Report: Classification of Financial Statements

Objective:

The objective of this project was to classify tables from financial statements into five categories: Income Statements, Balance Sheets, Cash Flows, Notes, and Others. This classification task was performed using machine learning techniques on tabular data extracted from HTML files.

Data Extraction and Preprocessing:

- HTML files containing tabular data were processed to extract tables using BeautifulSoup.
- The extracted tables were converted into pandas DataFrames for further processing.
- Text data from these DataFrames was cleaned by removing unnecessary characters and converting numeric columns to float data type.

Feature Extraction:

Use of TF-IDF Vectorizer:

- We utilized the Term Frequency-Inverse Document Frequency (TF-IDF) technique for feature extraction.
- TF-IDF is effective for text data as it highlights the importance of words in a document relative to their frequency in the entire corpus.
- This vectorization technique helps in capturing the unique characteristics of each financial statement table, enabling better discrimination between categories.

Model Selection and Training:

- Five classification models were considered: Logistic Regression, Support Vector Machine (SVM), Random Forest, Gradient Boosting, and Neural Network (Multi-layer Perceptron).
- Each model was trained on the TF-IDF transformed training data and evaluated using the testing data.
- Accuracy was chosen as the key criterion for model performance evaluation.

Model Evaluation:

Here are the accuracy scores for each model:

- **Logistic Regression: 92%**
- **SVM: 94%**
- **Random Forest: 92%**

- **Gradient Boosting: 94%**
- **Neural Network: 93%**

Out of all the models, SVM has slightly high accuracy when observed up to 4 decimal places.

Reason for SVM's Highest Accuracy:

- Support Vector Machine (SVM) demonstrated the highest accuracy among the considered models.
- SVM is effective in high-dimensional spaces and is suitable for text classification tasks where the feature space can be large.
- SVM maximizes the margin between classes, making it robust to overfitting and capable of handling non-linear decision boundaries.
- Additionally, SVM is less affected by irrelevant features, which can be advantageous when dealing with text data with a large number of features (words).

Best Model Selection and Saving:

- The SVM model, which achieved the highest accuracy, was selected as the best model for classifying financial statements.
- The SVM model and the TF-IDF vectorizer used for feature extraction were saved for future use.

Classifying New Files:

- A function was developed to classify new CSV files based on the trained SVM model and TF-IDF vectorizer.
- This function successfully classified a new CSV file into one of the predefined categories.

Conclusion:

In conclusion, the project successfully achieved its objective of classifying financial statements tables into five categories. The TF-IDF vectorizer was chosen for feature extraction due to its ability to capture the unique characteristics of text data. The SVM model demonstrated the highest accuracy, benefiting from its effectiveness in high-dimensional spaces and robustness to irrelevant features in text classification tasks.

Future Work:

Potential future work includes:

- Fine-tuning the SVM model parameters to further improve classification accuracy.
- Exploring additional feature extraction techniques to enhance model performance.
- Scaling the system to handle larger datasets and real-time classification tasks.

- This concludes the report summarizing the approach, model selection, and results of the financial statements classification project.