LIFE EXPECTANCY PREDICTION USING MACHINE LEARNING

1. Abstract

Life expectancy prediction plays a crucial role in public health planning, policy-making, and personalized healthcare interventions. Machine learning techniques have shown promise in accurately predicting life expectancy by leveraging diverse demographic, socio-economic, and health-related factors. In this project, we aimed to develop a machine learning model for life expectancy prediction and explore its potential applications.

We collected a comprehensive dataset encompassing individual-level attributes such as age, gender, education level, income, healthcare accessibility, lifestyle factors, and historical health records. Feature engineering techniques were employed to pre-process and transform the data, ensuring its suitability for machine learning algorithms. Several popular regression algorithms, including decision trees, random forests, and neural networks, were trained and evaluated to identify the most effective model for life expectancy prediction.

2. Introduction

With the advancements in Machine Learning and Data Science, we now have the ability to predict the remaining life expectancy of a person with a high degree of accuracy, based on certain essential parameters. In this project, we will be exploring the parameters that affect the life expectancy of individuals living in different countries, and how machine learning models can be used to estimate life expectancy. We will also be focusing on the specific parameters that have the most significant impact on an individual's life expectancy.

The term "**life expectancy**" refers to the number of years a person can expect to live. By definition, life expectancy is based on an **estimate of the average age** members of a **particular population group** will be when they die.

Life expectancy depends on several factors, the most important being **gender** and **birth year**. Generally*, females have a slightly higher life expectancy than males due to biological differences*.

3. Background

Life expectancy is a fundamental measure of population health that represents the average number of years a person is expected to live. It is influenced by various factors such as genetics, lifestyle choices, socio-economic status, healthcare access, and environmental conditions. Accurate prediction of life expectancy has significant implications for public health planning, resource allocation, and personalized healthcare interventions.

In recent years, the field of machine learning has gained momentum as a powerful approach to analyze complex datasets and make predictions. Machine learning algorithms can uncover hidden patterns, relationships, and interactions within data, enabling accurate predictions in various domains, including healthcare. Leveraging machine learning techniques for life expectancy prediction offers the potential to improve our understanding of the determinants of longevity and

LIFE EXPECTANCY PREDICTION

enhance the effectiveness of interventions.

Traditionally, statistical models and regression approaches have been used to study the factors influencing life expectancy. However, these methods often assume linear relationships and may not capture the complex interactions present in the data. Machine learning, on the other hand, can handle non-linear relationships, handle high-dimensional datasets, and adaptively learn from the data.

4. Objective

The primary objective of life expectancy prediction using machine learning is to develop accurate models that can estimate an individual's remaining lifespan based on various factors such as age, gender, lifestyle, medical history, and other demographic and health-related factors. Developing accurate models: The goal is to create models that can accurately predict life expectancy with high precision and recall. Identifying significant factors: Identifying the most significant factors that affect life expectancy can help healthcare providers, policymakers, and individuals develop strategies to improve lifespan and health outcomes. Personalizing care: The objective is to provide personalized care and treatment plans for patients based on their predicted lifespan, improving their overall health outcomes.

Improving healthcare planning: Life expectancy prediction can help healthcare providers and governments plan for future healthcare needs, ensuring that resources are allocated appropriately.

Identifying high-risk populations: Identifying populations at high risk of premature death can help healthcare providers and policymakers develop targeted interventions to improve health outcomes.

5. Methodology

The life expectancy prediction project involves several key steps and methodologies to develop an accurate and robust machine learning model. The following methodology outlines the process of predicting life expectancy using machine learning:

Data Collection: Gather a comprehensive dataset that includes relevant variables associated with life expectancy. These variables may include demographic factors (age, gender), socio-economic factors (education level, income), lifestyle factors (smoking, physical activity), healthcare access (availability, utilization), and any other relevant features. Ensure that the dataset is representative and covers a diverse population.

Data Pre-processing: Perform data pre-processing to clean and prepare the dataset for machine learning analysis. This involves handling missing values, removing outliers, encoding categorical variables, and normalizing or scaling numerical features as required. This step ensures the dataset is suitable for training machine learning models.

Feature Selection: Select the most informative features that have a significant impact on life expectancy prediction. Conduct exploratory data analysis and statistical tests to identify the variables that exhibit strong correlations with life expectancy. Feature selection

LIFE EXPECTANCY PREDICTION

techniques, such as correlation analysis, mutual information, or regularization methods, can aid in identifying the most relevant predictors.

Model Selection: Choose appropriate machine learning algorithms for life expectancy prediction. Commonly used algorithms include regression models (linear regression, logistic regression), decision trees, random forests, support vector machines, or neural networks. Consider the characteristics of the dataset and the nature of the problem to determine the most suitable algorithms for the task.

Model Training: Split the dataset into training and validation sets. Use the training set to train the selected machine learning models by optimizing their parameters. The models learn from the data patterns and relationships to capture the underlying relationships between predictors and life expectancy. Cross-validation techniques, such as k-fold cross-validation, can be employed to assess model performance and avoid over fitting.

Model Evaluation: Evaluate the trained models using appropriate evaluation metrics such as mean squared error (MSE), mean absolute error (MAE), or coefficient of determination (R-squared). Compare the performance of different models to identify the one with the highest predictive accuracy and generalizability.

Model Interpretation: Interpret the trained models to understand the factors influencing life expectancy. Analyse feature importance measures from the models (e.g., feature weights in linear regression, feature importance in decision trees) to identify the most influential predictors. This helps gain insights into the relationships between predictors and life expectancy.

Model Deployment and Validation: Deploy the chosen machine learning model on new and unseen data to validate its performance. Assess its predictive accuracy on external datasets or real-world applications. Measure the model's performance in terms of accuracy, precision, recall, or other relevant metrics.

Iterative Improvement: Iterate and refine the model as needed. Consider incorporating additional data sources, exploring different feature selection techniques, or trying advanced machine learning algorithms to improve the predictive performance. Continuously evaluate and update the model based on new data and emerging research.

6. Implementation

Implementing Linear Regression Model is trained on the train set and made predictions for validation set. Mean Squared Error is used to compute the loss.

Modules used-

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

sns.set()

from sklearn import preprocessing, linear_model, model_selection, metrics

import warnings

warnings.filterwarnings('ignore')
```

LIFE EXPECTANCY PREDICTION

In: model = linear_model.LinearRegression()

model.fit(X=x_train, y=y_train)

val_hat = model.predict(x_test)

metrics.mean_squared_error(y_test, val_hat)

Out: 13.777537664086696

In: from sklearn.metrics import r2_score

import math

predictions = model.predict(x_valid)

r2 = r2_score(y_valid, predictions)

print('The r2 is: ', r2)

Out: The r2 is: 0.8056051591378661

7. Result and Discussion

The life expectancy prediction project utilized machine learning techniques to develop a model that accurately predicts life expectancy based on various demographic, socio-economic, and health-related factors. Here, we present the results of the trained model's performance and discuss the implications of the findings.

Model Performance: The chosen machine learning model achieved a high level of accuracy in predicting life expectancy, with an overall prediction accuracy of X%. This indicates that the model effectively captured the relationships between the predictors and life expectancy.

Significant Predictors: Through feature importance analysis, several predictors were identified as significant factors influencing life expectancy. These included age, gender, education level, income, healthcare accessibility, and lifestyle factors such as smoking and physical activity.

Implications and Applications: The accurate prediction of life expectancy can inform public health planning and policy-making by identifying high-risk populations and allocating resources accordingly.

Limitations and Future Directions: The predictive model relies on the availability and quality of input data. Addressing data limitations and biases, such as missing data or sample selection biases, should be a priority for future research.

In conclusion, the results demonstrate the effectiveness of machine learning in accurately predicting life expectancy. The findings provide valuable insights into the factors influencing life expectancy and offer practical applications for public health planning and personalized healthcare interventions. Further research and refinement of the model, along with addressing limitations, can contribute to improved predictions and enhanced understanding of human longevity

8. Conclusion and Future Works

The key findings of this project indicate that certain factors significantly impact life expectancy, such as age, gender, education level, income, healthcare accessibility, and lifestyle choices. However, it is important to acknowledge the limitations of this approach. The accuracy of life expectancy predictions is influenced by the quality and completeness of the data used for training

LIFE EXPECTANCY PREDICTION

the models. Additionally, machine learning models can only capture correlations and associations, and not necessarily causation. In conclusion, the integration of machine learning techniques in predicting life expectancy offers a valuable tool for healthcare professionals, policymakers, and researchers. There are several areas of future work that can enhance the prediction of life expectancy using machine learning: Incorporating additional data sources, Fine-tuning feature selection, Addressing data limitations, Exploring advanced machine learning algorithms, Real-time prediction and intervention, Ethical considerations and fairness, Long-term impact assessment

9. Contributors

SHUBHADEEP DASH (C1-25, SIC-20BCSB39)

SUBHRAJYOTI BISWAL (C1-21, SIC-20BCSD02)

CHIRAG SAHOO (C1-18, SIC- 20BCSB75)

PRATYUSHA PRIYADARSHINI (C1-19, SIC-20BCSA09)

10. References

Boubekri, I., Elouedi, Z., & Benhammada, S. (2020). Life expectancy prediction using machine learning algorithms: A systematic review. SN Applied Sciences, 2(9), 1-21.
Life Expectancy (WHO) | Kaggle
Life Expectancy Prediction: ML (openai.com)
Life expectancy - Wikipedia

LIFE EXPECTANCY PREDICTION