectancy data dimensions are very high. It is important to sift through the data to find the important features. In this paper we are using Decision tree regression , KNearest Neighbor Regression and Linear regression algorithms are used in combination of two feature selection methods correlation coefficient and Mutual information. The model's performance is evaluated using RMSE,MAE and R^2 error[4].

Proposal:

There are multiple factors that affect life expectancy or mortality rate. In this project we are collecting a dataset that is publicly available in kaggle the original source of the dataset is. This project is mainly focused on the immunization factor. The features include Hepatitis, economic factors, social factors in total 22 features to predict the life expectancy of the country. In this project we are proposing several machine learning techniques and feature selection methods to predict the life expectancy number. For this experiment we are proposing Linear Regressor, Random Forest regressor and XGBoost Regressor algorithms.

Implementation:



In the first step data is collected. After collecting the data it is cleaned by imputing or replacing the null values with imputation methods.After cleaning the raw data, the underlying structure of the data is understood using data visualizations. Various feature selection methods are implemented to select the important features. After the feature selection data is split into 3 parts: train,test and validation set. Train set used to train the model using validation set Cross Validation is performed for 5 folds for each fold accuracy of the model is recorded. In the end model predictions are recorded from the test dataset.

Contributions:

1.For data analysis pandas library is used.

2.To preprocess the data KNNImputer is used to fill the missing values.This method is imported from scikit-learn library.

3.For data visualization matplotlib and seaborn libraries are used.
4. For Cross Validation and training is performed using scikit-learn
5.The execution of the program is conducted in the PyCharm environment.
6.Numpy library is used for mathematical operations

Results:

Since the problem type is regression.The evaluation metrics used are:
RMSE(Root Mean Squared Error), Mean Absolute Error(MAE), RMSLE(Root Mean Squared Logarithmic Error),Mean Squared Error(MSE), R squared and Adjusted R Squares metrics.

References:

[1].M. M. Biltawi and R. Qaddoura, "The Impact of Feature Selection on the Regression Task for Life Expectancy Prediction," 2022 International Conference on Emerging Trends in Computing and Engineering Applications (ETCEA), Karak, Jordan, 2022, pp. 1-5, doi: 10.1109/ETCEA57049.2022.10009674.
[2].X. He, J. Hu, C. Liu and Y. Zhang, "Analysis on Relevant Factors Affecting Life Expectancy," 2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), Dalian, China, 2022, pp. 569-572, doi: 10.1109/IPEC54454.2022.9777372.
[3].V. Bali, D. Aggarwal, S. Singh and A. Shukla, "Life Expectancy: Prediction & Analysis using ML," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2021, pp. 1-8, doi: 10.1109/ICRITO51393.2021.9596123.
[4].S. S. Meshram, "Comparative Analysis of Life Expectancy between Developed and Developing Countries using Machine Learning," 2020 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 2020, pp. 6-10, doi: 10.1109/IBSSC51096.2020.9332159.