Samkit Shah (B20CS059)

CSL2050 – Bonus Project REPORT

Table of Contents

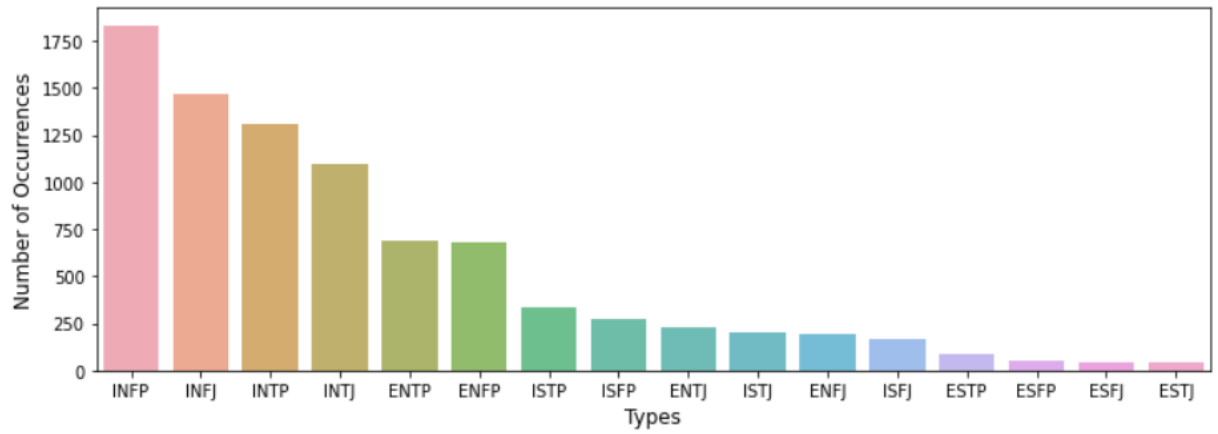
Main

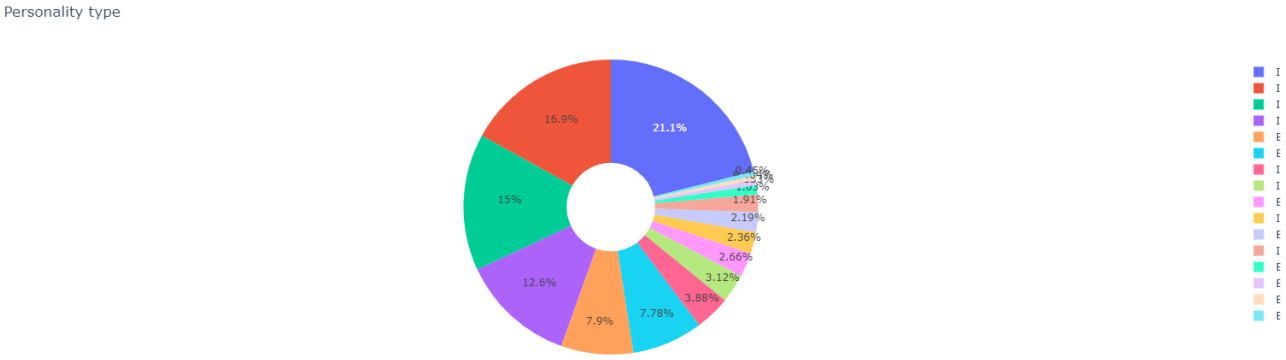
Aim

The dataset used to train the model was the (MBTI) Myers-Briggs Personality Type Dataset. The model makes a prediction based on data and classifies them into one of the 16 MBTI types. The webapp is built on Flask, Bootstrap.

Exploratory Data Analysis

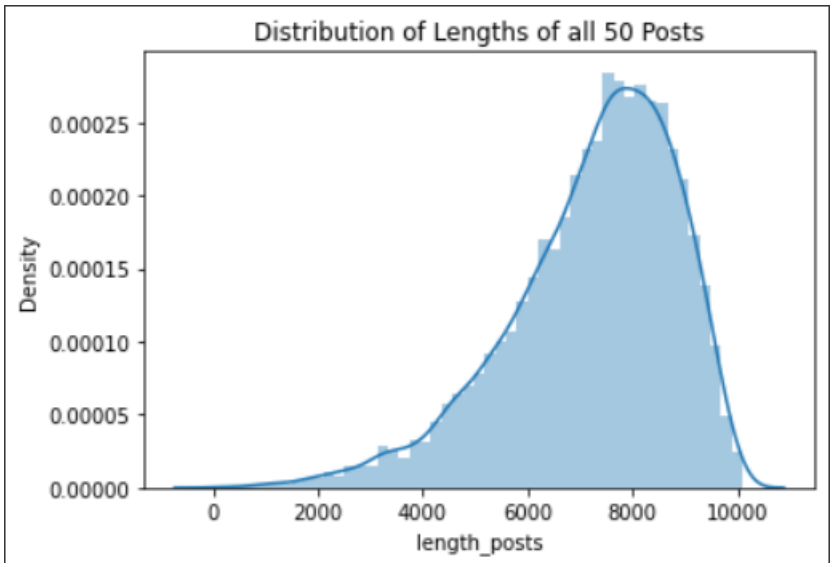
- **Imported** the data
- Explored Class disparity with the following graph





We can see that distribution is highly skewed and class wise disparity is very large.

- Next visualised density of lengths of posts in order to find some relation.



Distribution is almost perfect gaussian.

- Found the most frequent words. Top 15 words are:

	word	count
0	I	387957
1	to	290168
2	the	270699
3	a	230918
4	and	219498
5	of	177853
6	is	128804
7	you	128750
8	that	127221
9	in	117263
10	my	104561
11	it	93101
12	for	83057
13	have	79784
14	with	77131

- Made Word Cloud for every class:

```
•     Word Cloud is a data visualization technique used for representing text data in
which the size of each word indicates its frequency or importance.
```

Pre-processing

1) Label-Encoding

   Encoded the classes from 0-15 with help of Label Encoder

2) Cleaning the Text

   Since the data set comes from an Internet forum where individuals communicate strictly via written text, some word removal was clearly necessary. For example, there were several instances of data points containing links to websites. Since we want our model to generalize to the English language, we removed any data points containing links to websites. Next, since we want every word in the data to be as meaningful as possible, we removed so-called" stop words" from the text (e.g., very common filler words like" a"," the"," or", etc.) using python's NLTK. Finally, since the particular data set, we are working with comes from a website intended for explicit discussion of personality

models, especially MBTI, we removed types themselves (e.g., 'INTJ', 'INFP', etc.), so as to prevent the model from" cheating" by learning to recognize mentions of MBTI by name.

- ➔ Removed "|||" marks that separated the posts
- ➔ Used BeautifulSoup to scrape tables and text from '. lxml' URLs
- ➔ Removed URLs
- ➔ Removed digital numbers
- ➔ Removed white spaces
- ➔ Removed newlines
- ➔ Kept Punctuations
- ➔ Removed Stop words

3) Lemmatisation

Used nltk. stem. WordNetLemmatizer to lemmatize the text, meaning that inflected forms of the same root word were transformed into their dictionary form (e.g.," walking"," walked"," walk" all become" walk"). This will allow us to make use of the fact that inflected forms of the same word still carry one shared meaning.

4) Tokenization

TfidfVectorizer - Transformed text to feature vectors that can be used as input to estimator. vocabulary_ Is a dictionary that converts each token (word) to feature index in the matrix, each unique token gets a feature index.

- ➔ Split dataset into stratified splits of 0.8:0.2.

Training Models

- Modelled the data on following models:
  1) Linear Regression

train classification report

	precision	recall	f1-score	support
ENFJ	0.83	0.13	0.23	152
ENFP	0.81	0.64	0.71	540
ENTJ	0.94	0.26	0.41	185
ENTP	0.81	0.65	0.72	548
ESFJ	0.00	0.00	0.00	33
ESFP	0.00	0.00	0.00	38
ESTJ	0.00	0.00	0.00	31
ESTP	1.00	0.04	0.08	71
INFJ	0.73	0.82	0.77	1176
INFP	0.65	0.93	0.76	1466
INTJ	0.74	0.80	0.77	873
INTP	0.67	0.87	0.76	1043
ISFJ	0.89	0.25	0.39	133
ISFP	0.85	0.21	0.33	217
ISTJ	0.84	0.25	0.38	164
ISTP	0.88	0.50	0.64	270
accuracy			0.71	6940
macro avg	0.66	0.40	0.43	6940
weighted avg	0.73	0.71	0.68	6940

test classification report

	precision	recall	f1-score	support
ENFJ	1.00	0.08	0.15	38
ENFP	0.76	0.53	0.62	135
ENTJ	0.60	0.13	0.21	46
ENTP	0.63	0.50	0.56	137
ESFJ	0.00	0.00	0.00	9
ESFP	0.00	0.00	0.00	10
ESTJ	0.00	0.00	0.00	8
ESTP	0.00	0.00	0.00	18
INFJ	0.63	0.72	0.68	294
INFP	0.55	0.87	0.68	366
INTJ	0.61	0.63	0.62	218
INTP	0.65	0.82	0.72	261
ISFJ	0.75	0.09	0.16	33
ISFP	0.80	0.15	0.25	54
ISTJ	0.50	0.05	0.09	41
ISTP	0.70	0.34	0.46	67
accuracy			0.62	1735
macro avg	0.51	0.31	0.33	1735
weighted avg	0.62	0.62	0.58	1735

```
CV Accuracy: 0.6003 (+/- 0.0114)
CV F1: 0.6003 (+/- 0.0114)
CV Logloss: 1.5063 (+/- 0.0183)
```

  2) Linear SVC

train classification report

	precision	recall	f1-score	support
ENFJ	0.90	0.46	0.61	152
ENFP	0.85	0.76	0.80	540
ENTJ	0.92	0.62	0.74	185
ENTP	0.84	0.80	0.82	548
ESFJ	0.83	0.30	0.44	33
ESFP	1.00	0.08	0.15	38
ESTJ	1.00	0.26	0.41	31
ESTP	0.91	0.44	0.59	71
INFJ	0.81	0.84	0.83	1176
INFP	0.76	0.93	0.84	1466
INTJ	0.82	0.85	0.84	873
INTP	0.79	0.89	0.84	1043
ISFJ	0.91	0.64	0.75	133
ISFP	0.88	0.56	0.68	217
ISTJ	0.89	0.66	0.76	164
ISTP	0.89	0.81	0.85	270
accuracy			0.81	6940
macro avg	0.88	0.62	0.68	6940
weighted avg	0.82	0.81	0.81	6940

test classification report

	precision	recall	f1-score	support
ENFJ	0.67	0.21	0.32	38
ENFP	0.76	0.59	0.66	135
ENTJ	0.60	0.33	0.42	46
ENTP	0.58	0.53	0.56	137
ESFJ	1.00	0.33	0.50	9
ESFP	0.00	0.00	0.00	10
ESTJ	1.00	0.12	0.22	8
ESTP	0.67	0.22	0.33	18
INFJ	0.68	0.71	0.70	294
INFP	0.61	0.86	0.71	366
INTJ	0.62	0.62	0.62	218
INTP	0.70	0.82	0.75	261
ISFJ	0.62	0.24	0.35	33
ISFP	0.77	0.31	0.45	54
ISTJ	0.75	0.29	0.42	41
ISTP	0.70	0.52	0.60	67
accuracy			0.65	1735
macro avg	0.67	0.42	0.48	1735
weighted avg	0.66	0.65	0.63	1735

3) Decision Tree Classifier

train classification report

	precision	recall	f1-score	support
ENFJ	0.82	0.47	0.60	152
ENFP	0.93	0.77	0.84	540
ENTJ	0.74	0.77	0.75	185
ENTP	0.89	0.81	0.84	548
ESFJ	0.78	0.42	0.55	33
ESFP	0.75	0.32	0.44	38
ESTJ	1.00	0.32	0.49	31
ESTP	0.78	0.55	0.64	71
INFJ	0.85	0.84	0.84	1176
INFP	0.67	0.94	0.78	1466
INTJ	0.87	0.83	0.85	873
INTP	0.87	0.81	0.84	1043
ISFJ	1.00	0.46	0.63	133
ISFP	0.91	0.67	0.77	217
ISTJ	0.70	0.63	0.66	164
ISTP	0.87	0.72	0.79	270
accuracy			0.81	6940
macro avg	0.84	0.65	0.71	6940
weighted avg	0.82	0.81	0.80	6940

test classification report

	precision	recall	f1-score	support
ENFJ	0.14	0.08	0.10	38
ENFP	0.55	0.37	0.44	135
ENTJ	0.23	0.26	0.24	46
ENTP	0.49	0.45	0.47	137
ESFJ	0.00	0.00	0.00	9
ESFP	0.00	0.00	0.00	10
ESTJ	0.00	0.00	0.00	8
ESTP	0.14	0.06	0.08	18
INFJ	0.58	0.60	0.59	294
INFP	0.50	0.69	0.58	366
INTJ	0.54	0.56	0.55	218
INTP	0.61	0.60	0.61	261
ISFJ	0.38	0.15	0.22	33
ISFP	0.42	0.28	0.33	54
ISTJ	0.43	0.29	0.35	41
ISTP	0.59	0.54	0.56	67
accuracy			0.52	1735
macro avg	0.35	0.31	0.32	1735
weighted avg	0.51	0.52	0.51	1735

```
CV Accuracy: 0.4762 (+/- 0.0110)
CV F1: 0.4762 (+/- 0.0110)
CV Logloss: 11.5810 (+/- 0.4364)
```

4) Random Forest Classifier

train classification report

	precision	recall	f1-score	support
ENFJ	1.00	0.02	0.04	152
ENFP	0.98	0.44	0.61	540
ENTJ	1.00	0.09	0.16	185
ENTP	0.97	0.56	0.71	548
ESFJ	0.00	0.00	0.00	33
ESFP	0.00	0.00	0.00	38
ESTJ	0.00	0.00	0.00	31
ESTP	0.00	0.00	0.00	71
INFJ	0.79	0.82	0.80	1176
INFP	0.45	0.99	0.62	1466
INTJ	0.92	0.74	0.82	873
INTP	0.78	0.86	0.82	1043
ISFJ	1.00	0.11	0.19	133
ISFP	1.00	0.06	0.10	217
ISTJ	1.00	0.03	0.06	164
ISTP	1.00	0.19	0.32	270
accuracy			0.66	6940
macro avg	0.68	0.31	0.33	6940
weighted avg	0.78	0.66	0.62	6940

test classification report

	precision	recall	f1-score	support
ENFJ	0.00	0.00	0.00	38
ENFP	0.70	0.19	0.30	135
ENTJ	0.00	0.00	0.00	46
ENTP	0.67	0.16	0.26	137
ESFJ	0.00	0.00	0.00	9
ESFP	0.00	0.00	0.00	10
ESTJ	0.00	0.00	0.00	8
ESTP	0.00	0.00	0.00	18
INFJ	0.59	0.61	0.60	294
INFP	0.36	0.95	0.52	366
INTJ	0.72	0.38	0.49	218
INTP	0.62	0.63	0.62	261
ISFJ	0.00	0.00	0.00	33
ISFP	0.00	0.00	0.00	54
ISTJ	0.00	0.00	0.00	41
ISTP	1.00	0.06	0.11	67
accuracy			0.48	1735
macro avg	0.29	0.19	0.18	1735
weighted avg	0.50	0.48	0.42	1735

```
CV Accuracy: 0.4628 (+/- 0.0113)
CV F1: 0.4628 (+/- 0.0113)
CV Logloss: 1.9037 (+/- 0.0061)
```

5) XGBoost Classifier

train classification report

	precision	recall	f1-score	support
ENFJ	1.00	0.93	0.97	152
ENFP	0.96	0.90	0.93	540
ENTJ	0.99	0.93	0.96	185
ENTP	0.94	0.92	0.93	548
ESFJ	1.00	0.94	0.97	33
ESFP	1.00	0.97	0.99	38
ESTJ	1.00	0.90	0.95	31
ESTP	1.00	0.96	0.98	71
INFJ	0.92	0.90	0.91	1176
INFP	0.89	0.95	0.92	1466
INTJ	0.93	0.93	0.93	873
INTP	0.90	0.92	0.91	1043
ISFJ	1.00	0.94	0.97	133
ISFP	0.99	0.92	0.95	217
ISTJ	0.99	0.93	0.96	164
ISTP	0.97	0.95	0.96	270
accuracy			0.93	6940
macro avg	0.97	0.93	0.95	6940
weighted avg	0.93	0.93	0.93	6940

test classification report

	precision	recall	f1-score	support
ENFJ	0.72	0.34	0.46	38
ENFP	0.71	0.59	0.65	135
ENTJ	0.71	0.37	0.49	46
ENTP	0.62	0.62	0.62	137
ESFJ	1.00	0.11	0.20	9
ESFP	0.50	0.10	0.17	10
ESTJ	1.00	0.38	0.55	8
ESTP	0.56	0.28	0.37	18
INFJ	0.70	0.74	0.72	294
INFP	0.64	0.81	0.72	366
INTJ	0.69	0.67	0.68	218
INTP	0.66	0.76	0.71	261
ISFJ	0.62	0.45	0.53	33
ISFP	0.64	0.33	0.44	54
ISTJ	0.68	0.32	0.43	41
ISTP	0.67	0.64	0.66	67
accuracy			0.67	1735
macro avg	0.70	0.47	0.52	1735
weighted avg	0.67	0.67	0.65	1735

```
CV Accuracy: 0.6405 (+/- 0.0127)
CV F1: 0.6405 (+/- 0.0127)
CV Logloss: 1.1671 (+/- 0.0361)
```

6) Multinomial Naïve Bayes

train classification report

	precision	recall	f1-score	support
ENFJ	0.00	0.00	0.00	152
ENFP	0.91	0.02	0.04	540
ENTJ	0.00	0.00	0.00	185
ENTP	0.92	0.06	0.12	548
ESFJ	0.00	0.00	0.00	33
ESFP	0.00	0.00	0.00	38
ESTJ	0.00	0.00	0.00	31
ESTP	0.00	0.00	0.00	71
INFJ	0.53	0.62	0.57	1176
INFP	0.35	0.95	0.51	1466
INTJ	0.79	0.42	0.55	873
INTP	0.59	0.63	0.61	1043
ISFJ	0.00	0.00	0.00	133
ISFP	0.00	0.00	0.00	217
ISTJ	0.00	0.00	0.00	164
ISTP	1.00	0.00	0.01	270
accuracy			0.46	6940
macro avg	0.32	0.17	0.15	6940
weighted avg	0.53	0.46	0.38	6940

test classification report

	precision	recall	f1-score	support
ENFJ	0.00	0.00	0.00	38
ENFP	0.71	0.04	0.07	135
ENTJ	0.00	0.00	0.00	46
ENTP	0.75	0.02	0.04	137
ESFJ	0.00	0.00	0.00	9
ESFP	0.00	0.00	0.00	10
ESTJ	0.00	0.00	0.00	8
ESTP	0.00	0.00	0.00	18
INFJ	0.39	0.41	0.40	294
INFP	0.31	0.92	0.46	366
INTJ	0.68	0.19	0.29	218
INTP	0.52	0.52	0.52	261
ISFJ	0.00	0.00	0.00	33
ISFP	0.00	0.00	0.00	54
ISTJ	0.00	0.00	0.00	41
ISTP	0.00	0.00	0.00	67
accuracy			0.37	1735
macro avg	0.21	0.13	0.11	1735
weighted avg	0.41	0.37	0.29	1735

```
CV Accuracy: 0.3562 (+/- 0.0101)
CV F1: 0.3562 (+/- 0.0101)
CV Logloss: 2.3181 (+/- 0.0137)
```

7) MLP Classifier

train classification report

	precision	recall	f1-score	support
ENFJ	1.00	1.00	1.00	152
ENFP	1.00	1.00	1.00	540
ENTJ	1.00	1.00	1.00	185
ENTP	1.00	1.00	1.00	548
ESFJ	1.00	1.00	1.00	33
ESFP	1.00	1.00	1.00	38
ESTJ	1.00	1.00	1.00	31
ESTP	1.00	1.00	1.00	71
INFJ	1.00	1.00	1.00	1176
INFP	1.00	1.00	1.00	1466
INTJ	1.00	1.00	1.00	873
INTP	1.00	1.00	1.00	1043
ISFJ	1.00	1.00	1.00	133
ISFP	1.00	1.00	1.00	217
ISTJ	1.00	1.00	1.00	164
ISTP	1.00	1.00	1.00	270
accuracy			1.00	6940
macro avg	1.00	1.00	1.00	6940
weighted avg	1.00	1.00	1.00	6940

test classification report

	precision	recall	f1-score	support
ENFJ	0.74	0.37	0.49	38
ENFP	0.55	0.53	0.54	135
ENTJ	0.42	0.28	0.34	46
ENTP	0.50	0.45	0.47	137
ESFJ	1.00	0.22	0.36	9
ESFP	0.00	0.00	0.00	10
ESTJ	1.00	0.25	0.40	8
ESTP	0.56	0.28	0.37	18
INFJ	0.54	0.61	0.58	294
INFP	0.57	0.68	0.62	366
INTJ	0.48	0.52	0.50	218
INTP	0.56	0.65	0.60	261
ISFJ	0.73	0.33	0.46	33
ISFP	0.62	0.33	0.43	54
ISTJ	0.63	0.29	0.40	41
ISTP	0.65	0.52	0.58	67
accuracy			0.55	1735
macro avg	0.60	0.39	0.45	1735
weighted avg	0.56	0.55	0.54	1735

```
CV Accuracy: 0.5589 (+/- 0.0107)
CV F1: 0.5589 (+/- 0.0107)
CV Logloss: 1.9749 (+/- 0.0536)
```

8) Poly SVC

train classification report

	precision	recall	f1-score	support
ENFJ	1.00	0.99	1.00	152
ENFP	1.00	1.00	1.00	540
ENTJ	1.00	1.00	1.00	185
ENTP	1.00	0.99	1.00	548
ESFJ	1.00	0.97	0.98	33
ESFP	1.00	0.95	0.97	38
ESTJ	1.00	1.00	1.00	31
ESTP	1.00	0.99	0.99	71
INFJ	1.00	0.99	1.00	1176
INFP	0.99	1.00	0.99	1466
INTJ	0.99	1.00	1.00	873
INTP	1.00	1.00	1.00	1043
ISFJ	1.00	1.00	1.00	133
ISFP	1.00	0.99	1.00	217
ISTJ	1.00	1.00	1.00	164
ISTP	1.00	0.99	0.99	270
accuracy			1.00	6940
macro avg	1.00	0.99	0.99	6940
weighted avg	1.00	1.00	1.00	6940

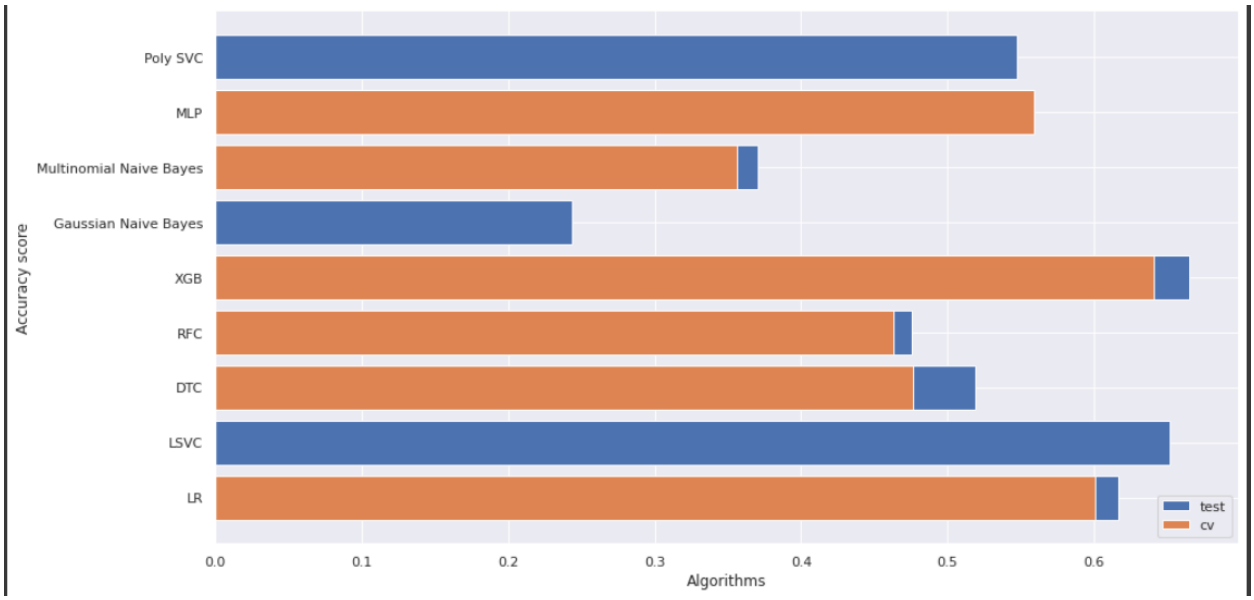
test classification report

	precision	recall	f1-score	support
ENFJ	0.75	0.08	0.14	38
ENFP	0.76	0.36	0.48	135
ENTJ	0.67	0.04	0.08	46
ENTP	0.65	0.30	0.41	137
ESFJ	0.50	0.11	0.18	9
ESFP	0.00	0.00	0.00	10
ESTJ	0.00	0.00	0.00	8
ESTP	0.00	0.00	0.00	18
INFJ	0.66	0.63	0.64	294
INFP	0.44	0.89	0.59	366
INTJ	0.64	0.54	0.59	218
INTP	0.57	0.81	0.67	261
ISFJ	0.75	0.09	0.16	33
ISFP	0.50	0.04	0.07	54
ISTJ	0.33	0.02	0.05	41
ISTP	0.57	0.12	0.20	67
accuracy			0.55	1735
macro avg	0.49	0.25	0.27	1735
weighted avg	0.58	0.55	0.50	1735

Comparison

Test Accuracies of different models:

	Models	Test accuracy
0	XGB	0.665130
1	LSVC	0.651297
2	LR	0.616138
3	MLP	0.551009
4	Poly SVC	0.546974
5	DTC	0.519308
6	RFC	0.475504
7	Multinomial Naive Bayes	0.370605
8	Gaussian Naive Bayes	0.243804



Hyperparameter Tuning

- Tuning Hyperparameters of 2 Models LSVC and XGBoost:

➔ LSVC

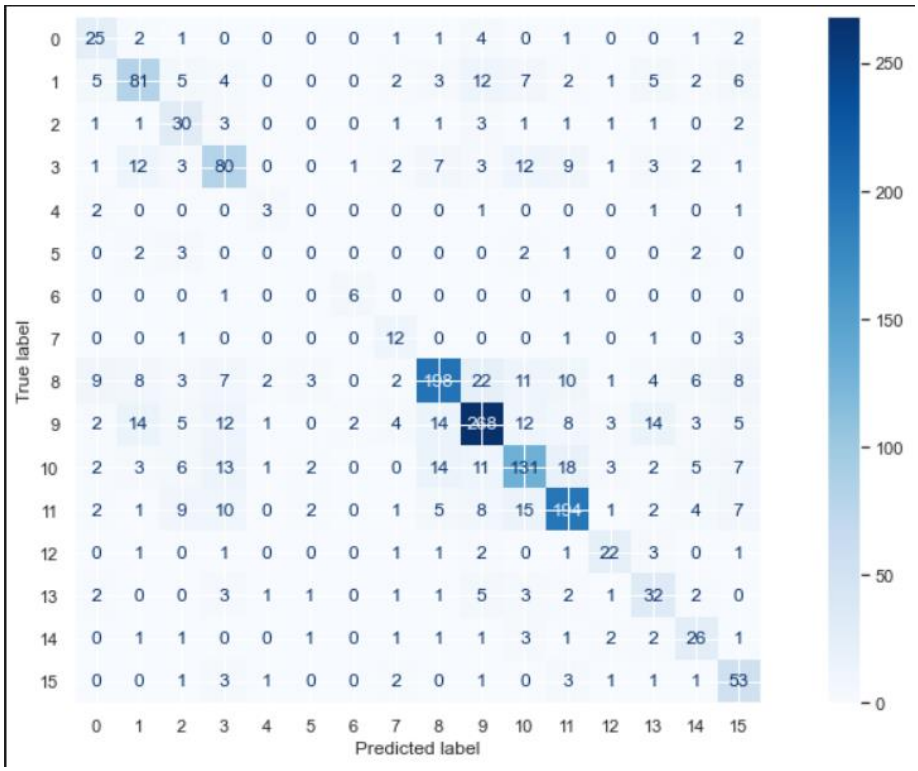
```
Best parameters found:
 {'C': 0.2, 'class_weight': 'balanced', 'max_iter': 1000, 'multi_class': 'ovr', 'penalty': 'l2'}
```

➔ XGBoost

Took too long to run.

Conclusions

- XGBoost Trees
  Image link: https://drive.google.com/file/d/1SdiQ-7pMtY7t2ME3UuKwMzgKE0VVJhfF/view?usp=sharing
- Confusion Matrix for LSVC



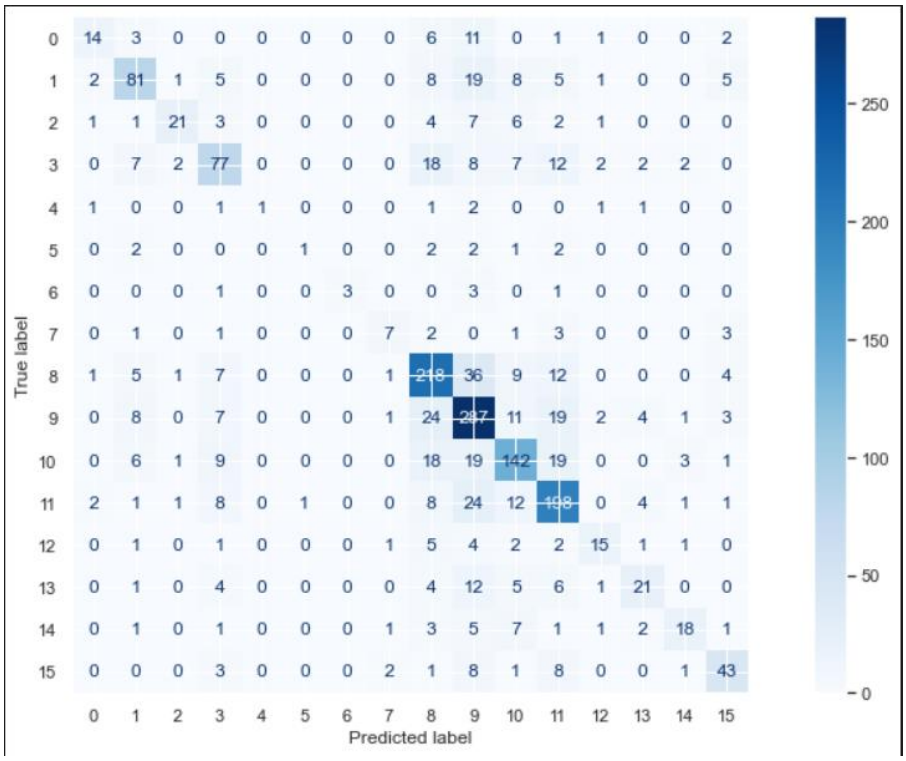
- Classification Report for LSVC

```
test classification report
              precision    recall  f1-score   support

        ENFJ       0.49      0.66      0.56        38
        ENFP       0.64      0.60      0.62       135
        ENTJ       0.44      0.65      0.53        46
        ENTP       0.58      0.58      0.58       137
        ESFJ       0.33      0.38      0.35         8
        ESFP       0.00      0.00      0.00        10
        ESTJ       0.67      0.75      0.71         8
        ESTP       0.40      0.67      0.50        18
        INFJ       0.80      0.67      0.73       294
        INFP       0.79      0.73      0.76       367
        INTJ       0.66      0.60      0.63       218
        INTP       0.77      0.74      0.75       261
        ISFJ       0.59      0.67      0.63        33
        ISFP       0.45      0.59      0.51        54
        ISTJ       0.48      0.63      0.55        41
        ISTP       0.55      0.79      0.65        67

    accuracy                           0.67      1735
   macro avg       0.54      0.61      0.57      1735
weighted avg       0.69      0.67      0.67      1735
```

- Confusion Matrix for XGB



- Classification Report for XGB

```
test classification report
              precision    recall  f1-score   support

        ENFJ       0.67      0.37      0.47        38
        ENFP       0.69      0.60      0.64       135
        ENTJ       0.78      0.46      0.58        46
        ENTP       0.60      0.56      0.58       137
        ESFJ       1.00      0.12      0.22         8
        ESFP       0.50      0.10      0.17        10
        ESTJ       1.00      0.38      0.55         8
        ESTP       0.54      0.39      0.45        18
        INFJ       0.68      0.74      0.71       294
        INFP       0.64      0.78      0.71       367
        INTJ       0.67      0.65      0.66       218
        INTP       0.68      0.76      0.72       261
        ISFJ       0.60      0.45      0.52        33
        ISFP       0.60      0.39      0.47        54
        ISTJ       0.67      0.44      0.53        41
        ISTP       0.68      0.64      0.66        67

    accuracy                           0.66      1735
   macro avg       0.69      0.49      0.54      1735
weighted avg       0.66      0.66      0.65      1735
```

- LSVC and XGBoost both performed equally well.
- Classes were highly disproportional hence the comparatively low accuracy.
- We can observe highest F-1 score for INFP class.
- Overall Model performed very well considering weak dataset and high no. of classes.
-

End-to-End & Pickle

Created a Class Pipeline for end-to-end implementation

- Loaded the final model with pickle

```
Accuracy of the Loaded Model is :  0.669164265129683
```

Extra-Effort

Transformer (BERT)

- BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its vanilla form, Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task. Since BERT's goal is to generate a language model, only the encoder mechanism is necessary.
- As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once. Therefore, it is considered bidirectional, though it would be more accurate to say that it's non-directional. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word).
  - ➔ Pre-processing was same but instead of vectorised we use a transformer to encode the sentences.
  - ➔ Model Summary

```
Model: "model_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_word_ids (InputLayer)  [(None, 1500)]            0
_________________________________________________________________
tf_bert_model (TFBertModel)  ((None, 1500, 1024), (Non 335141888
_________________________________________________________________
tf_op_layer_strided_slice (T [(None, 1024)]            0
_________________________________________________________________
dense_10 (Dense)             (None, 16)                16400
=================================================================
Total params: 335,158,288
Trainable params: 335,158,288
Non-trainable params: 0
_________________________________________________________________
```

  - ➔ Train Accuracy 94%

```
Epoch 11/20
305/305 [==============================] - 228s 749ms/step - loss: 0.2404 - accuracy: 0.9293 - val_loss: 1.6267 - val_accuracy: 0.6577
```

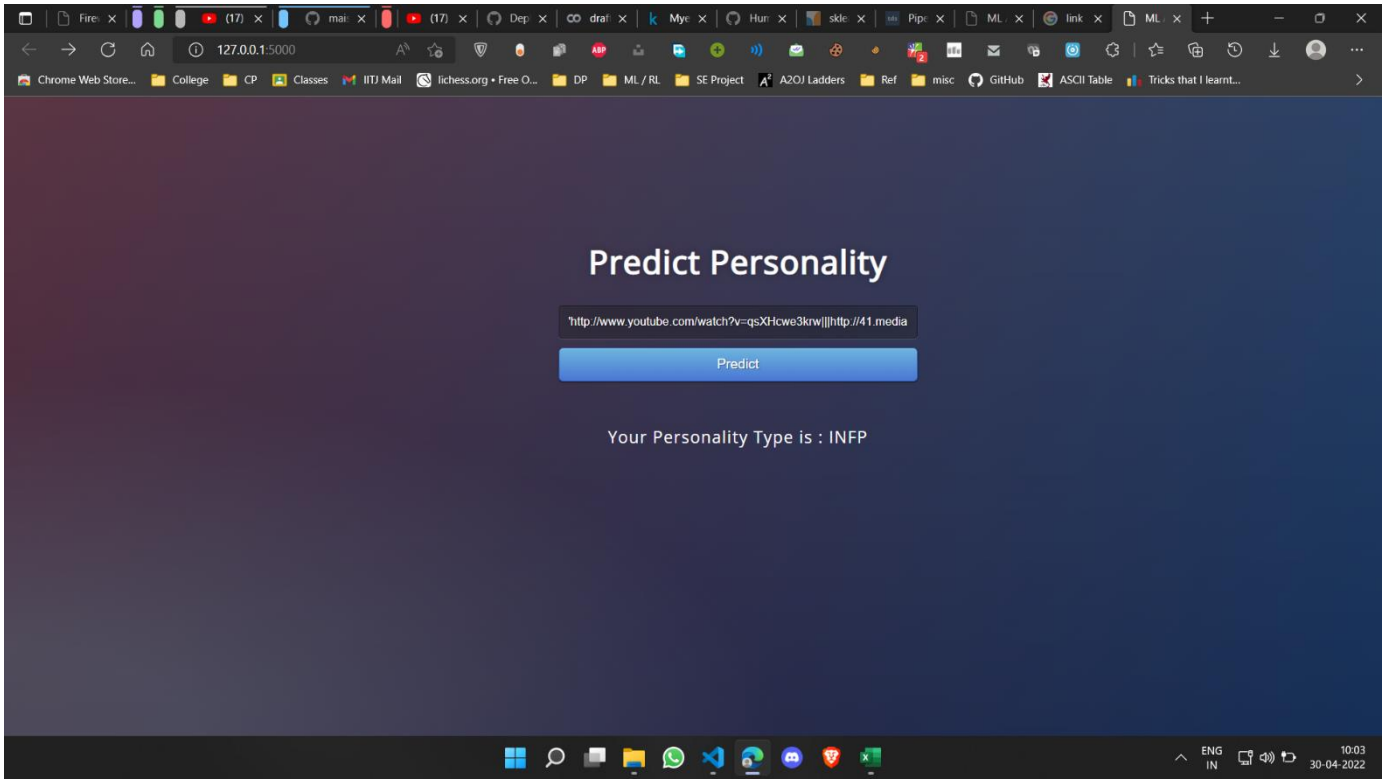
  - ➔ Test Accuracy 68%

```
[1.5094053745269775, 0.6809589862823486]
```

Website

- ➔ Made a website with Flask Framework to predict the personality by getting raw text input from the user.
- ➔ Will host it on Heroku in future.

Screenshot of the site



GitHub Link

Link: samkitshah1262/Personality-Predictor: NLP Project (github.com)