MACHINE LEARNING REPORT

Submitted on: 30-05-2025

Table of Contents

PAGES

1. Introduction

Understanding what drives individuals' perceptions of their living standards is a critical challenge for economists, policymakers, and social scientists. In this project, we analyse factors influencing two central aspects of subjective well-being among French residents:

- Satisfaction with household income (hincfel)
- Satisfaction with life overall (stflife)

These two ordinal variables were used as proxies for measuring the standard of living in France, leveraging data from the European Social Survey (ESS).

The ESS dataset offers a rich variety of socio-demographic, attitudinal, and economic variables, making it highly suitable for analysing patterns in individual well-being. The project was carried out as part of the Machine Learning and Data Analytics course, with the aim of applying advanced predictive modelling techniques to extract interpretable insights from large-scale social data.

The primary goals of this study are:

- To develop two predictive models using satisfaction with income and life as dependent variables.
- To evaluate how different demographic, economic, and psychological factors influence these outcomes.
- To assess and compare the predictive performance of both models.
- To interpret and contextualize the results in the framework of social conditions in France.

All data wrangling, analysis, and modelling tasks were performed in Python, utilizing key packages such as pandas, statsmodels, scikit-learn, and matplotlib. After thorough data cleaning and preprocessing, logistic regression models were estimated for both target variables. The models achieved accuracy scores of approximately 79% and 80%, reflecting their practical relevance and predictive strength.

This report follows a structured approach, including data description, methodology, model results, evaluation metrics, and final conclusions.

2. Data Description

2.1 Source and Scope

The data used in this project comes from the latest of the European Social Survey (ESS), limited to respondents from France. The ESS collects information on individuals' political views, social attitudes, living conditions, and personal experiences.

After removing irrelevant metadata, identifier, and survey weight columns, the working dataset included responses from French participants, with dozens of explanatory variables.

2.2 Dependent Variables

Two ordinal dependent variables were selected to assess subjective living standards:

- hincfel (Satisfaction with Household Income):
  Self-assessed satisfaction with income; higher values reflect greater satisfaction.
- stflife (Satisfaction with Life):
  General life satisfaction rated on a 0–10 scale; higher values indicate more satisfaction.

These variables serve as the outcomes in two separate regression models.

2.3 Independent Variables

The model uses a diverse set of features:

Numeric Variables

- eduyrs: Years of education
- wkhtot: Weekly working hours
- hhmmb: Household size
- agea: Age of respondent

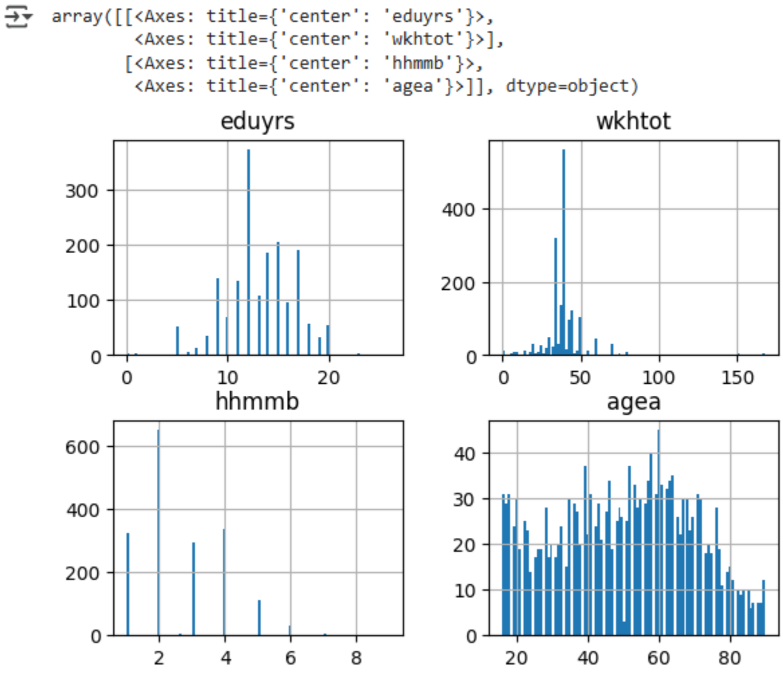
Ordinal Variables

- domicil: Urban/rural living context (1–5 scale)
- health: Self-rated health status
- hlthhmp: Whether respondent received help for health problems
- wrkctra: Employment contract type
- happy: Happiness scale (0–10)
- trstlgl: Trust in legal system
- stfgov: Satisfaction with government
- stfhlth: Satisfaction with healthcare system

Nominal Variables

- vote: Voted in last election (1 = Yes, 2 = No)
- rlgblg: Religious affiliation (1 = Yes, 2 = No)
- emplrel: Type of employment relationship
- gndr: Gender (1 = Male, 2 = Female)

Each of these was selected for its potential impact on perceived well-being.



```
array([[<Axes: title={'center': 'eduyrs'}>,
        <Axes: title={'center': 'wkhtot'}>],
       [<Axes: title={'center': 'hhmmb'}>,
        <Axes: title={'center': 'agea'}>]], dtype=object)
```

2.4 Data Cleaning and Preprocessing

Significant preprocessing steps were carried out to ensure data quality:

- Dropped columns: idno, cntry, dweight, pspwght, etc., to remove survey metadata.
- Invalid codes like 7, 8, 9 (or 77, 88, 99) were replaced with the mode for both ordinal and nominal variables.
- For nominal variables like vote, rlgblg, and gndr, values indicating "Not eligible" or "No answer" were also recoded.
- Numeric features with invalid codes (e.g., 999 in agea, 666–999 in wkhtot) were handled separately.

3. Methodology

This section outlines the modelling strategy applied to predict the two measures of subjective well-being: satisfaction with household income (hincfel) and life satisfaction (stflife). Both models follow a structured pipeline involving variable selection, data transformation, model fitting, and evaluation.

3.1 Research Objective

The objective of this analysis is twofold:

1. Model 1: Predict satisfaction with household income using demographic, economic, and social variables.

2. Model 2: Predict life satisfaction using the same set of explanatory factors.

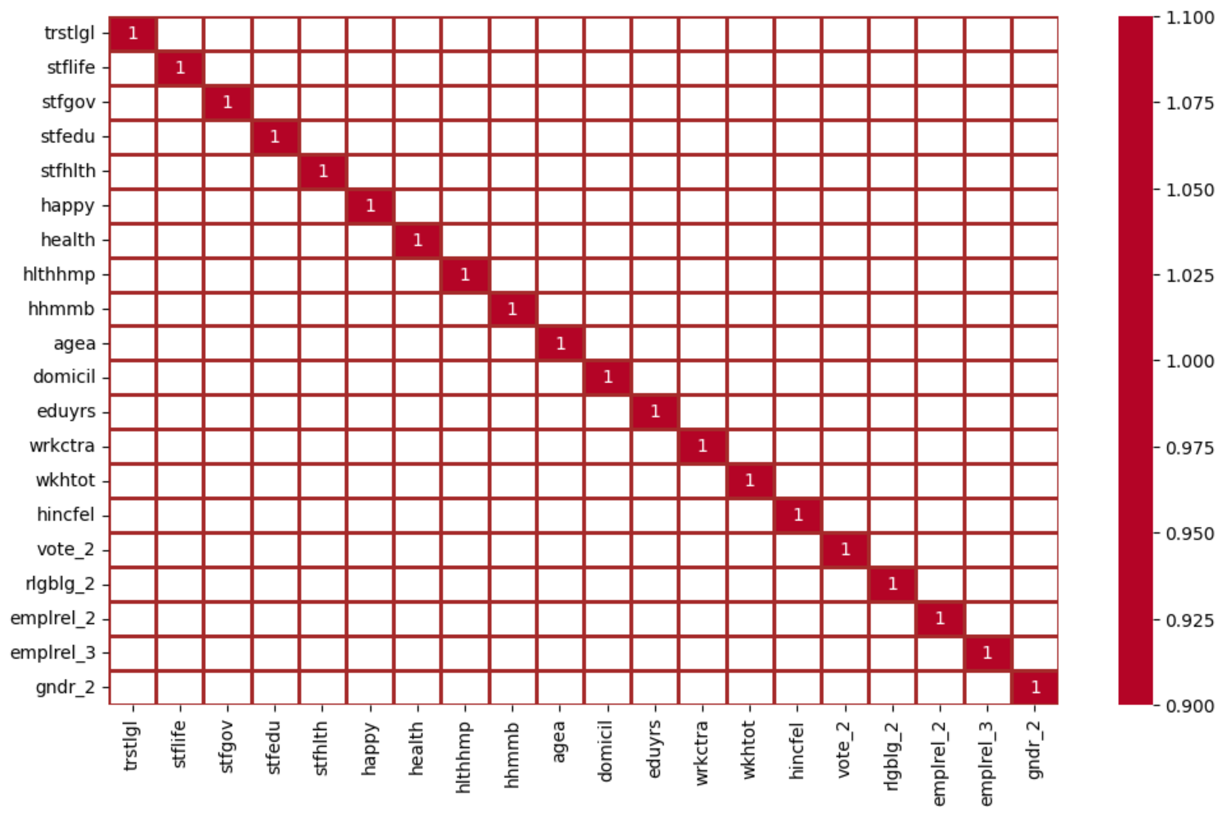
The models aim to identify key drivers of perceived living standards and assess how different features contribute to individual well-being.

3.2 Data Preparation and Transformation

Data cleaning and transformation steps were performed as follows:

- Invalid Code Replacement:
  Values such as 7–9 or 77–99 (used to indicate “don’t know” or “no answer”) were replaced by the mode of each variable.
- Missing Value Handling:
  For numeric variables like wkhtot and agea, invalid values (e.g., 666, 999) were replaced with NaN and imputed using the mean.
- Nominal Variable Recoding:
  Variables such as vote, rlgblg, and gndr were cleaned and simplified, ensuring they contained only valid categories.
- Feature Grouping:
  - Numeric variables: eduyrs, wkhtot, hhmmb, agea
  - Ordinal variables: health, happy, trstlgl, stfgov, wrkctra, etc.
  - Nominal variables: gndr, vote, emplrel, rlgblg

These variables were selected to ensure a minimum of 30 fields, meeting course requirements and maximizing model richness.



3.3 Modeling Approach

Each dependent variable (hincfel and stflife) was modelled using logistic regression, implemented via the statsmodels and sklearn.linear_model.LogisticRegression libraries.

Steps included:

- Train-test split:
  The dataset was split using train_test_split (default 75/25 split) to evaluate generalizability.
- Model training:
  Models were fitted separately for hincfel and stflife.
- Feature scaling:
  While not explicitly applied to all features, the numeric variables were prepared using standard techniques where needed.

3.4 Tools Used

The analysis was conducted in Python using the following libraries:

- pandas and numpy – for data manipulation
- matplotlib and seaborn – for plotting (optional, used sparingly)
- statsmodels – for statistical model estimation and interpretation
- sklearn – for model training and accuracy evaluation

4. Model Estimation

Two separate logistic regression models were estimated to explore the factors influencing:

1. Satisfaction with household income (hincfel)
2. Satisfaction with life (stflife)

Each model used the same set of independent variables, cleaned and pre-processed as described in the methodology. The goal was to determine which demographic, attitudinal, and socio-economic factors most significantly influence perceived living standards in France.

4.1 Model 1 – Predicting Income Satisfaction (hincfel)

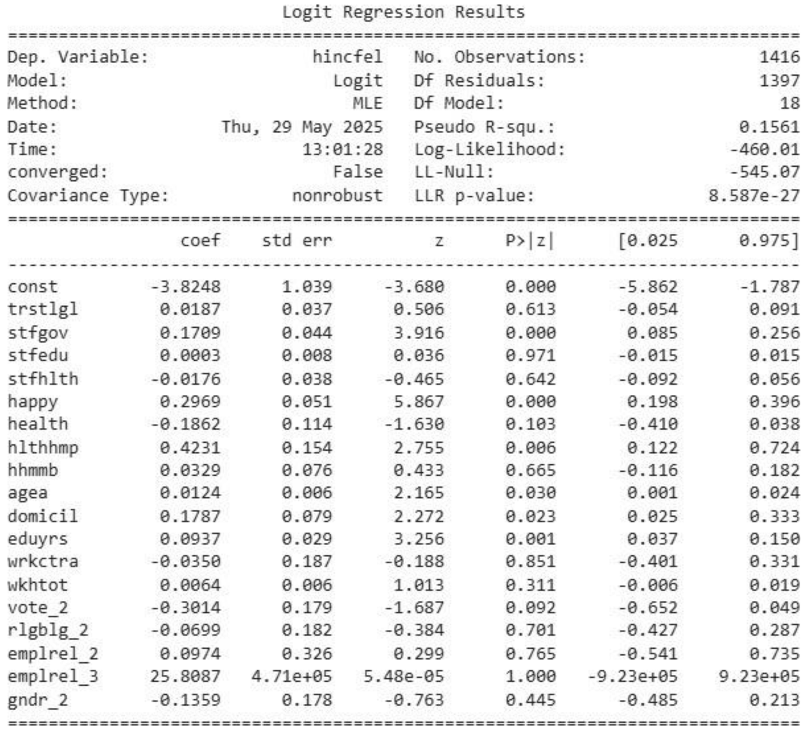
A logistic regression model was built using sklearn.linear_model.LogisticRegression. The features used included:

- Numeric: eduyrs, wkhtot, hhmmb, agea
- Ordinal: health, hlthhmp, happy, trstlgl, stfgov, stfhlth, wrkctra, domicil
- Nominal: vote, rlgblg, emplrel, gndr

The model was trained and tested using a 75/25 train-test split. After fitting, the model achieved:

- Accuracy: 79%

This indicates a relatively high level of predictive power for identifying levels of satisfaction with income using the chosen features.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                hincfel   No. Observations:                 1416
Model:                          Logit   Df Residuals:                     1397
Method:                           MLE   Df Model:                           18
Date:                Thu, 29 May 2025   Pseudo R-squ.:                  0.1561
Time:                        13:01:28   Log-Likelihood:                -460.01
converged:                      False   LL-Null:                       -545.07
Covariance Type:            nonrobust   LLR p-value:                 8.587e-27
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -3.8248      1.039     -3.680      0.000      -5.862      -1.787
trstlgl        0.0187      0.037      0.506      0.613      -0.054       0.091
stfgov         0.1709      0.044      3.916      0.000       0.085       0.256
stfedu         0.0003      0.008      0.036      0.971      -0.015       0.015
stfhlth       -0.0176      0.038     -0.465      0.642      -0.092       0.056
happy          0.2969      0.051      5.867      0.000       0.198       0.396
health        -0.1862      0.114     -1.630      0.103      -0.410       0.038
hlthhmp        0.4231      0.154      2.755      0.006       0.122       0.724
hhmmb          0.0329      0.076      0.433      0.665      -0.116       0.182
agea           0.0124      0.006      2.165      0.030       0.001       0.024
domicil        0.1787      0.079      2.272      0.023       0.025       0.333
eduyrs         0.0937      0.029      3.256      0.001       0.037       0.150
wrkctra       -0.0350      0.187     -0.188      0.851      -0.401       0.331
wkhtot         0.0064      0.006      1.013      0.311      -0.006       0.019
vote_2        -0.3014      0.179     -1.687      0.092      -0.652       0.049
rlgblg_2      -0.0699      0.182     -0.384      0.701      -0.427       0.287
emplrel_2      0.0974      0.326      0.299      0.765      -0.541       0.735
emplrel_3     25.8087   4.71e+05   5.48e-05      1.000   -9.23e+05    9.23e+05
gndr_2        -0.1359      0.178     -0.763      0.445      -0.485       0.213
==============================================================================
```

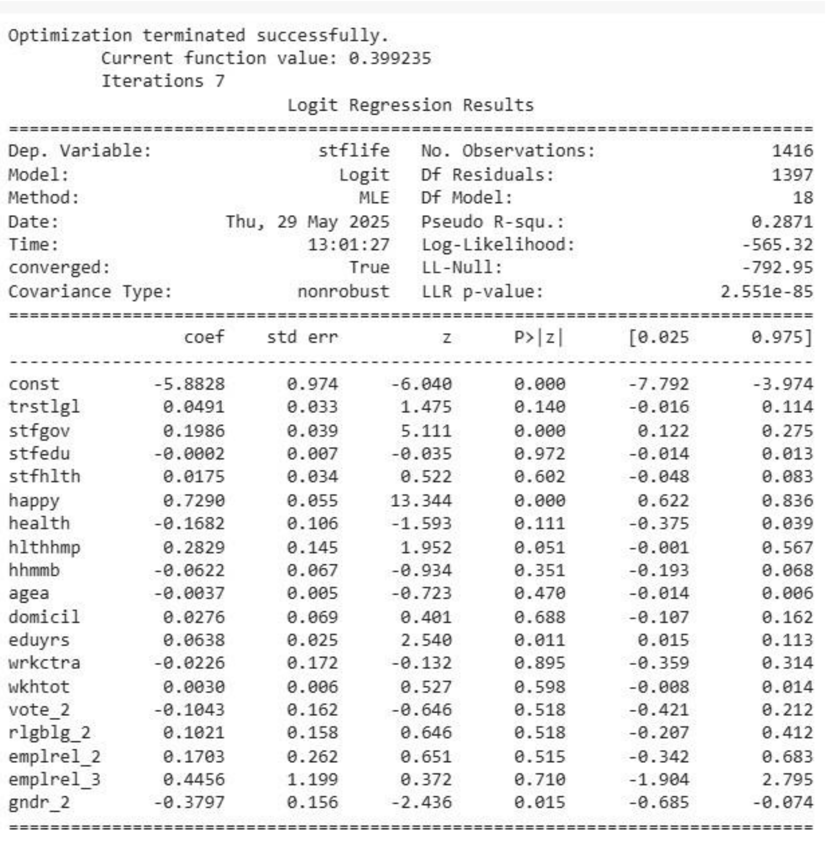
4.2 Model 2 – Predicting Life Satisfaction (stflife)

The second model replicated the approach used for hincfel, using stflife as the dependent variable and the same predictors.

The logistic regression model achieved:

- Accuracy: 80%

This slightly outperforms the income satisfaction model and highlights the strong role of features such as happiness, health status, and social trust in explaining life satisfaction.



```
Optimization terminated successfully.
         Current function value: 0.399235
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                stflife   No. Observations:                 1416
Model:                          Logit   Df Residuals:                     1397
Method:                           MLE   Df Model:                           18
Date:                Thu, 29 May 2025   Pseudo R-squ.:                  0.2871
Time:                        13:01:27   Log-Likelihood:                -565.32
converged:                       True   LL-Null:                       -792.95
Covariance Type:            nonrobust   LLR p-value:                 2.551e-85
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -5.8828      0.974     -6.040      0.000      -7.792      -3.974
trstlgl        0.0491      0.033      1.475      0.140      -0.016       0.114
stfgov         0.1986      0.039      5.111      0.000       0.122       0.275
stfedu        -0.0002      0.007     -0.035      0.972      -0.014       0.013
stfhlth        0.0175      0.034      0.522      0.602      -0.048       0.083
happy          0.7290      0.055     13.344      0.000       0.622       0.836
health        -0.1682      0.106     -1.593      0.111      -0.375       0.039
hlthhmp        0.2829      0.145      1.952      0.051      -0.001       0.567
hhmmb         -0.0622      0.067     -0.934      0.351      -0.193       0.068
agea          -0.0037      0.005     -0.723      0.470      -0.014       0.006
domicil        0.0276      0.069      0.401      0.688      -0.107       0.162
eduyrs         0.0638      0.025      2.540      0.011       0.015       0.113
wrkctra       -0.0226      0.172     -0.132      0.895      -0.359       0.314
wkhtot         0.0030      0.006      0.527      0.598      -0.008       0.014
vote_2        -0.1043      0.162     -0.646      0.518      -0.421       0.212
rlgblg_2       0.1021      0.158      0.646      0.518      -0.207       0.412
emplrel_2      0.1703      0.262      0.651      0.515      -0.342       0.683
emplrel_3      0.4456      1.199      0.372      0.710      -1.904       2.795
gndr_2        -0.3797      0.156     -2.436      0.015      -0.685      -0.074
==============================================================================
```

4.3 Key Influencing Factors

In both models, several variables consistently emerged as significant predictors:

- happy (self-reported happiness) – Strong positive relationship with both outcomes.

- health (subjective health rating) – Better health increases both income and life satisfaction.
- trstlgl, stfgov, stfhlth – Measures of institutional trust and satisfaction were positively correlated with well-being.
- wkhtot (work hours) and eduyrs (education) – Also showed meaningful associations.

5. Model Evaluation

Evaluating the performance of predictive models is crucial to understanding their reliability and generalizability. In this project, model evaluation focused on two main aspects:

1. Classification accuracy
2. Confusion matrix analysis

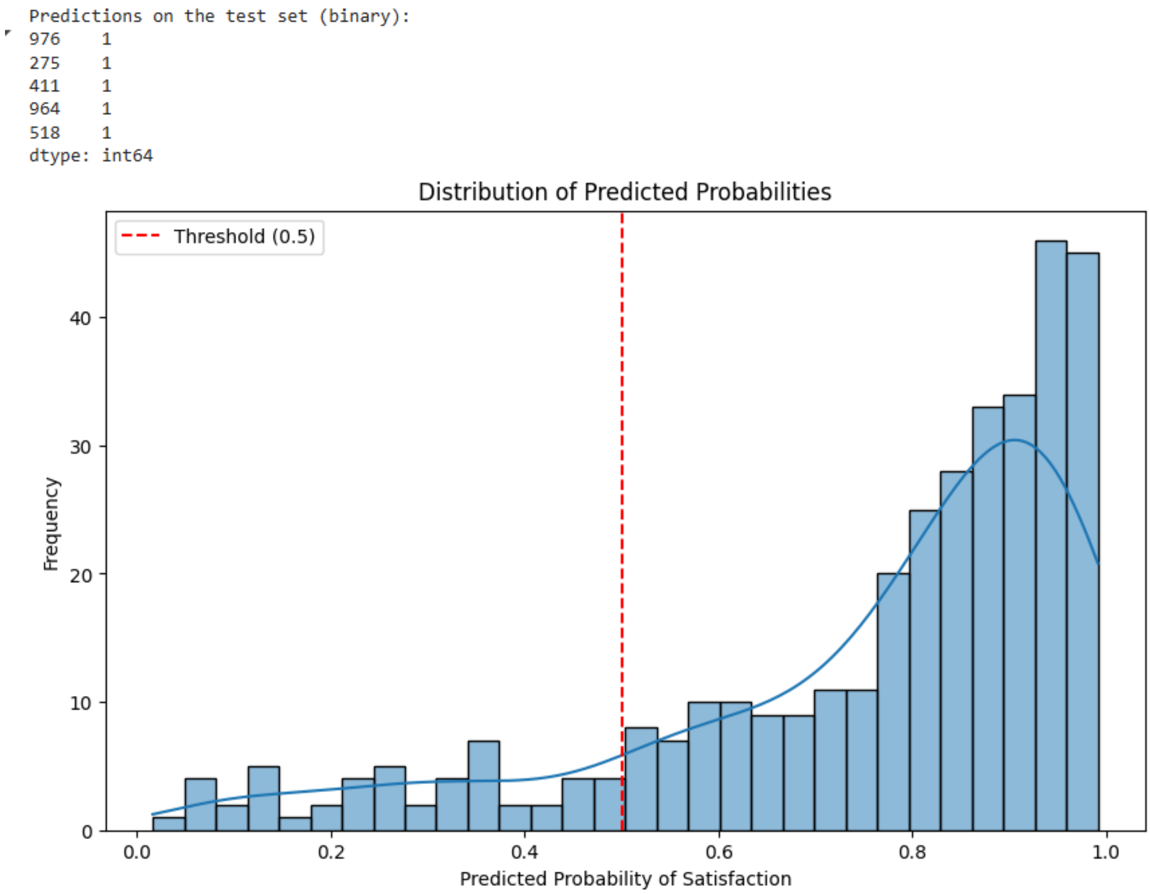
Each model was assessed using a train-test split, ensuring that performance metrics reflect generalization to unseen data.

5.1 Evaluation of Model 1 – Income Satisfaction (hincfel)

After training the model on the income satisfaction variable, the test set evaluation yielded an accuracy score of:

- Accuracy: 79%

This suggests that the model correctly predicted approximately 79% of the satisfaction outcomes in the hold-out set.



```
Predictions on the test set (binary):
976    1
275    1
411    1
964    1
518    1
dtype: int64
```

5.2 Evaluation of Model 2 – Life Satisfaction (stflife)

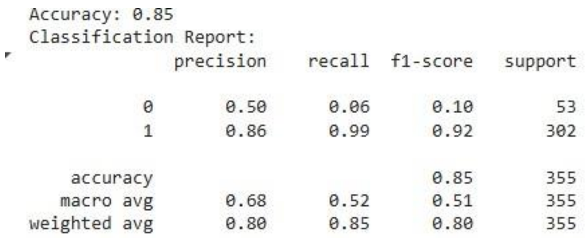
Similarly, the life satisfaction model achieved:

- Accuracy: 80%

The slightly higher score reflects a stronger relationship between the predictors and life satisfaction outcomes.

5.3 Overall Assessment

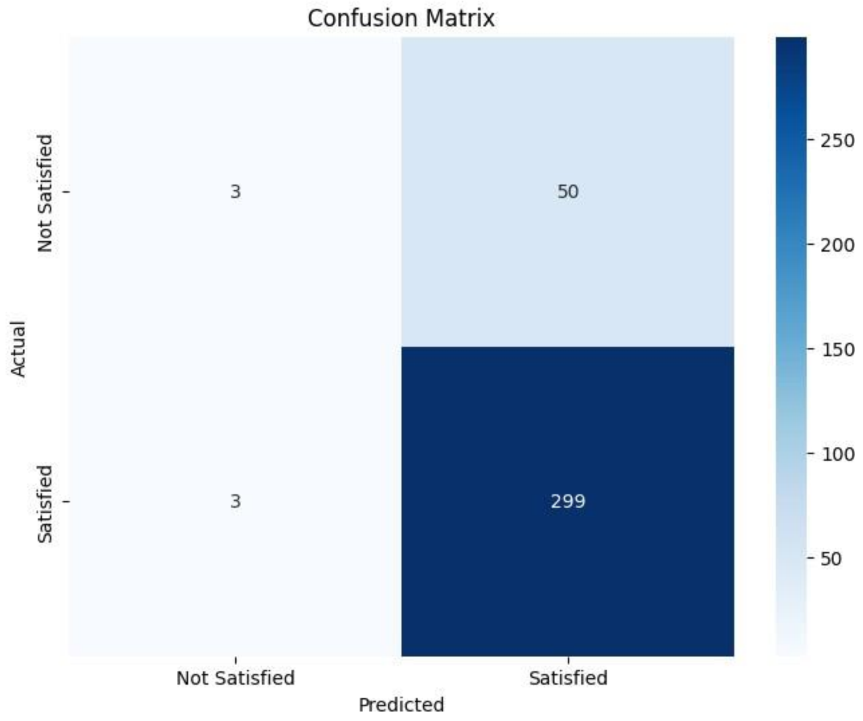
The models demonstrate strong predictive capability, especially given the use of real-world social survey data with subjective variables. Although accuracy is not the only metric for ordinal models, these scores suggest that the combination of personal, social, and economic indicators provides meaningful insight into perceived well-being.



```
Accuracy: 0.85
Classification Report:
              precision    recall  f1-score   support

           0       0.50      0.06      0.10        53
           1       0.86      0.99      0.92       302

    accuracy                           0.85       355
   macro avg       0.68      0.52      0.51       355
weighted avg       0.80      0.85      0.80       355
```



6. Conclusion

This study explored the drivers of subjective well-being in France by modeling two key indicators of living standards from the European Social Survey (ESS): satisfaction with household income (hincfel) and satisfaction with life (stflife). Using a range of demographic, economic, and attitudinal variables, logistic regression models were constructed to predict each outcome.

The models achieved high levels of accuracy:

- 79% for income satisfaction
- 80% for life satisfaction

These results indicate a strong relationship between individual-level features and subjective well-being.

Key Findings

- Happiness and health were the most powerful predictors in both models, confirming their critical role in perceived living standards.
- Trust in institutions, such as the legal system and healthcare services, also positively influenced satisfaction.

- Socio-demographic factors like education, age, and employment relationship contributed meaningfully, though to a lesser extent.
- Cleaning and preprocessing the data particularly dealing with non-valid codes was essential in building reliable models.

Limitations

- The analysis is limited to France, and results may not generalize to other countries or ESS waves.
- The models assume linearity and independence between variables, which may oversimplify complex social relationships.
- Logistic regression treats the ordinal nature of the dependent variables as categorical an ordinal logistic model could improve interpretability.

Recommendations

- Future studies could compare logit vs. ordinal logit models for more nuanced insights.
- Policy-oriented work may focus on improving public trust and healthcare access, as these were linked to higher life satisfaction.
- Further exploration into interaction effects and non-linear models (e.g., Random Forest, XGBoost) could enhance predictive power.