UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Faculty of Computing

SECB3203 PROGRAMMING FOR BIOINFORMATICS

REPORT

Semester 1 2025/2026

Section 02 Group 6

NAME	MATRIC NUMBER
NURUL SYASYAWAFA BINTI AMRAN	A23CS0167
ANGELA LEE SU ING	A23CS0047
KAREN YAM VEI XIN	A23CS0093

Lecturer: DR. SEAH CHOON SEN

Date Of Submission:

Table Of Content

1.0 Introduction

The sound generation and perception systems of animals have evolved to help them to survive in their environment. From an evolutionary perspective, the intentional sounds that are generated by animals should be distinct from random sounds of the environment. Some animals have special sensory capabilities such as vision, sights, feeling and awareness of natural changes as compared to human beings. Animal sounds can be helpful for human beings in terms of security, prediction of natural disasters and intimate interactions if we are able to recognize them properly.

1.1 Problem Background

The cat vocalizations form a significant mode of communication between cats and people. The meows can have different meanings; they can be a sign of hunger, a sign of distress, an attention seeking behavior, or any other emotional state. Historically, these vocalizations can be interpreted based on human observation and experience which can be subjective and unstable.

As large biological datasets are growing and machine learning is advancing, automated sound classification is gaining popularity. Computational models are also very popular in the analysis of any complex biological data in bioinformatics research. The use of automated classification of animal vocalizations is less developed than the use of genomic or image-based research, despite this advancement.

Biological audio data presents the chance to use machine learning methods on the data of the Cat Meow Classification dataset available on Kaggle. This data has audio recordings that are labeled and can be utilized to train and test sound classification models.

1.2 Problem Statement

The sounds of cat vocalizations can be manually classified with time and subject to human prejudice. This makes it hard to regularly recognize and examine various forms of cat meows because of the absence of an automated system. With the growth in the amount of bio audio data, manual methods are ineffective and unfeasible.

Thus, a need exists to have an automated classification system that is capable of analyzing and classifying cat vocalization sounds in an accurate manner. The present project will use machine learning to solve this problem by classifying cat meow sounds by extracted audio features.

1.3 Objectives

The primary goal of the project is to create a machine learning machine to categorize the sounds of cat vocalization.

The specific objectives are:

- To gather and process cat vocation audio information on a publicly available dataset.
- To obtain audio features, which can be analyzed and classified.
- To conduct data analysis, exploratory data analysis, on the extracted features.
- To create and model machine learning to classify sounds.
- To measure and adjust model performance by the means of relevant evaluation measures.

1.4 Scopes

The given project is devoted to the computational study of cat vocalization sounds with the help of machine learning. The scope is covered by open-source dataset data collection, audio preprocessing, feature extraction, exploratory data analysis, model development and evaluation.

The given analysis is restricted to offline audio information and lacks the real-time sound detection and development of mobile applications. Algorithms of machine learning that are frequently utilized are also targeted by the project, but not complex hardware-based implementation.

2.0 Data Collection and Pre-processing

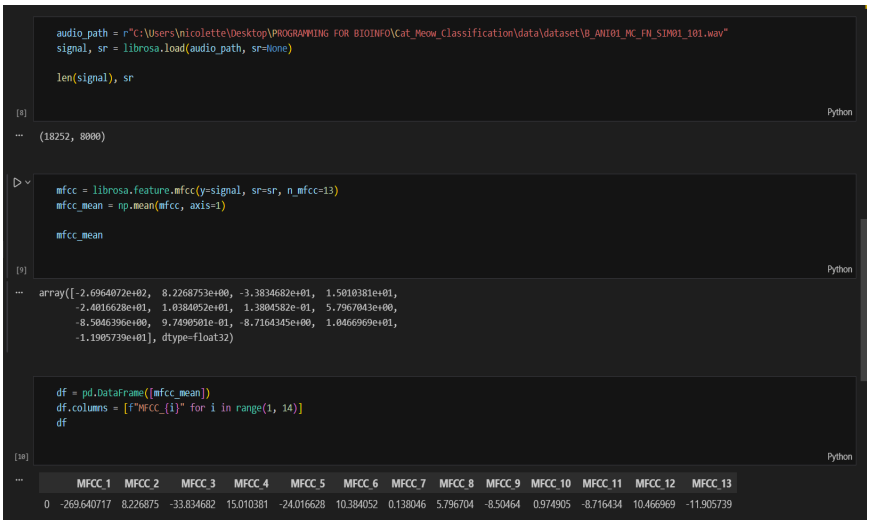
The data collection, dataset import, data cleaning and preparation of bioinformatics applications and machine learning models are described in detail in this section. The successful collection and processing of data prior to the application of machine learning techniques can have a significant impact on both the quality and performance, as well as improving the reliability and accuracy of results derived therefrom.

2.1 Importing Data

The data employed in this project was accessed at Kaggle and it comprises audio files (WAV format) of cat vocalization. These sound files are biological sound signals and can be considered the raw data that is supposed to be analyzed. All audio files were then downloaded onto the disk and grouped in a structured folder before analysis.

Data processing was done in Python which is the main language of programming. The librosa library, the library used for audio and signal processing, was used to import audio files. All the audio files were loaded and sampled in their original sampling rate to maintain the quality of the signals. Extraction of the features was then carried out to transform the raw audio signal to numerical form so that it is then analyzed.

Mel-Frequency Cepstral Coefficients (MFCCs) were derived from every audio file since they are useful in capturing spectral properties of audio signals and have been widely applied in the analysis of bio-signals and speech. The features that have been extracted on all audio files were collected into a structured dataset and saved as a CSV file. This allowed manipulation and analysis of data with ease, and reuse in further project phases.



2.2 Data Wrangling

The extracted MFCC features underwent data wrangling in order to make them ready to be analyzed and modelled. The data were put into a Pandas DataFrame with a possibility to work with structured data with ease. Preliminary analysis of the data was made to ensure the integrity, structure and consistency of data.

The checks of missing values were done and there were no missing or invalid values in the dataset. Types of data were confirmed so that all characteristics of MFCC were in numerical form, which is needed to perform statistical analysis and machine learning algorithms. The feature formatting was standardized so that there is consistency among all the samples.

The fundamental normalization and scaling issues were taken into account so that the MFCC features could be similar in their magnitude. This step is used to avoid features that have greater number spaces to dominate model training. There was also the exploration of the new derived attributes using feature grouping and categorization to facilitate further analysis.

All in all, data wrangling guaranteed the cleanness, structure and suitability of the dataset in the context of the exploratory data analysis and model development, which would serve as a trustworthy base of the further project stages.

2.3 Software and hardware requirements

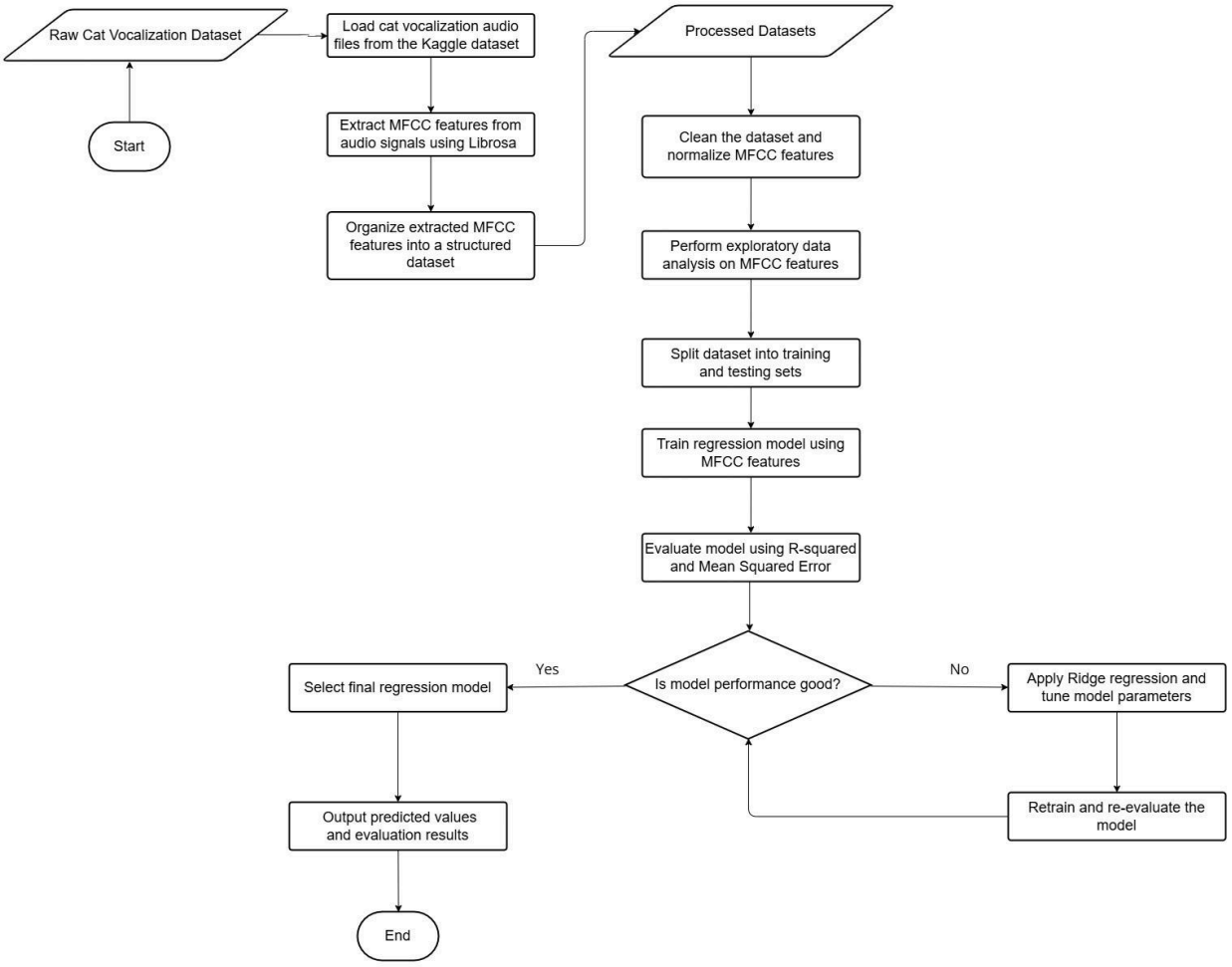
Software Requirements

Category	Tools
Programming Language	Python 3.9+
IDE/Editor	VS Code / Jupyter Notebook
Version Control	Git & Github
Dataset Source	Kaggle
Libraries	Numpy, Pandas, Matplotlib, Seaborn, Scikit-learn, TensorFlow / PyTorch, Librosa
Optional Cloud	Google Colab / Azure / AWS EC2

Hardware Requirements

Component	Minimum
CPU	Intel i5 / Ryzen 5
RAM	8 GB (16 GB recommend)
Storage	10 GB free
GPU (Optional)	NVIDIA GPU for CNN
OS	Windows / macOS / Linux

3.0 Flowchart of the proposed approach



3.1 Exploratory data analysis

Exploratory Data Analysis (EDA) was performed in order to have a general knowledge about the data and audio features extracted. The data is in the form of MFCC (Mel-Frequency Cepstral Coefficient) features of audio files of cat vocalizations. All audio samples were quantified with several MFCC coefficients which are employed in audio analysis and bio-signal analysis.

The central tendency and variability of the MFCC features, their mean, standard deviation, minimal, and maximum values were summarized with the help of the descriptive statistical analysis. This assisted in determining the overall distribution and magnitude of the features. There were no missing values in the dataset, which means that there was no necessity to impute the data before proceeding with its analysis.

Correlation analysis was performed also, to investigate the relationships among various MFCC features. The level of strength and direction of these relationships were visualized by creating a correlation heatmap. The findings indicated that a few MFCC features had a moderate correlation, implying that there may be linear correlation that can be examined using regression analysis. On the whole, EDA was useful in understanding the dataset structure and its properties, which would be used in the further model development.

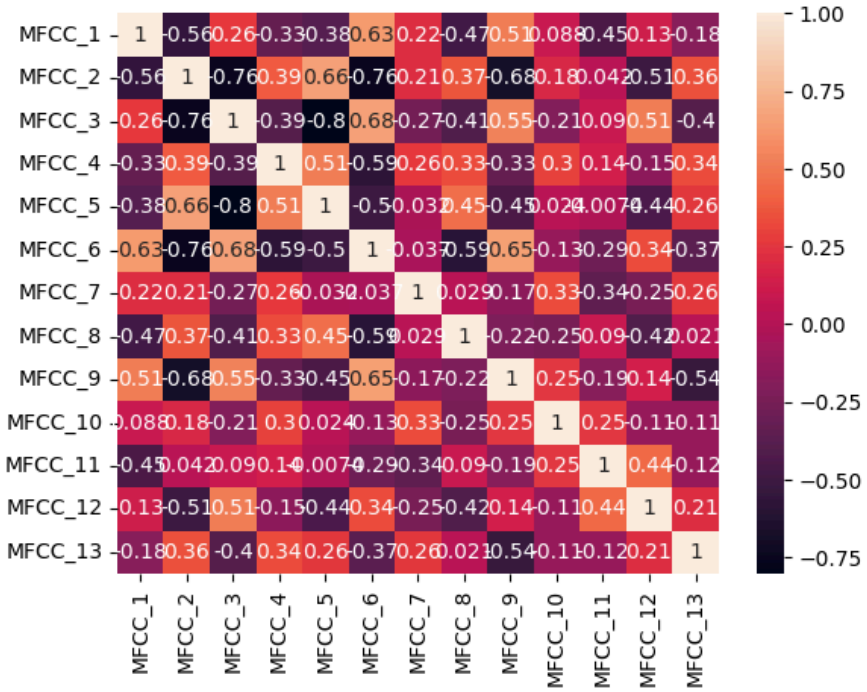


Figure 3.1: Correlation heatmap of MFCC features

3.2 Model development

During this step, machine learning models were designed using regression to determine the relationship between the extracted MFCC features. We have chosen one MFCC coefficient as the target variable and the rest of MFCC features as input variables. In this method, predictive relations between acoustical features of biological audio signals can be studied.

A simple and interpretable regression was the initial application of linear regression as the baseline model. Prediction of the target MFCC coefficient was done using the model which was being trained on several MFCC features. Furthermore, the use of polynomial regression was used to ensure that a non-linear relationship between the features may be undertaken. The formation of poly features was done to increase the capability of the model to approximate more and more complex patterns in the data.

R-squared (R^2) was used to estimate the model performance and mean squared error (MSE). Other visualization tools like actual versus predicted plot were also employed to identify the level at which the models fit the data. These models formed a basis of evaluation and improvements at the next stage.

3.3 Model evaluation

A fixed ratio between the training and testing set was used to determine the performance of the models and their generalization ability. This made sure that the models were tested on unseen data, which eliminated the chances of having bias performance estimation.

The trained regression models were tested through the R-squared (R^2) and the Mean Squared Error (MSE). The results of the training and testing were compared to determine possible overfitting and underfitting. Although linear regression gave a realistic baseline, training and testing results showed discrepancies, which showed that the model should be refined.

Ridge regression has been proposed to overcome overfitting because it implements regularization, which prohibits large coefficients. The optimal regularization parameter (alpha) was found by grid search. The final model featured better generalization of the test data, meaning that it is better optimized in terms of bias and variance.

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score


model = LinearRegression()
model.fit(X_train, y_train)


y_pred = model.predict(X_test)


print("R2 score:", r2_score(y_test, y_pred))
print("Mean Squared Error:", mean_squared_error(y_test, y_pred))
```

```
R2 score: 0.6656439392011622
Mean Squared Error: 1906.7554797238324
```

4.0 Testing and validation

The reliability and robustness of the developed models were also tested and validated. A train-test split was employed to test the model performance on untested data. The method emulates the application of models to new samples which is real world simulation.

Predictions based on the model on the test set were compared to the real values to assess the accuracy and consistency. The metrics of quantitative evaluation were further supported by visualization of prediction results. Ridge regression and hyperparameter tuning helped to achieve the best model stability and minimise overfitting.

On the whole, the testing and validation procedure demonstrated that the regression models could learn the significant relationships between the MFCC features and could reasonably predict the relationships with the unknown data. This shows that the proposed data analysis and modeling pipeline is effective in biological audio signal analysis.

5.0 Conclusion

In conclusion, this project implies that machine learning needs to be used to classify the sounds of cat vocalization. It introduces the vision of bioinformatics out of the framework of the conventional genetic analysis by means of the implementation of the computational methods to the biological audio data. It is hoped that the proposed outcome will be a working classification model that will validate machine learning in the processing of biological signals.