BANK DOCUMENT CLASSIFICATION

TEAM MEMBERS

- N190422 - P.BUJJI
- N190897 - B.SRUTHI
- N190201 - B.NAGARANI
- N190396 - B.PRAVALLIKA
- N191106 - R.UMADEVI
- N190392 - G.SIDDARDHA

TABLE OF CONTENTS

- Abstract
- Introduction
- Theoretical Framework for Document Classification
- Proposed Methodology
- Model selection
- Results and Discussions
- Observations
- Challenges And Limitations
- Conclusion
- Future work
- References

ABSTRACT

Bank Document Classification using Machine Learning Techniques

Bank Document Classification using ML is a process of automatically categorizing financial documents such as bank statements, invoices, credit card statements, and tax returns using machine learning algorithms. ML Techniques used for bank document classification can improve the accuracy and speed of document categorization, helping organizations to save time, resources, and money. This can help banks to automate the process of document classification, reducing manual effort and improving accuracy.The Project uses Machine Learning to classify bank documents into four categories: bank statements, invoices, credit card statements, and tax returns. It combines text and image analysis to achieve this. Key techniques include Computer Vision for image data, Natural Language Processing for text data. These technologies work together to automate the document classification process.

INTRODUCTION

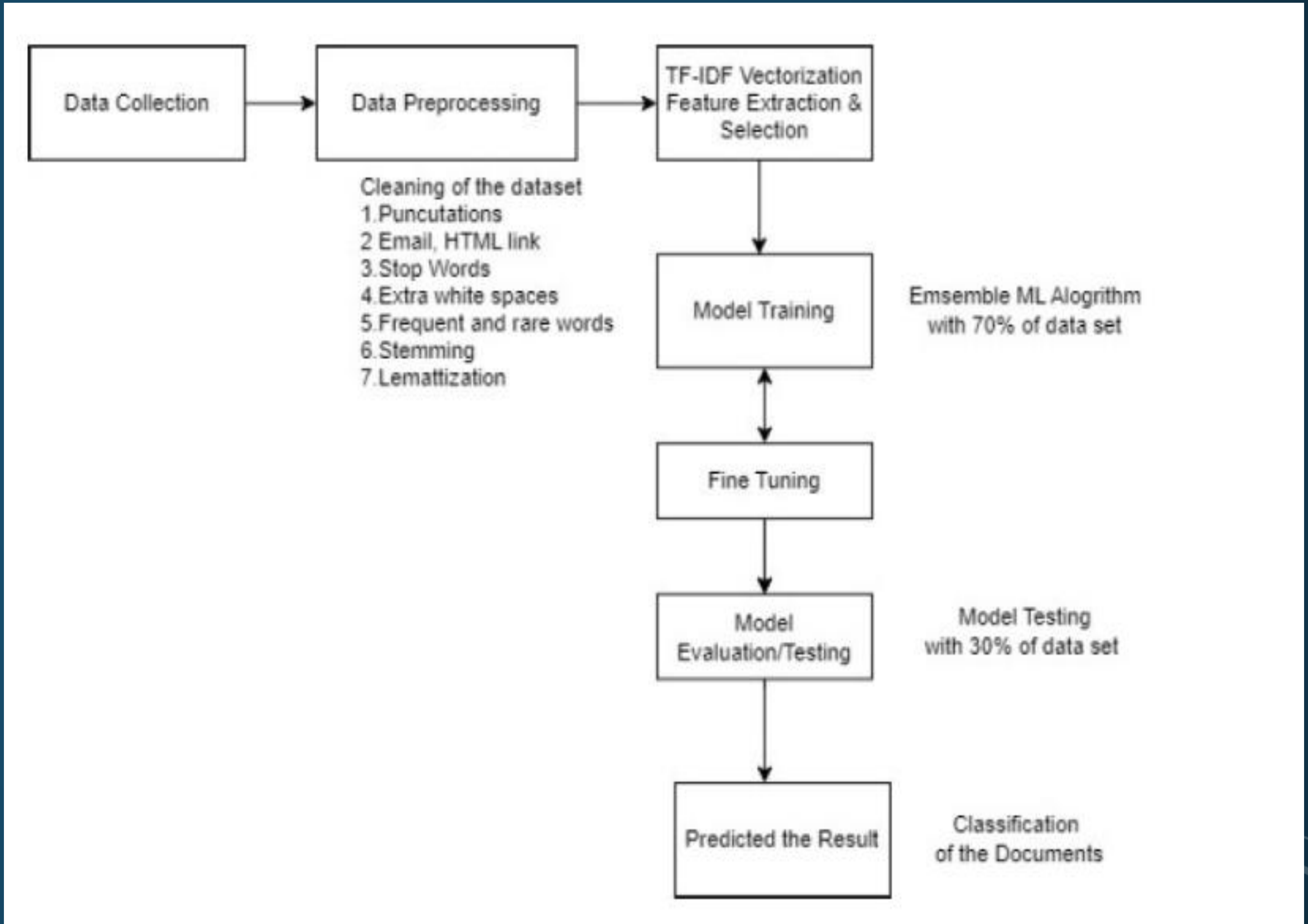
Classification of the bank documents/images such as bank statements, credit card statements, invoices and tax returns using ensemble Machine Learning algorithms. The most effective and efficient way to determine the best models to classify the bank documents using machine learning algorithms. The collected data sets are pre-processed and convert into machine readable format using Natural Language Processing (NLP) and Optical Character Recognition (OCR) techniques. Extract relevant features from preprocessed using techniques such as TF-IDF and word embeddings. An appropriate machine learning models are used to classify the documents, Random Forest, Cat Boost, XG Boost and Voting Classifier to achieve better accuracy. The model performance is evaluated by using the evaluation metrics such as Accuracy, Recall, Precision, and F1 score.

THEORETICAL FRAMEWORK FOR DOCUMENT CLASSIFICATION

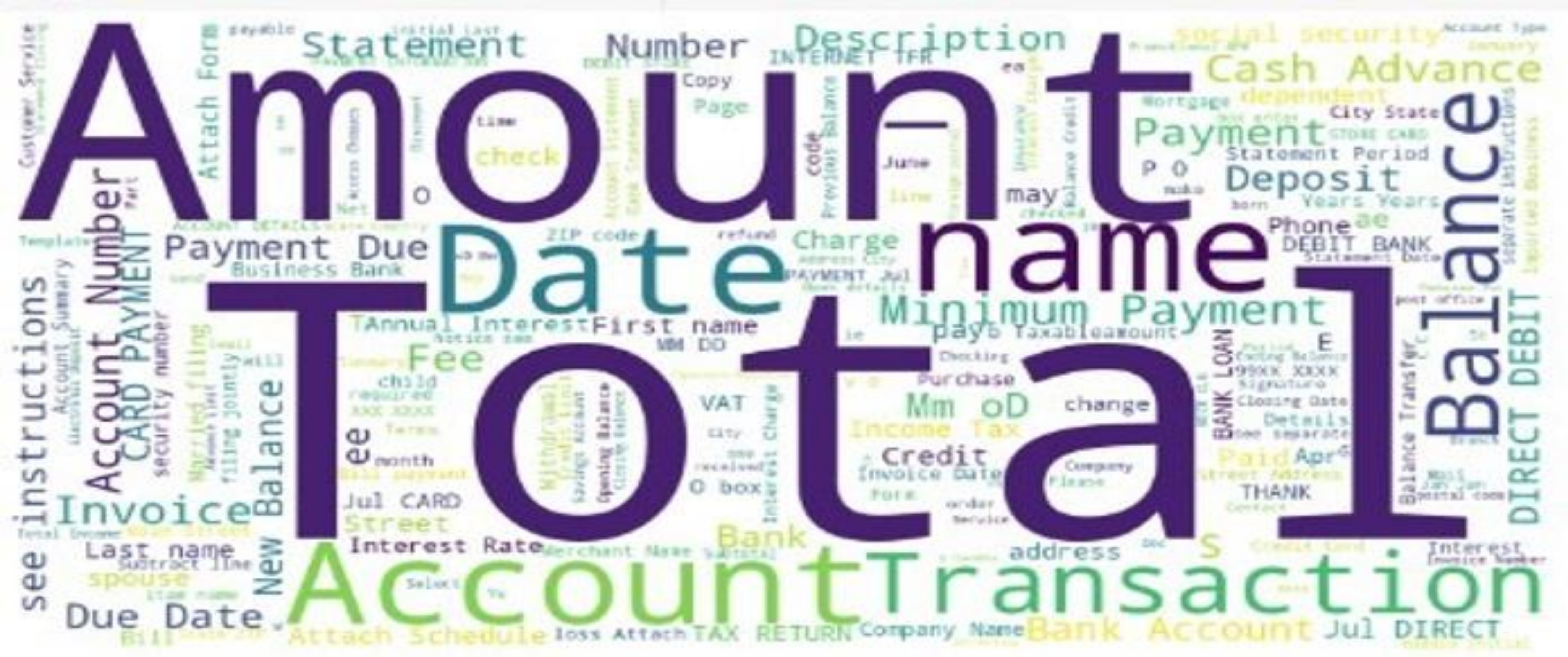
The purpose of document classification in banking is to organize and categorize documents such as bank statements, invoices, credit card statements, and tax returns, available in various formats (PDFs, Docs, Images), to improve processing efficiency, accuracy, compliance, and data security. Documents contain structured and unstructured data requiring preprocessing to remove unwanted elements. The text extraction process involves preparing and scanning documents, converting them to digital formats using OCR, preprocessing the text to clean and normalize it, and classifying it using machine learning algorithms. Online classification uses ML and NLP for quick, automated handling of digital documents, while offline classification involves manual processing of physical documents, essential for non-digital records.

PROPOSED METHODOLOGY :

The document classification process involves determining document types, collecting and preprocessing data, extracting features using TF-IDF vectorization , training and evaluating the model, and predicting the best-performing model.



WORD CLOUD REPRESENTATION OF KEYWORDS



DATA ACQUISITION AND PREPROCESSING :

Optical Character Recognition (OCR):

Utilize OCR technology to extract text from scanned or photographed bank documents, converting images into machine-readable data.

Text Extraction:

Apply advanced text extraction techniques to parse and isolate relevant information from the OCR output, preparing the data for classification.

Data Cleaning:

Implement robust data cleaning and normalization processes to address any inconsistencies, typos, or formatting issues in the extracted text.

MODEL SELECTION AND TRAINING

1. **Random Forest**

   Robust algorithm that creates multiple decision trees to classify documents. Handles both numerical and categorical features well.

2. **XGBoost**

   Highly optimized gradient boosting method known for its speed and accuracy on a wide range of document classification tasks.

3. **CatBoost**

   Gradient boosting framework that can handle categorical variables effectively, making it well-suited for mixed data types in bank documents.

4. **Voting Classifier**

   Ensemble method that combines predictions from multiple models to improve overall classification performance and robustness.

EVALUATION PARAMETERS

Confusion Matrix:

A confusion matrix, also known as an error matrix, is a table that provides a summary of the performance of a classification model on a set of test data. It allows us to visualize the performance of a model by showing the number of correct and incorrect predictions made for each class. The confusion matrix is particularly useful when dealing with multi-class classification problems.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Evaluation Parameters :

A confusion matrix typically has four cells representing different outcomes:

- **True Positives (TP)**: The number of samples that are correctly predicted as positive (belonging to the positive class).
- **True Negatives (TN)**: The number of samples that are correctly predicted as negative (not belonging to the positive class).
- **False Positives (FP)**: The number of samples that are incorrectly predicted as positive (misclassified as belonging to the positive class when they actually don't).
- **False Negatives (FN)**: The number of samples that are incorrectly predicted as negative (misclassified as not belonging to the positive class when they actually do).

VARIOUS PERFORMANCE METRICS

From this confusion matrix, various performance metrics can be calculated, including:

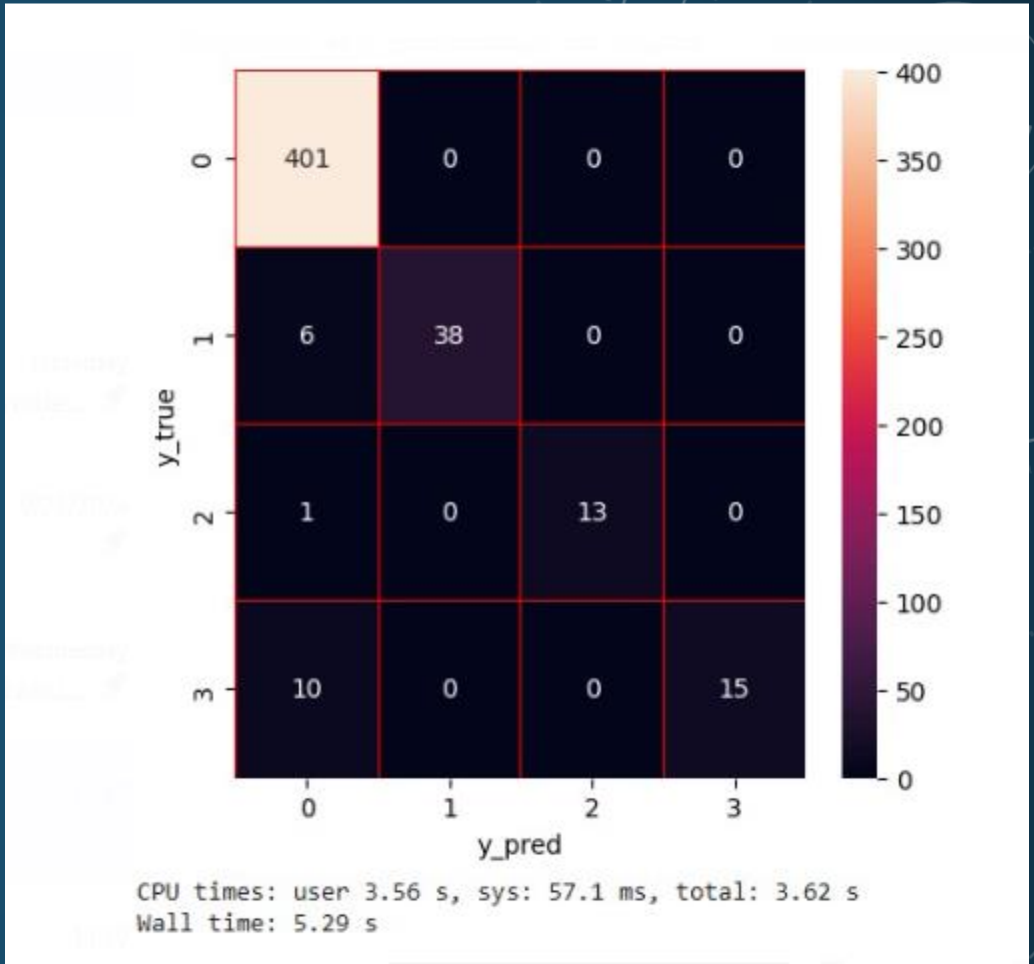
- **Accuracy**: The overall accuracy of the model, calculated as (TP + TN) / (TP + TN + FP + FN).
- **Precision**: Also known as the positive predictive value, it measures the proportion of correctly predicted positive samples out of all samples predicted as positive, calculated as TP / (TP + FP).
- **Recall**: Also known as sensitivity or true positive rate, it measures the proportion of correctly predicted positive samples out of all actual positive samples, calculated as TP / (TP + FN).
- **F1-score**: The harmonic mean of precision and recall, providing a balanced measure of a model's performance, calculated as 2 * (Precision * Recall) / (Precision + Recall).

RANDOM FOREST ALGORITHM

Random Forest algorithm is highly beneficial for bank document classification projects due to its capability to effectively manage large and varied datasets. By utilizing ensemble learning, Random Forest combines multiple decision trees, thereby improving classification accuracy while mitigating overfitting. It identifies influential features within documents, such as words, phrases, and metadata, essential for precise classification in banking contexts. Moreover, its robustness against noise and scalability to handle extensive datasets make it particularly suitable for real-world applications within banking environments.

CLASSIFIER-1 RANDOM FOREST

```
[[401   0   0   0]
 [  6  38   0   0]
 [  1   0  13   0]
 [ 10   0   0  15]]

Accuracy: 0.96

Micro Precision: 0.96
Micro Recall: 0.96
Micro F1-score: 0.96

Macro Precision: 0.99
Macro Recall: 0.85
Macro F1-score: 0.90

Weighted Precision: 0.97
Weighted Recall: 0.96
Weighted F1-score: 0.96

Classification Report

              precision    recall  f1-score   support

     Class 1       0.96      1.00      0.98       401
     Class 2       1.00      0.86      0.93        44
     class 3       1.00      0.93      0.96        14
     class 4       1.00      0.60      0.75        25

    accuracy                           0.96       484
   macro avg       0.99      0.85      0.90       484
weighted avg       0.97      0.96      0.96       484
```



1.a Classifier-1 Evaluation Metrics values

1.b Classifier -1 Heat Map

The Random Forest gives Accuracy (0.96), Precision(0.97), Recall(0.96),F1-Score(0.96)

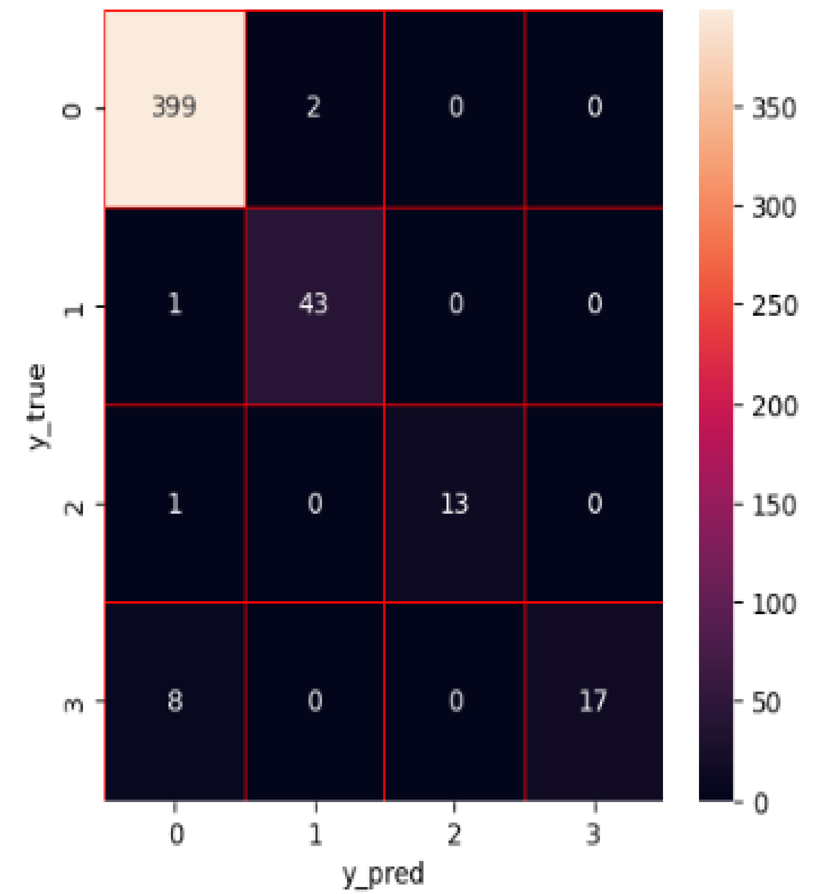
XGBOOST ALGORITHM(GRADIENT BOOSTING)

XGBoost, known as Extreme Gradient Boosting, is a powerful machine learning method specifically designed for handling structured data tasks efficiently. It works by building a series of simple predictive models, often decision trees, in a step-by-step manner. Each new model corrects errors made by the previous ones, leading to improved accuracy. What sets XGBoost apart is its ability to manage large datasets swiftly, handle missing data effectively, and deal with complex relationships between features. In our bank document classification project, XGBoost played a crucial role in achieving high accuracy by integrating with customized data preprocessing and feature engineering steps.

XGBOOST ALGORITHM

```
Confusion Matrix

[[399   2   0   0]
 [  1  43   0   0]
 [  1   0  13   0]
 [  8   0   0  17]]

Accuracy: 0.98

Micro Precision: 0.98
Micro Recall: 0.98
Micro F1-score: 0.98

Macro Precision: 0.98
Macro Recall: 0.90
Macro F1-score: 0.93

Weighted Precision: 0.98
Weighted Recall: 0.98
Weighted F1-score: 0.97

Classification Report

              precision    recall  f1-score   support

     Class 1       0.98      1.00      0.99       401
     Class 2       0.96      0.98      0.97        44
     class 3       1.00      0.93      0.96        14
     class 4       1.00      0.68      0.81        25

    accuracy                           0.98       484
   macro avg       0.98      0.90      0.93       484
weighted avg       0.98      0.98      0.97       484
```

2.a Classifier-2 Evaluation Metrics values




```
CPU times: user 6.37 s, sys: 45.3 ms, total: 6.42 s
Wall time: 8 s
```

2.b.Heat Map

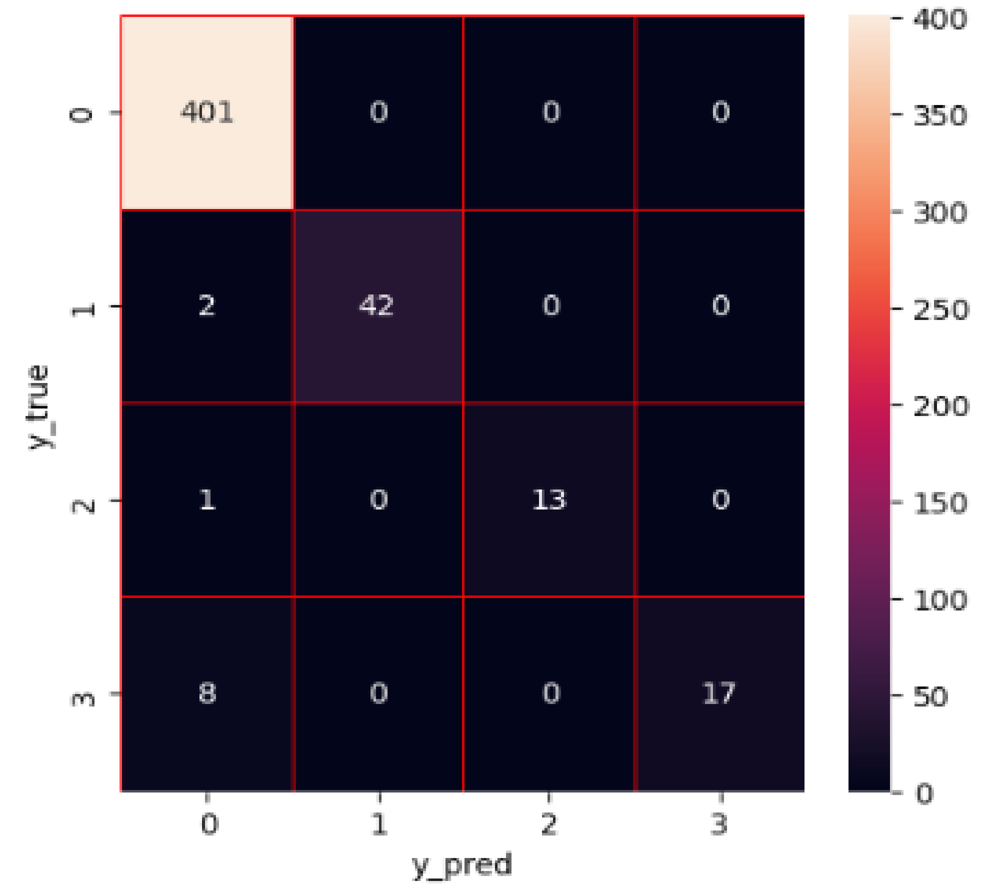
XGBoost gives the Accuracy (0.98), Precision (0.98), Recall(0.98),F1-Score (0.98)

CATEGORICAL BOOSTING ALGORITHM

CatBoost, short for Categorical Boosting, is a state-of-the-art machine learning algorithm specifically designed to handle categorical data efficiently. It uses gradient boosting on decision trees, similar to other boosting algorithms, but stands out due to its ability to directly process categorical features without extensive preprocessing. This results in more accurate and faster models. In our bank document classification project, CatBoost excelled by automatically handling the categorical nature of the data, leading to improved accuracy and efficiency. This made it an ideal choice for categorizing various bank documents, enhancing the overall performance and decision-making process within the financial institution.

CLASSIFIER-3: CATBOOST

```
Confusion Matrix

[[401   0   0   0]
 [  2  42   0   0]
 [  1   0  13   0]
 [  8   0   0  17]]

Accuracy: 0.98

Micro Precision: 0.98
Micro Recall: 0.98
Micro F1-score: 0.98

Macro Precision: 0.99
Macro Recall: 0.89
Macro F1-score: 0.93

Weighted Precision: 0.98
Weighted Recall: 0.98
Weighted F1-score: 0.98

Classification Report

              precision    recall  f1-score   support

     Class 1       0.97      1.00      0.99       401
     Class 2       1.00      0.95      0.98        44
     Class 3       1.00      0.93      0.96        14
     Class 4       1.00      0.68      0.81        25

    accuracy                           0.98       484
   macro avg       0.99      0.89      0.93       484
weighted avg       0.98      0.98      0.98       484
```




```
CPU times: user 5min 29s, sys: 2.12 s, total: 5min 31s
Wall time: 4min 10s
```

3.a. Classifier-3 Evaluation Metrics values

3.b Heat Map

CatBoost gives the Accuracy (0.98), Precision(0.98), Recall(0.98),F1-Score (0.98)

CLASSIFIER:4-VOTING CLASSIFIER

A Voting Classifier is an ensemble method that combines the predictions of multiple models, such as XGBoost, CatBoost, and Random Forests, to improve accuracy. By leveraging the strengths of each algorithm, it enhances robustness and performance. In our bank document classification project, the Voting Classifier significantly boosted accuracy and reliability, improving document categorization efficiency within the financial institution.

```
Confusion Matrix

[[401   0   0   0]
 [  2  42   0   0]
 [  1   0  13   0]
 [  9   0   0  16]]

Accuracy: 0.98

Micro Precision: 0.98
Micro Recall: 0.98
Micro F1-score: 0.98

Macro Precision: 0.99
Macro Recall: 0.88
Macro F1-score: 0.93

Weighted Precision: 0.98
Weighted Recall: 0.98
Weighted F1-score: 0.97

Classification Report

              precision    recall  f1-score   support

     Class 1       0.97      1.00      0.99       401
     Class 2       1.00      0.95      0.98        44
     class 3       1.00      0.93      0.96        14
      class4       1.00      0.64      0.78        25

    accuracy                           0.98       484
   macro avg       0.99      0.88      0.93       484
weighted avg       0.98      0.98      0.97       484
```

Voting Classifier gives the Accuracy (0.98), Precision (0.99), Recall(0.98),F1-Score (0.97)

TECHNOLOGIES AND LIBRARIES USED

Jupyter Notebook:

Jupyter Notebook is an open-source web-based application that allows you to create and share documents containing live code, visualizations, explanatory text, and more. Jupyter Notebook provides an interactive environment where you can write and execute code in cells, view the output, and document your work.

Python Libraries:

Numpy:

NumPy is a popular Python library for numerical computations. It stands for "Numerical Python." NumPy provides a powerful and efficient way to work with arrays, matrices, and multi-dimensional data in Python. It is a fundamental library for scientific computing and data analysis in Python.

TECHNOLOGIES AND LIBRARIES USED

Pandas:

Pandas is a powerful and popular open-source Python library for data manipulation and analysis. It provides data structures and functions that make it easier to work with structured data, such as CSV files and more. Pandas is built on top of NumPy and is widely used in data science, machine learning, and data analysis workflows.

Matplotlib:

Matplotlib is a widely used Python library for creating static, animated, and interactive visualizations. It provides a flexible and comprehensive set of tools for generating various types of plots, charts, and graphs. Matplotlib is often used in data analysis, scientific research, and data visualization tasks.

Seaborn:

Seaborn is a Python data visualization library built on top of Matplotlib. It provides a high-level interface for creating attractive and informative statistical graphics.

TEXT PREPROCESSING TECHNIQUES

Tokenization: Tokenization is the process of separating the text into words, sentences using the NLTK library. These tokens are useful for understanding context or developing NLP models.

Stemming: Stemming is a process which reduces a word to its base or root form. It is used to normalize the text.

Lemmatization: Lemmatization is a method of normalizing text documents. The main purpose of text normalization is to keep the vocabulary small and to remove noise, which helps improve the accuracy of many language modeling tasks

Vectorization: In general, Machines can't understand or processed text data in a raw form. The text is converted into numerical format (Vector) that easily readable by the Machine. TF-IDF**(Term frequency — Inverse document frequency)**

HARDWARE AND SOFTWARE REQUIREMENTS

Minimum Hardware requirement:

- RAM : 4GB
- Storage: HDD or SSD with sufficient storage
- System type: 64-bit

Minimum Software requirements:

- Operating System: Windows 8 and above
- Programming Language : Python 3.9
- IDE: Jupyter-notebook

RESULTS

Accuracy Analysis:

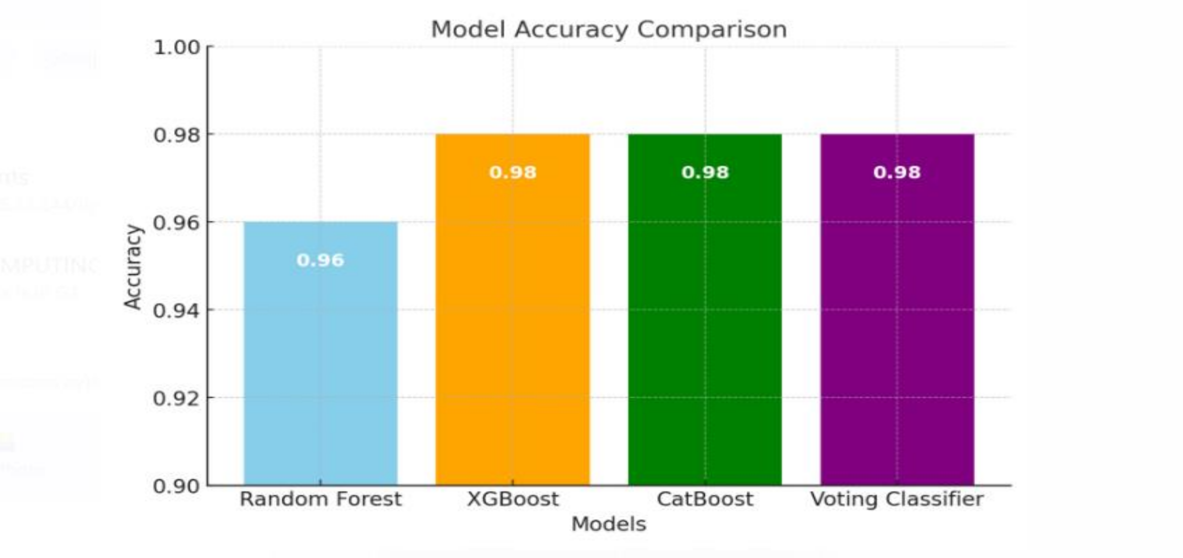
The classification models performed well, with accuracies between **96% and 98%.**

Models like Random Forest, XGBoost, CatBoost, and Voting Classifier were tested.

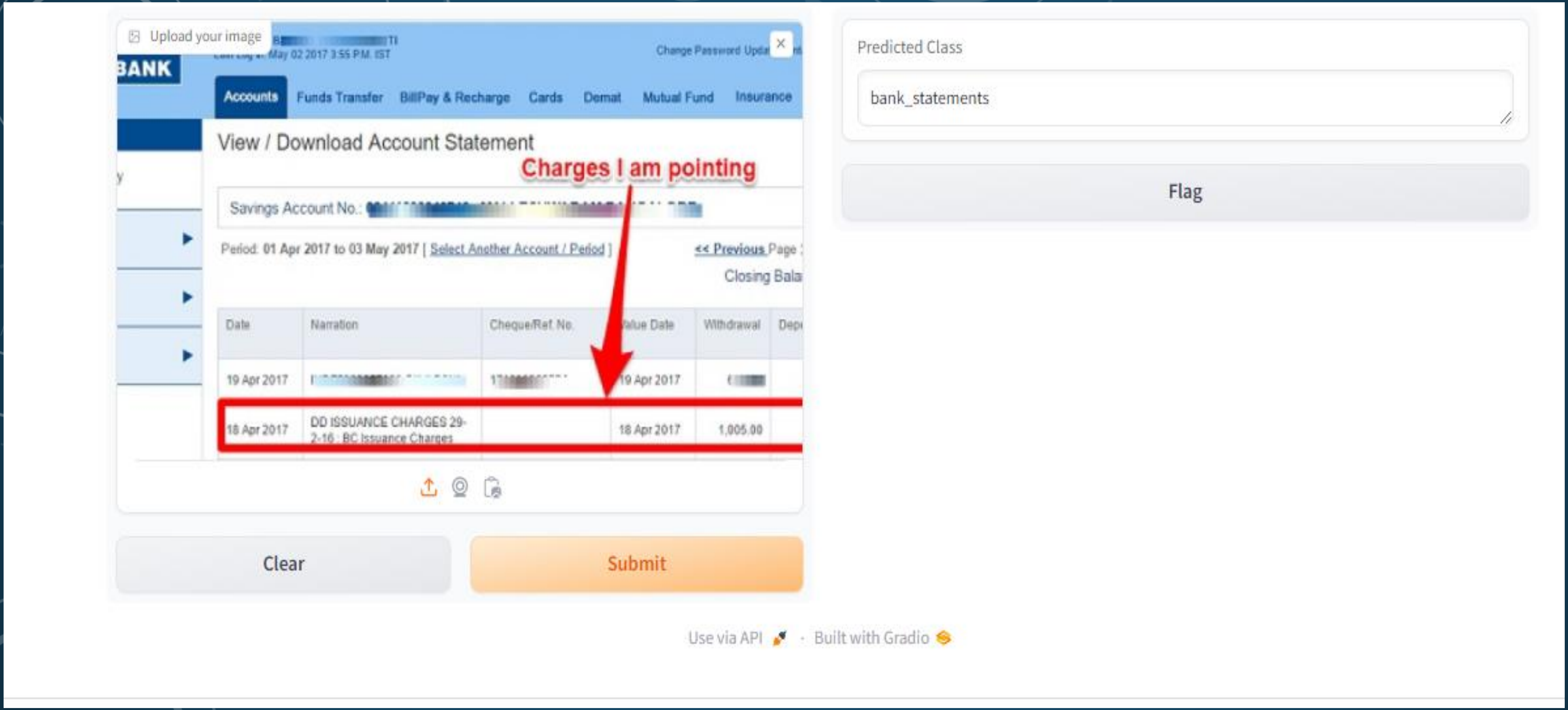
Individual Classifier Results:

Random Forest: 96% accuracy.
XGBoost, CatBoost, and Voting Classifier: 98% accuracy.

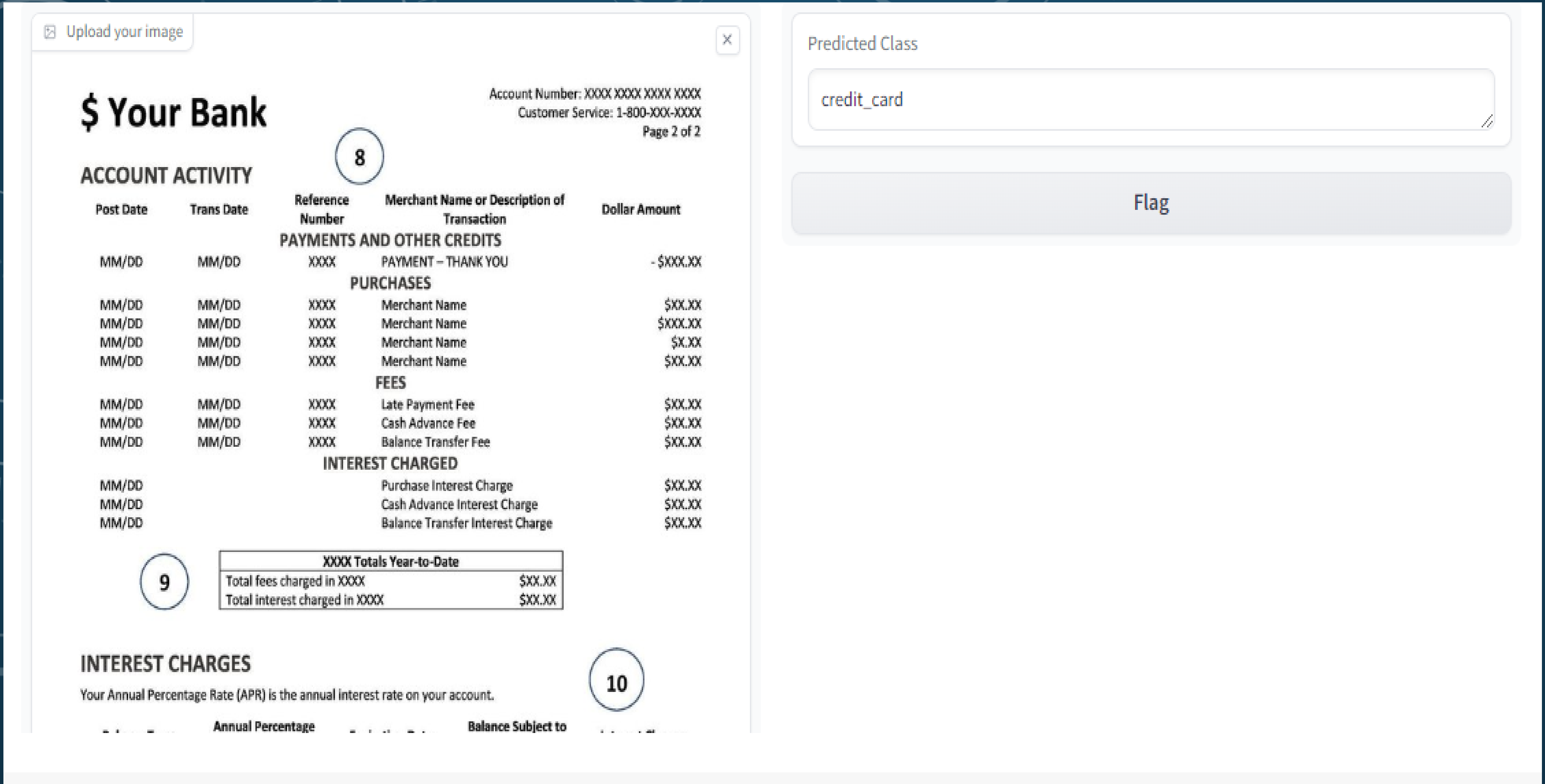
Model/Classifier	Accuracy	Precision	Recall	F1-Score
Random Forest	0.96	0.97	0.96	0.96
XG Boost	0.98	0.98	0.98	0.98
Cat Boost	0.98	0.98	0.98	0.98
Voting Classifier	0.98	0.99	0.98	0.97



Bank_Statements Output:



Credit_card Output:



$ Your Bank

Account Number: XXXX XXXX XXXX XXXX
Customer Service: 1-800-XXX-XXXX
Page 2 of 2

8

ACCOUNT ACTIVITY

Post Date	Trans Date	Reference Number	Merchant Name or Description of Transaction	Dollar Amount
		PAYMENTS AND OTHER CREDITS		
MM/DD	MM/DD	XXXX	PAYMENT – THANK YOU	- $XXX.XX
		PURCHASES		
MM/DD	MM/DD	XXXX	Merchant Name	$XX.XX
MM/DD	MM/DD	XXXX	Merchant Name	$XXX.XX
MM/DD	MM/DD	XXXX	Merchant Name	$X.XX
MM/DD	MM/DD	XXXX	Merchant Name	$XX.XX
		FEES		
MM/DD	MM/DD	XXXX	Late Payment Fee	$XX.XX
MM/DD	MM/DD	XXXX	Cash Advance Fee	$XX.XX
MM/DD	MM/DD	XXXX	Balance Transfer Fee	$XX.XX
		INTEREST CHARGED		
MM/DD			Purchase Interest Charge	$XX.XX
MM/DD			Cash Advance Interest Charge	$XX.XX
MM/DD			Balance Transfer Interest Charge	$XX.XX

9

XXXX Totals Year-to-Date	
Total fees charged in XXXX	$XX.XX
Total interest charged in XXXX	$XX.XX

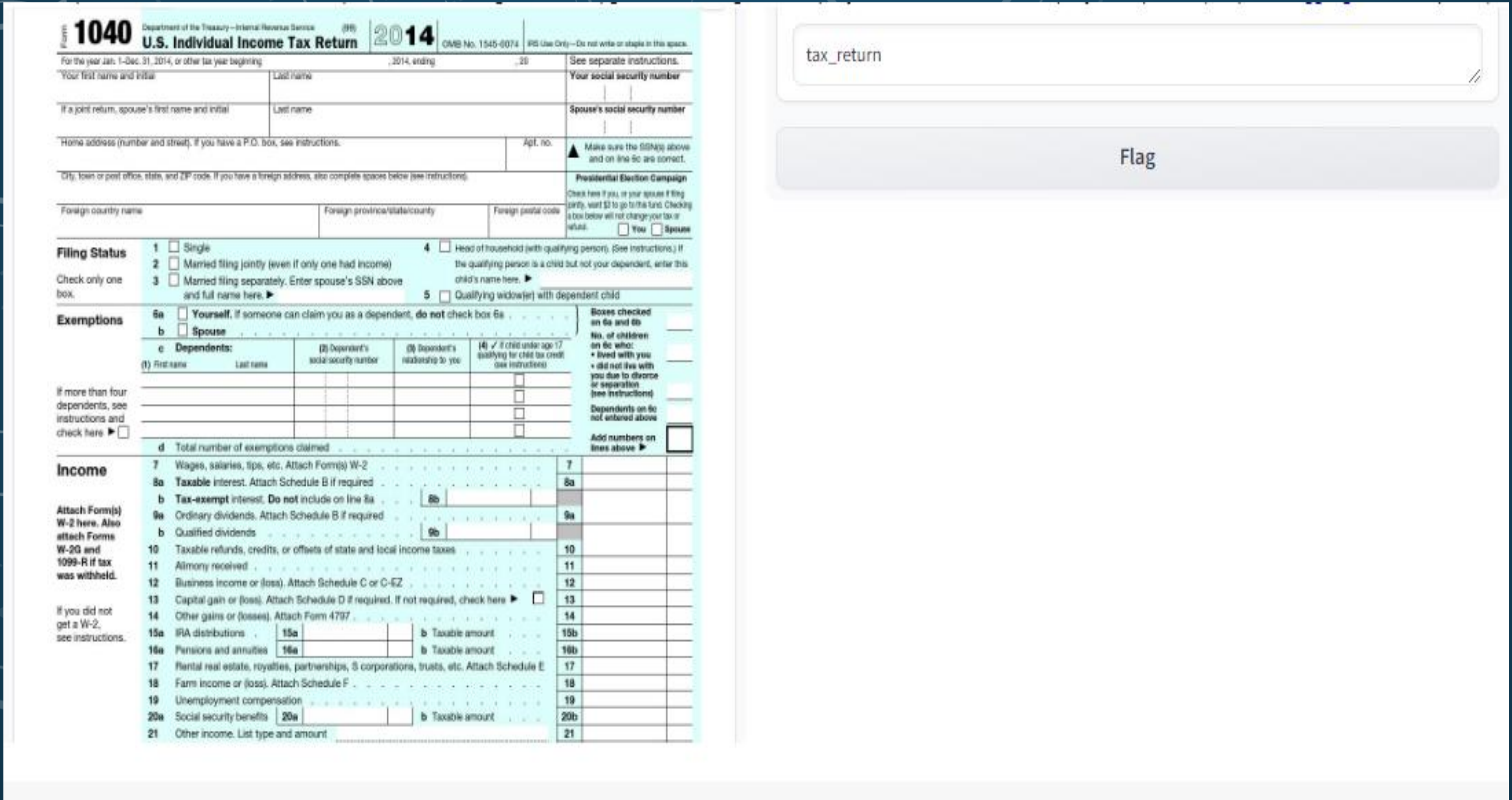
INTEREST CHARGES

10

Your Annual Percentage Rate (APR) is the annual interest rate on your account.

Annual Percentage … Balance Subject to

Tax_Return Output:


Form 1040 Department of the Treasury—Internal Revenue Service (99)
U.S. Individual Income Tax Return 2014 OMB No. 1545-0074 IRS Use Only—Do not write or staple in this space.

For the year Jan. 1–Dec. 31, 2014, or other tax year beginning , 2014, ending , 20 See separate instructions.

Your first name and initial | Last name | Your social security number

If a joint return, spouse's first name and initial | Last name | Spouse's social security number

Home address (number and street). If you have a P.O. box, see instructions. | Apt. no. | ▲ Make sure the SSN(s) above and on line 6c are correct.

City, town or post office, state, and ZIP code. If you have a foreign address, also complete spaces below (see instructions).

Presidential Election Campaign
Check here if you, or your spouse if filing jointly, want $3 to go to this fund. Checking a box below will not change your tax or refund. ☐ You ☐ Spouse

Foreign country name | Foreign province/state/county | Foreign postal code

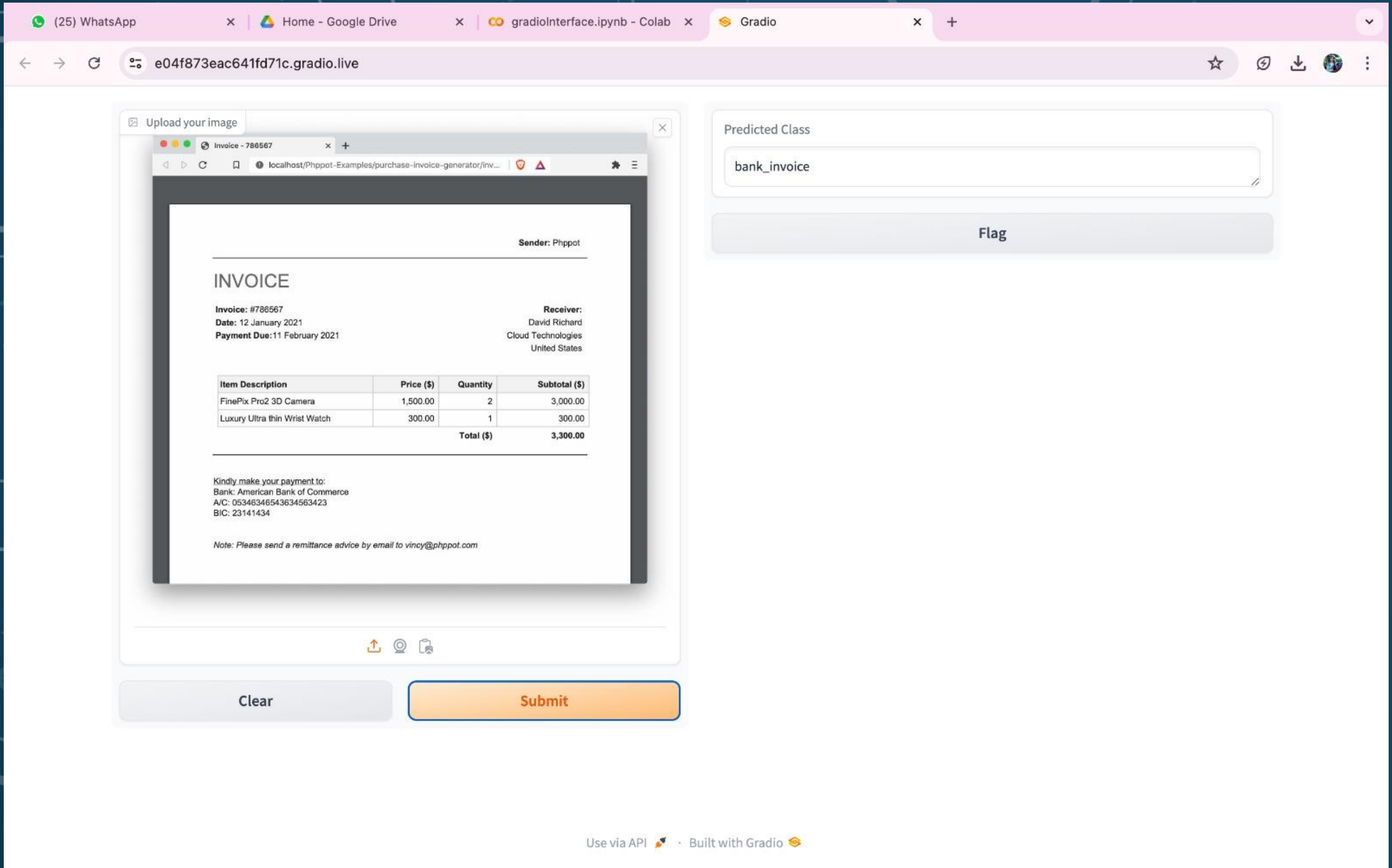
Filing Status
Check only one box.
1 ☐ Single
2 ☐ Married filing jointly (even if only one had income)
3 ☐ Married filing separately. Enter spouse's SSN above and full name here. ▶
4 ☐ Head of household (with qualifying person). (See instructions.) If the qualifying person is a child but not your dependent, enter this child's name here. ▶
5 ☐ Qualifying widow(er) with dependent child

Exemptions
6a ☐ **Yourself.** If someone can claim you as a dependent, **do not** check box 6a
b ☐ **Spouse**
c **Dependents:** (1) First name Last name | (2) Dependent's social security number | (3) Dependent's relationship to you | (4) ✓ if child under age 17 qualifying for child tax credit (see instructions)
If more than four dependents, see instructions and check here ▶ ☐
d Total number of exemptions claimed

Boxes checked on 6a and 6b
No. of children on 6c who:
• lived with you
• did not live with you due to divorce or separation (see instructions)
Dependents on 6c not entered above
Add numbers on lines above ▶

Income
Attach Form(s) W-2 here. Also attach Forms W-2G and 1099-R if tax was withheld.
If you did not get a W-2, see instructions.
7 Wages, salaries, tips, etc. Attach Form(s) W-2 — 7
8a **Taxable** interest. Attach Schedule B if required — 8a
b **Tax-exempt** interest. **Do not** include on line 8a — 8b
9a Ordinary dividends. Attach Schedule B if required — 9a
b Qualified dividends — 9b
10 Taxable refunds, credits, or offsets of state and local income taxes — 10
11 Alimony received — 11
12 Business income or (loss). Attach Schedule C or C-EZ — 12
13 Capital gain or (loss). Attach Schedule D if required. If not required, check here ▶ ☐ — 13
14 Other gains or (losses). Attach Form 4797 — 14
15a IRA distributions 15a — b Taxable amount — 15b
16a Pensions and annuities 16a — b Taxable amount — 16b
17 Rental real estate, royalties, partnerships, S corporations, trusts, etc. Attach Schedule E — 17
18 Farm income or (loss). Attach Schedule F — 18
19 Unemployment compensation — 19
20a Social security benefits 20a — b Taxable amount — 20b
21 Other income. List type and amount — 21

Bank Invoice Output:



COMPARATIVE ANALYSIS OF CLASSIFIER'S RESULTS

XGBoost, CatBoost, and Voting Classifier achieved the highest accuracy of 91%, outperforming Random Forest.

The Voting Classifier combined predictions from multiple models, making it the most reliable and robust.

Model/Classifer	Classification	Accuracy	Precision	Recall	F1-Score
Random Forest	bank_invoice	0.96	0.96	1.00	0.98
	bank_statements		1.00	0.86	0.93
	credit_card		1.00	0.93	0.96
	tax_return		1.00	0.6	0.75
XG Boost	bank_invoice	0.98	0.98	1.00	0.99
	bank_statements		0.96	0.98	0.97
	credit_card		1.00	0.93	0.96
	tax_return		1.00	0.68	0.81
Cat Boost	bank_invoice	0.98	0.97	1.00	0.99
	bank_statements		1.00	0.95	0.98
	credit_card		1.00	0.93	0.96
	tax_return		1.00	0.68	0.81
Voting Classifier	bank_invoice	0.98	0.97	1.00	0.99
	bank_statements		1.00	0.95	0.98
	credit_card		1.00	0.93	0.96
	tax_return		1.00	0.64	0.78

OBSERVATION

1. It is observed that the classification of the bank documents achieved better results with different ensemble Machine Learning Models (i.e Random Forest, XG Boost and Cat Boost as model and Voting Classifier is bench mark model

2. The performance of the individual classifier is measured based on the output values of evaluation metrics (Accuracy, Precision, Recall and F1-Score)
   i. The Random Forest gives Accuracy (0.96), Precision(0.96), Recall(0.96),F1-Score (0.96)
   ii. XGBoost gives the Accuracy (0.98), Precision (0.98), Recall(0.91),F1-Score (0.98)
   iii. CatBoost gives the Accuracy (0.98), Precision(0.98), Recall(0.98),F1-Score (0.98)
   iv. Voting Classifier gives the Accuracy (0.98), Precision (0.98), Recall(0.98),F1- Score (0.98)

Based the results obtained from the above mentioned 4 classifiers , It is observed that Voting Classifier, XGBoost and CatBoost algorithms gives highest accuracy(98%) among all four classifiers and equally performed.

CHALLENGES:

- Cleaning the datasets (Scanned documents)
- Bank documents can come in a variety of formats, including PDF files, scanned images, and handwritten notes, making it difficult to extract useful information. Some files may contain information belonging to more than one category, making it difficult to categorize them correctly.
- Banks may update their document formats, which requires retraining of classification patterns to recognize new patterns.

LIMITATIONS:

- Extracting handwritten signatures and converting to computer generated text
- While converting PDF documents(text and images) into text only text information is retrieved.

FUTURE WORK:

- As part of future work , finance and Banking sectors are tremendously developed globally integrating the Deep Learning and Natural Language Processing techniques into document classification can enhance accuracy and efficiency by reducing errors.

CONCLUSION :

- The proposed system uses advanced ensemble machine learning algorithms and natural language processing techniques, which helps to automatically categorize and classify the bank documents that contains both structured and unstructured data. The salient features extracted from the ensemble machine learning model are fed into output classification.
- The proposed ensemble machine learning models XGBoost(98%) and Voting Classifier(98%) achieved highest accuracy in comparison with Random Forest (96%) and CatBoost(98%).

REFERENCES

1. Sarosh Dandoti,(2022), Text Document Classification System,International Research Journal of Engineering and Technology (IRJET), 9(6),2847-2850
2. Arslan, Ö., & Uymaz, S. A. (2022). Classification of Invoice Images by Using Convolutional Neural Networks. Journal of Advanced Research in Natural and Applied Sciences, 8(1), 8-25.
3. Sheth, V., Tripathi, U., & Sharma, A. (2022). A Comparative Analysis of Machine Learning Algorithms for Classification Purpose. Procedia Computer Science, 215, 422-431.
4. Ghumade, T. G., & Deshmukh, R. A. (2019). A document classification using NLP and recurrent neural network. Int. J. Eng. Adv. Technol, 8(6), 632-636.
5. Engin, D., Emekligil, E., Oral, B., Arslan, S., & Akpınar, M. (2019). Multimodal deep neural networks for banking document classification. In International Conference on Advances in Information Mining and Management (pp. 21-25).

THANK YOU