Happiness Prediction from Survey Data: Project Report

---

1. Introduction

Understanding the factors that influence individual happiness has become a significant area of interest across policy, economics, and social sciences. In this project, we used machine learning models to predict self-reported happiness levels using data from the Caucasus Barometer 2024 survey (Armenia). The project aimed to clean and transform raw survey data, engineer meaningful features, and evaluate predictive models to understand which variables most strongly influence happiness.

---

2. Dataset Overview

- **Source**: Caucasus Barometer 2024 (Armenia)
- **File Format**: `.dta` file (Stata format), converted to `.csv` using a custom Python script
- **Sample Size**: ~1,500 respondents
- **Target Variable**: `RATEHAP` – Self-reported happiness
- **Features**: Demographic, socio-economic, health, trust, religion, media behavior, and political attitude questions

---

3. Data Cleaning and Preprocessing

- Removed system-generated or irrelevant columns: `ID`, `INDWT`, `PSU`, `HHWT`, etc.
- Mapped non-numeric responses (e.g., "Very satisfied", "Don’t know") to numeric values or `NaN`

- Dropped or imputed rows with missing target or essential feature values
- Applied one-hot encoding to categorical features
- Developed a custom `replace_survey_words()` function to standardize and convert over 50 survey response phrases into numeric equivalents

---

4. Feature Engineering

- Performed encoding with `pd.get_dummies()` for categorical variables without ordinal scale
- Created numeric mappings for ordinal scales (e.g., level of agreement, trust, importance)
- Aligned training, validation, and test sets to have consistent columns using `DataFrame.align()`
- Selected top 40 features based on Random Forest feature importance to reduce dimensionality
- Applied `StandardScaler` to scale numeric features for KNN model compatibility

---

5. Target Variable Transformation

The original `RATEHAP` scale ranged from 1 (extremely unhappy) to 10 (extremely happy). To simplify classification, we grouped responses into three categories:

- **Low Happiness**: 1–2
- **Mid Happiness**: 3–5
- **High Happiness**: 6–10

This transformation enabled a balanced and interpretable multi-class classification problem.

---

6. Models Implemented

1. K-Nearest Neighbors (KNN)

- Selected optimal `k` using elbow method (k = 9)
- Applied feature scaling with `StandardScaler`
- **Test Accuracy**: 70.59%

2. Decision Tree Classifier

- Captured non-linear patterns without requiring feature scaling
- Prone to overfitting but interpretable
- **Test Accuracy**: 58.82%

3. Logistic Regression

- Initially considered, but excluded from final evaluation due to unmet assumptions (linearity, low multicollinearity)

---

7. Results and Insights

Top Predictive Features (via Random Forest):

- **LIFESAT** – Life satisfaction
- **HLTH / RATEHLTH** – Self-rated health
- **GALLTRU / TRUST1** – General trust in others
- **FATE** – Belief in fate vs. personal control

- **RLGNIMPT, RLGNBELIEVE** – Religious importance and belief
- **Education level, internet usage, political orientation, and fairness perception** also contributed meaningfully

Model Comparison

Model	Test Accuracy
KNN (k=9)	70.59%
Decision Tree	58.82%
Logistic Regression	Not suitable

---

8. File Structure

```
Happiness Classification/
├── Happiness_Prediction_Model.ipynb      # Full modeling pipeline
├── Data/
│   ├── clean_data.xlsx                   # Cleaned dataset
│   ├── Data_Cleaning.ipynb               # Cleaning and transformation logic
│   └── transform to CSV.py               # Script to convert .dta to .csv
└── README.md                             # Project overview and setup instructions
```

---

9. Conclusion

This project demonstrated that survey-based happiness prediction is feasible using well-structured preprocessing and modeling. KNN showed the best performance for this dataset, and feature analysis offered valuable insights into what drives subjective well-being in Armenia. The methodology can be extended to other countries, survey waves, or predictive targets like trust or satisfaction with democracy.

Author: Kima Badalyan, Armine Hakobyan, Nare Sargsyan, Ruben Galoyan

Institution: American University of Armenia

Date: May 2025