Machine Learning Project

Proposal

DTSC691: Applied Data Science

Project Overview

The early identification of cognitive decline among individuals suffering from dual diseases of type 2 diabetes and hearing loss, while factoring in a host of lifestyle variables, represents the emerging healthcare challenge this project will be attempting to address by the implementation of a machine learning model. This need comes to greater prominence given the outcry for predicting tools by early interventions aimed at bettering patient outcomes, against the turning tide of diabetes prevalence from 537 million afflicted adults globally in 2021 to 783 million in 2045 (International Diabetes Federation, 2021).

Diabetes is linked to hearing loss by various comorbidities, which together currently affect 466 million individuals globally (World Health Organization, 2021). Another number emerging to change the lay of the world is one showing cognitive decline and dementia impinged on the mind of a little over 55 million people and projected to rise in the presence of aging populations; consequently, laid as quite a burden on public health (Livingston et al., 2020). Studies have shown that augmenting hearing loss after being diagnosed as diabetic increases cognitive impairment, but the various possible mechanisms for the seemingly intertwined situation are underexplored (Biessels & Despa, 2018).

The BRFSS, the Behavioral Risk Factor Surveillance System, is a large, state-based surveillance system for the collection of health-related risk factor information, which is to be used in furtherance of this project, especially self-reports of cognitive health, diabetes status, hearing ability, and other lifestyle habits (Centers for Disease Control and Prevention, 2022). Such data are extensive and informed enough to guide the building of predictive models to identify a patient at high risk for cognitive decline.

Machine learning models provide insight based on patterns and predictive factors that will enable healthcare providers to act. In particular, these insights will facilitate early cognitive health screenings, targeted lifestyle interventions, and a more proactive healthcare strategy that, together, promise yet higher quality of life while possibly lowering medically-related costs related to cognitive impairment and dementia.

Project Goals

Purpose:

The project mainly focuses on solving the challenge of early detection of cognitive decline in diabetes and hearing-loss patients. The cognitive decline, in many circumstances, is detected late in clinical settings and early prediction could be of immense help in formulating preventive strategies. The model would provide targeted screenings and healthcare interventions based on the modeled lifestyle factors causing the decline of cognitive health.

Project Focus:

This project will investigate how type 2 diabetes and hearing loss are related to cognitive decline, as well as how lifestyle factors may modify this risk. Specific research questions include:

1. To what extent does hearing loss, combined with type 2 diabetes, predict cognitive decline?

2. Which lifestyle factors (e.g., physical activity, smoking, diet) are associated with changes in cognitive health in this population?

Project Goals:

- To gather and process BRFSS data on diabetes, hearing loss, cognitive health, and other lifestyle factors.
- Perform exploratory data analysis to obtain insight into the prevalence and distribution of cognitive decline among participants with diabetes and hearing loss.
- Develop and refine a machine learning model for cognitive decline prediction.
- Provide a user interface for the model so that medical professionals can use it to evaluate risks.

Expected Outcomes:

This project aims to deliver several key outcomes:

1. **Predictive Model**: From the understanding developed surrounding insights on BRFSS data, a strong machine learning model has been providing appropriate predictions for the cognitive decline of diabetic individuals with hearing impairment.
2. **Risk Factor Insights**: These would provide the substantial clues to various researches with respect to relating diabetes, hearing loss, and brain health.
3. **Performance Evaluation**: Comprehensive reporting on model performance metrics, including accuracy and precision, to ensure reliability and validity of the findings.
4. **User Interface Development**: Creation of an intuitive user interface that enables healthcare providers and researchers to easily access and utilize the predictive model.
5. **Guidance for Interventions**: Insights that inform early screening protocols and interventions, targeting at-risk populations to mitigate cognitive decline.
6. **Contribution to Literature**: Documenting lessons learnt about potential publishing will help contribute to academic knowledge in both public health and machine learning applications.
7. **Educational Resources**: Produce materials summarizing methodology and analysis as a means of referencing future research.

These outcomes aim to translate technical findings into actionable insights that enhance cognitive health strategies for at-risk populations.

Project Description

Project Objective and Scope

This project mainly intends to develop a machine learning model to predict the cognitive decline of individuals with diseases including type 2 diabetes and hearing loss and is built on the data received by BRFSS. It reveals the relationships among these domains in public health, contributing to public health in terms of developing better strategies for disease prevention.

Objectives:

- **Predictive Modeling**: Create a reliable model that assesses the risk of cognitive decline based on diabetes, hearing loss, and relevant lifestyle factors, employing appropriate machine learning algorithms.
- **Understanding Interrelationships**: Investigate the connections between diabetes, hearing loss, and cognitive decline, utilizing exploratory data analysis to identify how these conditions impact each other.
- **Identification of Key Predictors**: Determine significant predictors of cognitive decline within the dataset, aiding in early intervention strategies.
- **Development of a User Interface**: Design a user-friendly interface to allow healthcare providers and patients to interact with the model, facilitating practical applications of the predictive insights.

Scope:

- **Data Source**: The project will utilize the BRFSS dataset, which provides extensive data on health-related risk factors and chronic conditions from a diverse population since 2011.
- **Target Population**: Focus will be on adults diagnosed with type 2 diabetes and hearing loss, exploring demographic factors such as age and socioeconomic status.
- **Methods**: Employ statistical analysis and machine learning techniques, including data cleaning, model training, and evaluation to ensure robust outcomes.
- **Deliverables**: Expected outputs include a predictive model, a comprehensive analysis report, a user interface, and a presentation summarizing the project’s findings.
- **Limitations**: The project will acknowledge constraints such as the nature of observational data and the need for further validation in clinical settings.

This project aims to establish a foundation for understanding the interplay between diabetes, hearing loss, and cognitive decline, ultimately enhancing health outcomes through predictive analytics.

Data Description:

- **Data Source**: The BRFSS dataset (2011-present), with a focus on the 2016 survey, combines landline and cellphone responses and contains information on health-related risk factors across U.S. states.
- **Collection Methods**: The BRFSS collects data on risk factors, such as lifestyle, demographic, and health behaviors.
- **Data Relevance**: BRFSS data is appropriate as it provides extensive information on cognitive health, diabetes prevalence, hearing ability, and related risk factors.

Exploratory Data Analysis (EDA)

The EDA will include:

- Statistical summaries and visualizations to examine data distribution and prevalence of cognitive decline, diabetes, and hearing loss.
- Correlation analysis to identify associations between lifestyle factors and cognitive decline.
- Subgroup analyses by age, gender, and other demographics to assess variation in risk.

Data Preparation and Cleaning:

- Handle missing data, possibly using imputation techniques for continuous data and categorical approaches where appropriate.
- Remove outliers or data inconsistencies in preparation for model training.
- Feature engineering to enhance model accuracy, such as categorizing age groups and binning income levels.

Model Training:

In this project, the model training phase will focus on selecting appropriate machine learning algorithms, optimizing their performance, and validating the results. The steps involved in this phase include:

1. **Model Selection**: Various machine learning algorithms will be considered to identify the best fit for the prediction of cognitive decline. Potential candidates include:

   - **Logistic Regression**: A foundational method for binary classification tasks, suitable for understanding the influence of predictor variables on cognitive decline.
   - **Random Forest**: An ensemble learning method that combines multiple decision trees, which is effective for handling non-linear relationships and interactions between features.
   - **Support Vector Machines (SVM)**: A powerful classification technique that can manage high-dimensional data and find the optimal hyperplane for separating classes.
   - **Gradient Boosting Machines (GBM)**: A robust method that builds models sequentially, allowing for improved predictive accuracy and performance.

2. **Training Procedures**: The selected models will be trained using a subset of the BRFSS data, with careful attention to balancing the dataset to address any class imbalances (e.g., between cognitive decline and non-decline cases). The training process will involve:

   - **Splitting the Data**: The dataset will be divided into training (70%), validation (15%), and test (15%) sets to ensure unbiased evaluation of model performance.
   - **Feature Selection**: Relevant features will be identified based on exploratory data analysis, considering demographic variables, health conditions, and lifestyle factors. Techniques like recursive feature elimination or LASSO regression may be applied to enhance model interpretability and reduce overfitting.

3. **Parameter Tuning**: Hyperparameter optimization will be performed using techniques such as grid search or random search to identify the best settings for each model. This step is crucial for improving model accuracy and reducing overfitting.

4. **Validation Strategies**: Cross-validation will be employed to assess model stability and generalizability. K-fold cross-validation will help evaluate how the model performs on unseen data by partitioning the training set into K subsets, iteratively training the model on K-1 subsets, and validating it on the remaining subset.

5. **Model Evaluation Metrics**: The performance of the trained models will be evaluated using various metrics, including:

  - **Accuracy**: The proportion of correctly predicted instances among the total instances.
  - **Precision and Recall**: To assess the model's effectiveness in identifying positive cases of cognitive decline.
  - **F1 Score**: The harmonic mean of precision and recall, providing a balanced measure of the model's performance.
  - **ROC-AUC**: The area under the Receiver Operating Characteristic curve will be calculated to evaluate the model's discriminative ability across different thresholds.

The training phase aims to establish a robust predictive model that accurately identifies individuals at risk for cognitive decline, setting the stage for further evaluation and practical application.

Model Evaluation:

- Use metrics such as accuracy, F1 score, AUC, and confusion matrix to evaluate model performance.
- Interpret results to assess model effectiveness in identifying predictive factors for cognitive decline.
- Perform feature importance analysis to identify the most influential factors contributing to cognitive decline.

User Interface Integration:

- A simple user interface will be developed in Jupyter notebooks, where users can input health data and receive cognitive decline risk predictions. This interface will provide healthcare practitioners with actionable insights and visualizations.

Capstone Complexity

This project is designed to meet master's level complexity by integrating multiple advanced concepts in machine learning, data analysis, and healthcare research. The complexity of the capstone project will be established through several key dimensions:

1. Data complexity: The principal dataset will be from the BRFSS because this dataset offers an extensive range of variables related to health behavior and chronic health conditions and demographic characteristics. This database in particular offers a very interesting opportunity for an exhaustive analysis regarding interaction effects between other lifestyle factors with type 2 diabetes and hearing loss. This interaction is also likely to include a lot of other strenuous ways of data manipulation and analysis. Analysis involves handling huge amounts of data, which shall, among other things, include handling missing values, taking note of outliers, and establishing some sort of consistency in diverse forms, therefore requiring thorough data collation, cleaning, and preprocessing.

2. Modeling techniques: A number of algorithms that would be applied to the project include logistic regressions, decision trees, and ensemble methods like random forests and gradient boosting. Each of the models shall be selected based on identifiable criteria such as applicability and interpretability for classification tasks, measured by performance scores. Complexity will also be further enhanced through processes such as hyper-parameter tuning and model optimization for an in-depth look at how each model performs in terms of predicting cognitive decline.

3. Statistical analysis: Complementing machine learning, the project will involve sophisticated statistical tests. Other than that, other analyses to be involved in the project include exploratory data analysis for correlations and trends within the data; multivariate analysis in examining how the three variables being analyzed interact among themselves, namely, diabetes, hearing loss, and cognitive health; and logistic regression to see the effect of lifestyle factors on the status of a person. All this, in turn, will give way to applying sophisticated statistical techniques able to dig deeper into the relations of variables and mechanisms of cognitive decline.

4. Integration of User Interface: The project will culminate in the development of a user-friendly interface to facilitate interactions with the machine learning model. This interface will allow users to input their data and receive predictions regarding their risk of cognitive decline. The integration of a user interface adds a layer of complexity as it requires an understanding of front-end and back-end development, ensuring seamless communication between the model and the end-user.

5. Real-world Application and Impact: Other than building a predictive model, the project would be developed in terms of effective contribution toward public health initiatives. By finding the at-risk subjects, it looks for actionable insights that drive impact in the field of healthcare policy and practice. Similarly, real-world applicability infuses a variety of rigorous validation and testing into the already complicated goals of the project.

6. Interdisciplinary Approach: This project integrates knowledge from various fields, including healthcare, statistics, and computer science, reflecting the interdisciplinary nature of modern research. By synthesizing these domains, the project exemplifies the complexity expected at the master's level and demonstrates an ability to navigate and apply concepts across disciplines effectively.

Through these various dimensions of complexity, the project will not only fulfill academic requirements but also provide a robust framework for understanding the interplay between type 2 diabetes, hearing loss, and cognitive decline, ultimately contributing valuable insights to the field of public health.

Software

1. **Python & Jupyter Notebooks**: For data analysis, preprocessing, model training, and result visualization.
2. **scikit-learn & XGBoost**: For implementing machine learning models and handling hyperparameter tuning.
3. **Pandas & NumPy**: For data manipulation and statistical analysis.
4. **Matplotlib & Seaborn**: For data visualization in EDA and result interpretation.
5. **Streamlit** (or similar) for the user interface, depending on feasibility.

Project Completion Plan

- **Week 1-2**: Finalize proposal, gather data from BRFSS, initial data preprocessing. Conduct exploratory data analysis, identify relevant features.
- **Week 3**: Clean and preprocess data, handle missing values, and create engineered features.
- **Week 4**: Implement initial models, conduct hyperparameter tuning.
- **Week 5**: Evaluate models, select the best-performing model, and begin UI development.
- **Week 6**: Finalize user interface and integrate the model.
- **Week 7**: Prepare final presentation and refine the project for submission.

Presentation Plan

The presentation will include a 30-minute video walkthrough covering:

1. Project objectives and the impact of diabetes and hearing loss on cognitive health.
2. Data exploration and feature selection process.
3. Model training, evaluation, and optimization.
4. Demonstration of the user interface and interpretability features.
5. Concluding remarks and potential real-world applications.

Resources

- **BRFSS Dataset**: https://catalog.data.gov/dataset/behavioral-risk-factor-surveillance-system-brfss-prevalence-data-2011-to-present/resource/6bb82759-3ca7-4e1d-a17d-89e5e63aba1c
- **Machine Learning Textbooks**: "Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow" by Aurélien Géron.
- **scikit-learn and XGBoost Documentation**: For model development and tuning.
- **Streamlit Documentation**: For creating an interactive interface.
- **Python Data Analysis Tutorials**: Relevant tutorials and guides for data analysis.

- Biessels, G. J., & Despa, F. (2018). Cognitive decline and dementia in diabetes mellitus: Mechanisms and clinical implications. *Nature Reviews Endocrinology*, 14(10), 591-604. https://doi.org/10.1038/s41574-018-0048-7
- Centers for Disease Control and Prevention. (2022). *Behavioral Risk Factor Surveillance System (BRFSS): Prevalence data & data analysis tools*. https://www.cdc.gov/brfss/
- International Diabetes Federation. (2021). *IDF Diabetes Atlas, 10th Edition*. https://diabetesatlas.org/
- Livingston, G., et al. (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *The Lancet*, 396(10248), 413-446. https://doi.org/10.1016/S0140-6736(20)30367-6
- World Health Organization. (2021). *World report on hearing*. https://www.who.int/publications/i/item/world-report-on-hearing