**Mona Albarqi , Albatool Moathen**

**Final Report**

**Introduction**

The goal of this project was to analyze factors influencing individual happiness and develop a predictive model to estimate happiness levels using socio-demographic and lifestyle variables. By leveraging a dataset with features such as income, health, and social contact, the project aimed to answer: *What are the primary predictors of happiness, and how accurately can these levels be predicted?*

This report documents the methodology, analysis, and findings, including key insights and challenges encountered.

**Data**

**Dataset Description**

* The dataset contains 6000 rows and 11 columns, representing socio-demographic and lifestyle information.
* The target variable is HAPPINESS, a categorical variable with three levels: "Very Happy," "Pretty Happy," and "Not Too Happy."
* Predictors include:
  + **Numeric Features**: AGE, INCOME, EDUCATION, SOCIAL\_CONTACT.
  + **Categorical Features**: MARITAL\_STATUS, GENDER, EMPLOYMENT\_STATUS, HEALTH, RELIGION, POLITICAL\_VIEWS.

**Data Preprocessing**

1. **Handling Missing Values**:
   * Numeric columns were imputed using mean values.
   * Categorical columns were imputed using mode values.
2. **Encoding Categorical Variables**:
   * Label encoding was applied to transform non-numeric features into numeric values.
3. **Feature Scaling**:
   * Numeric features were standardized using StandardScaler to ensure compatibility with the models.
4. **Exploratory Data Analysis**:
   * Distributions of numeric variables were visualized.
   * Correlation analysis revealed relationships between predictors.

**Methodology**

**Models**

1. **Logistic Regression**:
   * A baseline classification model for its simplicity and interpretability.
2. **Random Forest Classifier**:
   * An ensemble model selected for its ability to handle non-linear relationships and provide feature importance.

**Evaluation Metrics**

* **Accuracy**: Proportion of correctly classified instances.
* **Precision, Recall, F1-Score**: Evaluated for each class to measure predictive quality.
* **Confusion Matrix**: Visualized true vs. predicted classifications.

**Train-Test Split**

* Data was split into training (80%) and testing (20%) subsets.

**Results**

**Model Performance**

* **Logistic Regression**:
  + Accuracy: 0.68
  + Classification Report: Precision, Recall, and F1-scores were moderate across classes.
* **Random Forest Classifier**:
  + Accuracy: 0.74
  + Classification Report: Higher scores compared to Logistic Regression, particularly for minority classes.

**Feature Importance**

* Top predictors of happiness (Random Forest):
  1. INCOME
  2. HEALTH
  3. SOCIAL\_CONTACT
* These results suggest that financial stability, physical well-being, and social engagement significantly influence happiness levels.

**Visualizations**

* Confusion matrices and feature importance plots highlighted key insights and model performance.

**Discussion**

**Key Findings**

* Random Forest outperformed Logistic Regression in both accuracy and handling class imbalances.
* INCOME, HEALTH, and SOCIAL\_CONTACT emerged as the most influential predictors of happiness.

**Practical Implications**

* Policies aimed at improving income levels and healthcare access could significantly enhance societal happiness.
* Encouraging social interactions and community engagement also appears vital.

**Limitations**

* The dataset was synthetic and may not fully capture real-world complexities.
* Missing values were imputed, which might introduce bias.
* Further analysis with a more diverse dataset is recommended.

**Conclusion**

This project successfully identified key drivers of happiness and demonstrated the utility of machine learning in socio-demographic analyses. While Random Forest proved to be the superior model, future work should focus on refining data collection and exploring additional predictors.

**References**

* General Social Survey (GSS): [gss.norc.org](http://gss.norc.org/)
* Scikit-learn Documentation: [scikit-learn.org](https://scikit-learn.org/)