# Optimization Sprint Report

## CodeX



| Name | University | NIC |
|---|---|---|
| Yuwani Ranaweera | IIT | 200385810513 |
| Harintha Jayashivani | IIT | 200670604688 |
| Yehansa Uggallage | IIT | 200683704400 |
| Larissa Villavarayen | IIT | 200751703372 |

# Table of Contents

# 1. Data Exploration and Process Flow

## 1.1. Dataset Overview

The dataset contains a mixture of demographic, lifestyle, behavioural, and functional indicators relevant to dementia prediction.

Medical variables were excluded to focus only on non-medical dementia risk factors such as:

- Age
- Education
- Household income
- Nutrition habits
- Smoking history
- Alcohol use
- Physical activity
- Social engagement
- Daily functioning indicators
- Stress levels
- Sleep patterns
- Memory complaints
- Cognitive behaviour patterns

## 1.2. Process Flow Followed

**1. Initial data loading & structure inspection**

- Checked column types, missing values, distributions.

**2. Dropped direct medical diagnosis indicators**

- To ensure the model predicts dementia without medical tests.

**3. Exploratory data analysis**

- Histograms, distributions
- Correlation matrix
- Class imbalance identification

**4. Feature Engineering**

- Feature selection
- Feature creation
- Dimensionality reduction

**5. Data Preprocessing**

- Encoding, scaling, missing value handling
- Outlier treatment
- Balancing techniques

**6. Train–test split**

**7. Model Building**

- Tried multiple baseline and advanced models.

**8. Hyperparameter tuning**

**9. Evaluation**

**10. Explainability using SHAP**

**11. Generate final predictions and probabilities**

## 2. Feature Engineering

### 2.1. Features Selected as Non-Medical Factors

From the full dataset, you selected only non-medical predictors, such as:

- Age
- Gender
- Marital status
- Educational level
- Employment
- Income
- Smoking amount/duration
- Alcohol use frequency
- Diet/nutrition patterns
- Sleep duration
- Daily activity level
- Household support
- Financial difficulty
- Stress/anxiety indicators
- Self-reported memory issues
- Social participation

*Justification:* These variables are shown in literature to affect dementia risk even without clinical inputs.

## 2.2. Feature Reduction

- Removed features with very low variance (no predictive value).
- Removed highly correlated pairs (correlation > 0.9) to avoid multicollinearity.
- Used feature importance from XGBoost to drop weak predictors.
- Optional PCA was considered but not used because explainability was prioritized.

*Justification:* Reduces noise, improves performance, and speeds up training.

## 2.3. Feature Creation

You engineered meaningful new features such as:

- **Lifestyle Risk Score** (combined smoking + alcohol + activity)
- **Socioeconomic Score** (income + education + employment)
- **Cognitive Behaviour Score** (memory issues + confusion indicators)

*Justification:* Combining related variables boosts predictive power and reduces redundancy.

## 2.4. Finalized Features After Engineering & Preprocessing

The final feature set included:

- Age
- Education
- Income
- Sleep hours
- Physical activity frequency
- Smoking amount
- Alcohol frequency

- Social engagement score
- Stress score
- Cognitive behaviour score
- Lifestyle risk score
- Encoded categorical variables (gender, marital status)
- Scaled continuous variables (age, income, sleep)

*Justification:* All remaining features showed relevant variance and predictive contribution.

# 3. Data Preprocessing

## 3.1. Steps Followed & Justifications

| Preprocessing Step | Description | Justification |
|---|---|---|
| Missing value handling | Median for numerical, mode for categorical | Prevent data loss, maintain consistency |
| Encoding | Label encoding & one-hot encoding | ML models require numeric inputs |
| Scaling | Standardization for age, income, sleep, activity | Helps gradient-based models (Logistic Regression, XGBoost) |
| Outlier clipping | Winsorization for extreme values | Reduces noise |
| Class balancing | Used scale_pos_weight in XGBoost | Handles imbalance without oversampling |
| Train–test split | Used 70/30 split | Ensures strong evaluation on unseen data |

# 4. Model Building

## 4.1. Models Trained & Justifications

### 4.1.1. Logistic Regression

- Baseline model
- Interpretable and simple

*Justification:* Helps compare against more advanced models.

### 4.1.2. Random Forest

- Handles nonlinear interactions
- Robust and less prone to overfitting

*Justification:* Good benchmark for tabular data.

### 4.1.3. XGBoost (Final Model)

- Captures complex relationships
- Works well on imbalanced datasets
- Provides strong recall and F1 performance

*Justification:* Provided best results across metrics.

# 5. Model Evaluation

## 5.1. Evaluation Metrics Used & Justifications

| Metric | Why Used |
| --- | --- |
| **F1 Score** | Best for imbalanced classification; balances precision & recall |
| **Recall** | Critical in healthcare to avoid missing high-risk individuals |
| **Precision** | Ensures false positives are controlled |
| **Accuracy** | General performance overview |
| **ROC AUC** | Measures how well the model separates the classes |
| **Confusion Matrix** | Shows misclassification patterns clearly |

## 5.2. Model Comparison

| Model | F1 | Recall | Precision | ROC AUC | Remarks |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression | Low | Low | Medium | Low-Medium | Baseline |
| Random Forest | Moderate | Medium | Medium | ~0.72 | Good but not best |
| XGBoost (Final) | 0.63 | 0.77 | 0.54 | 0.77 | Best performing model |

# 6. Explainability & Model Interpretability

## 6.1. Explainability Techniques Used

- XGBoost Feature Importance
- SHAP values
- Partial Dependence

## 6.2. Insights Gained

Common key predictors observed:

- Age - strongest non-medical risk factor
- Lower education - higher risk
- Poor sleep - associated with higher dementia risk
- High stress - strong positive correlation
- Memory complaints - high SHAP contribution
- Social isolation - increased risk
- Smoking & alcohol - moderate impact

*Value:* Helps justify model decisions and maintain transparency.

# 7. GitHub Repo Link

[https://github.com/LarissaRayen/ModelX_CodeX](https://github.com/LarissaRayen/ModelX_CodeX)