Capstone Project Concept Note and Implementation Plan

Project: Life expectancy prediction using Machine Learning.

Team Members

1. Abdhine Ben-Ali Mahamat Zenal

Concept Note

1. Project Overview

- This capstone project aims to predict life expectancy using machine learning techniques, utilizing data provided by the World Health Organization (WHO). It aligns with several Sustainable Development Goals (SDGs), including SDG 3 (Good Health and Well-being) and SDG 10 (Reduced Inequalities).
- The project addresses the critical issue of understanding factors influencing life expectancy globally, thereby potentially aiding policymakers in targeted interventions to improve public health outcomes.

2. Objectives

- Develop machine learning models to predict life expectancy based on demographic, socio-economic, and health-related factors.
- Evaluate the performance of various machine learning algorithms to identify the most accurate model.
- Provide insights into the significant predictors of life expectancy and their relative importance.
- Contribute to public health research by offering a predictive model that can assist in understanding and potentially improving life expectancy outcomes globally.

3. Background

- Life expectancy is a crucial indicator of a population's overall health and well-being. Factors influencing life expectancy include access to healthcare, economic stability, education levels, and environmental conditions. Previous studies have utilized regression and machine learning approaches to predict life expectancy based on these variables. A machine learning approach is advantageous due to its ability to handle complex interactions among predictors and potentially uncover non-linear relationships that traditional regression models might miss.

4. Methodology

The project will employ supervised machine learning techniques, focusing on regression models such as Linear Regression, Decision Trees, Random Forest, and Gradient Boosting. Feature selection and engineering will be crucial to enhance model performance and interpretability. Cross-validation and hyperparameter tuning will optimize model accuracy and generalizability.

5. Architecture Design Diagram

| Input Data | Preprocessing | ANN Model | Prediction | Training | Output |
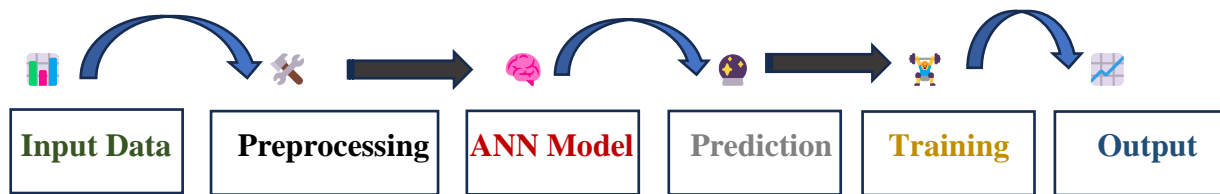
6. Data Sources

The primary data source will be datasets curated by the World Health Organization (WHO), containing country-level statistics on healthcare, socio-economic factors, and health outcomes. Data preprocessing steps will include handling missing values, normalization of numerical features, encoding categorical variables, and possibly feature scaling to ensure optimal model performance.

7. Literature Review

Recent studies have shown that machine learning models can effectively predict life expectancy using various datasets, including those from WHO. These approaches have highlighted the importance of socio-economic factors, healthcare accessibility, and environmental quality in determining life expectancy outcomes. This project builds upon existing research by applying a comprehensive set of machine learning algorithms to WHO data, aiming to provide accurate predictions and valuable insights.

Implementation Plan

1. Technology Stack

- Programming **Languages:** Python
- Libraries**:** Scikit-learn, Pandas, NumPy, Matplotlib, Seaborn
- Frameworks**:** Flask (for API deployment)
- Other **Tools:** Jupiter Notebook, Git.

2. Timeline

- o Data collection and preprocessing
  - Features Metadata
  - Exploratory Data Analysis
  - DataFrame Shape
  - Data Cleaning
  - Data Visualization
  - Data Preprocessing
  - Building ANN Model.
- o Deployment, documentation.

3. Milestones

- ✓ Completion of data collection and preprocessing
- ✓ Successful development and validation of machine learning models

✓ Deployment of the predictive model via a web interface or API.

4. Challenges and Mitigations

- o **Data Quality:** Rigorous preprocessing and validation procedures.
- o **Model Performance:** Experimentation with multiple algorithms and tuning hyperparameters.
- o **Technical Constraints:** Scalability considerations and efficient model deployment strategies.

5. Ethical Considerations

Ensuring data privacy compliance, mitigating bias in model predictions, and considering the potential socio-economic impact of the project are paramount. Transparency in model development and addressing biases in data sources will be prioritized.

6. References

[1] Sabri, R. (February, 2024). Life expectancy prediction with ANN. Kaggle. Retrieved June 20, 2024, from https://www.kaggle.com/code/ranasabrii/life-expectancy-regression-with-ann#Features-Metadata

[2] Sk, Khader Basha. (2022). MACHINE LEARNING METHOD FOR LIFE EXPECTANCY PREDICTION. Journal of Modern Physics. XI. 140-145.

[3] Jarshi, K. (2019). Life expectancy (WHO). Kaggle. Retrieved June 15, 2024, from https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who