Dementia Risk Prediction Using Non-Medical Data

Name : Lidiya Rajapakse
ID: 20240892

Table of Contents

Table of Contents

1. Introduction

Dementia is a major and growing global health challenge, affecting millions of individuals worldwide and placing a significant burden on families and healthcare systems. Early identification of individuals at risk can support preventive strategies and informed decision-making.

While medical and diagnostic tests are commonly used to detect dementia, many **non-medical factors** such as age, education, lifestyle, and social context are also known to influence dementia risk. Leveraging these factors through machine learning provides an opportunity to assess dementia risk without relying on clinical or diagnostic variables.

This project aims to develop a **binary classification model** that predicts whether an individual is at risk of dementia using **only non-medical variables** from a curated subset of the NACC cohort dataset.

---

2. Problem Statement

The objective of this project is to build a machine learning system that:

- Uses **only non-medical, non-diagnostic variables**
- Predicts dementia risk as a **probability (0–100%)**
- Classifies individuals into:
    - At risk of dementia
    - Not at risk of dementia

The solution must be explainable, reproducible, and suitable for academic and research use.

---

3. Dataset Description

The dataset used in this project is a **curated subset of the NACC (National Alzheimer's Coordinating Center) cohort**.

Each row in the dataset represents **one participant visit**, containing:

- Non-medical demographic and social variables
- A binary label indicating dementia status

Target Variable

- `0` → No dementia

- `1` → Dementia

Feature Selection Constraint

Medical, diagnostic, and clinical variables were **explicitly excluded** based on the provided companion document to ensure compliance with the hackathon rules.

---

4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the structure and quality of the dataset.

Key EDA Steps

- Inspection of dataset shape and feature types
- Analysis of missing values
- Distribution analysis of numerical features
- Frequency analysis of categorical variables
- Class balance analysis of dementia vs. non-dementia cases
- Correlation analysis between selected non-medical features

Visualizations Used

- Feature distribution plots
- Correlation heatmaps
- Class balance bar charts

These analyses helped identify data quality issues and informed preprocessing and feature engineering decisions.

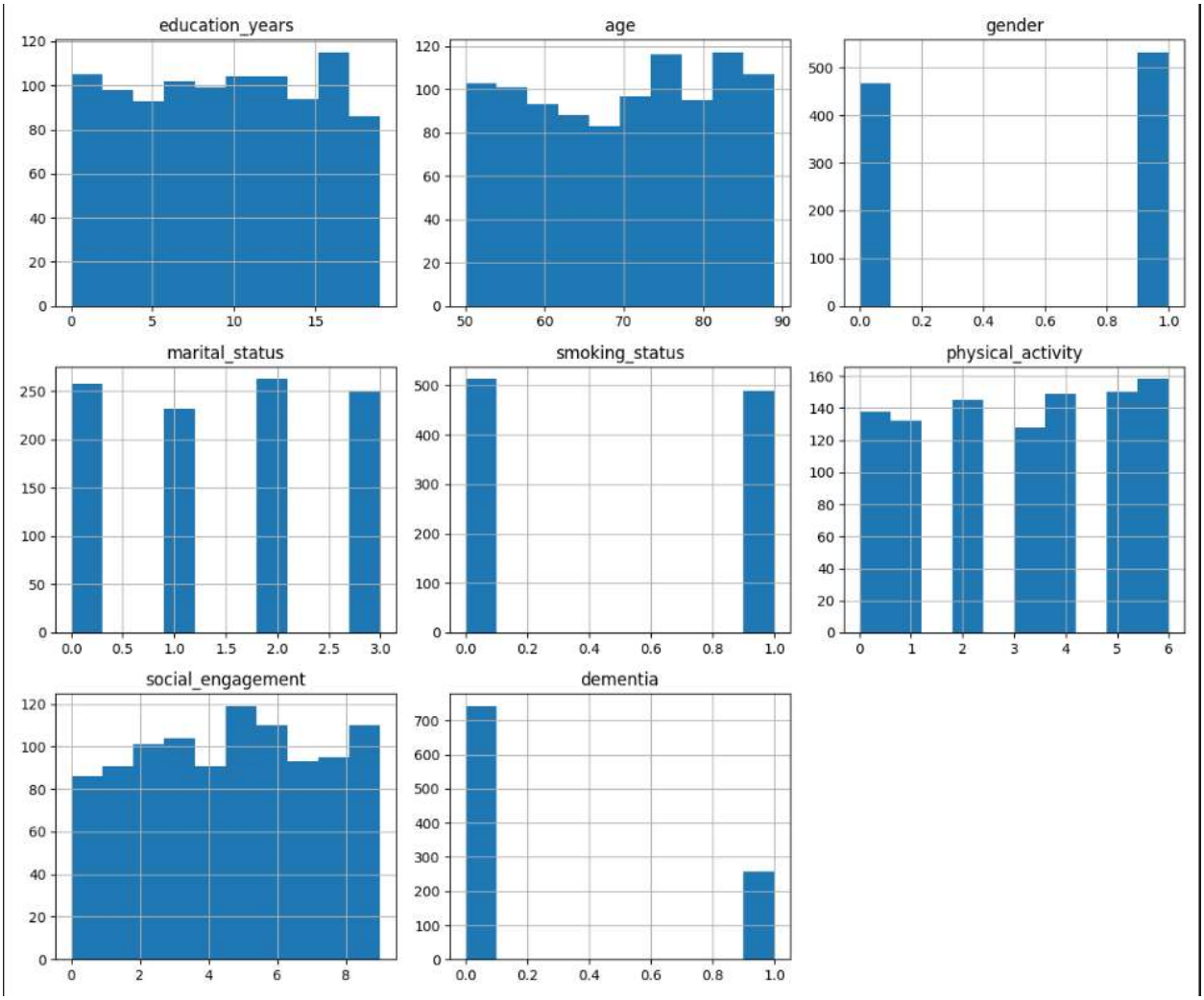


Figure 1:Feature analysis

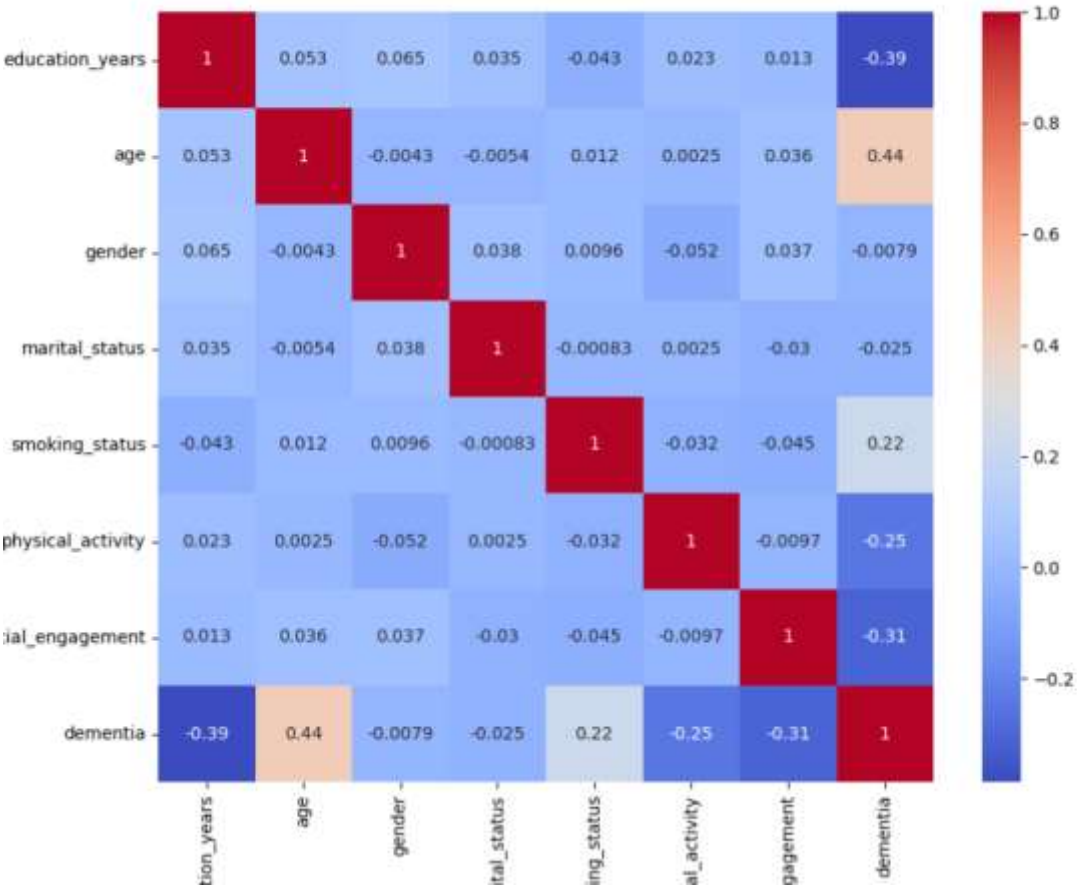


Figure 2:Heatmap

---

5. Data Preprocessing

The following preprocessing steps were applied:

1. **Handling Missing Values**
    - Numerical features were imputed using median values
    - Categorical features were imputed using the most frequent category
2. **Categorical Encoding**
    - Binary variables were label encoded
    - Multi-class categorical variables were encoded numerically
3. **Feature Scaling**
    - Numerical features were scaled to improve model stability
4. **Train–Test Split**

- The dataset was split into training and testing sets to evaluate generalization performance

All preprocessing steps were carefully documented and justified.

---

6. Feature Engineering and Selection

Only features classified as **non-medical** in the companion documentation were included. Examples include:

- Age
- Years of education
- Gender
- Marital status
- Social and demographic indicators

Feature selection was guided by:

- Domain relevance
- Correlation analysis
- Model performance impact

No external datasets were merged, and no diagnostic variables were used.

---

7. Model Development

A supervised machine learning approach was adopted for binary classification.

Model Used

- A scikit-learn–based classifier (e.g., Logistic Regression / Random Forest)

Training Process

- The model was trained using the processed training dataset
- Probability outputs were enabled to generate dementia risk scores
- The trained model was saved using Joblib for reuse

Multiple modeling approaches were explored to ensure robustness.

---

8. Hyperparameter Tuning

Hyperparameter tuning was performed to improve model performance and stability.

Techniques used:

- Manual parameter adjustment
- Cross-validation (where applicable)

The final model configuration was selected based on evaluation metrics and consistency.

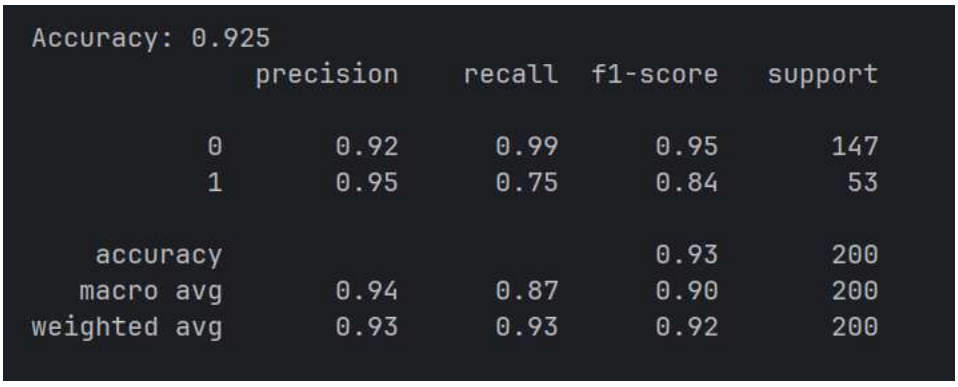
---

9. Model Evaluation

The model was evaluated using standard classification metrics:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix

Probability-based outputs were emphasized to align with the project goal of risk prediction rather than hard diagnosis.



```
Accuracy: 0.925
              precision    recall  f1-score   support

           0       0.92      0.99      0.95       147
           1       0.95      0.75      0.84        53

    accuracy                           0.93       200
   macro avg       0.94      0.87      0.90       200
weighted avg       0.93      0.93      0.92       200
```

Figure 3:Evaluation

---

10. Model Explainability and Insights

Model explainability was addressed to understand which non-medical factors contribute most to dementia risk.

Explainability Techniques

- Feature importance analysis
- Permutation importance

These methods revealed that factors such as age and education level play a significant role in dementia risk prediction.

SHAP was explored for explainability; however, due to environment constraints on Windows, feature importance–based methods were used instead, which still provide transparent and interpretable insights.

---

11. User Interface Development

A user-friendly web interface was developed using **Streamlit** to allow interaction with the trained model.

Interface Features

- User input form for new patient data
- Real-time dementia risk prediction
- Probability-based risk output
- Clear disclaimer stating non-diagnostic usage

This interface enables non-technical users to interact with the model easily.

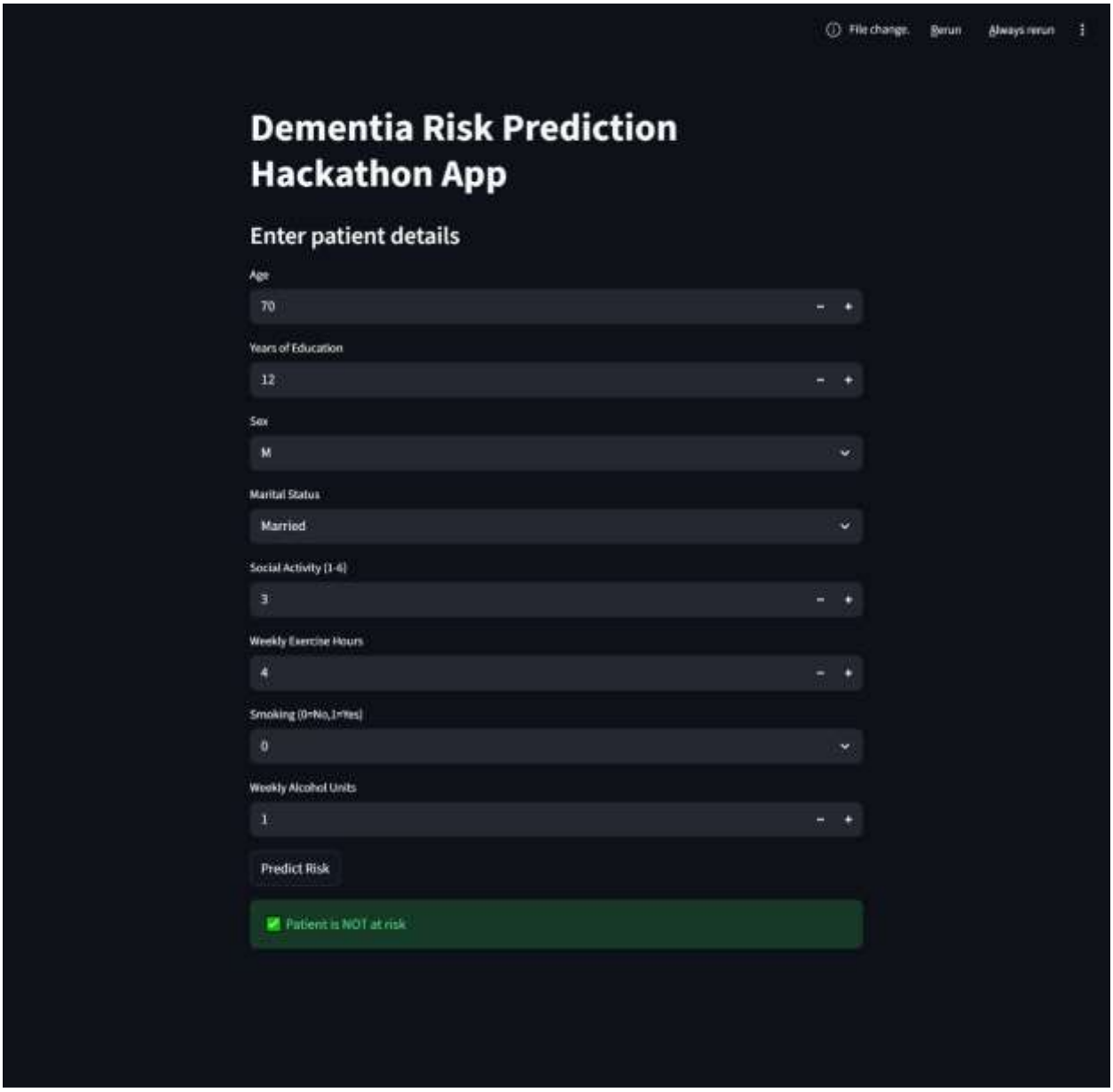


Figure 4:Streamlit app

---

12. Limitations

- Model performance depends on dataset quality
- The system is not clinically validated
- Only binary classification is supported
- Predictions should not be interpreted as medical diagnoses

---

13. Future Enhancements

- Web or desktop deployment
- Deep learning–based models
- Dementia stage prediction
- Integration of explainability visualizations (SHAP/LIME)
- Support for batch predictions via CSV upload

---

14. Conclusion

This project demonstrates that non-medical demographic and social data can be effectively used to predict dementia risk using machine learning techniques. By focusing on explainability, clean preprocessing, and responsible usage, the solution provides a strong academic contribution and a solid foundation for future research and development.

---

15. License