Group

- Dubai CW PG Thursday Group 6

Group Members

- Mohamed Aman
- Faizan Watare
- Fardeen Khan
- Pratibha Yadubanshi

F21DL Coursework Part 1 – Data Analysis and Bayes Nets

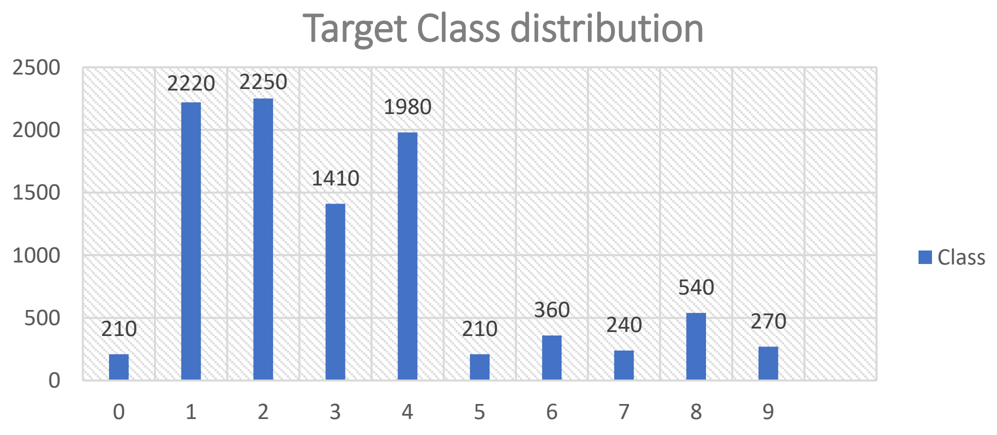
In this report, we present the results and findings from Part 1 of our assignment, focusing on data analysis and the use of Naïve Bayes classifiers on the dataset. This part of the assignment allowed us to conduct experiments based on the lectures from week 1 – 5.

Exploratory Data Analysis (EDA) and Data Visualization

During the data analysis and visualization phase, we found the following insights in our dataset:

- The dataset consists of 9690 images, each represented by 2304 features (48x48 pixels), emphasizing the size and dimensionality of the dataset.
- The dataset doesn't have any null values.
- The dataset consists of 10 distinct classes. However, this is an imbalanced dataset, with class distributions showing significant disproportions between classes. This was found as a result of data visualization by matplotlib and seaborn libraries.

We visualized a subset of images to understand the quality and characteristics of the data, then used MinMax Scaler to scale the pixel values of the dataset between 0 and 1. This ensures that all features (pixels) are on a common scale.



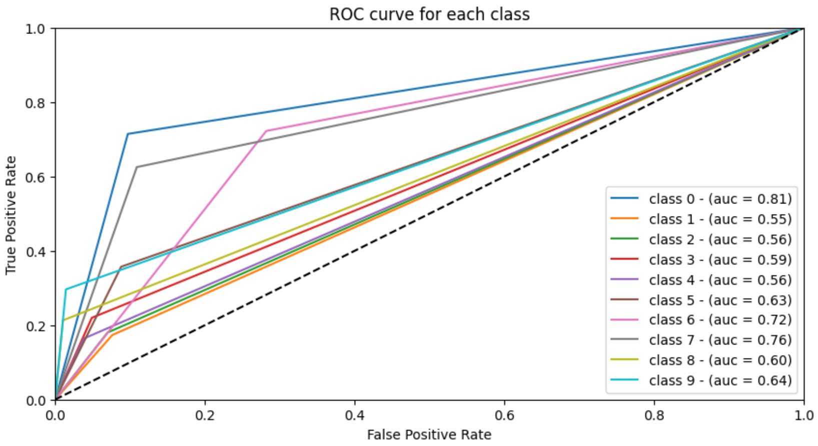
Naïve Bayes Classification and Performance Metrices

To attain the best model performance, we explored various data split strategies. Initially, we did a regular train-test split with different test sizes (0.2 and 0.3) to assess the Gaussian Naïve Bayes classifier's performance. We chose to run the GaussianNB as it is used for continuous data. Then, we transitioned to stratified train-test splitting, once again testing with test sizes (0.2 and 0.3). The following table presents the model's performance metrics across these different approaches:

	Regular Split		**Stratified Split**	
	Test size = 0.2	**Test size = 0.3**	**Test size = 0.2**	**Test size = 0.3**
Accuracy	23.79%	23.98%	23.27%	22.5%
Precision (weighted Average)	40%	39.95%	41.64%	40.5%
Recall (weighted Average)	23.79%	23.98%	23.27%	22.5%
F1-Score (weighted Average)	25.78%	26.21%	25.38%	24.75%

Our comparison revealed that the stratified train-test split with a test size of 0.2 offered the best balance between accuracy, precision, recall, and F1-score, with a notable performance improvement in precision compared to the regular split.

The ROC curve analysis and auc score showed more insights on our dataset. Class 0 – (auc = 0.81), Class 7 – (auc = 0.76) and Class 6 – (auc = 0.72) has the highest auc scores in our dataset.



We conducted StratifiedKFold cross-validation with five splits to evaluate the performance of the Naïve Bayes classifier. The key performance metrics, including accuracy, precision, recall, and F1 score were calculated for each fold. The results for each metric across the five folds are as follows:

Accuracy	21.67%	24.66%	22.39%	21.41%	24.30%
Precision - (weighted Average)	36.14%	40.17%	39.55%	38.44%	41.19%
Recall - (weighted Average)	21.67%	24.66%	22.39%	21.41%	24.30%
F1-Score - (weighted Average)	23.37%	26.84%	25.19%	23.29%	26.44%

Correlation and Naïve bayes classification on reduced dataset

To analyse the most correlating features for each class, the dataset was copied ten times, once for each class 0 to 9. For each class-specific dataset, we calculated the correlation between features and the target class. We utilized One Vs Rest training dataset provided for this task. Then we identified the top 5, 10, and 20 features with the highest correlation values for each and created 3 datasets with reduced features.

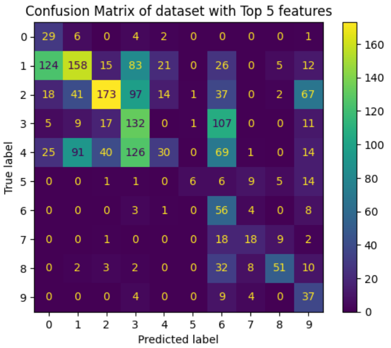
- Data Set 1 (Top 5 features): 45 unique features.
- Data Set 2 (Top 10 features): 82 unique features.
- Data Set 3 (Top 20 features): 148 unique features.

Next, we split the data using stratified train-test split with 80% for training and reserving 20% for testing set. Finally, we ran the naïve bayes classifier to get the major metrices. Following table shows the metrics results.

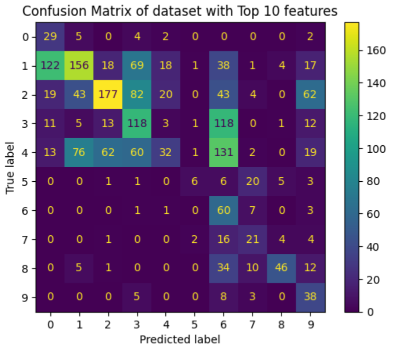
	Dataset_1 (top 5 features)	**Dataset_2 (top 10 features)**	**Dataset_3 (top 20 features)**
Accuracy	35.60%	35.24%	33.18%
Precision - (weighted Average)	49.19%	48.68%	46.72%
Recall - (weighted Average)	35.60%	35.24%	33.18%
F1-Score - (weighted Average)	36.02%	36.16%	34.09%

The reduced datasets, when subjected to Naïve Bayes classification, gave higher scores compared to the original dataset comprising all features. This affirms that the inclusion of additional, less relevant features (noise) tends to diminish accuracy. Also, it is evident that a selection of only the most highly correlated columns is adequate for achieving higher classification scores.

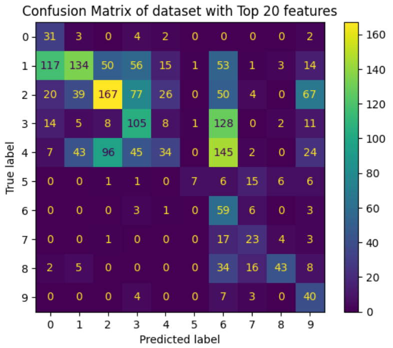
Confusion matrix for 3 datasets with top5, top 10 and top 20 features are shown below.



	0	1	2	3	4	5	6	7	8	9
0	29	6	0	4	2	0	0	0	0	1
1	124	158	15	83	21	0	26	0	5	12
2	18	41	173	97	14	1	37	0	2	67
3	5	9	17	132	0	1	107	0	0	11
4	25	91	40	126	30	0	69	1	0	14
5	0	0	1	1	0	6	6	9	5	14
6	0	0	0	3	1	0	56	4	0	8
7	0	0	1	0	0	0	18	18	9	2
8	0	2	3	2	0	0	32	8	51	10
9	0	0	0	4	0	0	9	4	0	37

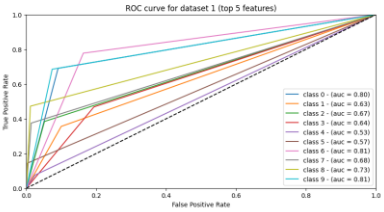


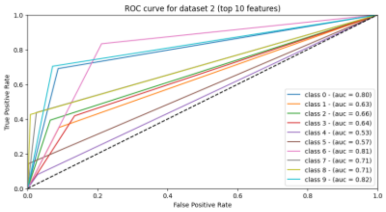
	0	1	2	3	4	5	6	7	8	9
0	29	5	0	4	2	0	0	0	0	2
1	122	156	18	69	18	1	38	1	4	17
2	19	43	177	82	20	0	43	4	0	62
3	11	5	13	118	3	1	118	0	1	12
4	13	76	62	60	32	1	131	2	0	19
5	0	0	1	1	0	6	6	20	5	3
6	0	0	0	1	1	0	60	7	0	3
7	0	0	1	0	0	2	16	21	4	4
8	0	5	1	0	0	0	34	10	46	12
9	0	0	0	5	0	0	8	3	0	38

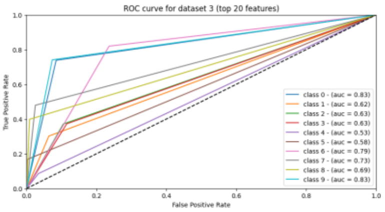


	0	1	2	3	4	5	6	7	8	9
0	31	3	0	4	2	0	0	0	0	2
1	117	134	50	56	15	1	53	1	3	14
2	20	39	167	77	26	0	50	4	0	67
3	14	5	8	105	8	1	128	0	2	11
4	7	43	96	45	34	0	145	2	0	24
5	0	0	1	1	0	7	6	15	6	6
6	0	0	0	3	1	0	59	6	0	3
7	0	0	1	0	0	0	17	23	4	3
8	2	5	0	0	0	0	34	16	43	8
9	0	0	0	4	0	0	7	3	0	40

ROC Curve and AUC score for 3 datasets with top5, top 10 and top 20 features are shown below.







Additional Analysis

For extra research, we did the dimensionality reduction using PCA retaining 95% data integrity. It became evident that only 71 features required to have 95% data integrity and we were able to get a better score than correlation as well. Following scores were obtained using PCA.

- Accuracy – 41.54%
- Precision – 54.68%
- Recall – 41.54%
- F1 Score – 43.81%

Introducing PCA with a variance retained at 95% led to an improvement in accuracy and precision compared to running naïve bayes on Data Sets 1, 2, and 3. This suggests that the dimensionality reduction applied by PCA allowed for more efficient representation of the data without significantly sacrificing classification performance.

Alternatively, we tried running the dataset 2 (top 10 correlating features as it almost matches with the number of columns selected by PCA) through other classifiers which support multi class like Logistic Regression, Decision tree and Random forest. The accuracy of these classifiers with this kind of image dataset was much better than naïve bayes. Highlighting scores for some of the classifiers.

	Logistic Regression	**Decision Tree**	**Random Forest**
Accuracy	86.64%	81.42%	95.87%
Precision - (weighted Average)	86.95%	81.44%	95.92%
Recall - (weighted Average)	86.64%	81.42%	95.87%
F1-Score - (weighted Average)	86.62%	81.39%	95.87%

Conclusion

Our analysis unveiled important aspects of the dataset and the challenges of classification with different types of data(image) and class imbalance. It highlighted the significance of feature selection and the impact it can have on classification performance. Also, it highlighted that effective preprocessing, feature selection, and the choice of classification algorithm can significantly influence the quality of results.

➢ Github Link: https://github.com/amannuhman/F21DL-CW

Thank You