Spotify dataset ML model

PREDICTING SONG'S POPULARITY

Team Members

- Ahmed Adel
- Ahmed Ayman
- Abdalla Elmougi
- Mazen Abdallah
- Mohamed Al-Sha3rawy
- Omar Mohamed Ahmed

- **Project headlines**

  - Project Overview
  - Dataset Overview
  - Preprocessing
  - ML models
  - DL model

Project Overview

- **Objectives:** Predict song popularity on Spotify using audio features
- **Problem Type**: Multi-class classification (Low, Normal, Medium, High, Very High) popularity
- **Dataset**: Spotify dataset with 114,000 songs and 21 features

Dataset Overview

- **Source**: Spotify Dataset for ML Practice from Kaggle
- **Size**: 114,000 songs, 21 features
- **Key Features**:
  - Audio: danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo
  - Metadata: artist frequency, track name (target-encoded), genre (one-hot encoded)
- **Target**: Popularity binned into 5 categories (Low: 0-20, Normal: 20-40, Medium-Low: 40-60, Medium-High: 60-80, High: 80-100)

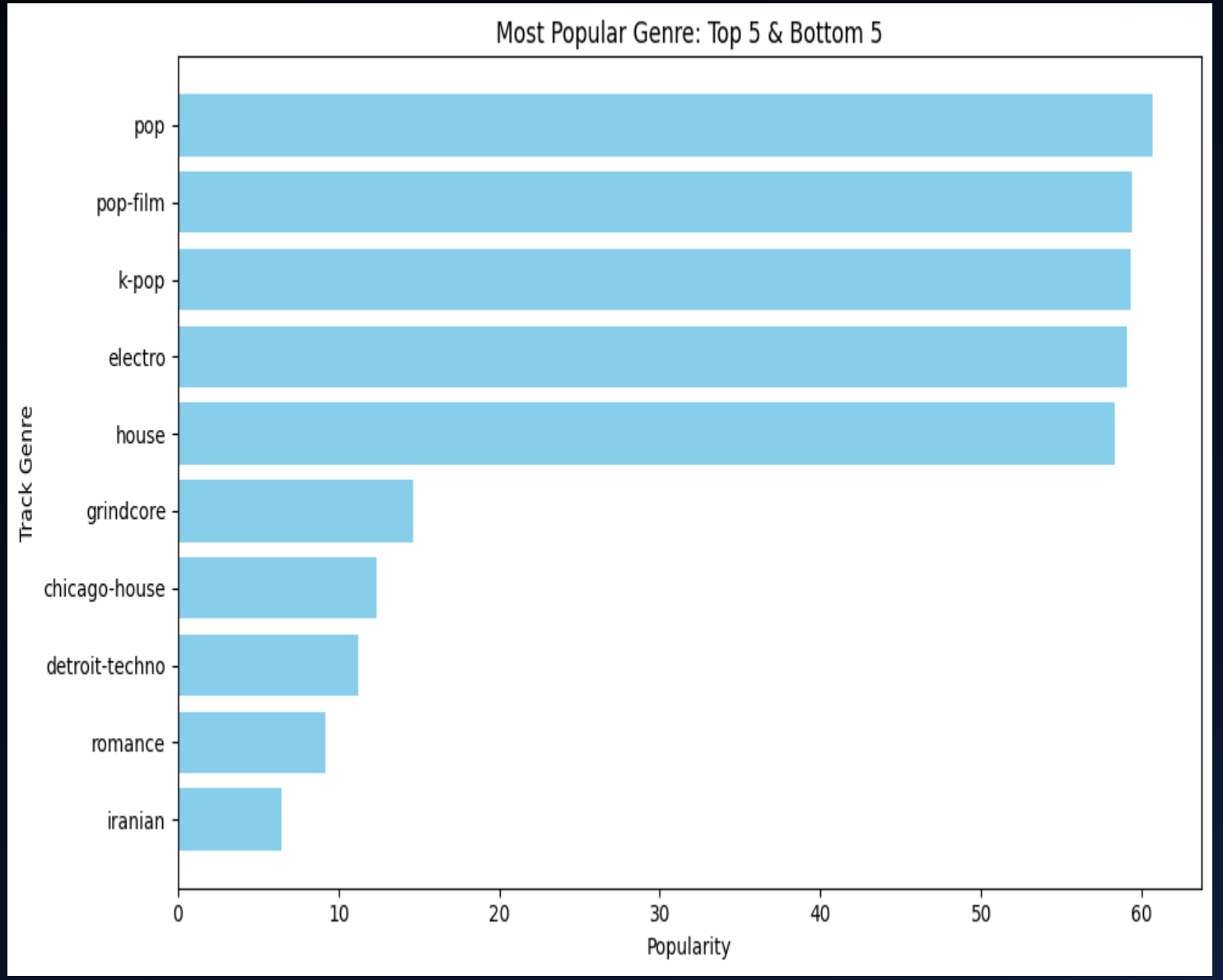
Features

1	track_id
2	artists
3	album_name
4	track_name
5	popularity
6	duration_ms
7	explicit
8	danceability
9	energy
10	key

11	loudness
12	mode
13	speechiness
14	acousticness
15	instrumentalness
16	liveness
17	valence
18	tempo
19	time_signature
20	track_genre

Relation between Genres and Popularity

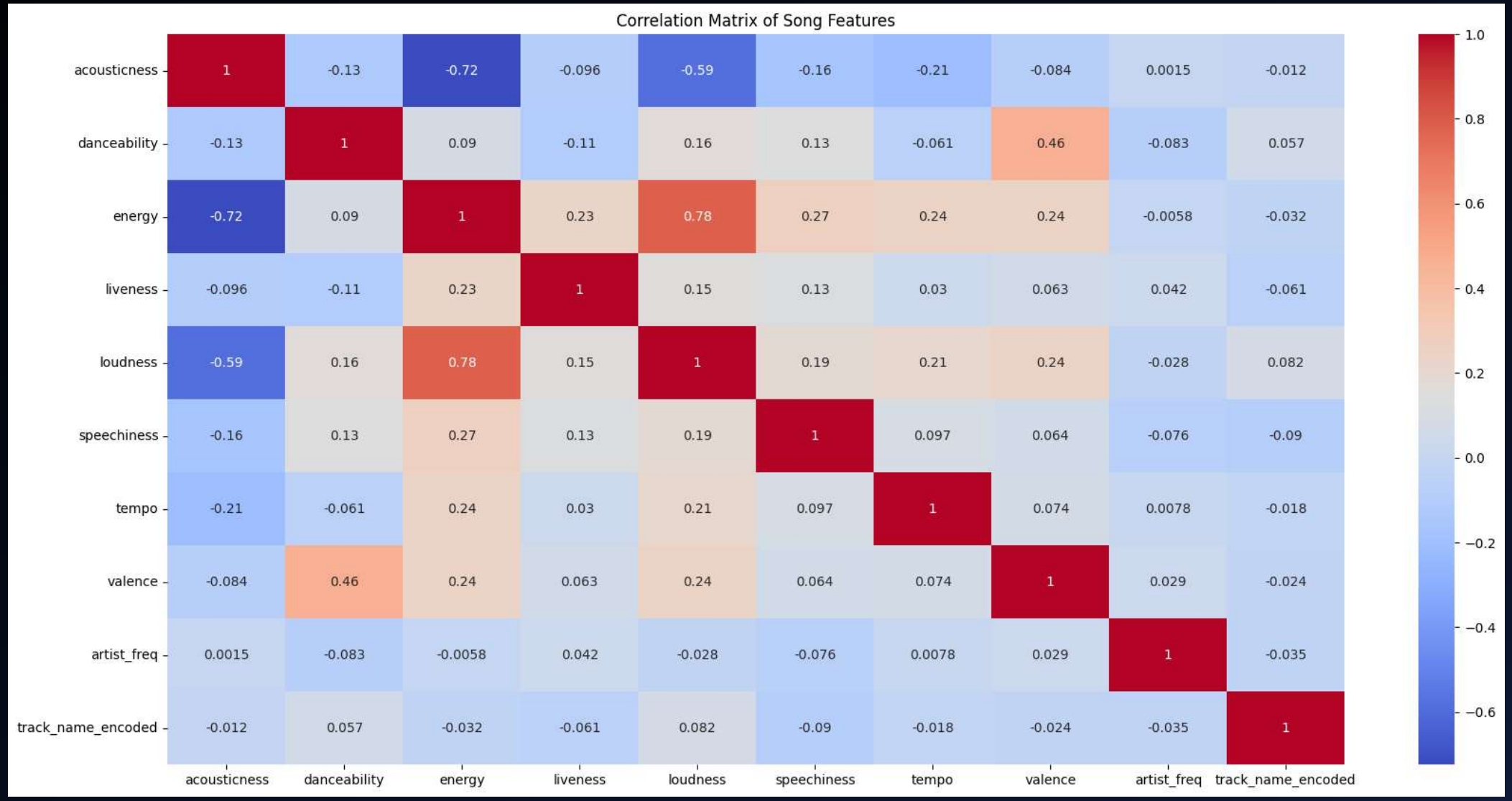
track_genre	popularity
pop	60.678161
pop-film	59.399198
k-pop	59.279458
electro	59.110016
house	58.364103
...	...
grindcore	14.673367
chicago-house	12.392137
detroit-techno	11.213280
romance	9.194842
iranian	6.464706



Preprocessing

- **Cleaning**: Dropped missing values, duplicates, and irrelevant columns
- **Outlier Handling**: Capped numerical features using IQR method
- **Scaling**: Standardized numerical features (e.g., danceability, tempo)
- **Class Imbalance**: Applied SMOTE to balance popularity classes
- **Feature Engineering**:
  - Converted duration_ms to min
  - Binned popularity into 5 categories
  - Target-encoded track_name, frequency-encoded artists
  - Binarized instrumentalness and liveness using mean thresholds

Correlation Matrix of Song Features



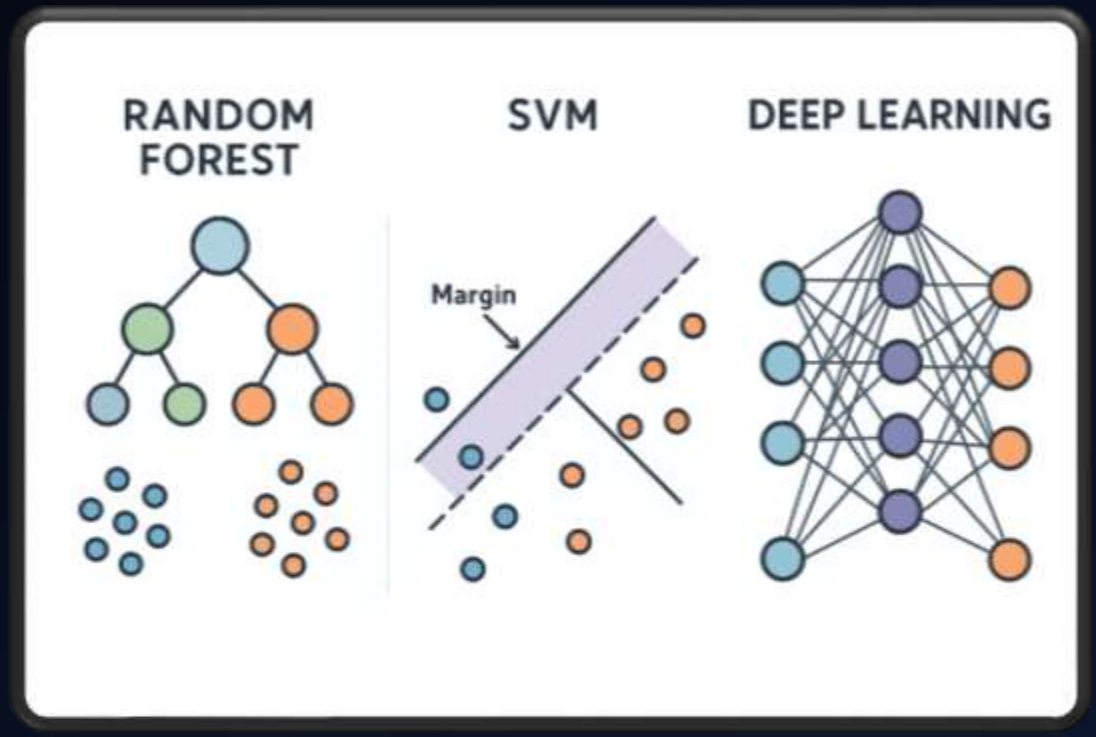
Before / After SMOTE

```
Popularity bin distribution:
popularity
Low          17474
Normal       33067
Medium       33061
High         12595
Very High      952
Name: count, dtype: int64
```

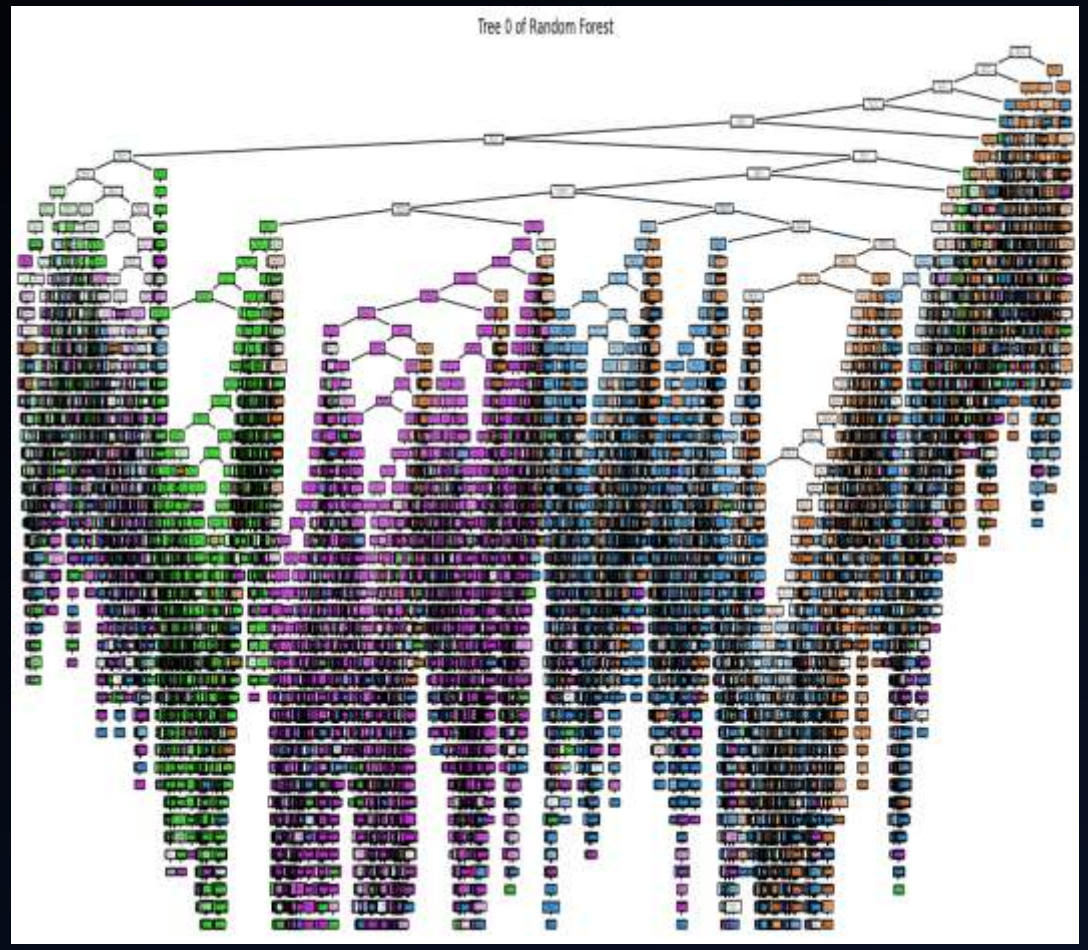
```
Class distribution after SMOTE:
popularity
Low          33067
Normal       33067
Medium       33067
High         33067
Very High    33067
Name: count, dtype: int64
```

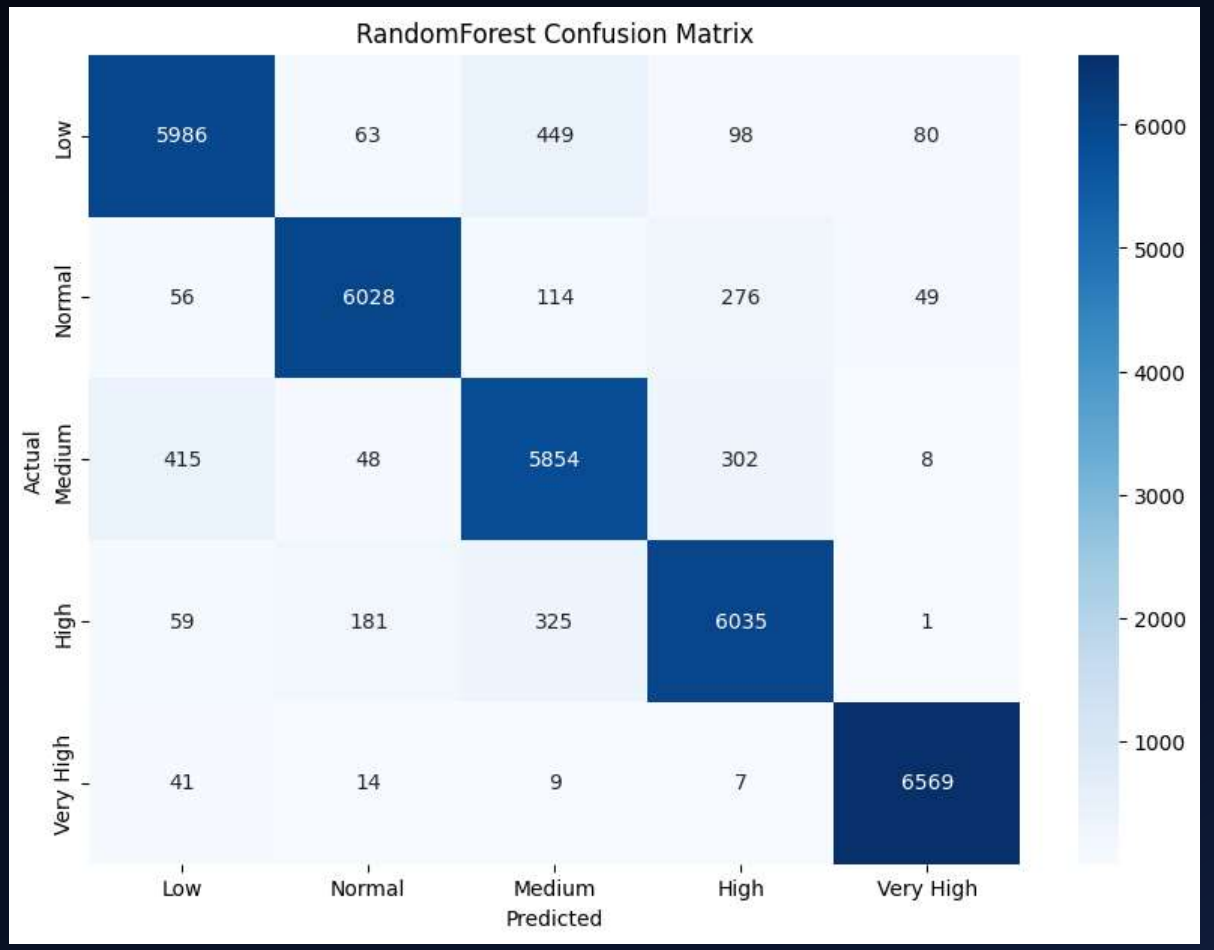
Models

- Random Forest
- Support Vector Machine
- Logistic Reg
- Deep Learning



Random Forest

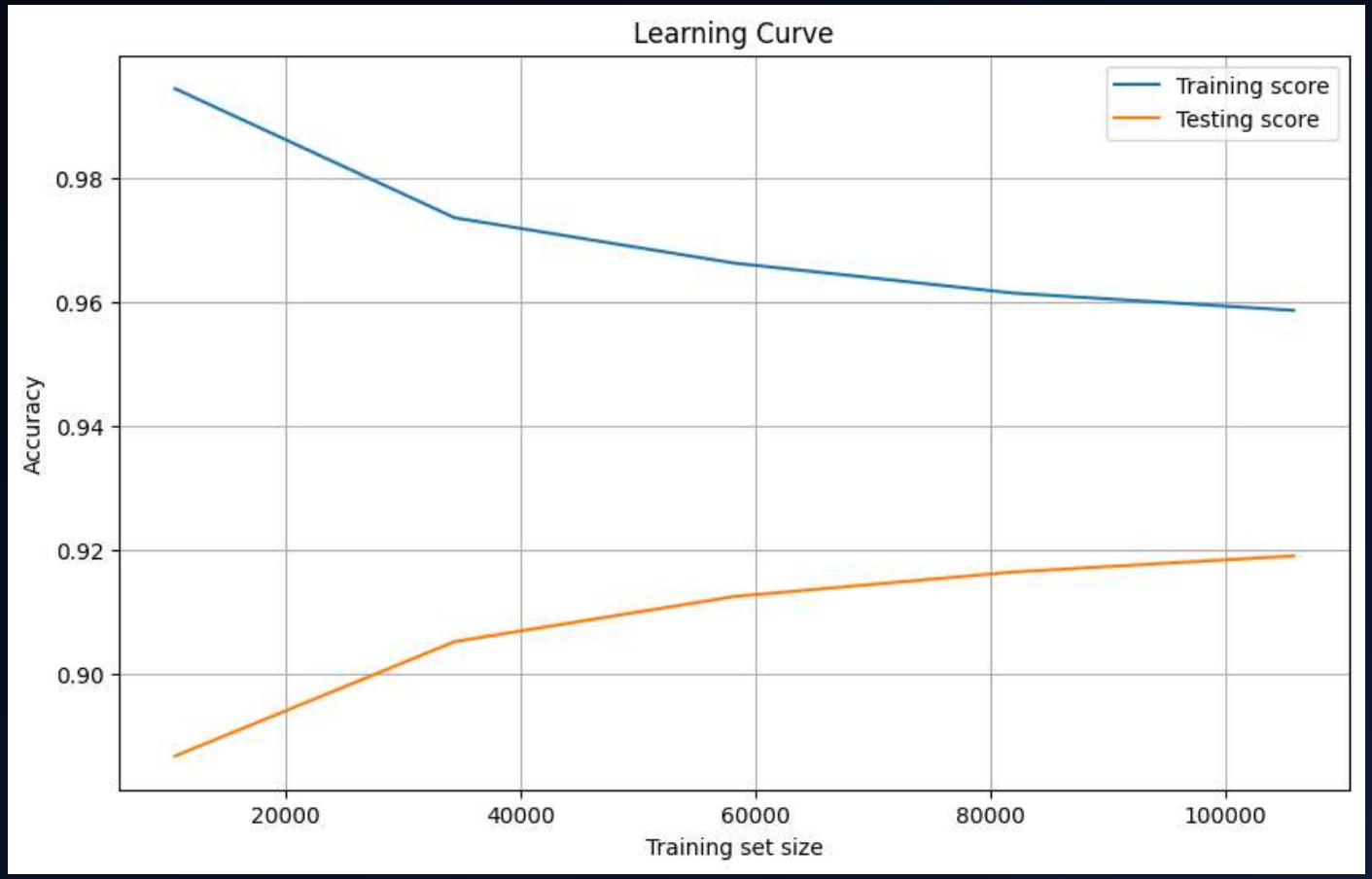




- Accuracy score

```
=== RandomForestClassifier ===

RandomForest Training Classification Report:
              precision    recall  f1-score   support

        High       0.96      0.95      0.95     26391
         Low       0.98      0.96      0.97     26544
      Medium       0.91      0.93      0.92     26440
      Normal       0.93      0.94      0.94     26466
   Very High       1.00      1.00      1.00     26427

    accuracy                           0.96    132268
   macro avg       0.96      0.96      0.96    132268
weighted avg       0.96      0.96      0.96    132268


RandomForest Testing Classification Report:
              precision    recall  f1-score   support

        High       0.91      0.90      0.90      6676
         Low       0.95      0.92      0.94      6523
      Medium       0.87      0.88      0.88      6627
      Normal       0.90      0.91      0.91      6601
   Very High       0.98      0.99      0.98      6640

    accuracy                           0.92     33067
   macro avg       0.92      0.92      0.92     33067
weighted avg       0.92      0.92      0.92     33067
```

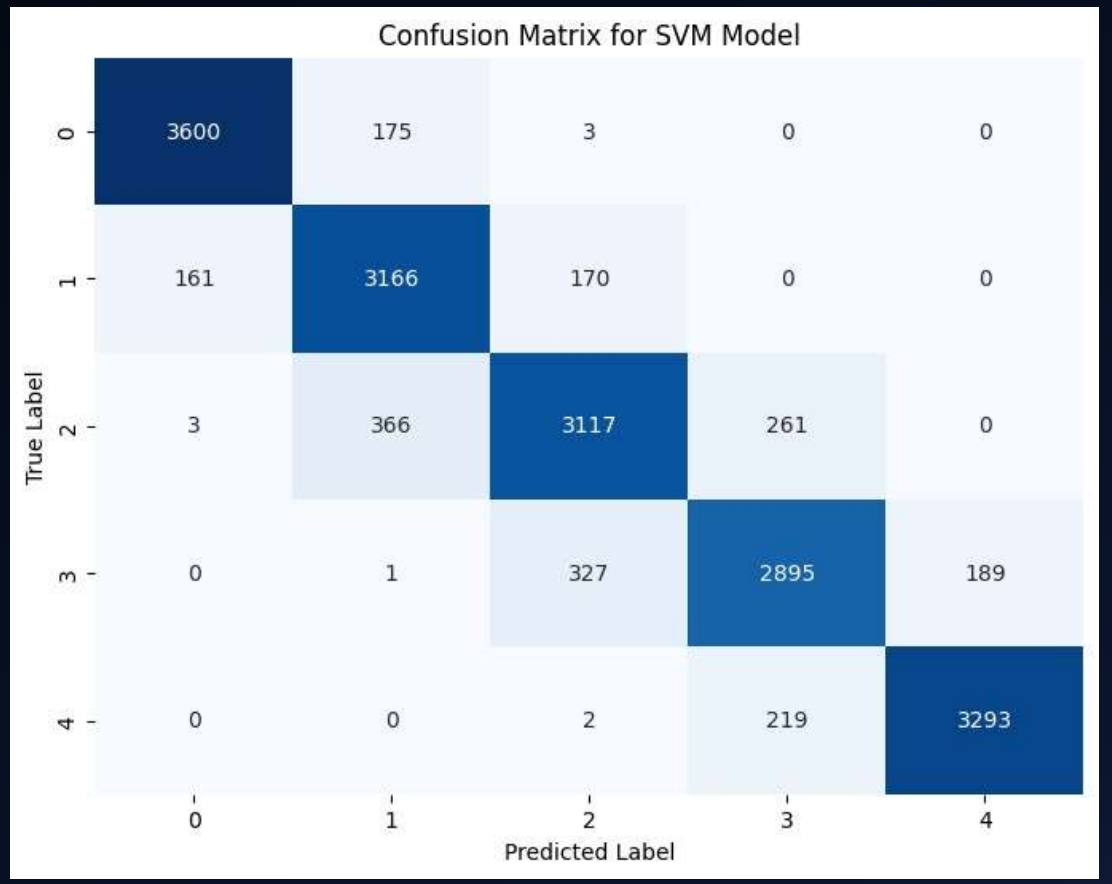
- Learning Curve



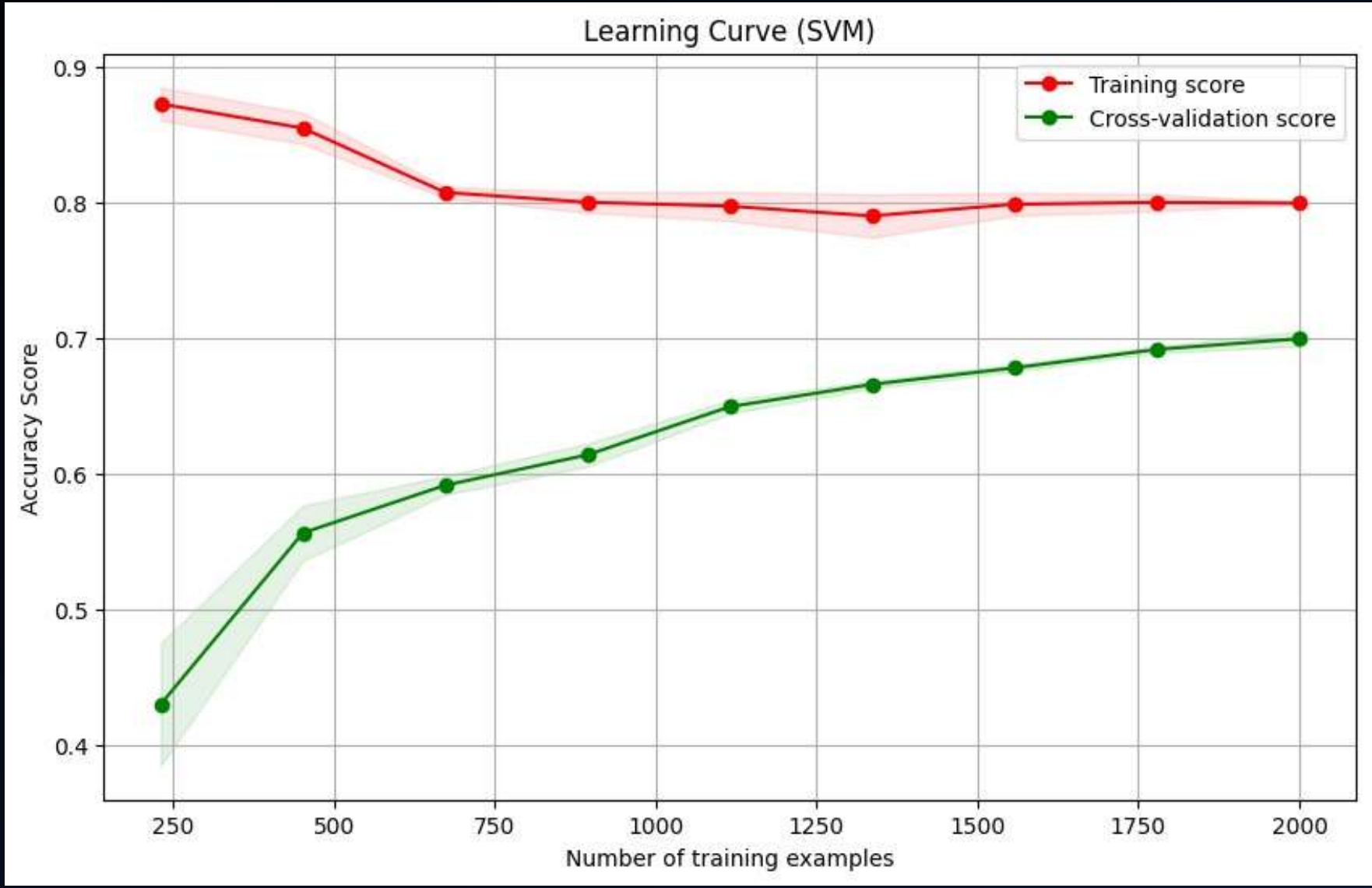
SVM

SVM Training Accuracy: 0.9138
SVM Test Accuracy: 0.8954

	precision	recall	f1-score	support
0	0.96	0.95	0.95	3778
1	0.85	0.91	0.88	3497
2	0.86	0.83	0.85	3747
3	0.86	0.85	0.85	3412
4	0.95	0.94	0.94	3514
accuracy			0.90	17948
macro avg	0.90	0.90	0.89	17948
weighted avg	0.90	0.90	0.90	17948



Learning curve

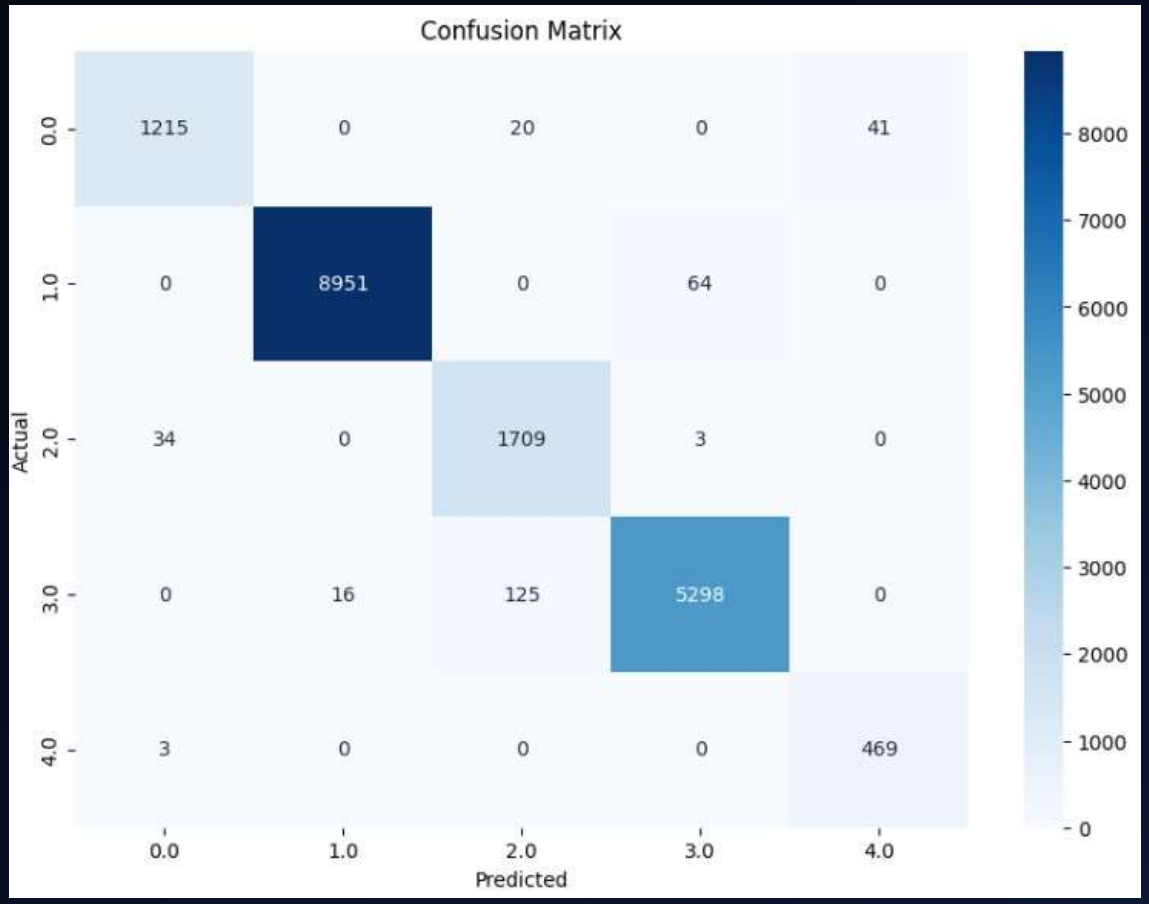


Logistic Reg

```
Train Accuracy: 0.9856
Test Accuracy: 0.9830
```

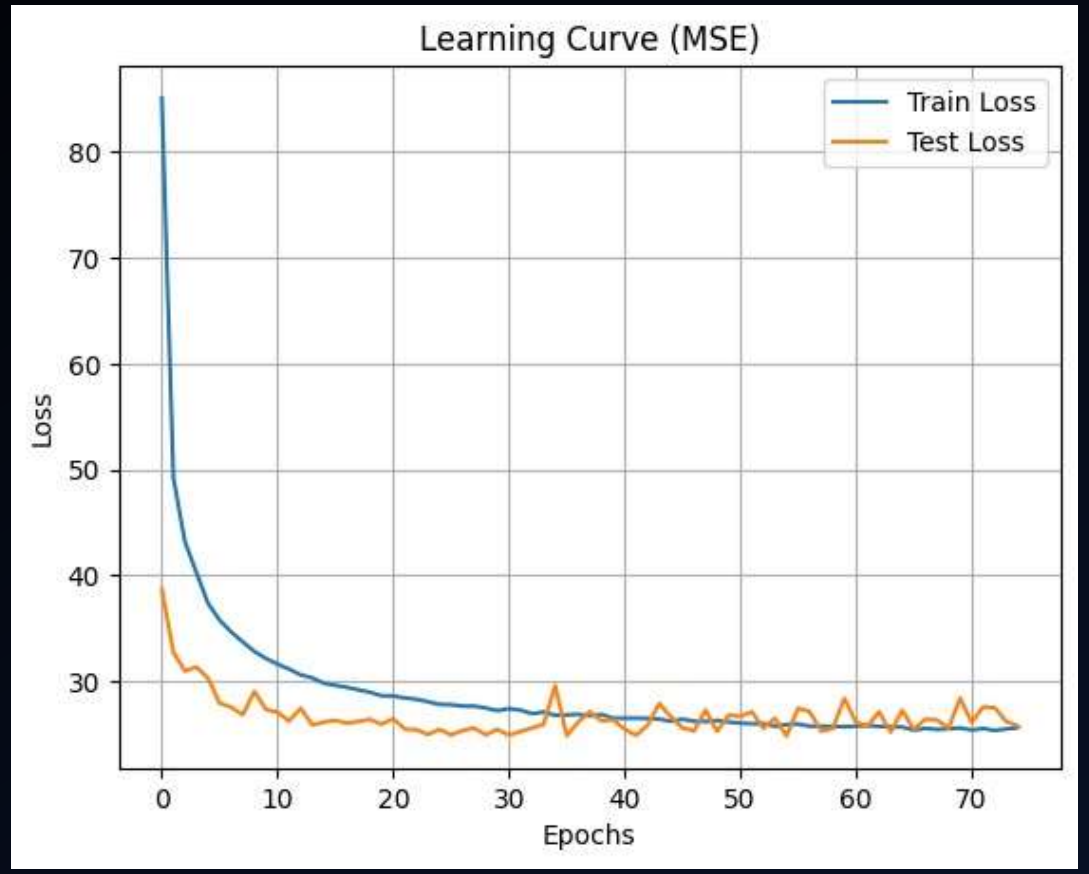
Testing Classification Report:

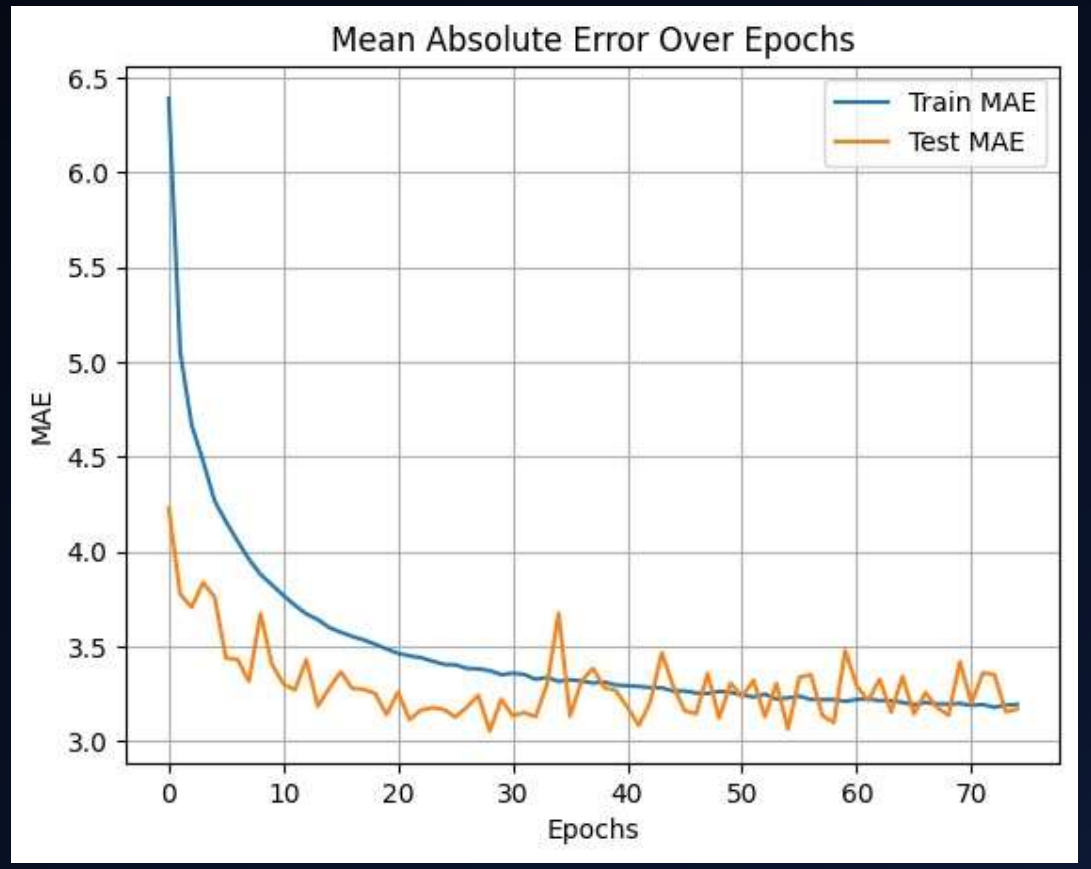
	precision	recall	f1-score	support
0.0	0.97	0.95	0.96	1276
1.0	1.00	0.99	1.00	9015
2.0	0.92	0.98	0.95	1746
3.0	0.99	0.97	0.98	5439
4.0	0.92	0.99	0.96	472
accuracy			0.98	17948
macro avg	0.96	0.98	0.97	17948
weighted avg	0.98	0.98	0.98	17948



Deep Learning (Reg)

```
Training MSE: 21.5954, MAE: 2.9920, R²: 0.9470
Testing MSE: 23.5067, MAE: 3.0616, R²: 0.9413
```



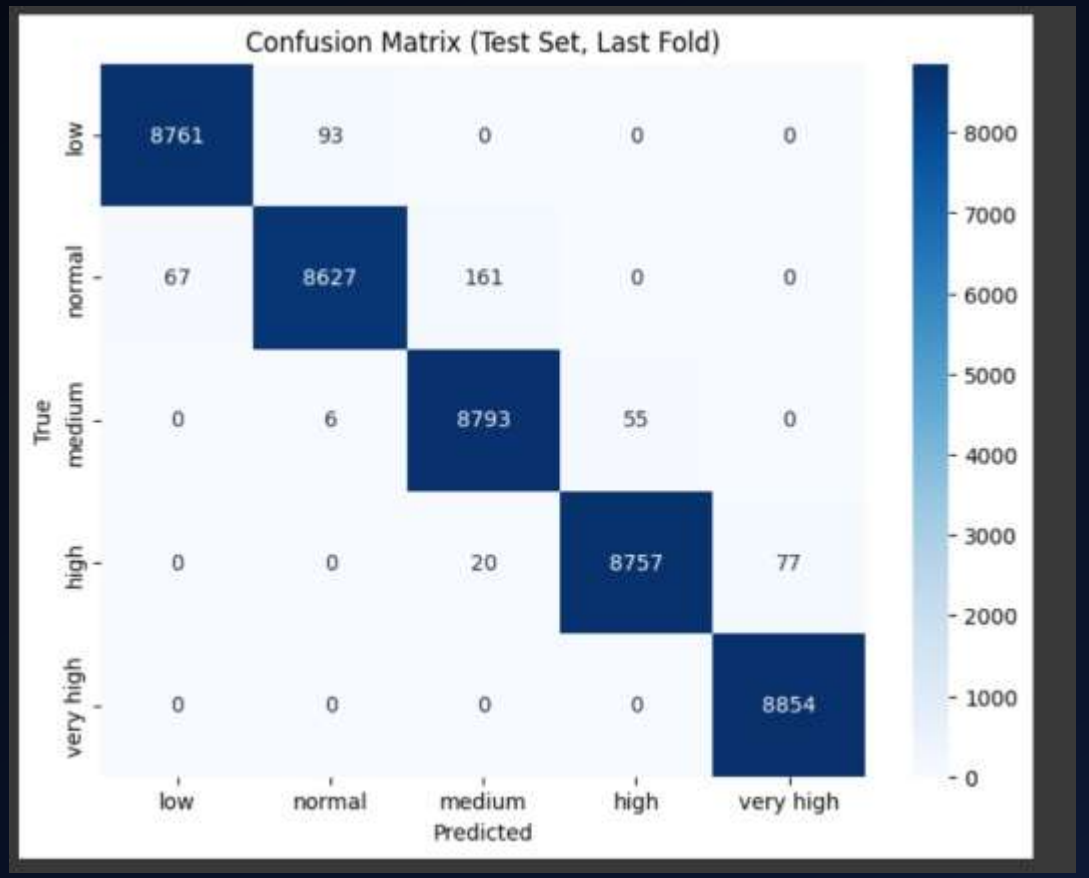


Deep Learning (Classification)

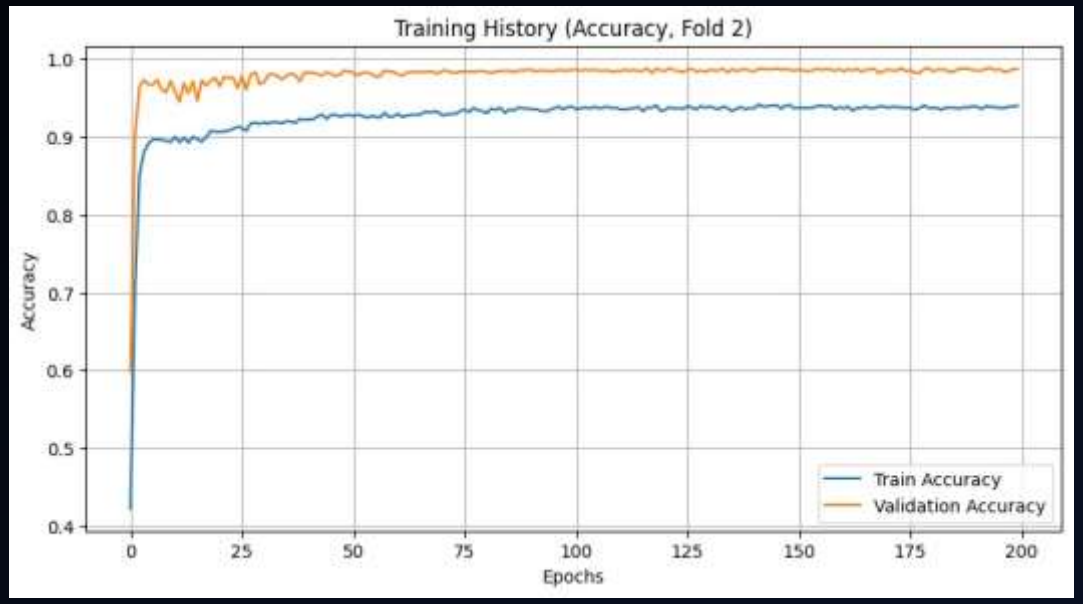
Model Train Accuracy: 0.9917
Model Test Accuracy: 0.9892

Classification Report (Test Set, Last Fold):

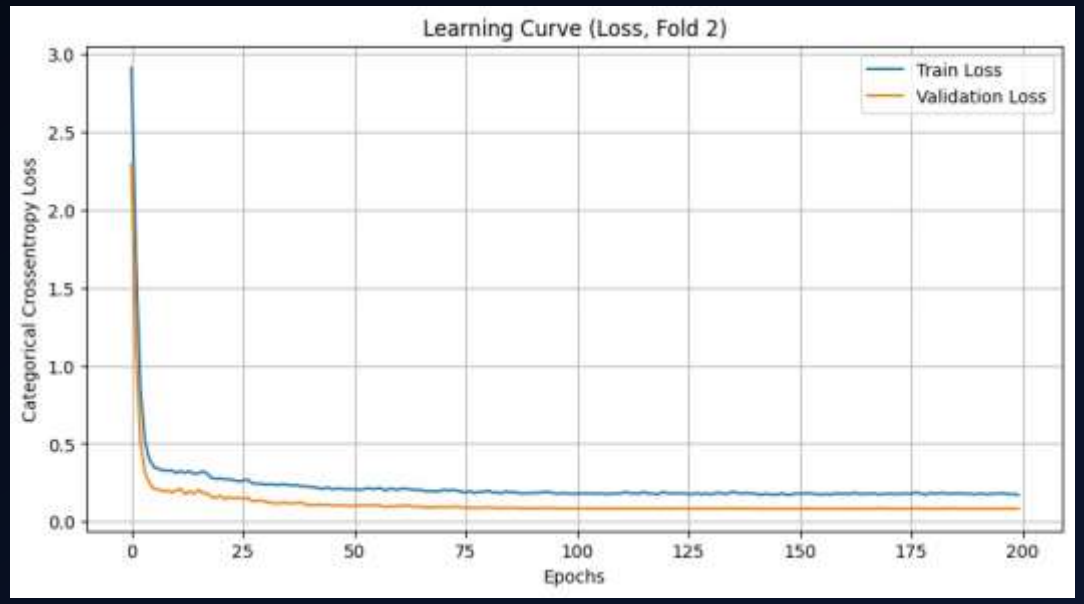
	precision	recall	f1-score	support
low	0.99	0.99	0.99	8854
normal	0.99	0.97	0.98	8855
medium	0.98	0.99	0.99	8854
high	0.99	0.99	0.99	8854
very high	0.99	1.00	1.00	8854
accuracy			0.99	44271
macro avg	0.99	0.99	0.99	44271
weighted avg	0.99	0.99	0.99	44271



- Accuracy plot



- Learning curve



Model’s Notebooks

- RF:
  - https://github.com/Mhmd-sh3rawy/Spotify_dataset_NTI/blob/main/RFspotify.ipynb
- SVM:
  - https://github.com/Mhmd-sh3rawy/Spotify_dataset_NTI/blob/main/Final_Model_In_Project.ipynb
- Logistic:
  - https://colab.research.google.com/drive/1v3sJWcRjtOCmrLisT4L_0w0rWiAHLgX0?usp=sharing
- DL:
  - https://github.com/Mhmd-sh3rawy/Spotify_dataset_NTI/blob/main/Project_Deep_Learning_Part.ipynb

LinkedIn Accounts

Ahmed Adel

Ahmed Ayman

Abdalla Elmougi

Mazen Abdallah

Mohamed Al-Sha3rawy

Omar Mohamed Ahmed

Contact Us

- Ahmed Adel - +20 112 989 6286
- Ahmed Ayman - +20 111 334 8379
- Abdallah Elmougi - +20 100 639 7016
- Mazen Abdallah - +20 127 329 5227
- Mohamed Al-Sha3rawy - +20 114 171 1354
- Omar Mohamed - +20 101 296 1123

REFERENCES

- **Spotify Dataset for ML Practice**
- **Spotify: Basic to ML Pipeline | Part 1 to 6**