# Puppet: Dementia Risk Prediction System

## Project Report - Team T74

Optimization Sprint Report

**Developer:** Bihara Malith
**University:** NIBM
**NIC:** 200204701757
**Team:** T74 - puppet
**Hackathon:** ModelX
**Date:** November 2025

---

## Executive Summary

This project presents **Puppet**, an intelligent dementia risk prediction system that predicts cognitive decline risk using only non-medical, self-reported information. By leveraging machine learning on the National Alzheimer's Coordinating Center (NACC) dataset, the system achieves high accuracy while remaining accessible to the general public without requiring medical tests or clinical expertise.

**Key Achievements:**

- Developed 3 machine learning models with hyperparameter optimization
- Achieved ROC-AUC scores above 0.85 on validation data
- Created interactive web-based risk prediction tool
- Identified 33 accessible non-medical predictive features
- Built production-ready application with real-time predictions

---

## 1. Problem Statement

### 1.1 Background

Dementia affects millions globally, with early detection being crucial for intervention and treatment planning. However, traditional diagnostic methods require expensive medical tests, specialized equipment, and clinical expertise, creating barriers to early screening.

### 1.2 Objective

Develop a machine learning system that:

- Predicts dementia risk using only self-reported, non-medical information
- Provides accessible screening without medical tests
- Offers interpretable results for general public understanding
- Achieves clinically relevant prediction accuracy

### 1.3 Success Criteria

- ROC-AUC score > 0.80 on test data
- Use only non-medical features (no lab tests, brain scans, etc.)
- Interactive web interface for real-time predictions
- Interpretable feature importance analysis

---

## 2. Dataset

### 2.1 Data Source

**National Alzheimer's Coordinating Center (NACC)** dataset containing comprehensive patient information from Alzheimer's Disease Research Centers across the United States.

### 2.2 Feature Selection

From the original dataset, **33 non-medical features** were carefully selected based on:

- Accessibility (self-reportable without medical tests)
- Clinical relevance to cognitive health
- Data quality and completeness

### 2.3 Feature Categories

#### Demographics (9 features)

- Age (NACCAGE, NACCAGEB)
- Sex
- Education level (EDUC)
- Marital status (MARISTAT)
- Living situation (NACCLIVS)
- Race and ethnicity (RACE, HISPANIC)
- Handedness (HANDED)

#### Lifestyle Factors (3 features)

- Smoking history (TOBAC100, SMOKYRS)
- Alcohol abuse (ALCOHOL)

#### Medical History (14 features)

- Cardiovascular: Heart attack (CVHATT), Atrial fibrillation (CVAFIB), Heart failure (CVCHF)
- Cerebrovascular: Stroke (CBSTROKE), TIA (CBTIA)
- Metabolic: Diabetes, Hypertension (HYPERTEN), High cholesterol (HYPERCHO)
- Other: Traumatic brain injury (NACCTBI), Sleep apnea (APNEA), Depression (DEP2YRS)

#### Physical Measurements (7 features)

- BMI (NACCBMI)
- Height and Weight
- Blood pressure (BPSYS, BPDIAS)
- Hearing status (HEARING, HEARAID)

#### Family History (3 features)

- Family cognitive issues (NACCFAM)
- Mother's cognitive status (NACCMOM)
- Father's cognitive status (NACCDAD)

### 2.4 Target Variable

**Binary Classification:**

- Class 0: No dementia risk
- Class 1: At risk for dementia

### 2.5 Data Statistics

- **Total Features:** 33 non-medical predictors
- **Numeric Features:** 9 (continuous measurements)
- **Categorical Features:** 24 (binary and ordinal)
- **Class Distribution:** Handled through stratified sampling and cross-validation

---

## 3. Methodology

### 3.1 Data Preprocessing

#### 3.1.1 Exploratory Data Analysis (EDA)

- Generated comprehensive profiling report using ydata-profiling
- Analyzed feature distributions and correlations
- Identified missing value patterns
- Examined class imbalance

#### 3.1.2 Data Cleaning

- Handled missing values appropriately per feature type
- Validated feature ranges and data types
- Removed or imputed outliers based on clinical plausibility
- Ensured data consistency across features

#### 3.1.3 Feature Engineering

- Calculated BMI from height and weight
- Standardized numeric features using StandardScaler
- Maintained categorical features in original encoding
- Created train-test splits with stratification

### 3.2 Machine Learning Models

#### 3.2.1 Model Selection

Three algorithms were chosen for comparison:

**1. Logistic Regression**

- **Rationale:** Baseline interpretable model
- **Advantages:** Fast training, interpretable coefficients
- **Use Case:** Benchmark for comparison

**2. Random Forest Classifier**

- **Rationale:** Ensemble learning with feature importance
- **Advantages:** Handles non-linear relationships, robust to outliers
- **Use Case:** Primary production model

**3. XGBoost Classifier**

- **Rationale:** Gradient boosting for high performance
- **Advantages:** State-of-the-art accuracy, handles imbalanced data
- **Use Case:** Performance comparison

#### 3.2.2 Training Strategy

- **Cross-Validation:** 5-fold stratified CV to ensure robust evaluation
- **Metric Selection:** ROC-AUC as primary metric (handles class imbalance)
- **Train-Test Split:** 80-20 split with stratification
- **Random State:** Fixed seed for reproducibility

### 3.3 Hyperparameter Tuning

#### 3.3.1 Random Forest Optimization

Used RandomizedSearchCV with 20 iterations to optimize:

- `n_estimators` : [100, 200, 300, 500]
- `max_depth` : [10, 20, 30, None]
- `min_samples_split` : [2, 5, 10]
- `min_samples_leaf` : [1, 2, 4]
- `max_features` : ['sqrt', 'log2']

**Best Parameters Found:**

- n_estimators: 300
- max_depth: 20
- min_samples_split: 5
- min_samples_leaf: 2
- max_features: 'sqrt'

#### 3.3.2 Feature Scaling

- Applied StandardScaler to numeric features
- Preserved categorical feature encoding
- Saved scaler for production use

---

## 4. Results

### 4.1 Model Performance Comparison

| Model | Accuracy | ROC-AUC | CV Mean | CV Std |
|---|---|---|---|---|
| Logistic Regression | 0.6699 | 0.7290 | 0.7300 | 0.0023 |
| Random Forest | **0.7757** | **0.8528** | **0.8395** | 0.0019 |
| XGBoost | 0.7037 | 0.7769 | 0.7743 | 0.0032 |

reports/model_comparison_results.csv

### 4.2 Best Model Selection

**Tuned Random Forest** was selected as the production model based on:

- Highest ROC-AUC score
- Balanced precision-recall trade-off
- Robust cross-validation performance
- Interpretable feature importance

### 4.3 Feature Importance Analysis

**Complete Feature Importance Ranking (All 33 Features):**

| Rank | Feature | Description | Importance |
|---|---|---|---|
| 1 | NACCAGEB | Age at baseline visit | 10.43% |
| 2 | NACCLIVS | Living situation | 8.87% |
| 3 | DEP2YRS | Recent depression (last 2 years) | 8.47% |
| 4 | EDUC | Education level (years) | 7.41% |
| 5 | NACCBMI | Body Mass Index | 7.07% |
| 6 | WEIGHT | Body weight (pounds) | 6.92% |
| 7 | NACCAGE | Current age | 6.70% |
| 8 | HEIGHT | Body height (inches) | 6.68% |
| 9 | BPSYS | Systolic blood pressure | 5.40% |
| 10 | BPDIAS | Diastolic blood pressure | 4.80% |
| 11 | SEX | Biological sex | 3.47% |
| 12 | SMOKYRS | Years of smoking | 3.27% |
| 13 | MARISTAT | Marital status | 2.69% |
| 14 | RACE | Race/ethnicity | 1.79% |
| 15 | NACCMOM | Mother's cognitive issues | 1.50% |
| 16 | HYPERCHO | High cholesterol | 1.42% |
| 17 | HYPERTEN | Hypertension | 1.39% |
| 18 | NACCFAM | Family cognitive history | 1.21% |
| 19 | HANDED | Handedness | 1.02% |
| 20 | CBSTROKE | Stroke history | 1.00% |
| 21 | NACCDAD | Father's cognitive issues | 0.97% |
| 22 | TOBAC100 | Smoked 100+ cigarettes | 0.87% |
| 23 | HEARING | Hearing ability | 0.85% |
| 24 | HEARAID | Hearing aid use | 0.81% |
| 25 | DIABETES | Diabetes diagnosis | 0.80% |
| 26 | HISPANIC | Hispanic/Latino ethnicity | 0.72% |
| 27 | NACCTBI | Traumatic brain injury | 0.67% |
| 28 | ALCOHOL | Alcohol abuse history | 0.65% |
| 29 | CVAFIB | Atrial fibrillation | 0.64% |

| Rank | Feature | Description | Importance |
|---|---|---|---|
| 30 | CBTIA | TIA (mini-stroke) | 0.54% |
| 31 | CVHATT | Heart attack | 0.47% |
| 32 | CVCHF | Heart failure | 0.28% |
| 33 | APNEA | Sleep apnea | 0.24% |

#### Feature Category Analysis:

**Demographics & Social (9 features):** 28.68%

- Age factors (NACCAGEB, NACCAGE): 17.13%
- Living situation: 8.87%
- Education: 7.41%
- Marital status: 2.69%
- Race/ethnicity: 2.51%
- Sex: 3.47%
- Handedness: 1.02%

**Physical Measurements (7 features):** 31.87%

- BMI: 7.07%
- Weight: 6.92%
- Height: 6.68%
- Blood pressure (BPSYS, BPDIAS): 10.20%
- Hearing factors: 1.66%

**Mental Health (1 feature):** 8.47%

- Depression: 8.47%

**Lifestyle Factors (3 features):** 4.79%

- Smoking (SMOKYRS, TOBAC100): 4.14%
- Alcohol abuse: 0.65%

**Medical History (10 features):** 7.58%

- Cardiovascular (CVHATT, CVAFIB, CVCHF): 1.39%
- Cerebrovascular (CBSTROKE, CBTIA): 1.54%
- Metabolic (DIABETES, HYPERTEN, HYPERCHO): 3.61%
- Other (NACCTBI, APNEA): 0.91%

**Family History (3 features):** 3.68%

- General family history: 1.21%
- Mother's history: 1.50%
- Father's history: 0.97%

#### Key Insights:

- **Age** is the strongest single predictor (non-modifiable risk factor)
- **Social factors** (living situation) show significant predictive power
- **Mental health** (depression) strongly correlates with cognitive risk
- **Lifestyle factors** (BMI, blood pressure) are modifiable risk factors
- **Family history** provides important genetic/environmental context

---

# 5. Web Application

## 5.1 Technology Stack

- **Framework:** Streamlit (Python web framework)
- **Visualization:** Plotly (interactive charts)
- **ML Libraries:** Scikit-learn, XGBoost, Joblib
- **Data Processing:** Pandas, NumPy

## 5.2 Application Features

### 5.2.1 Interactive Pages

1. **Home Page**
    - Project overview and highlights
    - Model performance summary visualization
    - Quick navigation to all features
2. **Risk Predictor** ⭐ Core Feature

- User-friendly form with 33 input fields
- Real-time BMI and blood pressure calculation
- Risk score visualization with gauge chart
- Probability percentage display
- Medical disclaimer

3. **Project Deliverables**
    - Embedded EDA report (ydata-profiling)
    - Key visualizations (correlations, distributions)
    - Dataset summary statistics
4. **Model Performance**
    - Comparison tables and charts
    - Confusion matrices for all models
    - ROC curves visualization
    - Hyperparameter tuning results
5. **Feature Importance**
    - Interactive importance ranking
    - Top features visualization
    - Clinical implications discussion
6. **About Page**
    - Methodology documentation
    - Project statistics
    - Team information and disclaimers

#### 5.2.2 User Experience Design

- **Intuitive Navigation:** Sidebar with emoji icons
- **Responsive Layout:** Multi-column design for better space usage
- **Visual Feedback:** Color-coded risk levels (green/red)
- **Help Text:** Tooltips explaining each input field
- **Professional Styling:** Custom CSS for polished appearance

### 5.3 Production Deployment

- **Model Loading:** Cached loading using @st.cache_resource
- **Prediction Pipeline:** Standardized scaling → model inference
- **Error Handling:** Graceful fallbacks for missing files
- **Performance:** Optimized for fast response times

---

## 6. Technical Implementation

### 6.1 Code Structure

```
ModelX_Hackthon_team-puppet/
├── EDA.py                          # Exploratory data analysis script
├── modeling.py                     # Model training and evaluation
├── hyperparameter_tuning.py        # Random Forest optimization
├── config.py                       # Configuration and file paths
├── requirements.txt                # Python dependencies
├── frontend/
│   └── streamlit_app.py           # Web application
├── models/
│   ├── tuned_random_forest.joblib # Production model (530 MB)
│   ├── best_model.joblib          # Alternative model (559 MB)
│   └── scaler.joblib              # Feature scaler
└── reports/
    ├── feature_importance_tuned_rf.csv
    ├── model_comparison_results.csv
    ├── modelx_eda_report.html
    └── *.png                      # Visualizations
```

### 6.2 Key Technologies

- **Python 3.8+**
- **Scikit-learn:** Model training and evaluation
- **XGBoost:** Gradient boosting implementation
- **Streamlit:** Web application framework

- **Plotly:** Interactive visualizations
- **Pandas/NumPy:** Data manipulation
- **Joblib:** Model serialization
- **ydata-profiling:** Automated EDA

### 6.3 Reproducibility

- Fixed random seeds across all scripts
- Version-controlled dependencies in requirements.txt
- Saved models and scalers for consistent predictions
- Documented hyperparameter configurations

---

## 7. Validation & Testing

### 7.1 Model Validation

- **Cross-Validation:** 5-fold stratified CV for unbiased estimates
- **Hold-out Test Set:** 20% of data reserved for final evaluation
- **Stratification:** Maintained class distribution in all splits
- **Multiple Metrics:** Evaluated accuracy, ROC-AUC, precision, recall, F1

### 7.2 Application Testing

- **Unit Testing:** Individual component validation
- **Integration Testing:** End-to-end prediction pipeline
- **User Testing:** Interface usability verification
- **Edge Cases:** Handling of extreme input values

### 7.3 Performance Metrics

- **Prediction Speed:** < 1 second per inference
- **Model Size:** ~530 MB (acceptable for deployment)
- **Memory Usage:** Optimized for standard hardware
- **Scalability:** Handles concurrent users via Streamlit

---

## 8. Limitations & Future Work

### 8.1 Current Limitations

1. **Binary Classification:** Does not predict dementia severity or type
2. **Self-Reported Data:** Subject to recall bias and inaccuracies
3. **Dataset Bias:** May not generalize to all populations equally
4. **Temporal Aspect:** Single time-point assessment, not longitudinal
5. **Medical Validation:** Requires clinical trials for real-world deployment

### 8.2 Future Enhancements

#### Short-term (Next 3-6 months)

- **Multi-class Classification:** Predict dementia type (Alzheimer's, vascular, etc.)
- **Confidence Intervals:** Provide uncertainty estimates with predictions
- **Explainability:** Add SHAP values for individual prediction explanations
- **Mobile App:** Deploy as mobile application for wider accessibility
- **Language Support:** Multi-language interface for global reach

#### Medium-term (6-12 months)

- **Longitudinal Modeling:** Track risk changes over time
- **Feature Expansion:** Include additional validated predictors
- **Clinical Integration:** API for healthcare systems
- **Personalized Recommendations:** Risk-reduction action plans
- **Cloud Deployment:** Scale to handle larger user base

#### Long-term (1-2 years)

- **Clinical Trials:** Validate in real-world healthcare settings
- **Regulatory Approval:** Pursue FDA/CE certification if applicable
- **Insurance Integration:** Connect with health insurance systems
- **Research Platform:** Enable researchers to query model insights
- **Federated Learning:** Train on distributed data while preserving privacy

---

# 9. Ethical Considerations

## 9.1 Privacy & Security

- **No Data Storage:** User inputs not stored or transmitted
- **Local Processing:** All predictions computed client-side where possible
- **Anonymization:** No personally identifiable information required
- **GDPR Compliance:** Designed with data protection in mind

## 9.2 Medical Ethics

- **Clear Disclaimers:** Prominent warnings about limitations
- **Not Diagnostic:** Explicitly stated as screening tool, not diagnosis
- **Professional Guidance:** Encourages consultation with healthcare providers
- **Accessibility:** Free and open-source for equitable access

## 9.3 Bias & Fairness

- **Dataset Diversity:** Acknowledged need for diverse training data
- **Fairness Auditing:** Ongoing evaluation across demographic groups
- **Transparent Limitations:** Clear communication about model boundaries
- **Inclusive Design:** Interface accessible to various education levels

---

# 10. Conclusion

## 10.1 Key Achievements

This project successfully demonstrates that:

1. **Dementia risk can be predicted** with reasonable accuracy using only non-medical features
2. **Machine learning models** (Random Forest) outperform traditional statistical approaches
3. **Accessible screening tools** can be built without requiring expensive medical tests
4. **Feature importance analysis** reveals modifiable risk factors for intervention
5. **Web-based deployment** makes the tool available to the general public

## 10.2 Impact Potential

The Puppet system has potential to:

- **Democratize Early Screening:** Enable risk assessment without medical access
- **Support Healthcare Systems:** Reduce burden through preliminary screening
- **Empower Individuals:** Provide actionable insights on modifiable risk factors
- **Advance Research:** Identify new patterns in cognitive decline prediction
- **Guide Interventions:** Inform targeted prevention strategies

## 10.3 Final Remarks

While this tool represents a significant step forward in accessible dementia risk assessment, it is crucial to emphasize that it **complements, not replaces**, professional medical evaluation. The system serves as an educational and awareness tool, encouraging individuals to seek appropriate medical care when risk factors are identified.

The combination of machine learning, accessible features, and user-friendly interface demonstrates the potential for AI to address critical healthcare challenges while maintaining ethical responsibility and medical accuracy.

---

# 11. References


1. National Alzheimer's Coordinating Center (NACC) Dataset
2. Scikit-learn Documentation: https://scikit-learn.org/
3. XGBoost Documentation: https://xgboost.readthedocs.io/
4. Streamlit Documentation: https://docs.streamlit.io/
5. Alzheimer's Association Guidelines
6. WHO Dementia Risk Reduction Guidelines


---

# 12. Appendices

## Appendix A: Installation Instructions

```
# Clone repository
git clone https://github.com/biharamalith/ModelX_Hackthon_team-puppet.git
cd ModelX_Hackthon_team-puppet

# Install dependencies
pip install -r requirements.txt

# Generate models (if not included)
python modeling.py
python hyperparameter_tuning.py

# Run application
cd frontend
streamlit run streamlit_app.py
```

## Appendix B: Dependencies

See `requirements.txt` for complete list including:

- streamlit >= 1.28.0
- scikit-learn
- xgboost
- pandas
- numpy
- plotly >= 5.17.0
- joblib
- ydata-profiling

## Appendix C: Model Files

- `tuned_random_forest.joblib` : 530.16 MB
- `best_model.joblib` : 559.64 MB
- `scaler.joblib` : < 1 MB

## Appendix D: Contact Information

- **Developer:** Bihara Malith
- **Team:** T74
- **GitHub:** https://github.com/biharamalith/ModelX_Hackthon_team-puppet

---

**Document Version:** 1.0
**Last Updated:** November 17, 2025
**License:** [Specify license if applicable]

---