**BAN 540 Final Project**

**1. Business Understanding**

**Problem Statement**

The objective of this project is to predict customer acceptance of personal loans (Personal\_Loan) offered by AllLife Bank. By identifying key factors influencing acceptance, the bank can refine its marketing strategies to target high-probability customers effectively.

**Objectives**

1. Analyze and understand the dataset to uncover trends and insights.
2. Develop predictive models using machine learning to forecast loan acceptance.
3. Select the most accurate and interpretable model for deployment.

**2. Data Understanding**

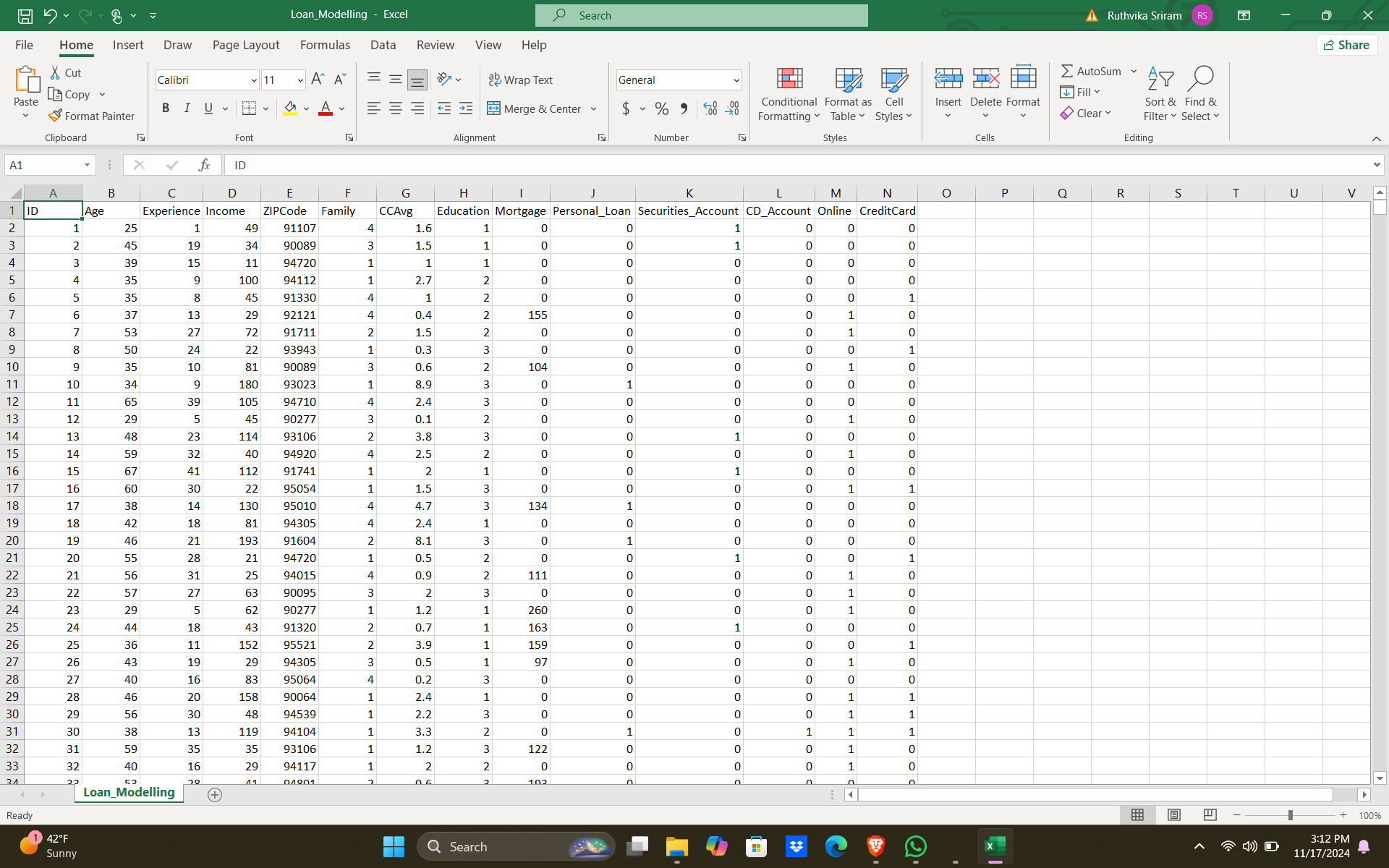
**Dataset Description**

* **Source:** AllLife Bank's customer data.
* **Structure:** 5000 rows and 14 columns.
* **Key Attributes:**
  + Income: Annual income of customers.
  + CCAvg: Average credit card spending per month.
  + Education: Customer education level (1=Undergrad, 2=Graduate, 3=Advanced/Professional).
  + Family: Family size.
  + Personal\_Loan: Target variable (0=No, 1=Yes).

**Key Observations**

* Majority (90%) of customers have not accepted a loan, indicating a significant class imbalance.
* Strong correlations observed between Income, CCAvg, and Personal\_Loan.
* Features like Mortgage and Family display distinct patterns.

**Screenshot of the Dataset:**



**3. Data Preparation**

**Steps Taken:**

1. **Missing Values:**
   * Verified the dataset had no missing values.
2. **Outlier Detection and Treatment:**
   * Identified outliers in Income and CCAvg using boxplots.
   * Applied Winsorization to cap extreme values.
3. **Feature Engineering:**
   * Created a new feature: Debt-to-Income Ratio.
4. **Encoding:**
   * Categorical variables (e.g., Education) were one-hot encoded.
5. **Normalization:**
   * Continuous features (Income, CCAvg, Mortgage) were normalized for consistency.
6. **Data Splitting:**
   * Split the dataset into training (80%) and testing (20%).

**4. Modeling**

**Models Developed**

1. **Decision Tree:** A straightforward model providing interpretable results.
2. **Logistic Regression:** A statistical approach focusing on probabilistic outputs.
3. **Artificial Neural Network (ANN):** A robust model offering superior performance.

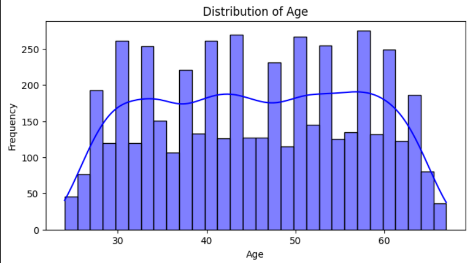
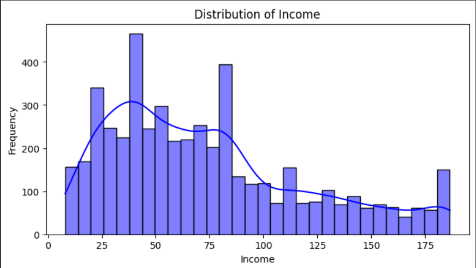
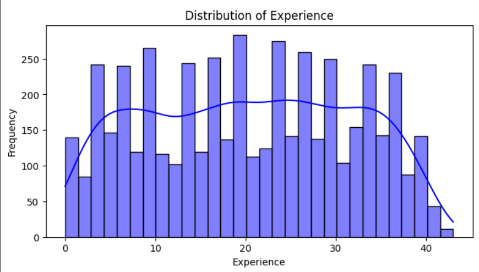
**Model Selection**

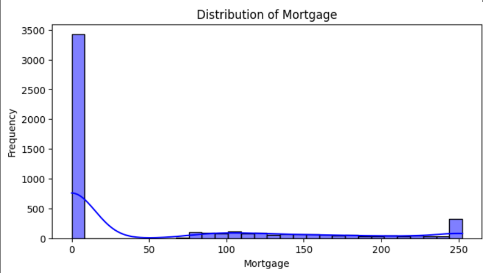
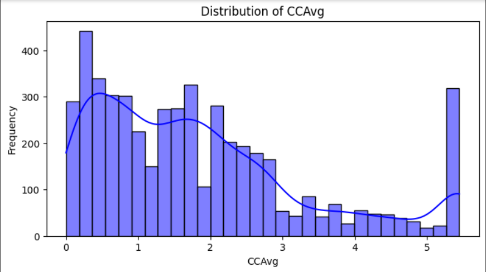
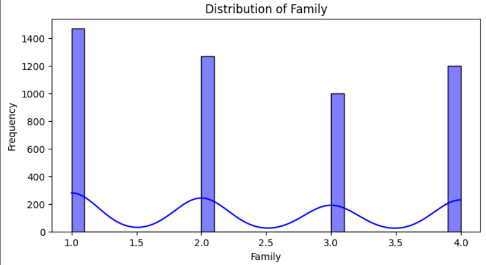
* ANN was selected as the final model due to its 90% accuracy and balanced precision-recall metrics.

**5. Main Process**

* Data flows through multiple stages: preprocessing → modeling → evaluation.
* ANN performed best in terms of accuracy and recall.

**Screenshots of the data visualization in histogram:**



 **Distribution of Mortgage:** This histogram shows that a significant number of customers have no mortgage (value = 0). The distribution is highly skewed, with a long tail of customers having mortgages up to 250. This suggests that many customers either do not hold mortgages or have paid them off.

 **Distribution of CCAvg (Credit Card Average Spending):** The distribution is right-skewed, with the majority of customers spending less than 2 units on average. A smaller segment spends above 5 units, indicating varying spending behaviors, with potential outliers at higher values.

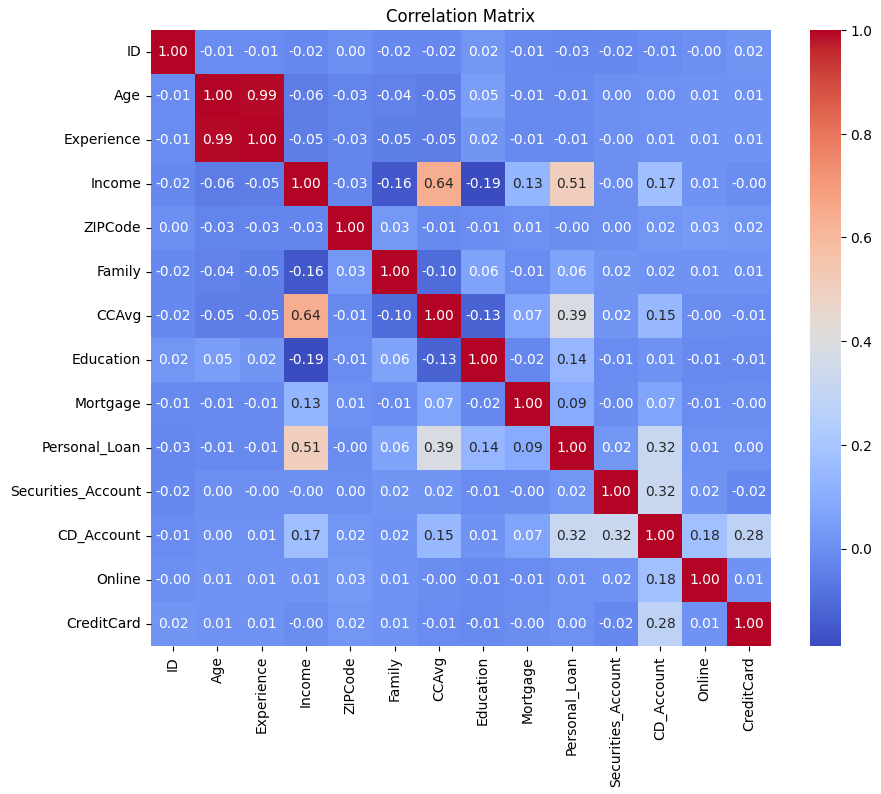
 **Distribution of Family Size:** Family size predominantly clusters at 1 and 2, suggesting that a large proportion of customers are either single or have small families. The distribution decreases as family size increases, reflecting fewer large households.

 **Distribution of Income:** Income is approximately normally distributed but with a slight positive skew. Most customers earn between 50 and 75 units, with fewer customers earning higher incomes, creating a tapering effect toward the upper end.

 **Distribution of Experience:** The histogram indicates a relatively uniform distribution across experience levels, with a slight decline in frequency at the highest levels. This suggests a diverse range of customer experience profiles.

 **Distribution of Age:** Age is uniformly distributed between 30 and 60, with a slight decline at the extremes. This reflects a consistent representation of customers in this age range, with a higher focus on the working-age population.

**Correlation Matrix:**

****

The correlation matrix provides a visual representation of the relationships between the features in the dataset. Key observations include:

1. **Income and CCAvg:** These variables exhibit a moderately strong positive correlation (0.64), indicating that as income increases, credit card spending also tends to rise.
2. **Income and Personal Loan:** A correlation of 0.51 suggests a significant positive relationship, meaning higher-income individuals are more likely to accept personal loans.
3. **Education and Personal Loan:** The correlation (0.14) shows that education levels have a smaller yet positive influence on loan acceptance.
4. **CCAvg and Personal Loan:** A correlation of 0.39 indicates that customers with higher credit card spending are more likely to accept loans.
5. **Age and Experience:** These two variables are highly correlated (0.99), as expected, since they are inherently related.

**Google Colab: Notebook Link:** [*https://colab.research.google.com/drive/17wpkSsKVm4JkSc3yGE2NkQ7RI5COfTHU?usp=sharing*](https://colab.research.google.com/drive/17wpkSsKVm4JkSc3yGE2NkQ7RI5COfTHU?usp=sharing)

**6. Deployment**

**Insights Gained**

* Income, credit card spending, and education level are critical predictors of loan acceptance.
* Customers with higher income and advanced education levels are more likely to accept loans.
* Class imbalance highlighted the need for robust evaluation metrics beyond accuracy.

**Challenges Faced**

1. **Class Imbalance:**
   * Mitigated using balanced class weights and sampling techniques.
2. **Model Selection:**
   * Extensive testing and evaluation helped in choosing ANN as the final model.

**Lessons Learned**

* Data preprocessing significantly impacts model performance.
* Addressing class imbalance is crucial for achieving meaningful results.
* Iterative model development and evaluation ensure better insights.