Financial Data Classification Report

1. Introduction

This report summarizes the approach, model selection, and results for a financial data classification task. The objective was to classify financial data extracted from HTML files into predefined categories using various machine learning and deep learning models.

2. Approach

The approach to this task was divided into several key steps:

Data Extraction and Preprocessing:

HTML files were parsed to extract text data from tables.
Extracted text was cleaned by removing stop words, punctuation, and applying lemmatization.
Categorization:

Tables were categorized into 'Income Statements', 'Balance Sheets', 'Cash Flows', 'Notes', and 'Others' using regex patterns and keyword matching.
Model Selection:

Multiple models were trained and evaluated: RNN, Decision Tree, Random Forest, SVM, and XGBoost.
The text data was tokenized and padded for input into the models.
Labels were one-hot encoded for classification tasks.
Training and Evaluation:

Data was split into training and testing sets (80-20 split).
Models were trained using the training set and evaluated on the testing set.
Metrics used for evaluation included accuracy and loss.
Results and Model Comparison:

Performance of each model was measured and compared.
Models demonstrated high accuracy, with RNN, Decision Tree, Random Forest, and SVM achieving 100% accuracy, and XGBoost achieving 98% accuracy.

3. Model Performance

The following models were trained and evaluated on the dataset:

Model Architecture	Model Type	Accuracy	Loss
Recurrent Neural Network (RNN)	Deep Learning	100%	0
Decision Tree	Machine Learning	100%	0
Random Forest	Machine Learning	100%	0
Support Vector Machine (SVM)	Machine Learning	100%	0
XGBoost	Machine Learning	98%	2

4. **Conclusion**

The analysis demonstrated that the selected models perform exceptionally well in classifying financial data, with RNN, Decision Tree, Random Forest, and SVM achieving perfect accuracy, and XGBoost closely following with 100% accuracy.

The high accuracy rates indicate that the models can reliably classify financial documents into the predefined categories. Future work could explore hyperparameter tuning and incorporating more diverse datasets to further validate the model's robustness.