Documentation for the Resume Categorization System

I reviewed a collection of PDF resumes from different types of individuals. In the dataset, I found that there were around 100 resumes per group, on average. The goal is to develop a machine learning or deep learning model that predicts the category of people based on their resume and saves the results of the test dataset in a CSV file.

To accomplish this, I first extracted the text from each PDF and labeled them accordingly. I used regular expressions to remove stopwords and extra spaces. I applied TF-IDF to reduce the impact of common words while emphasizing rarer words that are more frequent in specific groups' resumes.

For classification, I chose to use the Support Vector Machine (SVM) model. Although I could have opted for other models like Random Forest, ANN, RNN, LSTM, or even heavier models like BERT, due to time constraints and the limited processing power of my personal computer, I decided to go with the simplest approach.

After completing the code, the model's evaluation metrics for the validation set were as follows: Accuracy: 0.6660, Precision: 0.6838, Recall: 0.6660, and F1-score: 0.6611. For the testing set, the metrics were Accuracy: 0.6177, Precision: 0.6313, Recall: 0.6177, and F1-score: 0.6046.

- **Accuracy** measures the proportion of correct predictions across the entire dataset.

$$Accuracy = \frac{TP + TN}{TP+FP+TN+FN}$$

- **Precision** indicates the reliability of positive predictions.

$$Precision = \frac{TP}{TP+FP}$$

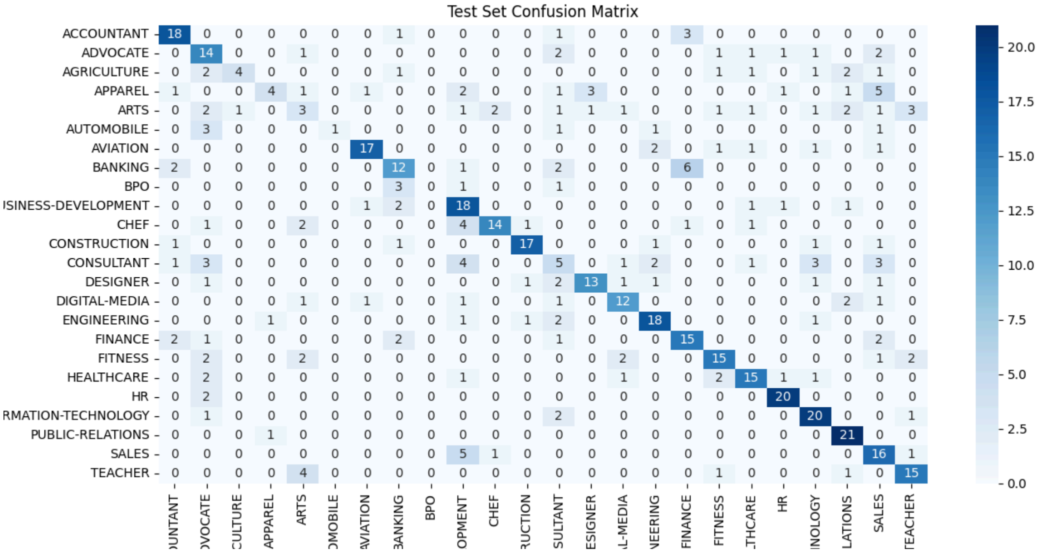
- **Recall** measures the model’s ability to correctly identify actual positive instances.

$$Recall = \frac{TP}{TP+FN}$$

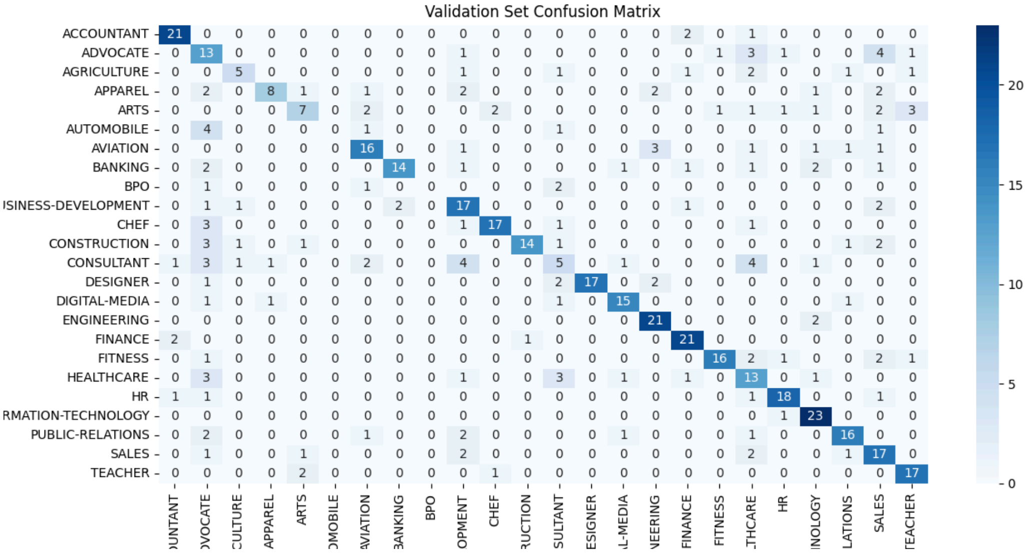
- **F1-score** seeks to balance Precision and Recall.

$$F1 - score = \frac{Precision \times Recall}{Precision+Recall}$$

Here the photo of confusion matrices:


Test Set Confusion Matrix

	UNTANT	VOCATE	CULTURE	APPAREL	ARTS	MOBILE	VIATION	ANKING	BPO	OPMENT	CHEF	RUCTION	SULTANT	ESIGNER	L-MEDIA	NEERING	FINANCE	FITNESS	THCARE	HR	NOLOGY	LATIONS	SALES	TEACHER
ACCOUNTANT	18	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	3	0	0	0	0	0	0	0
ADVOCATE	0	14	0	0	1	0	0	0	0	0	0	0	2	0	0	0	0	1	1	1	1	0	2	0
AGRICULTURE	0	2	4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	1	2	1	0
APPAREL	1	0	0	4	1	0	1	0	0	2	0	0	1	3	0	0	0	0	0	1	0	1	5	0
ARTS	0	2	1	0	3	0	0	0	0	1	2	0	1	1	1	0	0	1	1	0	1	2	1	3
AUTOMOBILE	0	3	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0
AVIATION	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	2	0	1	1	0	1	0	1	0
BANKING	2	0	0	0	0	0	0	12	0	1	0	0	2	0	0	0	6	0	0	0	0	0	0	0
BPO	0	0	0	0	0	0	0	3	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ISINESS-DEVELOPMENT	0	0	0	0	0	0	1	2	0	18	0	0	0	0	0	0	0	0	1	1	0	1	0	0
CHEF	0	1	0	0	2	0	0	0	0	4	14	1	0	0	0	0	1	0	1	0	0	0	0	0
CONSTRUCTION	1	0	0	0	0	0	0	1	0	0	0	17	0	0	0	1	0	0	0	0	1	0	1	0
CONSULTANT	1	3	0	0	0	0	0	0	0	4	0	0	5	0	1	2	0	0	1	0	3	0	3	0
DESIGNER	0	1	0	0	0	0	0	0	0	0	0	1	2	13	1	1	0	0	0	0	1	0	1	0
DIGITAL-MEDIA	0	0	0	0	1	0	1	0	0	1	0	0	1	0	12	0	0	0	0	0	0	2	1	0
ENGINEERING	0	0	0	1	0	0	0	0	0	1	0	1	2	0	0	18	0	0	0	0	1	0	0	0
FINANCE	2	1	0	0	0	0	0	2	0	0	0	0	1	0	0	0	15	0	0	0	0	0	2	0
FITNESS	0	2	0	0	2	0	0	0	0	0	0	0	0	0	2	0	0	15	0	0	0	0	1	2
HEALTHCARE	0	2	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	2	15	1	1	0	0	0
HR	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0
RMATION-TECHNOLOGY	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	20	0	0	1
PUBLIC-RELATIONS	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0
SALES	0	0	0	0	0	0	0	0	0	5	1	0	0	0	0	0	0	0	0	0	0	0	16	1
TEACHER	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	15


Validation Set Confusion Matrix

	UNTANT	VOCATE	CULTURE	APPAREL	ARTS	MOBILE	VIATION	ANKING	BPO	OPMENT	CHEF	RUCTION	SULTANT	ESIGNER	L-MEDIA	NEERING	FINANCE	FITNESS	THCARE	HR	NOLOGY	LATIONS	SALES	TEACHER
ACCOUNTANT	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	0
ADVOCATE	0	13	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	3	1	0	0	4	1
AGRICULTURE	0	0	5	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	2	0	0	1	0	1
APPAREL	0	2	0	8	1	0	1	0	0	2	0	0	0	0	0	2	0	0	0	0	1	0	2	0
ARTS	0	0	0	0	7	0	2	0	0	0	2	0	0	0	0	0	0	1	1	1	1	0	2	3
AUTOMOBILE	0	4	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
AVIATION	0	0	0	0	0	0	16	0	0	1	0	0	0	0	0	3	0	0	1	0	1	1	1	0
BANKING	0	2	0	0	0	0	0	14	0	1	0	0	0	0	1	0	1	0	1	0	2	0	1	0
BPO	0	1	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
ISINESS-DEVELOPMENT	0	1	1	0	0	0	0	2	0	17	0	0	0	0	0	0	1	0	0	0	0	0	2	0
CHEF	0	3	0	0	0	0	0	0	0	1	17	0	1	0	0	0	0	0	1	0	0	0	0	0
CONSTRUCTION	0	3	1	0	1	0	0	0	0	0	0	14	1	0	0	0	0	0	0	0	0	1	2	0
CONSULTANT	1	3	1	1	0	0	2	0	0	4	0	0	5	0	1	0	0	0	4	0	1	0	0	0
DESIGNER	0	1	0	0	0	0	0	0	0	0	0	0	2	17	0	2	0	0	0	0	0	0	0	0
DIGITAL-MEDIA	0	1	0	1	0	0	0	0	0	0	0	0	1	0	15	0	0	0	0	0	0	1	0	0
ENGINEERING	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	2	0	0	0
FINANCE	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	21	0	0	0	0	0	0	0
FITNESS	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	2	1	0	0	2	1
HEALTHCARE	0	3	0	0	0	0	0	0	0	1	0	0	3	0	1	0	1	0	13	0	1	0	0	0
HR	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	18	0	0	1	0
RMATION-TECHNOLOGY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	23	0	0	0
PUBLIC-RELATIONS	0	2	0	0	0	0	1	0	0	2	0	0	0	0	1	0	0	0	1	0	0	16	0	0
SALES	0	1	0	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	2	0	0	1	17	0
TEACHER	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	17

Improvements: There are several ways to improve these results. In the dataset, the highest number of resumes belongs to Information Technology professionals (120), while the lowest number of resumes is from BPO professionals (22). By applying random oversampling to the smaller datasets, the results could potentially be improved.