SCHOOL OF MATHEMATIS & COMPUTER SCIENCE

Part 1: Data Analysis and Bayes Nets

Prepared by:
Akshay Garg, Irfan Syed, Naveen Jain,
Nemr Aslam, Rithin Thomas

Group Number :
13

Prepared on:
20th October 2023

1. DATA ANALYSIS & VISUALIZATION

- **X-train and X-test:** Contains images of road signs in Germany.
- **Y-train and Y-test:** Labels for the provided data

- Shapes of the provided datasets:
  - **X-train:** 9690 rows and 2304 columns
  - **X-test:** 3090 rows and 2304 columns
  - **Y-train:** 9690 rows and 1 column
  - **Y-test:** 3090 rows and 1 column

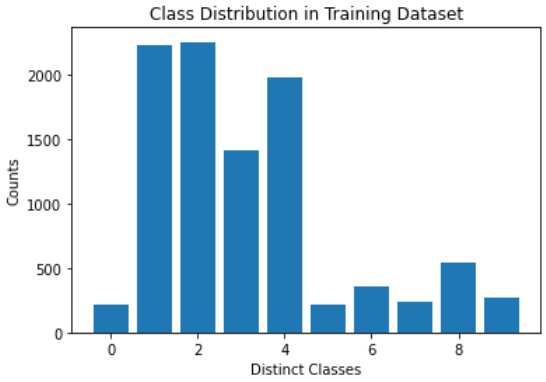
- All the features in the datasets are of numeric data types (int64 and float64), indicating that no additional data type conversions are required.

- There are no null values in any of the datasets, making them clean and ready for analysis.

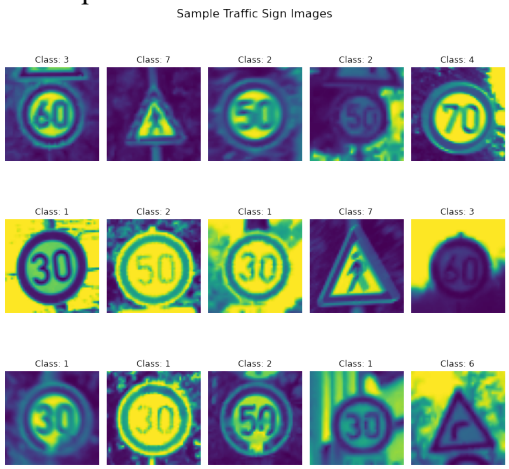
- The dataset was scaled to meet the 48x48 requirement of the images.

- The minimum/maximum values of the records ranges from 3.0 – 255.0 (X-train), indicating that there are no error outliers such as negative values or extremes.

- Plotted the counts of the number of records in each distinct class in the training set to understand the nature of the images in the dataset



- 15 of the sample images were plotted to ensure there was no error in the attributes

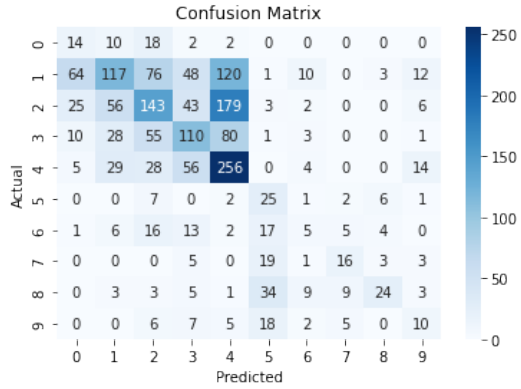


2. NAÏVE BAYES CLASSIFIER METRICS

- We split the data into a training set and a test set using a train-test split.
  Post-Split:
    - X-train: (7752, 2304)
    - X-test: (1938, 2304)
    - Y-train: (7752, 1)
    - Y-test: (1938, 1)

Upon running the **Multinomial Naïve Bayes Classifier**, we observed the following metrics:

- **Accuracy Score**: 0.3715170278637771
    - We observe an accuracy of **37%**, which is relatively low, which means the model is struggling to accurately classify a large majority of the instances in the dataset.
- **Confusion matrix**



- **True Positives, False Positives, False Negatives**

	FP	FN	TP
0	105	32	14
1	132	334	117
2	209	314	143
3	179	178	110
4	391	136	256
5	93	19	25
6	32	64	5
7	21	31	16
8	16	67	24
9	40	43	10

- **Precision, recall and f-measure:**

	Precision	Recall	F1 Score
0	0.117647	0.304348	0.169697
1	0.469880	0.259424	0.334286
2	0.406250	0.312910	0.353523
3	0.380623	0.381944	0.381282
4	0.395672	0.653061	0.492782
5	0.211864	0.568182	0.308642
6	0.135135	0.072464	0.094340
7	0.432432	0.340426	0.380952
8	0.600000	0.263736	0.366412
9	0.200000	0.188679	0.194175

- **Classification report.**

	precision	recall	f1-score	support
0	0.12	0.30	0.17	46
1	0.47	0.26	0.33	451
2	0.41	0.31	0.35	457
3	0.38	0.38	0.38	288
4	0.40	0.65	0.49	392
5	0.21	0.57	0.31	44
6	0.14	0.07	0.09	69
7	0.43	0.34	0.38	47
8	0.60	0.26	0.37	91
9	0.20	0.19	0.19	53
accuracy			0.37	1938
macro avg	0.33	0.33	0.31	1938
weighted avg	0.40	0.37	0.36	1938

- Class 3 and 4 have relatively high F1 scores, which shows they have better precision and recall.
- Class 0 and 6 have low F1 scores, meaning that the classifier struggles to correctly predict these classes.
- The classification report and confusion matrix show a very varied performance across the classes, indicating class imbalance.

3. FEATURE ANALYSIS AND SELECTION FOR IMPROVED CLASSIFICATION

- We calculated the correlation matrix for each class:

```
0       1.000000
2261    0.190060
2262    0.190014
2263    0.187942
2213    0.186962
2260    0.186465
2212    0.184727
2214    0.182960
2264    0.182396
2215    0.181238
Name: 0, dtype: float64
```

```
1       1.000000
1073    0.377106
1121    0.365046
1120    0.364920
1074    0.360629
1072    0.350379
1168    0.350274
1025    0.332318
1167    0.331780
1026    0.329635
Name: 1, dtype: float64
```

```
2       1.000000
1316    0.302777
1030    0.302559
1       0.299793
1317    0.299426
1268    0.293152
1269    0.289142
982     0.274233
1364    0.273505
1315    0.268385
Name: 2, dtype: float64
```

- Then for each of the 10 classes we find the 5, 10, 20 features that best correlate with classes.

50 unique features (5x10)

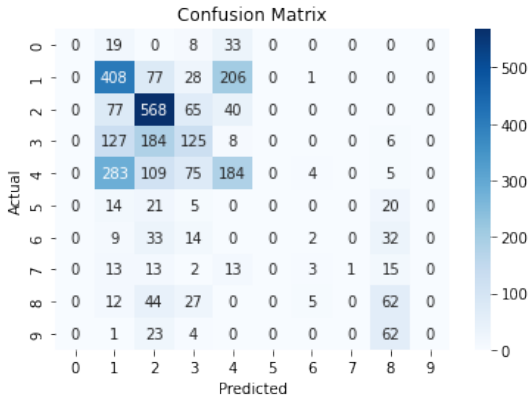
- **Accuracy Score**: 0.4368932038834951
  - The accuracy has improved to ~43.7%, which is still relatively low, indicating that the model, with these new correlated features is still having trouble predicting the classes.

- **Precision, recall, f-measure**

	precision	recall	f1-score	support
0	0.00	0.00	0.00	60
1	0.42	0.57	0.48	720
2	0.53	0.76	0.62	750
3	0.35	0.28	0.31	450
4	0.38	0.28	0.32	660
5	0.00	0.00	0.00	60
6	0.13	0.02	0.04	90
7	1.00	0.02	0.03	60
8	0.31	0.41	0.35	150
9	0.00	0.00	0.00	90
accuracy			0.44	3090
macro avg	0.31	0.23	0.22	3090
weighted avg	0.40	0.44	0.40	3090

  - Class 1 has the best **precision** of **42%**, while classes 0, 5 and 9 have **0%** precision, meaning that the model correctly predicts a lot of some classes, and null or near null for some classes.
  - Class 2 has the best **recall** of **76%**, while classes 0, 5, and 9 have **0%** recall, meaning that the model is effective at identifying most positive instances of some classes, while it struggles or is unable to identify positive instances for some.
  - Overall, some classes show balanced performance, while others pose significant challenges for the model.

- **Confusion Matrix:**

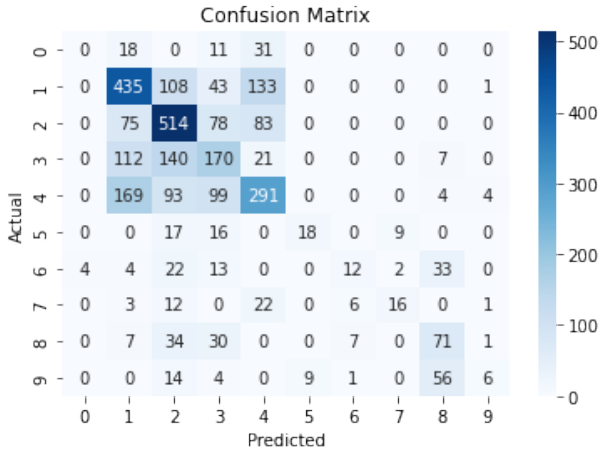


100 unique features (10x10)

- **Accuracy Score:** 0.4961165048543689
  - The accuracy has improved to **49.6%** . While this improvement isn't too significant, it helps us understand that adding features has had a positive impact on the model's performance.
- **Precision, recall, f-measure**

	precision	recall	f1-score	support
0	0.00	0.00	0.00	60
1	0.53	0.60	0.56	720
2	0.54	0.69	0.60	750
3	0.37	0.38	0.37	450
4	0.50	0.44	0.47	660
5	0.67	0.30	0.41	60
6	0.46	0.13	0.21	90
7	0.59	0.27	0.37	60
8	0.42	0.47	0.44	150
9	0.46	0.07	0.12	90
accuracy			0.50	3090
macro avg	0.45	0.33	0.36	3090
weighted avg	0.49	0.50	0.48	3090

  - The precision and recall still vary greatly and the weighted average f-measure remains 0.48, indicating that adding 50 features hasn't had great impact on overall model performance and we can still see class imbalance.
- **Confusion matrix**



- There is now an increase in true positives for some classes compared to 50-feature dataset (e.g., class 4), which means classification has seen some improvements.
- On the other hand, there is an increase in false positives compared to the 50-feature dataset (e.g., class 9)

200 unique features (20x10)

- **Accuracy Score:** 0.5074433656957928
  - The accuracy is now **50.7%,** showing a significant improvement compared to the 100-feature dataset.

- **Precision, recall and f-measure:**

	precision	recall	f1-score	support
0	0.03	0.02	0.02	60
1	0.56	0.55	0.55	720
2	0.56	0.58	0.57	750
3	0.40	0.46	0.43	450
4	0.47	0.50	0.48	660
5	0.81	0.42	0.55	60
6	0.56	0.33	0.42	90
7	0.89	0.40	0.55	60
8	0.51	0.52	0.52	150
9	0.58	0.42	0.49	90
accuracy			0.51	3090
macro avg	0.53	0.42	0.46	3090
weighted avg	0.51	0.51	0.51	3090

  - While precision and recall values still vary greatly across classes, there is an improvement in some classes compared to previous datasets (e.g., class 1 and 2).
  - While the model still struggles to predict some classes, there is some improvements in the values. The weighted f-measure has increased to **0.51**, showing that **addition of features has improved the model's performance.**

- **Confusion matrix**