LIFE EXPECTANCY PREDICTION USING MEDICAL PARAMETERS, SMOKING AND DRINKING HABITS – AN AI BASED APPROACH

BY:
PARUL BHUTANI WADHWA
SHEHAL KATHIRIYA

Table of Contents

Abstract

The goal of this project is to create an innovative AI based model that determines the life expectancy of an individual based on a comprehensive set of medical parameters along with their smoking and drinking habits. Predicting life expectancy can be of a significant importance for healthcare, insurance, and public health policies. This project uses a diverse dataset which has wide range of medical variables including gender, age, blood pressure, cholesterol levels and various medical conditions. Using a combination of data processing, feature engineering and model training, we aim to create a highly accurate predictive model capable of determining an individual's remaining years of life.

Introduction

In today's healthcare landscape, it is essential to understand the impact of a person's health history and lifestyle choices on life expectancy. This predictive information can be used by medical practitioners and healthcare organizations to provide personalized health plans and preventive measures to improve a patient's quality of life and life expectancy.

In our project, we consider an individual's health and lifestyle data in a data-driven manner. We analyze past medical records, amount of alcohol consumption, and smoking habits using advanced AI and ML algorithms to uncover patterns and correlations that are beyond the scope of conventional analysis.

We will guide you through the full process, from design to deployment, in an easy-to-use user interface in this report. We will also look at how healthcare practitioners might utilize this tool to better understand their patients and provide more responsive, tailored treatment. Although no one human life can be assessed, this technology can provide significant information that can help individuals make better decisions and live healthier lives.

This report is divided into sections that represent each phase of the project in detail. By adhering to this structure, you will be able to gain a better understanding of the project's execution from start to finish in an organized and comprehensive manner, allowing you to appreciate the breadth and depth of the work that went into achieving the result.

We hope that as you read this report, you will gain a better understanding of how this project uses the transformative power of artificial intelligence and machine learning to improve healthcare delivery and disease prevention.

Problem Statement

Life expectancy prediction is one of the most important challenges in healthcare, insurance, and public health. The goal of our project is to solve these problems by creating a Machine learning-based model that predicts life expectancy accurately, but also provides tailored health improvement suggestions, supports decision making, and creates marketing plans for insurance & Mediclaim providers.

Personal health Improvement: In today's fast-paced environment, people are often too busy to prioritize their health. They don't know what's wrong with them or how it affects their life expectancy, and they don't seek help until it's too late. This puts them at risk for chronic diseases and even premature death.

We're here to help. Our model analyzes a wide range of data, including your medical records, your lifestyle habits and even your genetic markers. Based on this information, you can make informed healthcare decisions, make changes to your lifestyle, seek preventative care, or take proactive steps to reduce your health risks.

Medical Decision Support: Poorly predicting life expectancy can result in poor diagnosis, ineffective treatment, and inefficient allocation of resources. With predictive analytics and actionable insights, the model helps healthcare professionals identify high risk patients, deliver timely interventions, and optimize patient outcomes, resulting in better healthcare outcomes and better life expectancy.

Insurance and Mediclaim Companies: Our AI-driven model analyzes different data points to deliver actionable insights & predictive analytics to help insurance & Mediclaim companies set more accurate life expectancy, set more accurate premiums & create better marketing strategies. By solving these issues, the model can dramatically improve insurance & Mediclaim operations resulting in better financial results and improved customer satisfaction.

Background

Historical Context

Predicting individual life expectancy has long been a focus in healthcare and medical study. Medical research and technological breakthroughs have permitted a more thorough knowledge of the elements impacting an individual's longevity throughout time. Traditional methods of estimating life expectancy frequently depended on demographic data, but modern approaches examine and interpret complicated statistics using advanced computing techniques such as artificial intelligence (AI).

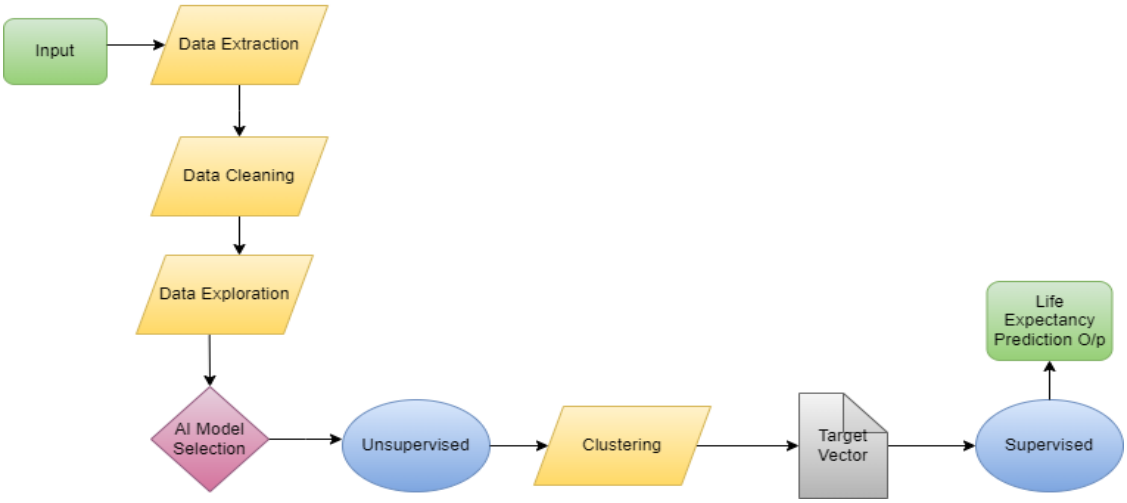
State of the Art

Recent advances in machine learning and artificial intelligence (AI) have opened novel possibilities for forecasting life expectancy with greater precision. The combination of modern algorithms and large medical databases enables a more tailored and nuanced assessment. Studies have shown that combining lifestyle indicators, such as smoking and drinking behaviors, into prediction models provides a more comprehensive picture of an individual's health trajectory.

With the advent of predictive analytics in healthcare, a paradigm changes from reactive to proactive healthcare management has occurred. Predictive models not only help with tailored patient treatment, but they also provide useful insights for public health measures.

Approach

In creating the Life Expectancy Predictor Model, we took a holistic and systematic approach to leveraging the power of AI and leveraging the right data sets. The methodology includes the use of medical, smoking and drinking data to build a strong predictive model.



Data Collection

The core of the project is a collection of unique and representative dataset. Our main raw data source for the supervised learning model is Kaggle. This dataset serves as the foundation for detecting underlying patterns and structures in data using unsupervised learning techniques.
A significant research effort was performed in response to the requirement for a specific dataset geared to forecast individuals' life expectancy. This requires gathering and organizing information from a variety of sources, including medical records and lifestyle markers. The dataset was meticulously developed to capture the many elements that influence life expectancy, assuring its uniqueness and representativeness. The dataset consists of around 900000 samples with features like sex, age, height, weight, waistline, sight, hear, SBP, DBP, HDL, LDL, total cholesterol, triglycerides, hemoglobin, urine protein, serum creatinine, AST, ALT, smoking frequency, and alcohol intake. Since the dataset was too large to be processed by the CPU, it was filtered for only chain smokers and alcoholic people. This generated a new dataset consisting around 150000 samples.

Data Exploration

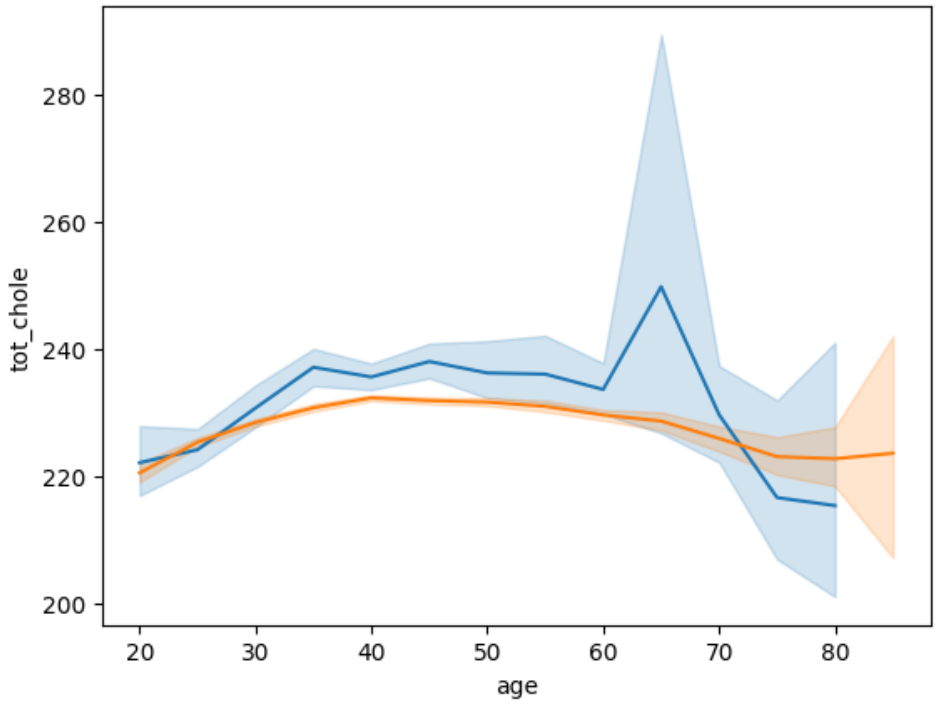
Prior to model creation, a deep exploratory phase was conducted to understand the properties of the collected data sets. We used descriptive statistics, visualizations, and correlation analysis to visualize the distribution of the variables, find outliers, and uncover correlations.

Based on the relationships various categories were created like BP, Diabetes, Triglyceride, Hemoglobin, Kidney, Liver, BMI. This transformation helped reducing features from 25 to 8.

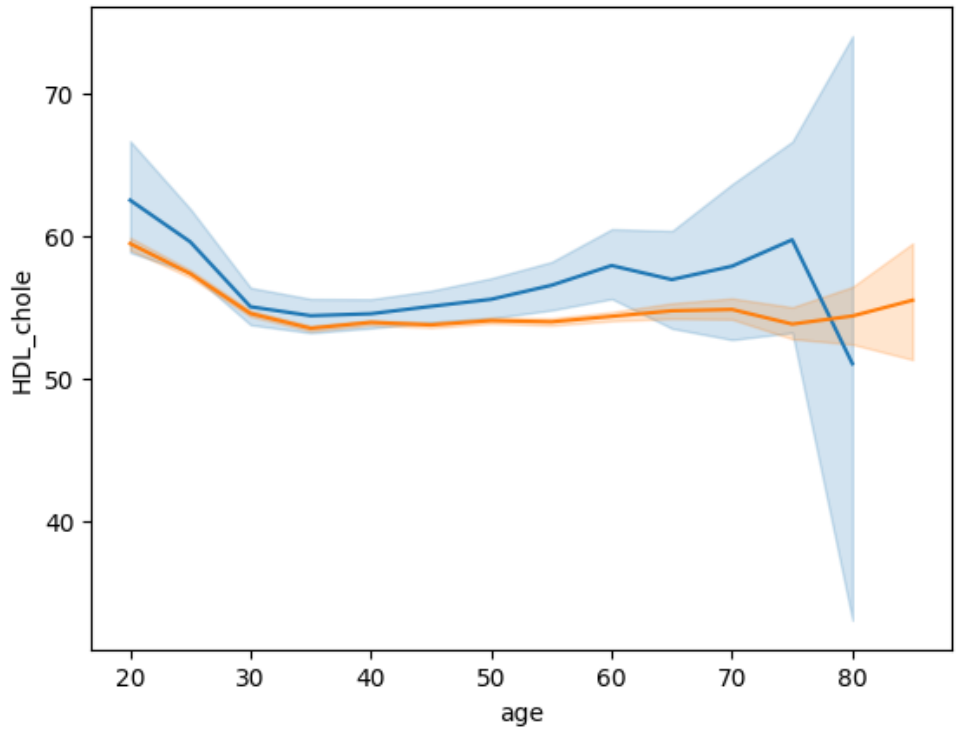
This exploratory phase not only informed the feature selection of the unsupervised learning model, but also provided us with a better understanding of relationships within the data set.

Moreover, data analysis was performed with the transformed feature set. Some interesting facts were discovered like:

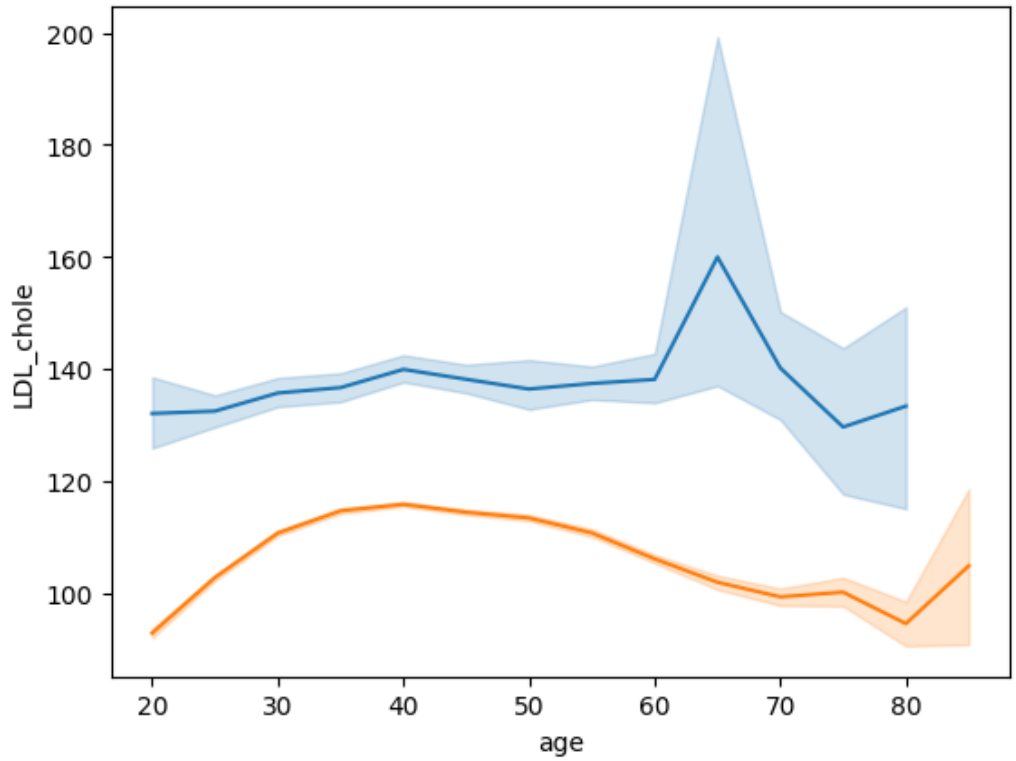
People who are diabetic, suffers from high total cholesterol, which can lead to many heart diseases compared to the people who are non-diabetic. This justifies that their life expectancy range decreases.



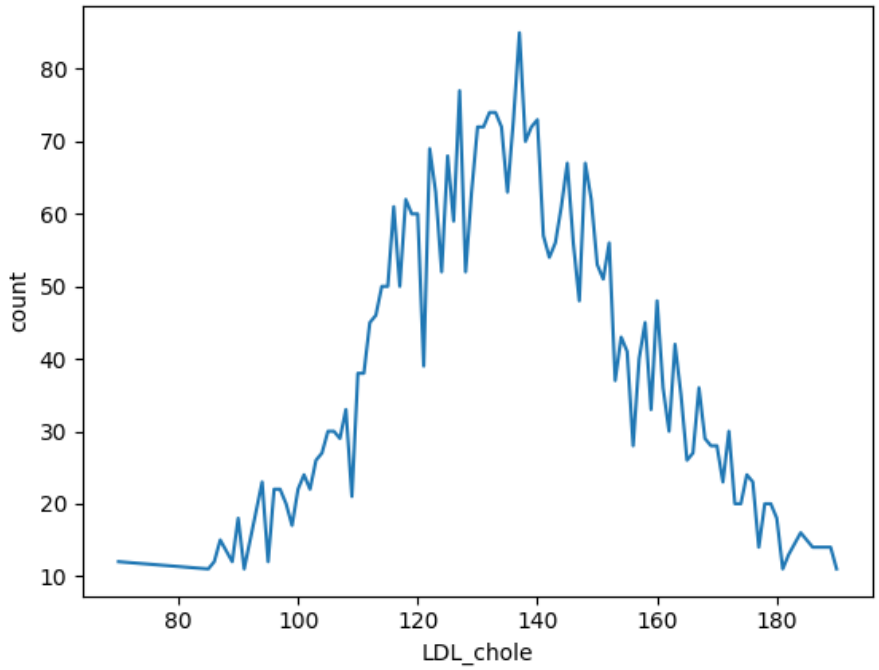
Secondly, diabetic people have high HDL levels compared to non-diabetic people. Since HDL is a good cholesterol, so even if the value increases the normal range, it will not have an adverse effect on the heart. Therefore, this parameter is not directly co-related to our research.



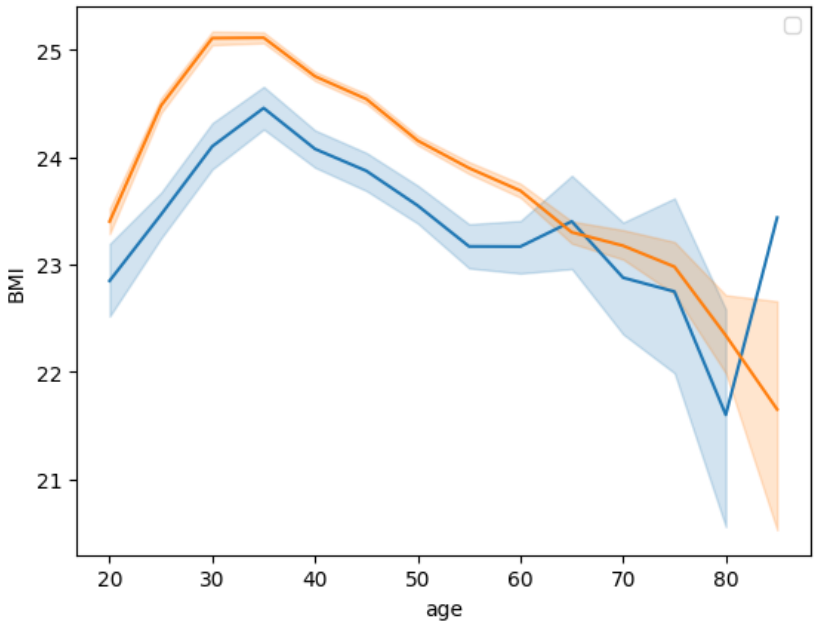
Also, below graph shows high LDL levels for diabetic people which is very harmful to the heart and is a major reason for heart-attacks compared to orange line for non-diabetic people. Hence this parameter is a major factor to be considered in our dataset.



The below graph represents the count of people having various LDL levels. It shows that maximum people have LDL values ranging above 120 which is a poor heart indicator. Hence, most of the samples in the dataset will suffer heart issues in future and their life expectancy will directly depend on one of these factors.



Higher BMI levels are indicators of obesity which is one of the main reasons of all the health issues today. From the graph we can see, non-diabetic people (orange line) have higher BMI index compared to diabetic people. So, we can conclude that BMI is not related to diabetes and may be concerned with other parameters.



Various other analysis showed that weight is directly proportional to serum creatinine, less or more creatinine value leads to chronic kidney disease or kidney failure respectively.

chronic kidney disease is also proportional to low values of ALT and people who suffer kidney issues may suffer from liver diseases in future and viz-a-viz. Hence, lowering life expectancy ratios.

All these visualizations justified the relationship between individuals medical record with respect to future diseases they may encounter and thereby defining their life expectancy.

Data Preprocessing
A thorough preprocessing phase was carried out to ensure the quality and consistency of the datasets. The preprocessing phase included normalization of numerical features using MinMaxScaler, and the encoding of categorical variables using oneHotEncoder to produce a standardized and homogeneous dataset for the training and testing of the predictive model.

Model Selection

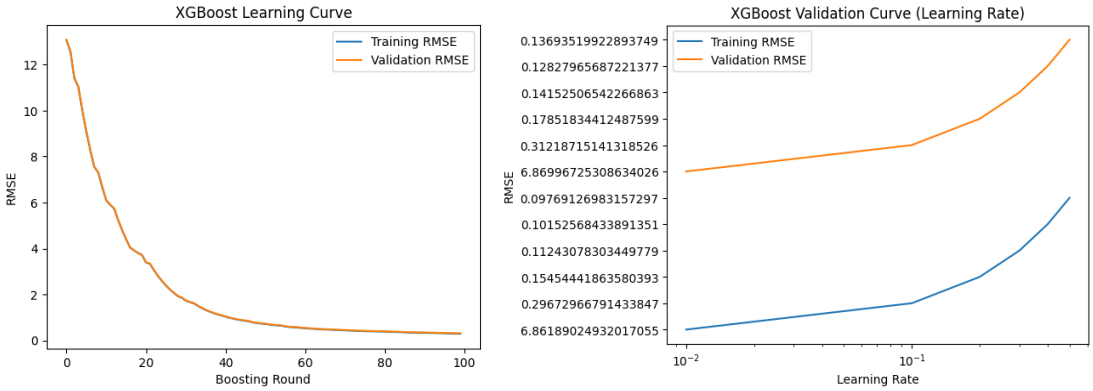
Unsupervised Learning Model
To better understand the structure of the data and define them in certain categories, decision was made to work with clustering-based algorithms. After trying various clustering models DBSCAN was chosen. It is a density-based clustering technique, which has ability to detect clusters of diverse forms while successfully dealing with data noise. This model facilitates in the identification of underlying patterns and structures in unsupervised learning datasets. 127 categories were created from this algorithm which displayed groups of people with similar medical records.

In the next stage outliers were understood from the output of the model and separate categories of clusters were created for it. Moreover, all the clusters that were originally named 0,1, 2...127 were renamed using configuration naming convention.

After in-depth research of all these clusters, the future life expectancy of a person along with the diseases one can encounter were formularised. Thereby, creating a new target vector “FinalAge” in the dataset. As a result, using unsupervised learning, the model identified clusters and assigned a predicted final age of an individual.

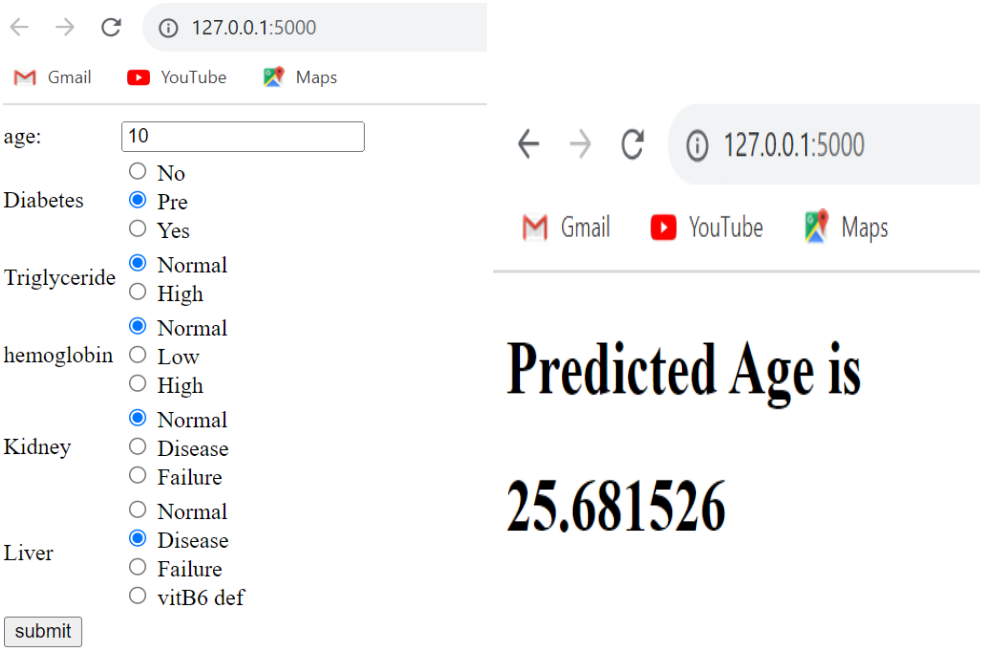
Supervised Learning Model
The next step is to build a model that predicts a future age of any individual with different values of the same parameters. For this a supervised learning algorithm was considered. The output generated from the DBSCAN model eventually served as the input for the supervised learning model. After checking the accuracy and evaluating various models, the ensemble learning method XGBoost was chosen for its durability and strong performance in predictive modeling applications. It is well-suited for our life expectancy prediction model due to its capacity to handle complicated interactions among features and produce reliable forecasts.

Before fitting the model, the imbalance of the minority clusters was adjusted using RandomOverSampler technique to achieve model's accuracy of 87%. The model was evaluated using mean squared error and then validated using Learning and Validation curve.



User-interface

Flask API was used to create a user interface. In the UI, user enters the data by clicking on the radio buttons and a response predicting user's future age is received after clicking at the submit button.



Ethical Considerations

Ethical considerations were at the forefront throughout the development process. Privacy and confidentiality of users in datasets, compliance with data protection legislation, and transparency of model limitations were key components of the approach.

Limitations

It's important to recognize the limitations of our approach. The fact that we rely on a limited dataset, due to confidentiality restrictions and limited inclusion criteria, may restrict our findings from being generalizable. The exclusion of physically active people and those on medication further limits our predictions.

Future Work

Data Diversity*:* The focus should be on obtaining more diverse data sets, while ensuring representativeness and confidentiality. This will enable the development of models that go beyond the current demographic limitations.

Inclusivity of Health Behaviors*:* Extending the scope to people who exercise or take medication will provide a more comprehensive view of life expectancy predictions. Such an inclusivity provides insights into a broader range of health dynamics.

Secondly, dataset of people who doesn't drink or smoke or do it occasionally will also be added to the model building which will enhance the scope of the project.

Conclusion

In Conclusion, our project on life expectancy prediction, based on a certain sample of data, has provided important insights into the relationship between medical data, lifestyle changes, and life expectancy. The limited nature of the data, collected through rigorous analysis and confidentiality, shapes our findings, and highlights the need for careful interpretation.

References

Smoking and Drinking Dataset with body signal. (2023, August 30). Kaggle. https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset

Sklearn.cluster.DBSCAN. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

Brownlee, J. (2021, March 6). *XGBOost for regression*. MachineLearningMastery.com. https://machinelearningmastery.com/xgboost-for-regression/

Home. (n.d.). PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/

Szili-Torok, T., Bakker, S. J. L., & Tietge, U. J. F. (2022). Normal fasting triglyceride levels and incident type 2 diabetes in the general population. *Cardiovascular Diabetology*, *21*(1). https://doi.org/10.1186/s12933-022-01530-8

DynaMed. (n.d.). https://www.dynamed.com/condition/hypertension#GUID-D034AC2E-FF8B-44AD-BAC4-FDF54AFF377D

https://www.medicinenet.com/creatinine_blood_test/article.htm

Daniel, C. (2023, November 9). *ALT and AST Liver Enzymes: What High Levels Mean*. Verywell Health. https://www.verywellhealth.com/liver-enzymes-1759916#:~:text=An%20AST%2FALT%20ratio%20higher,sign%20of%20alcoholic%20liver%20disease.

Body mass Index (BMI). (2022, June 3). Centers for Disease Control and Prevention. https://www.cdc.gov/healthyweight/assessing/bmi/index.html#:~:text=Body%20Mass%20Index%20(BMI)%20is,or%20health%20of%20an%20individual.

Professional, C. C. M. (n.d.). *Low hemoglobin*. Cleveland Clinic. https://my.clevelandclinic.org/health/symptoms/17705-low-hemoglobin