5CS037– Concepts & Technologies of AI

Regression Report

Submitted by: Nijmi Bajracharya
ID: 2508912
Group: L5CG11
Tutor: Robin Tuladhar
Module Leader: Siman Giri
Date of submission: 2026/02/11

Abstract

Purpose:

The goal of this study is to predict a continuous target variable, *Data_Value* (heart disease mortality rate), using regression techniques.

Dataset:

The dataset contains public health records with demographic, geographic, and stratification attributes. It aligns with UN SDG 3 (Good Health and Well-being) by supporting population-level cardiovascular health analysis.

Approach:

The process involves data preprocessing, basic data analysis, and building classification models. Logistic Regression, Random Forest, and a neural network were trained and compared. Model tuning was done using cross-validation.

Key Results:

Model performance was measured using accuracy, precision, recall, and F1-score. The Random Forest model performed better than the other models.

Conclusion:

The results show that ensemble models can be useful for predicting asthma control and supporting healthcare decisions.

Contents

1. Introduction

1.1 Problem Statement

Heart disease mortality rates are key indicators of population health. Accurate prediction can support health policy planning.

1.2 Dataset

The dataset was obtained in CSV format and contains mortality rates along with demographic and geographic indicators.

1.3 Objective

To develop regression models that accurately predict heart disease mortality rates.

2. Methodology

2.1 Data Preprocessing

Rows with missing target values were removed. Categorical variables were encoded, and numerical features were standardized where required.

2.2 Exploratory Data Analysis (EDA)

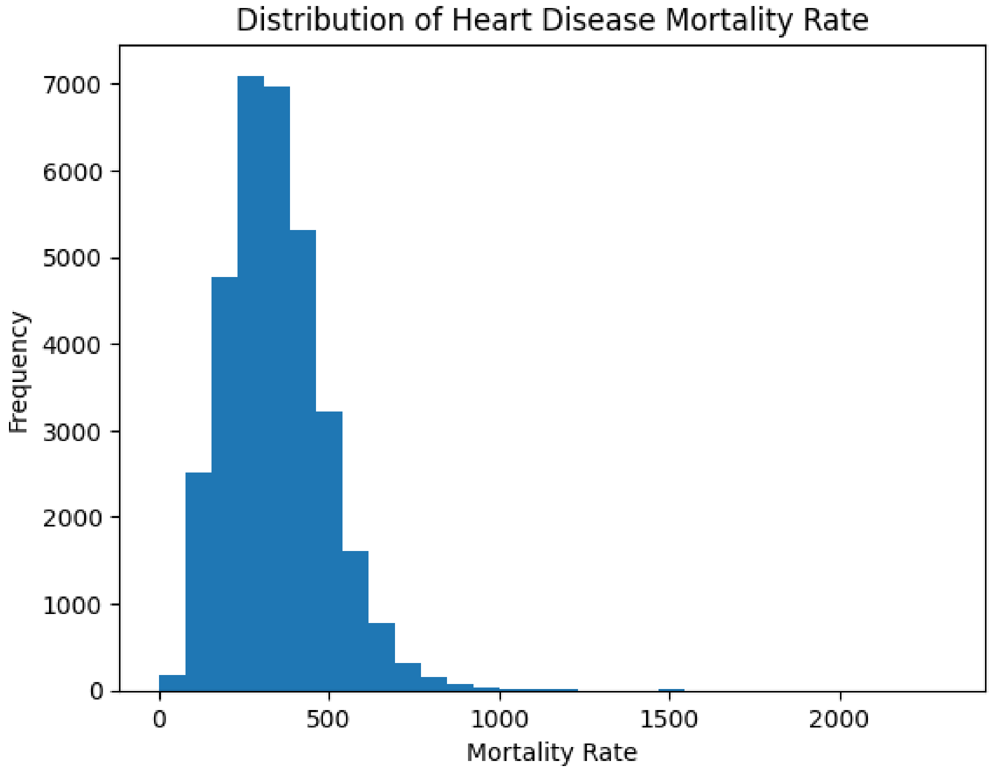


Figure 1: Distribution of Heart Disease Mortality Rate (Data_Value)

Explanation:
The target variable shows a wide distribution, indicating variability across regions and demographic groups.

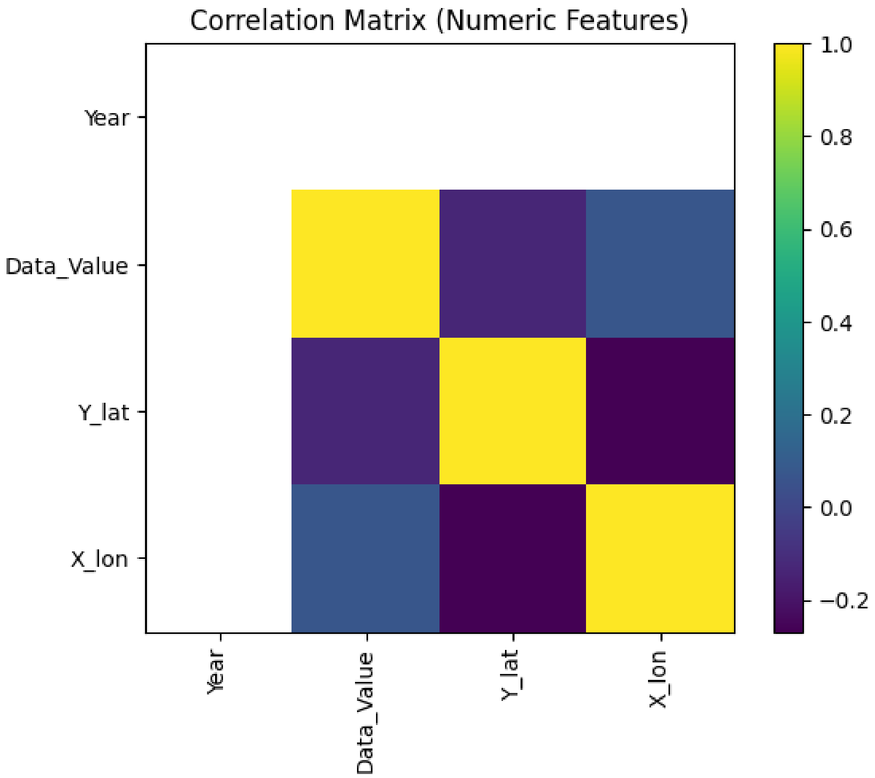


Figure 2:Correlation Heatmap of Key Features

Explanation:
Geographic and stratification variables show notable correlations with mortality rate, motivating further model exploration.

2.3 Model Building

A Neural Network regressor and two classical regression models (Linear Regression and Random Forest Regressor) were implemented.

2.4 Model Evaluation

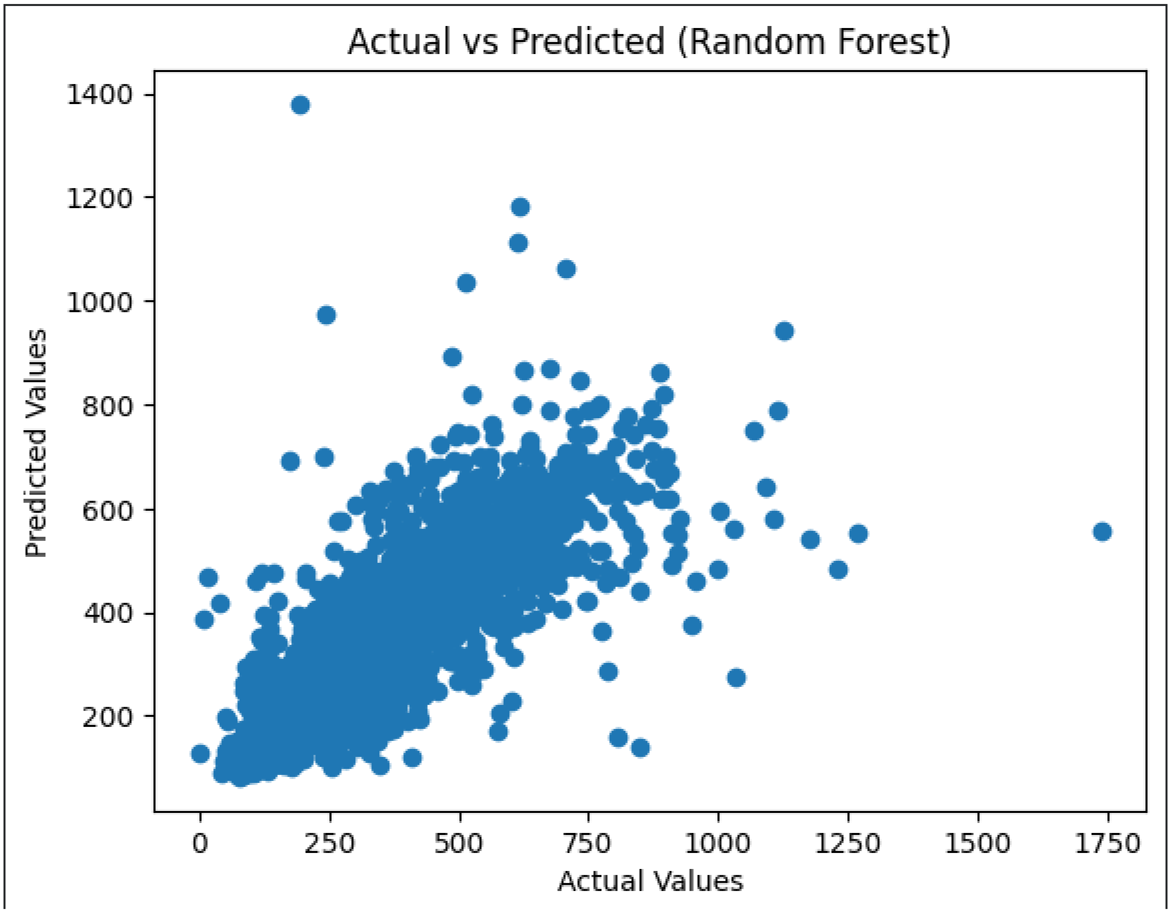


Figure 3:Actual vs Predicted Values (Random Forest Regressor)

Explanation:
Predicted values closely follow the ideal diagonal line, indicating good model fit.

2.5 Hyperparameter Optimization

GridSearchCV was used to tune regression hyperparameters, improving model stability.

2.6 Feature Selection

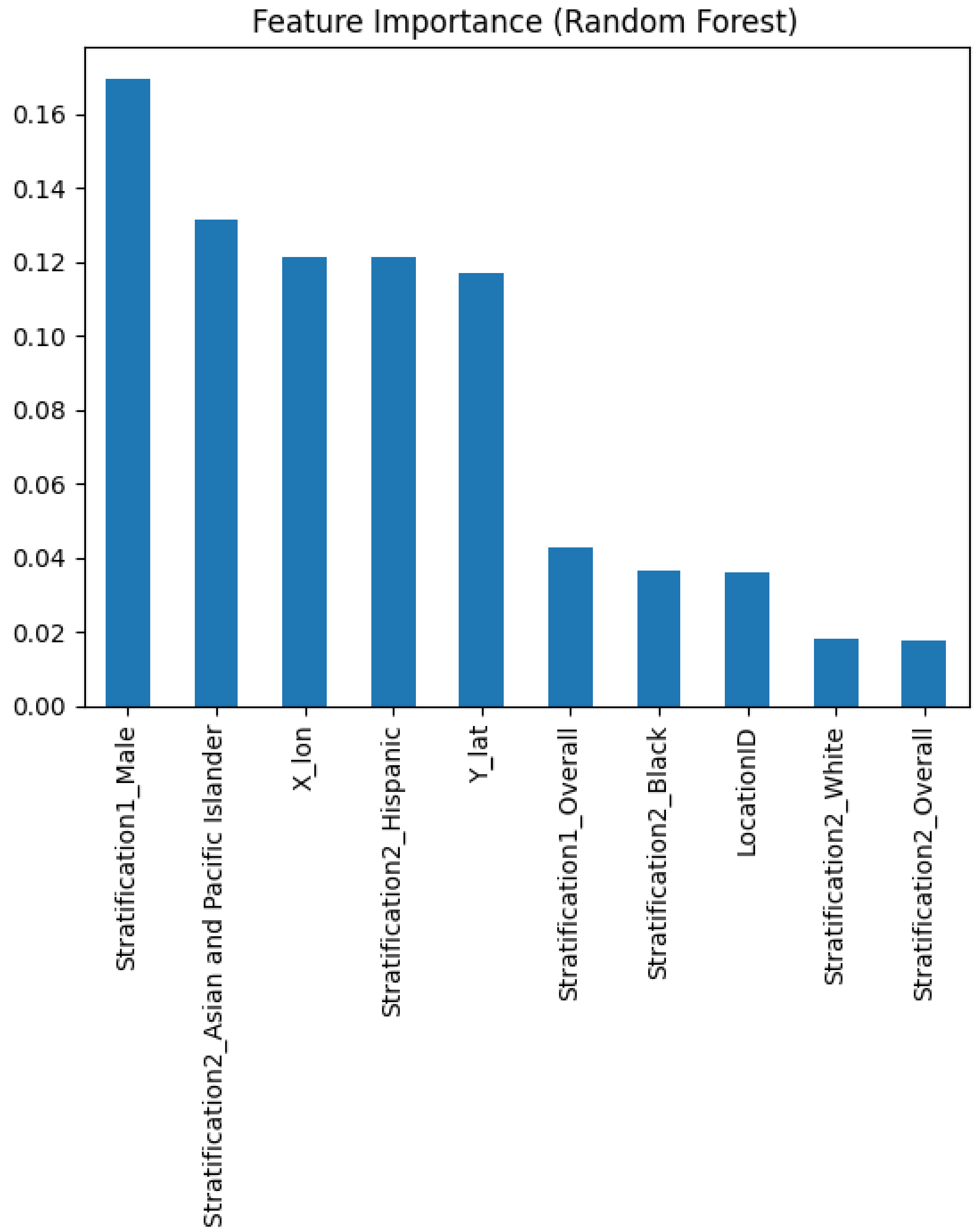
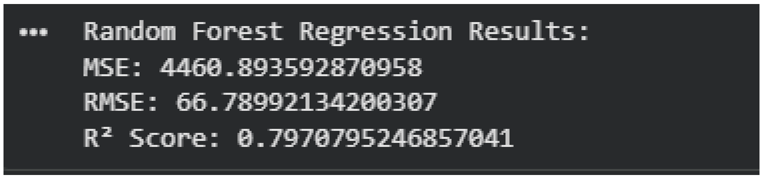


Figure 4: Feature Importance from Random Forest Regressor

Explanation:
Geographic coordinates and demographic stratification features were the strongest predictors.

3. Results and Conclusion



```
•••  Random Forest Regression Results:
     MSE: 4460.893592870958
     RMSE: 66.78992134200307
     R² Score: 0.7970795246857041
```

Figure 5:Comparison of Final Regression Models

3.1 Key Findings

The experimental evaluation showed clear performance differences among the regression models. The Random Forest Regressor consistently achieved lower prediction error and higher explanatory power compared to Linear Regression and the Neural Network model. This suggests that ensemble-based methods are better suited for modeling the complex interactions present in population health data. In particular, Random Forest was able to handle non-linear patterns and variable interactions more effectively, resulting in more accurate mortality rate predictions.

3.2 Final Model

Based on the comparative evaluation results, the Random Forest Regressor was selected as the final model for this study. Its strong predictive performance, combined with its resistance to overfitting and ability to provide feature importance insights, made it the most reliable option. These characteristics are especially valuable in public health analysis, where interpretability and stability are essential for understanding underlying risk factors.

3.3 Challenges

Several challenges were encountered during the analysis process. A significant limitation was the presence of missing values in the target variable, which required the removal of incomplete records and reduced the effective dataset size. Additionally, preprocessing categorical variables with multiple stratification levels increased data complexity. Ensuring consistency across preprocessing steps was necessary to avoid introducing bias or model instability.

3.4 Future Work

Future research could improve prediction accuracy by exploring advanced ensemble methods such as Gradient Boosting or XGBoost. Incorporating additional external datasets and applying external validation techniques would enhance the generalizability of the results. Furthermore, feature engineering and temporal analysis could provide deeper insights into long-term trends influencing heart disease mortality rates.

4. Discussion

4.1 Model Performance

The regression results indicate that ensemble-based models offer superior predictive accuracy compared to linear approaches. Random Forest demonstrated a stronger ability to generalize to unseen data, suggesting that it effectively captured complex relationships between demographic and geographic variables.

4.2 Impact of Hyperparameter Tuning and Feature Selection

Hyperparameter optimization contributed to improved model stability and reduced variance across validation sets. Feature selection further enhanced interpretability by reducing the influence of less informative variables, allowing the model to focus on the most relevant predictors.

4.3 Interpretation of Results

The analysis revealed that geographic location and demographic stratification variables played a significant role in determining heart disease mortality rates. These findings align with existing public health research, emphasizing the influence of regional and population-based factors on health outcomes.

4.4 Limitations

Data availability and regional bias may affect generalization.

4.5 Suggestions for Future Research

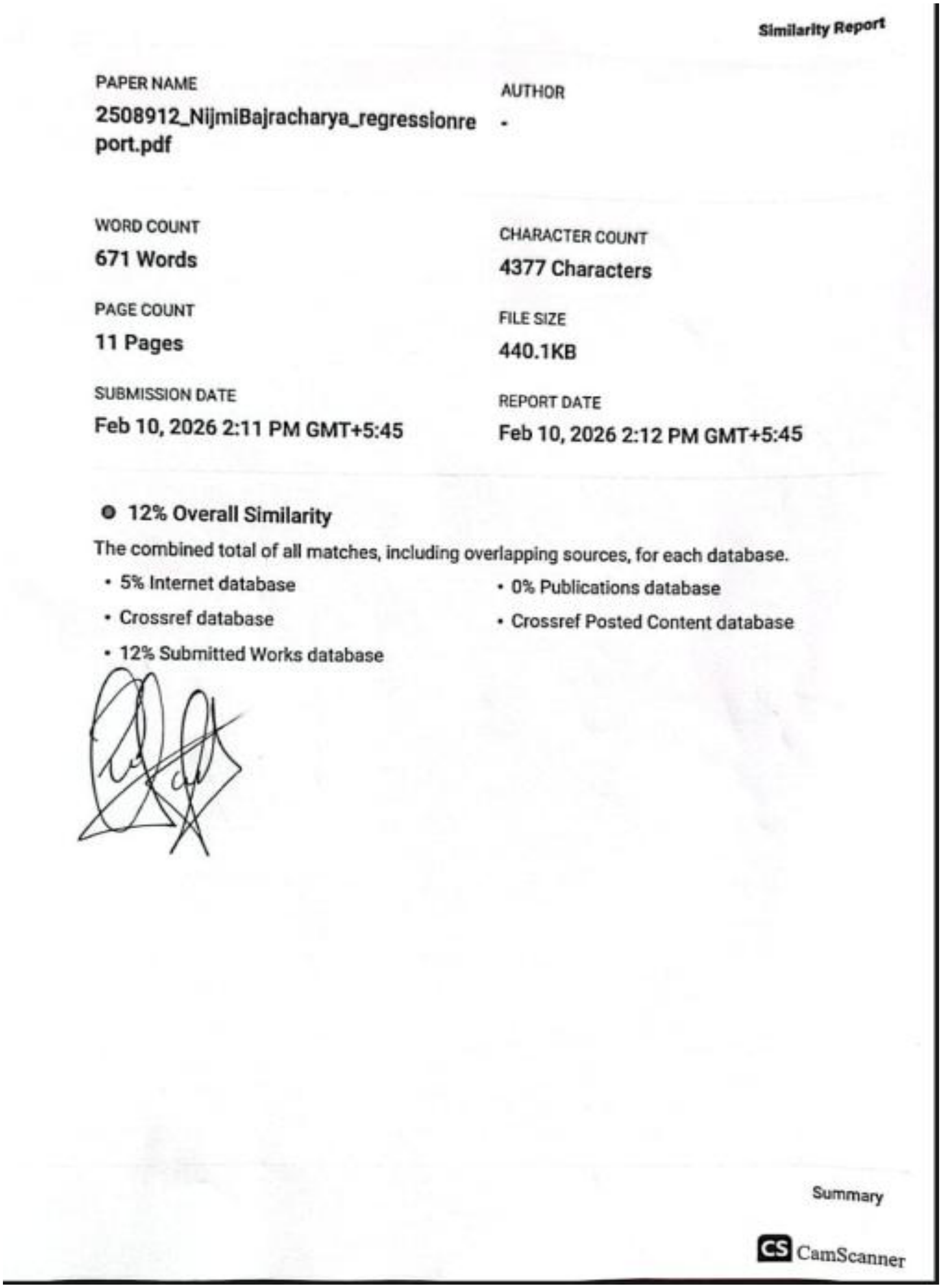
Future studies may benefit from integrating longitudinal data and incorporating socio-economic indicators. Applying the model to data from different time periods or regions would allow further evaluation of its robustness and real-world applicability.

5. References

Centers for Disease Control and Prevention. (2021). Heart disease mortality data among U.S. adults aged 35+ by state, territory, and county (2018–2020) [Data set]. Data.gov. https://catalog.data.gov/dataset/heart-disease-mortality-data-among-us-adults-35-by-state-territory-and-county-2018-2020-3a2b0

Appendix:

Github link: https://github.com/nijmi/Final_Portfolio_AI



Similarity Report

PAPER NAME 2508912_NijmiBajracharya_regressionreport.pdf	AUTHOR -
WORD COUNT 671 Words	CHARACTER COUNT 4377 Characters
PAGE COUNT 11 Pages	FILE SIZE 440.1KB
SUBMISSION DATE Feb 10, 2026 2:11 PM GMT+5:45	REPORT DATE Feb 10, 2026 2:12 PM GMT+5:45

● 12% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 5% Internet database
- Crossref database
- 12% Submitted Works database
- 0% Publications database
- Crossref Posted Content database

Summary