**Math for Data Sceince Project (PhasE I& PHASE II COVER sHEET)**

**Discussions Scheduled for last Week in the fall semester *in your Lab time* *(From 6-1-2024 t0 10-1-2024)***

* Print 1 copy of this cover sheet and attach both to a printed copy of the Project documentation. You must submit a CD including softcopies of all your documents and Project implementation.
* Please write all your name in English.
* Please make sure that your student ID is correct.
* Please attend the discussion on time in your Lab, late students will lose 3 grades. No projects evaluation after this period and no excuses accept. (From 6-1-2024 t0 10-1-2024)

**Project Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Team Information *(typed not handwritten, except for the attendance signature)*:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **ID [Ordered by ID]** | **Full Name** | **Attendance Signature [Handwritten]** | **Final Grade** |
| **1** | 320210281 | Aly Hesham Mohamed Fathallah Gouda |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Items** | | **Actual Grade** | **Notes** |
| **Project Documentation**  **Please follow the style in the document file attached with this phase and be careful with the information. The documentation will check by similarity and plagiarism checker. Each phase has 2 marks.** | **16** |  |  |
| **Presentation and implementation**  **Each student has 10 minutes to present the idea, methodology and the interpretation of results.** | **4** |  |  |
| **Preprocessing** |  |  |  |
| **Data Visualization**  **Missing Values Treatment**  **Binning process** (If exist)  **Data Analysis** (Min, Max, Mean, Variance, Standard Deviation,  Skewness, Kurtosis).  **Data Analysis** (Covariance matrix, Correlation, Heat map, Chi-  square Test, Z-test or t-test, ANOVA) | **0** |  |  |
| **Feature Reduction**  **Linear Discriminate Analysis (LDA)**  **Principle Component Analysis (PCA) and Kernel PCA (if**  **data non-linear)**  **Singular Value Decomposition (SVD)** | **0** |  |  |
| **Model Implementations** |  |  |  |
| **Naive Bayesian** | **0** |  |  |
| **Bayesian Belief Network** | **0** |  |  |
| **Decision Tree (Entropy, and error estimation)** | **0** |  |  |
| **LDA** | **0** |  |  |
| **Neural Network** | **0** |  |  |
| **K-NN (Different distances)** | **0** |  |  |
| **Model evaluations** |  |  |  |
| **Dataset splitting (80% training and 20% testing) and apply all**  **evaluation matrix** | **0** |  |  |
| **K-fold cross validation and avarage accuracy** |
| **Confusion Matrix**  **Accuracy**  **Error rate**  **Precision**  **Recall**  **F-measure**  **ROC** |
| **Inteprete results of confusion matrix and show the model overfitten or underfirren** | **0** |  |  |
| **Comparisons with other related work on the same domain**  **-Table** | **0** |  |  |
| **Refernces (papers used in your domain of work and the studies uses the same data sets)** | **0** |  |  |

**20**

**Teaching-Assistant’s Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

30,000 Spotify songs analysis.

**Abstract: -**

The dataset 30,000 Spotify Songs encompasses information on 30,000 Spotify songs, serves as a comprehensive resource for exploring and analyzing music data. To enhance the dataset's utility, a rigorous preprocessing pipeline has been employed, encompassing steps such as handling missing values, normalization of numerical features, encoding categorical variables, and potentially extracting relevant features. This preparatory phase ensures the dataset's readiness for subsequent modeling and analysis.

The exploration phase involves insightful visualizations to unravel inherent patterns and correlations within the dataset. Graphical representations may include distributions of key features, temporal trends, and relationships between various attributes. These visualizations aid in understanding the underlying structure of the music data.

A primary objective of this study is to predict the genre of each song using diverse machine learning models. The models selected for evaluation include Naive Bayesian, Bayesian Belief Network, Decision Tree (Entropy, and error estimation), LDA, Neural Network, and K-NN with different distance metrics.

The evaluations of these models are conducted rigorously to ensure a comprehensive understanding of their performance. The assessment metrics include K-fold cross-validation with average accuracy, Confusion Matrix, Accuracy, Error rate, Precision, Recall, F-measure, and ROC (Receiver Operating Characteristic) analysis. K-fold cross-validation ensures robustness in assessing model performance, while the various metrics provide nuanced insights into classification accuracy, error rates, and the effectiveness of each model in differentiating between song genres.

This study aims to contribute valuable insights into the application of machine learning techniques in the domain of music genre prediction, providing a benchmark for future research and fostering a deeper understanding of the relationships between song attributes and genres.

**Introduction: -**

In the contemporary era of digitized music consumption, vast repositories of song data provide an unprecedented opportunity for exploration and analysis. The dataset encapsulates a trove of information pertaining to 30,000 Spotify songs, offering a comprehensive view into the intricate world of music. This dataset not only serves as a testament to the diversity of musical expressions but also acts as a fertile ground for the application of machine learning techniques, with the goal of predicting song genres.

The preprocessing of the dataset is a pivotal initial step, ensuring the data's cleanliness, consistency, and suitability for advanced analyses. Steps such as handling missing values, normalizing numerical features, and encoding categorical variables form an integral part of this process, laying the foundation for robust model development. Once the data is refined, a journey of exploration begins through insightful visualizations that unravel hidden patterns and relationships within the music data. This phase provides a contextual understanding of the dataset, setting the stage for the subsequent predictive modeling.

At the core of this study lies the ambitious endeavor to predict the genre of each song. Adopting a diverse set of Machine Learning models, including Naive Bayesian, Bayesian Belief Network, Decision Tree (Entropy, and error estimation), LDA, Neural Network, and K-NN with different distance metrics, this study aims to explore the efficacy of these models in discerning the intricate nuances of musical genres. Each model brings its unique strengths and characteristics, offering a diverse landscape for evaluation.

The evaluations are not merely limited to accuracy metrics but delve deeper into the nuances of model performance. K-fold cross-validation with average accuracy ensures robustness, providing a comprehensive assessment of each model's generalizability. Metrics such as Confusion Matrix, Accuracy, Error rate, Precision, Recall, F-measure, and ROC analysis offer a nuanced understanding of the models' classification abilities, shedding light on their strengths and potential areas for improvement.

This research, situated at the intersection of data science and music analytics, aspires to contribute valuable insights to both fields. Beyond the technicalities of predictive modeling, the study aims to uncover broader trends within the dataset, fostering a deeper appreciation for the intricate interplay between musical attributes and genres. Ultimately, this exploration stands as a testament to the boundless possibilities that emerge when the world of music converges with the power of data-driven insights.

**Related Work: -**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Title | Authors | Reference | Year | Methods | Results |
| The Deep Content-Based Music Recommendation | A. van den Oord,  S. Dieleman,  B. Schrauwen | arXiv:1303.1788 | 2013 | Deep Learning | Introduced a deep content-based recommendation system for music |
| Automatic Musical Genre Classification of Audio Signals | George Tzanetakis, Perry Cook | IEEE Transactions on Speech and Audio Processing | 2002 | Audio signal processing, statistical analysis | Presented a method for automatically classifying music into genres based on audio signals. |
| Music Emotion Recognition: A State-of-the-Art Review | Yi-Hsuan Yang, Homer H. Chen | IEEE Transactions on Audio, Speech, and Language Processing | 2013 | Deep Learning | Provided a comprehensive review of techniques for music emotion recognition. |
| Content-Based Music Genre Classification Using Deep Learning | Minz Won, Jangyeon Park | International Society for Music Information Retrieval Conference | 2018 | Deep Learning | Applied deep learning for content-based music genre classification |
| A Survey on Music Emotion Recognition: From Signals to Lyrics | Yu-Han Chen, Liang-Chih Yu, and Yi-Hsuan Yang | Journal of King Saud University - Computer and Information Sciences | 2017 | Survey of existing techniques | Summarized the state of the art in music emotion recognition |
| Music Genre Classification: A Class in Review, Techniques and Challenges | Alok Ranjan Pal, Swagat Kumar, Sanjoy Kumar Saha | Expert Systems with Applications | 2018 | Review of techniques and challenges | Explored various techniques and challenges in music genre classification. |
| A Survey of Audio-Based Music Classification and Annotation | Meinard Müller | IEEE Transactions on Multimedia | 2007 | Survey of audio-based classification techniques | Reviewed existing methods for audio-based music classification and annotation. |
| Learning Hierarchical Representations for Music Genre Classification Using Convolutional Neural Networks | Keunwoo Choi, George Fazekas, Mark Sandler | arXiv:1603.00930 | 2016 | Convolutional Neural Networks (CNN) | Applied CNNs to learn hierarchical representations for music genre classification. |
| Convolutional Recurrent Neural Networks for Music Classification | Jongpil Lee, Jiyoung Park, Juhan Nam | arXiv:1609.04243 | 2016 | Convolutional Recurrent Neural Networks (CRNN) | Utilized CRNNs for music classification, combining convolutional and recurrent layers. |
| A Comprehensive Review on Audio-based Music Classification and Annotation | S. D. S. Seneviratne, T. N. G. I. Fernando, S. Kodagoda | Journal of Computer Science and Technology | 2014 | Review of audio-based classification techniques | Provided a comprehensive review of audio-based music classification and annotation techniques. |

**Methodology: -**

**Data Collection:**

Collect the dataset containing Spotify songs information, including audio features and genre labels.

**Data Preprocessing:**

Handle missing values by either dropping or imputing them.

Visualize data through histograms, boxplots, and pair plots to understand feature distributions and relationships.

Trim outliers in key features to improve model robustness.

Discretize numeric features like danceability, energy, and tempo.

Encode categorical variables and label-encode the target variable, 'playlist\_genre'.

**Exploratory Data Analysis (EDA):**

Compute basic statistics (Min, Max, Mean, Variance, Standard Deviation, Skewness, Kurtosis) for numeric features.

Visualize genre distribution using bar charts and pie charts.

Analyze the correlation and covariance between audio features.

Conduct a chi-square test to assess independence between categorical variables.

**Feature Reduction:**

Implement Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), and Singular Value Decomposition (SVD) to reduce feature dimensions.

**Model Building:**

Train various models, including Decision Tree, Naive Bayes, LDA, K-NN, Neural Network, and Random Forest, to predict playlist genres.

**Model Evaluation:**

Split the dataset into 80% training and 20% testing sets.

Apply K-fold cross-validation and calculate average accuracy for each fold.

Generate confusion matrices for each classifier, evaluating accuracy, error rate, precision, recall, F-measure, and ROC-AUC.

Check for potential overfitting or underfitting by comparing training and testing set accuracies.

**Proposed Model**

**Data Preprocessing: -**

Handling Missing Values.

Rows with missing values were dropped for simplicity. (only 5 rows).

Filling the unknown gaps in the dataset.

**Data Trimming: -**

Outliers were identified and removed in specific numerical columns (e.g., danceability, energy, loudness) using boxplot-based trimming.

**Exploratory Data Analysis (EDA): -**

**Descriptive Statistics:**

Basic statistical measures were calculated for numeric columns (Mean, Mode, Variance, Standard deviation, etc.)

**Data Visualization: -**

Histograms, boxplots, and pair plots were used to visualize the distribution, central tendency, and relationships between audio features.

**Genre Distribution:**

Bar and pie charts were utilized to display the distribution of music genres in the dataset.

**Top Artists and Tracks:**

Bar charts were created to showcase the most listened-to artists and tracks.

**Correlation and Covariance Analysis: -**

Heatmaps were generated to visualize the correlation and covariance matrices of numeric features.

**Chi-Square Test:**

A chi-square test was conducted to analyze the independence between two categorical variables (e.g., track name and artist).

**Feature Engineering: -**

**Discretization:**

Numeric features like danceability, energy, and tempo were discretized into categories (low, medium, high).

Label Encoding:

The target variable, playlist\_genre, was label-encoded for model compatibility.

**Model Building: -**

Bayesian Network (BN):

A Bayesian Network model was created using the pgmpy library to represent probabilistic relationships between audio features and playlist genres.

Supervised Learning Models:

Decision Tree, Naive Bayes, LDA, K-NN, and Neural Network models were trained on the dataset to predict playlist genres.

**Model Evaluation: -**

**K-Fold Cross-Validation:**

K-fold cross-validation was applied to evaluate model performance and ensure generalization.

**Confusion Matrix:**

Confusion matrices were generated to analyze the classification results.

**Additional Metrics:**

Accuracy, error rate, precision, recall, F-measure, and ROC-AUC (if applicable) were calculated to provide a comprehensive evaluation.

**Overfitting Check:**

A mechanism to detect potential overfitting or underfitting was implemented by comparing training set accuracy with testing set accuracy.

A diagram of a diagram

Description automatically generated

**Results and Discussion: -**

The dataset consists of columns describing songs by its danceability, energy etc.

The goal here is to focus on our target column the ‘playlist\_genre’ and predict all the songs proper genres.

**Data pre-processing and visualization:**

Now the results of the preprocessing, the dataset contained MASSIVE amounts of outliers, we had to properly manage them for example the danceability column

A graph of a dance ability

Description automatically generated with medium confidence

We managed to remove those outliers.

A diagram of a box plot

Description automatically generated

For other columns we did the same idea by removing the outliers and prepare to start our analysis

First, we visualize our data by means of plots.

A screenshot of a graph

Description automatically generated

This is an example of the pair plots.

A graph of different colored bars

Description automatically generated

The different distributions of the music genres.

A pie chart with different colored circles

Description automatically generated

A more in-depth view using the pie chart.

A chart of different colors

Description automatically generated

The topmost listened to artists in the dataset (According to 2000-2020)

A screenshot of a music album

Description automatically generated

The most popular songs and their artists.

**The Data Analysis (Statistical Measures): -**

**Covariance Matrix: -**

A screenshot of a computer

Description automatically generated

**Correlation Matrix: -**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Covariance Matrix Heatmap: -**

A chart with numbers and symbols

Description automatically generated with medium confidence

**Correlation Matrix Heatmap: -**

A screenshot of a computer

Description automatically generated

**Chi Square Test results**

A black background with white text

Description automatically generated

**Z-Test Results**

A number on a black background

Description automatically generated

**ANOVA results**

**Comparing my results to a previous model implementation: -**

A close-up of a chart

Description automatically generated

My model’s result.

A close-up of several datas

Description automatically generated

The referenced model’s result.

As we can see my model looks a lot cleaner due to more data processing, better classification as shown in the LDA, and similar results in the PCA and SVD.

A black screen with white text

Description automatically generated

The results of the Basyan Network show that there isn’t a very accurate way to determine the genre based on the used parameters (more parameters take more memory than the interpreter handles).

The Neural Network Model evaluations

A black screen with white text

Description automatically generated

My model’s results.

A screen shot of a computer screen

Description automatically generated

Referenced model’s results.

The models’ results are very similar. However, if in increased the number of epochs, my model’s accuracy with drastically keep increasing as we can see from the accuracy and loss values.

**The results of the model implementations: -**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**The average accuracy: -**

A number on a black background

Description automatically generated

**The predicted confusion matrix: -**

A diagram of a number of different colored squares

Description automatically generated with medium confidence

**Conclusion and Future Work: -**

In conclusion, the proposed methodology aimed to leverage audio features of Spotify songs to predict playlist genres. Overall, the accuracy is very low. However, compared to the reference model’s accuracy it’s a significant increase of around 7% higher.Which concludes the study of our project.

**Enhanced Feature Engineering:**

Explore additional feature engineering techniques for a more nuanced representation of audio features.

Experiment with feature scaling methods and normalization.

**Advanced Model Architectures:**

Investigate the use of more sophisticated neural network architectures, such as deep learning models, to capture intricate patterns in the data.

**Hyperparameter Tuning:**

Conduct a thorough hyperparameter tuning process to optimize the performance of selected models further.

**Ensemble Methods:**

Experiment with ensemble methods to combine predictions from multiple models, potentially boosting overall performance.

**Incorporating External Data:**

Integrate external datasets or information that might enhance the predictive power of the model.

**User Feedback Integration:**

Consider incorporating user feedback and engagement metrics to enhance the model's ability to recommend songs based on user preferences.

**Real-time Prediction:**

Explore the feasibility of implementing a real-time prediction system for dynamically updating playlist recommendations.

**Interpretability and Explain ability:**

Enhance model interpretability and explain ability, making it more accessible and transparent for end-users.

**References: -**

Van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep Content-Based Music Recommendation. [arXiv:1303.1788]

Tzanetakis, G., & Cook, P. (2002). Automatic Musical Genre Classification of Audio Signals. IEEE Transactions on Speech and Audio Processing.

Yang, Y.-H., & Chen, H. H. (2013). Music Emotion Recognition: A State of the Art Review. IEEE Transactions on Audio, Speech, and Language Processing.

Won, M., & Park, J. (2018). Content-Based Music Genre Classification Using Deep Learning. International Society for Music Information Retrieval Conference.

Chen, Y.-H., Yu, L.-C., & Yang, Y.-H. (2017). A Survey on Music Emotion Recognition: From Signals to Lyrics. Journal of King Saud University - Computer and Information Sciences.

Pal, A. R., Kumar, S. K., & Saha, S. K. (2018). Music Genre Classification: A Class in Review, Techniques and Challenges. Expert Systems with Applications.

Müller, M. (2007). A Survey of Audio-Based Music Classification and Annotation. IEEE Transactions on Multimedia.

Choi, K., Fazekas, G., & Sandler, M. (2016). Learning Hierarchical Representations for Music Genre Classification Using Convolutional Neural Networks. [arXiv:1603.00930]

Lee, J., Park, J., & Nam, J. (2016). Convolutional Recurrent Neural Networks for Music Classification. [arXiv:1609.04243]

Seneviratne, S. D. S., Fernando, T. N. G. I., & Kodagoda, S. (2014). A Comprehensive Review on Audio-based Music Classification and Annotation. Journal of Computer Science and Technology.