Context Aware Case Based Reasoning (KNN) for Auto-Generation of Song Playlist based on Weather Forecast (ARIMA, ES): Dublin, Ireland

MSc Research Project
Data Analytics

Shravani Sunil Navghare
x17124816

School of Computing
National College of Ireland

Supervisor: Dympna O'Sullivan, Pramod Pathak, Paul Stynes

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Shravani Sunil Navghare
Student ID:	x17124816
Programme:	Data Analytics
Year:	2017
Module:	MSc Research Project
Lecturer:	Dympna O'Sullivan, Pramod Pathak, Paul Stynes
Submission Due Date:	13/08/2018
Project Title:	Context Aware Case Based Reasoning (KNN) for Auto-Generation of Song Playlist based on Weather Forecast (ARIMA, ES): Dublin, Ireland
Word Count:	6004

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	12th August 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Context Aware Case Based Reasoning (KNN) for Auto-Generation of Song Playlist based on Weather Forecast (ARIMA, ES): Dublin, Ireland

Shravani Sunil Navghare
x17124816
MSc Research Project in Data Analytics

12th August 2018

Abstract

The ever increasing amount of music data on internet has led many opportunities in this field for implementing services to ease the music experience of the user. Recommendation domain in music area is the booming topic. The researches done so far were mostly concerned in improving the music experience of a particular user considering individual factors like emotional status, personal likings, and response of an individual user in particular context. Effects of external factors on music preferences of the user are still in research.
Objective: This work proposes a model to auto generate the playlist of songs suitable to the weather situations using context aware case based reasoning for Dublin.
Dataset: Three datasets are required for the implementation of the project, meteorological data having necessary affecting features, music library containing the name of songs and tags and audio fingerprints of the music library.
Methodology: Two approaches are considered for this study: to generate forecasting model using ARIMA for weather predictions and to generate case based reasoning classification model using KNN.
Results: This research concludes that the classification model under the constraint of weather classifies the music data accurately with an accuracy of 73.97%.
The model will inhibit a listener to scroll through millions of music tracks to find that one suitable track for the moment. It will improve user experience by recommending them tracks considering weather context.

1 Introduction

Music is a boon in today's stressful world. People engage themselves in music as they find it peaceful, rewarding or meaningful (Levitin et al.; 2018). It is proved by Bramwell-Dicks et al. (2013) that music leads one to relieve from stress and is beneficial for mental as well as physical health. Listening to music has helped people from centuries and invention like radio and Internet has served to fuel this trend. Radio has been the first and evergreen source through which music was transmitted from radio stations to users. The first radio transmission in Ireland was considered on 24th April 1916 from General Post Office of the morse code during the Easter Rising [1]. Almost 85% of people in Ireland listens to radio on regular basis [2]. Radio systems have proved to be a perk for music lovers. The improvement in the field of radio broadcasting is remarkable.

The sudden rise of technology has affected humans life to a great extent. Innovation in the field of music is an outcome of the revolution in technical world. This statement is justifiable with the development in different music portals and streaming services like Spotify (Wang; 2006), Pandora (Lee and Lee; 2007), Last.fm (Lee and Lee; 2007), etc. The development in audio compression techniques (Pachet; 2003) along with the radio systems and internet has made the huge amount of music easily available to billions of user at their fingertip. The extensive distribution of the music has accelerated difficulties for its storage, retrieval and caused information overload on users (Qin; 2013). It has opened new opportunities for researchers to invent new services that ease music search and provide as many features as possible to improve the music experience (Kaminskas and Ricci; 2012). Suchman 1987 has proposed a fact that user interact with the system differently according to the context. During research, it is observed that user responds to the music according to external variables like mood, atmosphere,etc. This characteristic is known as context awareness (Kaminskas and Ricci; 2012). This has changed the approach towards many problems and left an open research question in the field of context aware response of user to music.

New era radio service which combines the essence of evergreen radio system with the efficiency of Internet has been introduced in market in 1993 named Internet Radio (Silverstone; 2017). The audio signals are transmitted thorough internet in Internet Radio unlike the traditional radio systems. Hence, music has found its new and efficient way to reach out to people. To increase user experience and guarantee user satisfaction lot of strategies were implemented on internet broadcasting services. Hence, this research project proposes one such idea to improve customer experience.

Users usually prefer music according to the atmosphere, surrounding which adds up to the mood (Levitin et al.; 2018). It is proved by Howarth and Hoffman (1984) that weather affects human mood. It is hypothesized that mood is heavily influenced by weather and human behaviour (Howarth and Hoffman; 1984). This tendency of human mind has heavy influence on users listening habits like in rainy atmosphere people usually likes to listen to soft and romantic music whereas, sunny days songs are mostly fun to listen to, energetic, the dance music that makes you forget the chilling cold and feels lively (Howarth and Hoffman; 1984).

The aim of this project is an attempt to implement relation between music and weather, to classify music based on weather conditions and accordingly recommend music. The research question for this study is to effectively auto generate song using context aware case based reasoning (K- Nearest Neighbour) based on weather forecast (ARIMA), thereby increasing user experience.

The rest of the paper is organized as: Section 2 gives an overview of the state of art in the domain of context aware music recommendation and how weather data has been

[1] http://www.rte.ie/radio1/bowman-sunday-830/
[2] http://www.rte.ie/radio/advertising/

used for prediction outside weather domain. Section 3 enlights about the dataset and the methodologies used in this study along with the performance measures. Section 4 provides a guideline about the implementation of the models. Section 5 provides result of the analysis along with discussion of results with reference to past works performed on the same line.Section 6 concludes the whole work by providing guidelines for future work.

2 Related Work

Context aware recommendation is almost a decade old in the field of music recommendation and retrieval, introduced in the year 2006 by Lee and Lee (2006). Hence, since then study of various factors affecting music preferences took pace and significant amount of literature is published in the field. The literature review of this study has two approaches, one is the predictions using weather data outside the weather domain and the other is the context aware predictions in the music domain. Hence, the study is grouped in following two sections.

2.1 Predictions using weather data

The traditional weather prediction system has used limited factors and the time taken to predict was also large enough Murphy (1993). Hence, computer based machine learning approach (Regression) has been used to increase the efficiency of predicting weather. The application of weather prediction has opened the gate to various domains where the weather data has been used to predict stuff outside the domain of weather. Some of the literature along with their key findings and inferences are explained below which has inspired this project to use weather data in music domain.

Electricity
Sandels et al. (2015) has generated a regression model using stepwise model selection to predict electricity consumption in a Swedish office building. The aim of the research is to classify the usage in three classes: cooling load, ventilation load and appliance load. With the variation of predictor variables total 7 regression models are produced for the three dependent variables and. The optimal model is chosen amongst the 7 models to calculate the results. The results shows that the appliance load is predicted best for warm months data with Akaike Information Criterion (AIC) and Bayesian Infromation Criterion (BIC) and co-variates obtained at average 7.3% RMSE which is 3% less than other two dependent variables, while for cold months data ventilation load shows highest accuracy with AIC, BIC and co-variate at average 6.5% accuracy.

Travel
To tackle the air congestion problem Smith and Sherry (2008) has proposed a solution of using SVM to predict the aircraft arrival rates (AAR) using Terminal Aerodrome Forecast (TAF) as the independent variable which will be used to predict the future airport capacity. Several SVM models were developed at 4 peak timing of the day to predict the traffic of one airport. The average training accuracy of the model is 8% while the testing accuracy is 83%. While, the delay range for any given peak time is between 96 min to 13 min. This study was carried out in 2008 and it was predicted then that in 2012, the airports in United States will operate at 89% of their maximum capacity.

The disadvantage of SVM in this problem is that SVM is not able to clearly indicate the factor affecting the outcome. Hence, this method will not be able to answer which factor affects the aircraft arrival rates.

Cheng et al. (2017) has researched on the impact of weather conditions on the motorcycle crash rate. The author has developed five models using Full Bayesian hierarchical formulation to handle multivariate and temporal correlation among the factors of San Francisco motorcycle crash injury dataset. The models were compared on fitness and performance and found that the two models with time and severity differences had highest accuracy of predicting a crash. The MSPE (mean-squared predictive error) of the two good performing models are 0.25468 and 0.30811 respectively with air temperature, humidity, rainfall as predictors. Hence, this study can help the transportation companies to plan their transport and be aware of the risks for certain weather conditions.

Yang and Qian (2018) has attempted an experiment to predict the travel time by road on various factors including weather. The prediction model selects the suitable features amongst weather, incidents, number of vehicles, time and location of traffic speeds by LASSO, PCA and regression analysis. It is concluded from the research that the travel time through road is related to wind speed, time of travel, weather type and precipitation with root-mean-squared error of 16.6%.

Food

The first study in this domain was carried out by Corsi et al. (2001). The author has predicted the quality of wine from weather data in Italy. Because of the lack of price data, various measures of quality declared by several experts were used to determine the quality of wine. According to Corsi et al. (2001), quality of wine is greatly affected by the weather conditions during the harvesting and growing period. It is proved after several researches that weather is an important parameter in predicting the prices of long run wines. Ratings for wines from three different sources were compared to see the effect of weather on wines. The issue was addressed in two ways: First, by calculating linear regression of all the ratings on weather data and second, building an ordered probit model for each source that combines different measures of quality and role of weather, by calculating chi-square between the residuals. The result of the analysis is derived to be χ^2 of 12.757.

The food industry heavily depends on weather for the production of good quality of food. There are several researches carried out on the impact of weather on the yielding of various crops, fruits and vegetables. Li et al. (2015) has performed an experiment to reduce the uncertainty in predicting the rice yield. 13 models with variety of data at four different sites with different climatic conditions was evaluated. The response of crop yielding to change in temperature and carbon dioxide contain was also measured at all the sites. It is found out that the average predictions of all the 13 models are 10% less uncertain than the measures yields. Hence, an ensemble approach was tested which proved that the accuracy of models changed from 5% to 60% of predictions at any given site.

Similarly, Saeed et al. (2017) has also carried out a research on forecasting the wheat yield before harvesting for Punjab province by developing a model using Random forest method on 15 independent variables. The model uses weather data and normalized difference vegetation index (NDVI) data for 14 years (2001-2014). ARIMA is used as a forecasting model to forecast the meteorological data. The RMSE of the forecast model is found out to be 0.586. RMSE of the average and generic random forest model is 0.5.

On the same scale, recently De Villiers (2017) has used weather data to develop a statistical model for predicting the quantity of harvested tomatoes in South Africa. The data used for this experiment was of 7 years; the weather data collected was of everyday and between 10 day intervals during those seven years. Different models on multiple linear regression, lasso, regression tree, random forest, bagged regression tree and boosted regression tree was carried out with all the crop and weather variables. However, the random forest model seems to be more accurate model by 25% less error than the average error rate measured on mean absolute error.

Hence all the above research has shown a way of connecting two datasets, using relevant features from both the datasets and finding inference from it.

2.2 Factors affecting music preferences Context aware prediction in music domain

Context awareness by case based reasoning system is the new area of artificial intelligence where the items are processed by considering the past behaviour of the system for similar problems under the same context introduced by Schilit et al. (1994). Dey and Abowd (2000) explained context in computing systems as any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves. The context aware system has been initially proposed in the field of music recommendation and retrieval in the year 2006 by Lee and Lee (2006). Despite of having high potential in the idea, the progress in the music recommendation systems based on users actual situations is at its early stageKaminskas and Ricci (2012).

Throughout the study on music information retrieval and recommendation it is concluded that the choice of music is heavily influenced by external factors like mood, climatic conditions and surrounding in which the user is listening to music. The system retrieves and classifies the songs depending on the context like mood of the listener, climate, or any other situation which changes the perception of the listener.

Lee and Lee (2006) has designed a music recommendation system, M^3, for a particular user which will recommend songs based on the behaviour of that particular user in the context of weather and mood. This system is composed of three modules, Intention Module, which will predict the intention of the user to listen to music or not. The Mood Module, after confirming that the user has intention to listen, it will judge the mood of the user for the current weather condition based on previous responses. Finally the recommendation module will recommend similar songs to the user. Case based reasoning was applied on all the modules. This system is compared with two other recommendation systems, one with considering similar users listening history (Con_CBR) while the other with not considering context data M^3-C). It is concluded that M^3 model has the highest accuracy amongst the three of 89.8% followed by M^3-C model of 73.6% and lowest of Con_CBR model of 71% accuracy.

The extension of above mentioned study is done by Lee and Lee (2007) by demonstrating an experiment on similar line by building a music recommendation system, C^2_Music. This system responds under the context of users demographics and behavioural pattern. Similar users who listened to the music in same context is retrieved to build CBR model. K fold cross validation is implemented on the CBR model which resulted in the accuracy of 54.2% which is found out to be 8% greater than C_Music model which is 46.1%. The C_Music model is developed by Lee and Lee (2007) for comparison with contextual recommendation system (C^2_Music model) using only the past listening history of the user.

Kaminskas and Ricci (2012) has compared and documented the evolution of music recommendation. According to Kaminskas and Ricci (2012) after the demonstration of M^3 model by Lee and Lee (2006) the next step in the field of context aware music recommendation is taken by Kaminskas and Ricci (2011) and Ankolekar and Sandholm (2011) to study the effect of location on music preferences. The latest study in this field is of Li and Shan (2007) and Stupar and Michel (2011) analysing the effect of thumbnails of the music track that encourages the user to listen to the music.

Hence this study focuses on the work of Lee and Lee (2007) to improve the accuracy of the music recommendation model in the context of weather.

3 Methodology

In this research, study on auto generation of song playlist according to the weather condition is performed. This study is carried out by designing two models, one for weather forecast and the other for classification of songs based on the weather. CRoss Industry Standard Process for Data Mining (CRISP-DM) approach is used throughout the implementation of the project. The work flow diagram is mentioned in Figure 1 gives an insight about the implementation of this project.

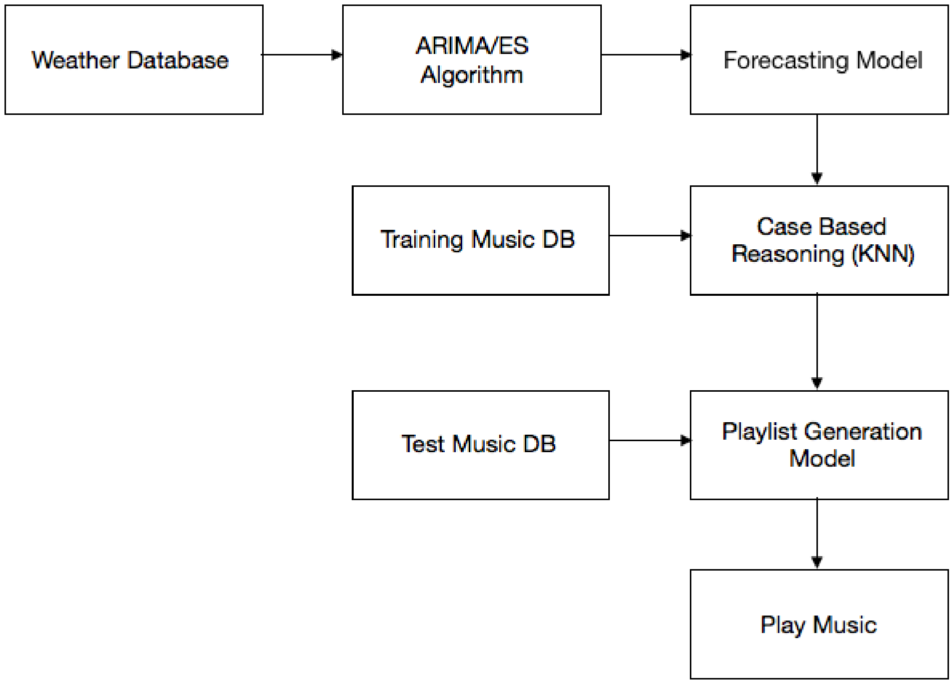

Figure 1: Work flow diagram

The meteorological data of Dublin [3] is used to train the weather forecasting model using machine learning algorithm for time series forecasting (ARIMA/ exponential smoothing). The forecasted weather (rainy or sunny) along with the music data is used to generate playlist of songs of the similar genres which were played in the similar condition of weather in the past using context aware case based reasoning. A classifier is built to classify the songs according to the context of weather. The classifier works on case based reasoning which has a policy that similar cases behave similarly in the same situation (Lee and Lee; 2006).

The general process to determine the connection between weather forecast model and music model is:

1. The weather is forecasted using weather model.
2. The forecasted weather is used as a tag to search for the music using last.fm rest API.
3. The music having the weather as a tag is retrieved and passed to music model.

Dublin is taken as a case study for this research as the uncertainty of weather is more in this area which will give this study a true test environment [4].

[3] `www.met.ie`

[4] `https://www.met.ie/climate/what-we-measure/rainfall`

3.1 Data acquisition

As mentioned before, this study requires historical data of weather of Dublin and music data containing appropriate tags related to weather. The historical daily weather data is collected from Met ireann website [5] which is freely available to download. The data has 961 records (daily weather data of past 3 years) containing essential factors useful to predict weather. The factors available in dataset are mentioned in Table 1.

Sr No.	Variable	Data Type	Meaning
1	year	int	Year
2	day	int	Number of day of the year
3	meant	num	Mean Air Temperature (C)
4	maxtp	num	Maximum Air Temperature (C)
5	mintp	num	Minimum Air Temperature (C)
6	mnmax	num	Mean Maximum Temperature (C)
7	mnmin	num	Mean Minimum Temperature (C)
8	rain	num	Precipitation Amount (mm)
9	gmin	num	Grass Minimum Temperature (C)
10	wdsp	num	Mean Wind Speed (knot))
11	maxgt	int	Highest Gust (knot)
12	sun	num	Sunshine duration (hours)

Table 1: Data Description of Weather Data

Last.fm is the largest music repository which collects data from multiple sources including online music services, users personal computer or any portable device (Kaminskas and Ricci; 2012). This data is scrobbled in the last.fm database when a user listens to it, hence building a profile of the user as well as generating playlist based on tags (Kaminskas and Ricci; 2012). Each and every song in the database has tag assigned to it by different users listing to the songs (Kaminskas and Ricci; 2012). These tags are assigned by the user depending on the feeling they get from the song. The tags of songs can be rock, smooth, cloudy, happy, sunny day, rocking, energetic, sad, nostalgic, etc representing either mood or genre of song or external conditions in which the song has been listened (Kaminskas and Ricci; 2012). The music data is scraped from last.fm website using rest API. The songs having tags related to weather condition (sunny, cloudy and rainy) are scraped and used to implement the music model. The balanced music dataset contains total 1043 songs having sunny, rainy and cloudy as tags. The features available after scrapping the data are mentioned in Table 2.

[5] www.met.ie

Sr No.	Variable	Data Type	Meaning
1	Song.name	factor	Name of the song
2	Duration	int	Duration of the song (sec)
3	mbid	factor	Unique ID
4	Artist	factor	Name of the artist
5	Tag	factor	Tag assigned to the song
6	Page	int	Number of page
7	Perpage	num	Songs present per page
8	Totalpage	num	Toal number of pages
9	Total	num	Total Songs

Table 2: Data Description of Music Data

Amongst the features mentioned in the above table, the features considered to build music model are song name, duration of song, mbid, artist, tag, page, per page, total page and total.

Audio fingerprinting has bought revolution in the world of music. The well-known application, Shazam, uses audio fingerprinting to identify the music playing around (Wang; 2006). Audio fingerprinting is the digital summary of audio signals representing frequency distribution of the audio data (Wang; 2006). It has been used in this study to classify the songs. Audio fingerprints of the music data is collected from Spotify. The attributes available in the data to represent the audio fingerprint are mentioned below in Table 3.

Sr No.	Variable	Data Type
1	danceability	num
2	energy	num
3	loudness	num
4	speechiness	num
5	acousticness	num
6	instrumentalness	num
7	liveness	num
8	valence	num
9	tempo	num

Table 3: Data Description of Audio Fingerprints

According to Pichl et al. (2016), all these features are useful in identification of songs.

3.2 Methodology of Models

Two models are developed for the implementation of this study as explained below.

3.2.1 Methodology of Weather forecast model (ARIMA/ES)

For generating weather forecast model, two time series forecasting models ARIMA and ES are compared and the best amongst them is chosen as the weather model. ARIMA (Box et al.; 2015) and ES (Koehler et al.; 2012) (Beaumont; 2014) are considered to be the most preferred algorithm for seasonal time series forecasting. RMSE is considered as the performance measure for the evaluation of forecast model. The model having the lowest RMSE value is considered as the best fit model having highest accuracy (Adhikari and Agrawal; 2013).

Autoregressive Integrated Moving Average (ARIMA) (p,d,q) (P,D,Q) is a machine learning model which is used the most for time series forecasting where p is order of AR model, d is order of differencing, q is order of MA model, P is order of seasonal AR model, D is order of seasonal differencing and Q is order of seasonal MA-model to forecast the temperatures (Adhikari and Agrawal; 2013). It is the generalization of ARMA model to involve non stationary case. For selecting the best fit ARIMA model, there is a function called auto.arima() in R which carries out following 4 phases (Box et al.; 2015) in background for a particular time series.

1. Model Identification.
2. Estimation of Model Parameters.
3. Diagnostic checking of appropriateness of the model.
4. If appropriate model is not found, repeat step 1.
5. If appropriate model found, declare the model as best fit model.
6. Use the best fit model for forecasting.

Exponential Smoothing is a forecasting technique that applies the difference of previous forecast and previous actual values to the next forecast. The smoothing equation for calculating the forecast for a period is Chiang (n.d.):

$$F_T = \alpha A_{t-1} + (1 - \alpha) F_{t-1} \tag{1}$$

where A is actual value, F is forecasted value and sym1 is smoothing factor. This method is considered for this experiment because ES is suitable for random behaviour of data (Box et al.; 2015).

Hence, the best fit ARIMA model will be compared with best fit exponential smoothing model to choose the best model for forecasting weather.

3.2.2 Methodology of Music Model (KNN)

The generation of playlist, based on weather conditions is always dependent on past knowledge provided by training data set. Hence, case based reasoning (CBR) is used for the implementation of this model. CBR is a data mining technique that uses previous solutions as reference to solve the new similar problem. It works on the policy that same problems occur frequently and same problems will have similar solutions (Aamodt and Plaza; 1994). CBR is a cyclic process, the famous 4 R cycle method designed by (Aamodt and Plaza; 1994): 1. Retrieve the similar cases, 2. Reuse the information to solve the new case 3. Revise the proposed solution and 4. Retain the solution for future reference.

The CBR system used in this project is K-nearest neighbour (KNN) algorithm. KNN is a classification algorithm which classifies the new case to one of the classified case, based on the similarity measure. 75% of the music dataset is used to train the model, while 25% is used to test the model.

3.3 Performance Measure

3.3.1 Evaluation of weather model

The best fit ARIMA and ES model are evaluated on Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) (Hyndman and Koehler; 2006). It is frequently used measure in the time series forecasting model to check models efficiency. Both the errors are measured by calculating the difference between the actual values and predicted values. RMSE is defined as the measure of accuracy to evaluate forecasting models by Hyndman and Koehler (2006).

3.3.2 Evaluation of music model

The KNN algorithm is implemented on multiple values of K to get a good value of K. Accuracy is used as a measure to check the goodness of the model on different flavours of K. The accuracy of music model is the accuracy of the project as this accuracy is influenced by the efficiency of forecast model. Confusion matrix is also used as a measure to get deeper insights of the results according to the business requirement.

4 Implementation

4.1 Data Preparation

The weather data collected from Met ireann website doesnt require much cleaning. The null values present in the dataset are replaced with NA or 0 according to the requirement. The data is converted to time series data by using ts() function in R. Last two months data is separated as testing dataset while the remaining data is assigned for training of the model by using window() function for dividing time series data.

The music data collected from last.fm has too many features which are not necessary for the study. Hence, such features are removed to increase the processing time. For collecting the audio fingerprints, a for lop is written which gets the audio fingerprint of the given tracks from spotify using get_track_audio_features() function of spotifyr package in R. These attributes are converted in numerical datatype as the knn classifier operates on numerical values only (Zhang; 2012).

These datasets are collected in a data frame and stored in csv files to execute the implementation of the study.

4.2 Models

4.2.1 Implementation of Weather Model

An experiment is performed to measure the performance of ARIMA and ES for forecasting temperature. The packages used to build these two forecasting models are forecast, stats, tseries and ggplot2. The evaluation of comparison of these two models is explained in detail in section 5. HoltWinters() function of stats package is used with forecast parameters value as α 0.2, β 0.1 and γ 0.1 for exponential smoothing where α is the extent of smoothness of level component, β is the extent of smoothness of trend component and γ is extent of smoothness of seasonal component (Mentzer; 1988). This fixed ES model is compared with estimated ES model where α, β, γ values are estimated by R and the best amongst them is compared with ARIMA. For the ARIMA model, the time series data is supplied to auto.arima() function of forecast package in R which selects the best ARIMA model for the data. It is evident from the results that ARIMA outperforms ES and hence, will be used to forecast precipitation amount, sunshine duration and temperature. Hence, three models are built to find the three predictions. It is found out that ARIMA(1,0,0)(2,1,0)[12] model is best for predicting the temperature, ARIMA(2,0,0)(2,1,0)[12] is best for predicting the sunshine duration while, ARIMA (1,0,1)(2,1,0)[12] model is best for predicting the precipitation amount. The results of these three models are explained in section 5.

After predicting the three values, viz. temperature, sunshine duration and precipitation amount; appropriate labels are assigned on the basis of the predicted values as rainy, cloudy or sunny. The first rule tested is sunshine duration According to [6], sunshine duration greater than 12 hours in a day is considered as sunny day. So, if the predicted value has sunshine duration greater than 12 hours it is considered as sunny day else the precipitation amount is check. If the precipitation amount indicates value greater than or equal to 15 mm then it is considered as rainy [7] or else the weather is considered as cloudy.

4.2.2 Implementation of Music model

After the generation of labels for the predicted weather, appropriate music is scraped from last.fm by considering those labels as tags using rest API. The audio fingerprints of the respective music is collected from spotify by using the track_uri feature from music dataset as the key to search for the audio fingerprints. The audio fingerprints are collected and stored in a data frame to generate the music model. The music model classifies the song based on the tags using KNN algorithm. All the values in audio fingerprint dataset are converted in numerical values by writing a normalize function. The dataset is randomly

[6] https://www.met.ie/climate/what-we-measure/sunshine
[7] https://www.met.ie/climate/what-we-measure/rainfall

divided in training and testing data in ratio 3:1. Confusion matrix and accuracy is used as a measure to evaluate performance of the model. The classifier is implemented using knn() function of caret package in R. To get a good value for k, the model is tested on multiple k values and the best k value is selected by examining the accuracies.

5 Evaluation

5.1 Comparison of ARIMA and ES

The ARIMA and the ES model are compared with each other to choose the best amongst them to forecast the weather. Temperature is chosen as a feature to test both the models as it is an important feature for predicting weather [8]. For an Exponential Smoothening (ES) model, two models are generated, one with fixed values of alpha, beta, gamma and the other with estimated values from R. The best model from these two is chosen to compare with the ARIMA model. RMSE of both the models is calculated which is shown in Table 4.

Sr No.	Model	RMSE
1	Fixed ES (alpha = 0.2, beta=0.1, gamma=0.1)	1.260027
2	Estimated ES	1.20565

Table 4: Evaluation of ES

It is clearly evident from the table that R estimated model of ES performs better as the interval is narrow. This model is now compared with the ARIMA model to choose the best model to forecast weather.

The auto.arima() function of forecast package has calculated that ARIMA (1,0,0)(2,1,0)[12] is the best fit model for predicting temperature. The diagnostic check of all the factors for selecting the best model is done by R in the background (Box et al.; 2015).Figure 2 shows the result generated by the ARIMA model.

```
ARIMA(1,0,0)(2,1,0)[12]

Coefficients:
         ar1     sar1     sar2
      0.2747  -0.6344  -0.3600
s.e.  0.0329   0.0318   0.0317

sigma^2 estimated as 0.01128:  log likelihood=711.55
AIC=-1415.1   AICc=-1415.05   BIC=-1396.03
```

Figure 2: Result of ARIMA

[8] https://www.fondriest.com/news/airtemperature.htm

The RMSE of the ARIMA model is found out to be 0.1053584 which is far less than the best ES model. Hence, ARIMA model is chosen for forecasting weather and considered as the weather model. The performance comparison of ES and ARIMA (1,0,0)(2,1,0) [12] on the measure of RMSE, MAE, MPE, MAPE and ACF1 is shown in Table 5.

Sr No.	Performance Measure	ES	ARIMA
1	RMSE	1.258174	0.1041081
2	MAE	0.9797966	0.06070996
3	MPE	-1.434447	0.9785288
4	MAPE	22.36235	9.94843
5	ACF1	0.1599501	-0.008958982

Table 5: Comparison of ES and ARIMA

5.2 Evaluation of weather model

ARIMA model is chosen as the weather model as it out performed ES in the comparison. The values of temperature, precipitation amount and sunshine duration are considered for the generation of weather label. Hence, three ARIMA models are built for the prediction of these three values. The plots of the actual and predicted value for sunshine duration and precipitation amount are shown in Figure 3 and Figure 4 respectively with blue line as the predicted values and black line as the actual values. The values of RMSE, MAE,

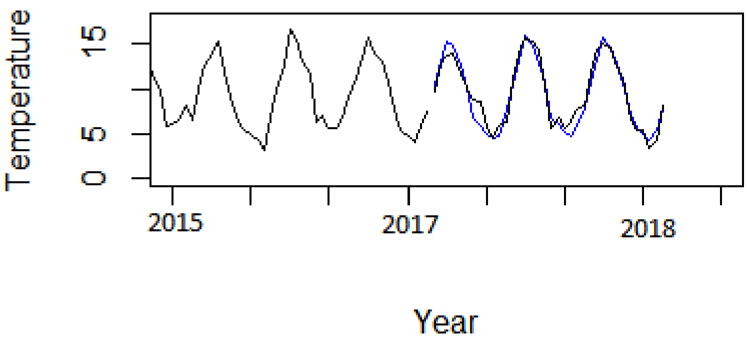

Figure 3: Prediction of Temperature using ARIMA

MPE, MAPE and ACF1 for all the three models are shown in Table 6.

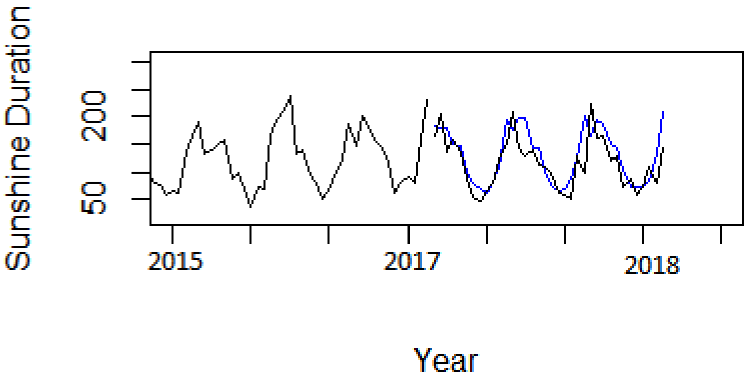

Figure 4: Prediction of Sunshine Duration using ARIMA

Perf. Measure	ARIMA(1,0,0)(2,1,0)[12]	ARIMA(2,0,0)(2,1,0)[12]	ARIMA(1,0,1)(2,1,0)[12]
RMSE	0.1041081	0.108035	0.105432
MAE	0.06070996	0.08268901	0.0756237801
MPE	0.9785288	-0.1572032	0.452380
MAPE	9.94843	0.8211699	5.23697
ACF1	-0.008958982	-0.003973276	-0.002476521

Table 6: Predictions using ARIMA

5.3 Evaluation of Music model

The music model classifies the music based on the audio fingerprints in three classes, viz. cloudy, rainy and sunny. KNN model is built for this classification as case based reasoning is dependent on KNN algorithm (Lee and Lee; 2006). Three different models are built on 3 values of K to choose the best amongst them. The three values of K are chosen as square root of number of instances, square root of number of attributes and the third value is random value. The accuracy of all three models are represented in Table 7.

KNN for K= 28 (sqrt of number of instances)	0.6173469
KNN for K= 3 (sqrt of number of attributes)	**0.7397959**
KNN for K= 8 (random value)	0.6594388

Table 7: Evaluation of Music Model

It is evident form table 4 that the KNN model for K=3 performs best with accuracy of nearly 74%. The accuracy of music model is the accuracy of this study as the performance of music model is influenced by the performance of weather model.

5.4 Discussion

This study classifies songs based on the audio fingerprints under the influence of weather. One of the observation made from the forecast models shows that temperature prediction shows more accuracy (RMSE = 0.1041081) than sunshine duration prediction (RMSE = 0.108035) as there is seasonality in temperature data. The trend in the data helped the ARIMA model to forecast temperature more effectively. An experiment performed by Saeed et al. (2017) to predict weather using ARIMA gave an accuracy of RMSE 0.586. The forecast model of this study clearly outperforms the work of Saeed et al. (2017) for weather forecast.

The comparison of classification models shows that the model for k=3 performs best with an accuracy of 73.97% which is justifiable with the number of classes that are actually present in the dataset hence providing a true test environment. The accuracy of research is better than the study of Lee and Lee (2007) (accuracy = 54.2%) under the same guideline. This proves that weather influences music preferences to a great extent along with demographics and behaviour of listener.

The aim of this study was to build a context aware case based reasoning model to generate music based on weather. A similar model was constructed by Lee and Lee (2006) to recommend music based on the context of listening history of similar users. This study shows an accuracy of 73.97% which outperforms the given study and proves that choice of music is dependent on various other features beside the listening history.

The music data and the labels used for the classification are the scope of this research. But this study is not limited to only music recommendation. The correlation defining context aware recommendation can be applied to correlation between human emotions and the shopping pattern (Lewis; 2017), influence of weather on labour relocation and the productivity of workers (Colmer; 2018). It can also be applied to various factors influencing the choice of music like demographical features (Lee and Lee; 2007), behavioural pattern of the listener (Lee and Lee; 2007) and many more.

6 Conclusion and Future Work

The contribution of this research proposal in the field of music is to use weather data in generating the list of songs according to the context of weather. The songs list is of the same genre which is played in the similar weather conditions in past by classifying them using case based reasoning The usage of weather data outside the weather domain is evident in Section 2 to predict energy consumption, crops yielding, crash rate of vehicles etc. The forecasting model has proved to be efficient than some past researches with RMSE of 0.10. The classification model has also proved to be efficient than past models with accuracy of 73.97%.

There are various perspectives that can be considered for future development of the current work. Hourly weather predictions can be made as the case study for this study if of Dublin where weather changes in hours. Due to the lack of features available in the hourly data, this study is concentrated towards daily predictions. The purpose of choosing KNN as a classifier for this study that KNN support case based reasoning but, an experiment can be performed using different classifiers to check for the accuracy. Connecting online music library (like spotify, pandora) with the model will provide the model with true test environment, better recommendation results hence increasing the music dataset which will promise novelty in the recommendation. The limitation of this project is that it did not include listening history of the user along with the weather context for recommending music. Further research can be done by inclusion of listeners feedback along the same lines.

As mentioned in Section 5.4, the scope of this project can be extended in correlated areas like human mood and shopping pattern which also provides way for future analysis. Also, to provide a new vision to the current work, this study can be experimented online to explore user satisfaction after listening to the recommended tracks.

References

Aamodt, A. and Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches, *AI communications* **7**(1): 39–59.

Adhikari, R. and Agrawal, R. (2013). An introductory study on time series modeling and forecasting, *arXiv preprint arXiv:1302.6613* .

Ankolekar, A. and Sandholm, T. (2011). Foxtrot: a soundtrack for where you are, *Proceedings of Interacting with Sound Workshop: Exploring Context-Aware, Local and Social Audio Applications*, ACM, pp. 26–31.

Beaumont, A. N. (2014). Data transforms with exponential smoothing methods of forecasting, *International Journal of Forecasting* **30**(4): 918–927.

Box, G. E., Jenkins, G. M., Reinsel, G. C. and Ljung, G. M. (2015). *Time series analysis: forecasting and control*, John Wiley & Sons.

Bramwell-Dicks, A., Petrie, H., Edwards, A. D. and Power, C. (2013). Affective musical interaction: Influencing users behaviour and experiences with music, *Music and Human-Computer Interaction*, Springer, pp. 67–83.

Cheng, W., Gill, G. S., Sakrani, T., Dasu, M. and Zhou, J. (2017). Predicting motorcycle crash injury severity using weather data and alternative bayesian multivariate crash frequency models, *Accident Analysis & Prevention* **108**: 172–180.

Chiang, D. T. C. (n.d.). Moving Averages and Exponential Smoothing.

Colmer, J. (2018). Weather, labor reallocation and industrial production: evidence from india.

Corsi, A., Ashenfelter, O. et al. (2001). Predicting italian wines quality from weather data and experts ratings,, *Cahier Scientifique de lOCVE* **4**(4).

De Villiers, M. (2017). *Predicting tomato crop yield from weather data using statistical learning techniques*, PhD thesis, Stellenbosch: Stellenbosch University.

Dey, A. K. and Abowd, G. D. (2000). Providing architectural support for building context-aware applications.

Howarth, E. and Hoffman, M. S. (1984). A multidimensional approach to the relationship between mood and weather, *British Journal of Psychology* **75**(1): 15–23.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy, *International journal of forecasting* **22**(4): 679–688.

Kaminskas, M. and Ricci, F. (2011). Location-adapted music recommendation using tags, *International Conference on User Modeling, Adaptation, and Personalization*, Springer, pp. 183–194.

Kaminskas, M. and Ricci, F. (2012). Contextual music information retrieval and recommendation: State of the art and challenges, *Computer Science Review* **6**(2-3): 89–119.

Koehler, A. B., Snyder, R. D., Ord, J. K. and Beaumont, A. (2012). A study of outliers in the exponential smoothing approach to forecasting, *International Journal of Forecasting* **28**(2): 477–484.

Lee, J. S. and Lee, J. C. (2006). Music for my mood: A music recommendation system based on context reasoning, *European Conference on Smart Sensing and Context*, Springer, pp. 190–203.

Lee, J. S. and Lee, J. C. (2007). Context awareness by case-based reasoning in a music recommendation system, *International Symposium on Ubiquitious Computing Systems*, Springer, pp. 45–58.

Levitin, D. J., Grahn, J. A. and London, J. (2018). The psychology of music: Rhythm and movement, *Annual review of psychology* **69**: 51–75.

Lewis, K. (2017). The influence of emotions on our shopping habits, *Siegel Institute Ethics Research Scholars* **1**(1): 3.

Li, C.-T. and Shan, M.-K. (2007). Emotion-based impressionism slideshow with automatic music accompaniment, *Proceedings of the 15th ACM international conference on Multimedia*, ACM, pp. 839–842.

Li, T., Hasegawa, T., Yin, X., Zhu, Y., Boote, K., Adam, M., Bregaglio, S., Buis, S., Confalonieri, R., Fumoto, T. et al. (2015). Uncertainties in predicting rice yield by current crop models under a wide range of climatic conditions, *Global change biology* **21**(3): 1328–1341.

Mentzer, J. T. (1988). Forecasting with adaptive extended exponential smoothing, *Journal of the Academy of Marketing Science* **16**(3-4): 62–70.

Murphy, A. H. (1993). What is a good forecast? an essay on the nature of goodness in weather forecasting, *Weather and forecasting* **8**(2): 281–293.

Pachet, F. (2003). Content management for electronic music distribution, *Communications of the ACM* **46**(4): 71–75.

Pichl, M., Zangerle, E. and Specht, G. (2016). Understanding playlist creation on music streaming platforms, *Multimedia (ISM), 2016 IEEE International Symposium on*, IEEE, pp. 475–480.

Qin, Y. (2013). *A Historical Survey of Music Recommendation Systems: Towards Evaluation*, PhD thesis, McGill University Libraries.

Saeed, U., Dempewolf, J., Becker-Reshef, I., Khan, A., Ahmad, A. and Wajid, S. A. (2017). Forecasting wheat yield from weather data and modis ndvi using random forests for punjab province, pakistan, *International journal of remote sensing* **38**(17): 4831–4854.

Sandels, C., Widén, J., Nordström, L. and Andersson, E. (2015). Day-ahead predictions of electricity consumption in a swedish office building from weather, occupancy, and temporal data, *Energy and Buildings* **108**: 279–290.

Schilit, B., Adams, N. and Want, R. (1994). Context-aware computing applications, *Mobile Computing Systems and Applications, 1994. Proceedings., Workshop on*, IEEE, pp. 85–90.

Silverstone, R. (2017). *Media, technology and everyday life in Europe: From information to communication*, Routledge.

Smith, D. A. and Sherry, L. (2008). Decision support tool for predicting aircraft arrival rates from weather forecasts, *Integrated Communications, Navigation and Surveillance Conference, 2008. ICNS 2008*, IEEE, pp. 1–12.

Stupar, A. and Michel, S. (2011). Picasso-to sing, you must close your eyes and draw, *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, ACM, pp. 715–724.

Wang, A. (2006). The shazam music recognition service, *Communications of the ACM* **49**(8): 44–48.

Yang, S. and Qian, Z. S. (2018). Understanding and predicting roadway travel time with spatio-temporal features of network traffic flow, weather conditions and incidents, *Technical report*.

Zhang, S. (2012). Nearest neighbor selection for iteratively knn imputation, *Journal of Systems and Software* **85**(11): 2541–2552.