**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Advanced Data Analytics, Big Data Storage and Processing |
| **Assessment Title:** | Ireland Future Years Rainfall Forecast with Deep Learning |
| **Lecturer Name:** | David McQuaid, Dr. Muhammad Iqbal |
| **Student Full Name:** | Deniz Ozsoy Ugur |
| **Student Number:** | 2023195 |
| **Assessment Due Date:** | 07/04/2024 |
| **Date of Submission:** | 07/04/2024 |
| **Word Count** | 4195 |
| **GitHub Repository** | CCT-Dublin Link:  <https://github.com/CCT-Dublin/adv-data-big-data-ft-ca1-denizozsoyugur>  The link used since the beginning of the project:  <https://github.com/denizozsoyugur/2023195_Integrated_CA> |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Ireland Future Years Rainfall Forecast with Deep Learning

Deniz Ozsoy Ugur  
MSc in Data Analytics  
CCT College DublinDublin,Ireland  
2023195@student.cct.ie

*Abstract*— *This article examines the prediction of rainfall for the coming years by examining a data set consisting of Ireland's monthly rainfall data between 1941 and 2023. These are stored in the database and scores were created using Apache Spark. Arima and LSTM were examined as two time series model records and future predictions were made; LSTM was chosen as the best performer and is noted as the basis for a Dash shared control panel in Python.*

Keywords— Apache Spark, Apache Hive, big data processing, time series analysis, ARIMA, LSTM

# Introduction

Since extraordinary weather events affect the course of our lives, being able to predict meteorological events will make our lives easier in many ways. Prediction of rainfall is very important in terms of its impact on water resources, water use areas and human life. Precipitation data is predictable and can be estimated almost accurately; This is a situation that will provide many advantages in terms of engineering. With the help of historical rainfall data, this prediction process can be made without specific mathematical equations, thanks to Artificial Neural Networks. Weather forecasting methods vary widely. Some of these are based on complex models that require high-capacity computers, using statistical and dynamic methods. In addition, it is used in predictions based on physics equations. However, it should not be forgotten that experience also plays an important role in practice.

Past experiences and acquired knowledge tell us a lot about how certain events may occur in the future. We can formulate these experiences by establishing a statistical relationship between the past and the future. In short, what happened yesterday and today can help us make statistical predictions about what will happen tomorrow. The resulting statistical estimates are often expressed as probabilities. For example: The most likely maximum temperature tomorrow will be 25˚C: There is a 90% chance that the actual maximum temperature will be between 20˚C and 30˚C. Such probability predictions are made using NWP (Numerical Weather Prediction) data.

Artificial neural networks look at examples of events and generalize about the relevant event. It also collects information and then makes decisions using this information when it encounters examples it has never seen before. In this study, the rainfall amount of Ireland in the coming years will be estimated. Monthly rainfall amounts between 1941 and 2023 were obtained from Data.Gov.ie. homogeneity analysis was performed before using these data. Using monthly precipitation data deemed appropriate in terms of homogeneity, various models were developed and compared to predict the annual precipitation values of Ireland with the artificial neural network’s method. It has been concluded that the developed models can be used in precipitation predictions when problems such as the measurement cannot be made, the measurement system is faulty, or the precipitation data are missing. This will be done as data is processed and moved from one database to another using big data storage and processing techniques such as Apache Spark and Hadoop Distributed File System.

This score will then be analyzed and used as the basis for time series analysis. This time series analysis will look at two different models, ARIMA and LSTM, where parameter tuning, and predictions are made with both. After the rainfall forecast for the coming years is made, the best forecasting model will be selected. These future predictions will then be plotted and displayed using an interactive control panel, allowing selection of the time period containing the predicted values.

# data mining (crisp-dm)

## Data Analysis Process

First, In Data Analysis process , a plan was made based on the CRISP-DM model. The Cross Industry Standard Process for Data Mining (CRISP-DM) represents the most common basic methodology used to standardise data mining processes in all sectors (Hotz, 2018).

1. Business/Research Understanding Phase

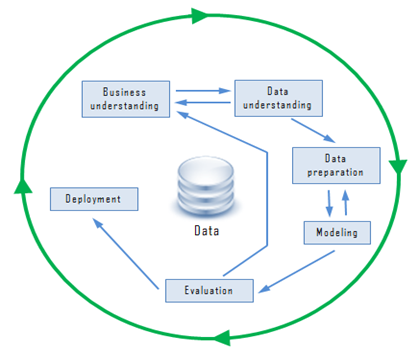
2. Data Understanding Phase

3. Data preparation Phase

4. Modelling Phase

5. Evaluation Phase

6. Deployment Phase



1. CRISP-DM

Understanding the goals of the project, defining the business problem, and determining success criteria. Relevant data sources are identified, data quality is evaluated and initial insights are obtained. Cleaning, transforming, and preparing data for modelling. The performance of the model is evaluated according to the success criteria defined in the Business Concept phase. The model is deployed in a production environment and its performance is monitored. Each phase builds on the previous one and is critical to the success of the project. The Business Understanding phase is critical to the success of the CRISP-DM methodology. Ensures that the project is aligned with business objectives, identifies relevant data sources, and establishes success criteria. A comprehensive Business Understanding phase can benefit the project, including improved outcomes, reduced risk, efficient use of resources, and improved communication.

## Database Benchmarking

The data.gov.ie open source dataset was used for this project. Care was taken to select the data set containing the most information among the data sets containing daily, monthly and annual precipitation information in Ireland. Testing was done on an Ubuntu virtual machine.

Structure of the data set; data is arranged by year and month. The size of the entire dataset was not very large. However, it was thought that using the months as patterns and specifying the months and years would be beneficial in terms of understanding the visualizations and the data set. Datasets were downloaded by year. The directory containing the files was made available to Ubuntu and Virtual Machine, which were used in part of the research. A Python file (read-filter-dataset.py) was created to read Pandas files and filter the data based on “Amount of Rain”, which is the keyword of the selected topic. This used Apache Spark to read the files, open them, filter them, and save them into a new file that was much smaller and contained only the rainfall data of interest. Certain steps were followed while using the Python program.

• Creating lists with files to analyse

• Reading the file in the list

• Writing the file to the file system

The use of Spark was not chosen at this stage because it provided few data frame management functions and had a built-in job optimization engine. Virtual Box was implemented using Jupiter Notebook.

# Exploratory data analysis

Exploratory data analysis (EDA) is an important step in the approach to data analysis and can "provide guidelines on how to look at and interpret data and is often a precursor to more advanced data analysis techniques (e.g. statistical modeling and machine learning)". (Azis, Fariz Darari and Muhammad Rizqy Septyandy, 2020) EDA was demonstrated by examining the shape, data types, and other typical features of the dataset and plotting predicted precipitation using a boxplot. At this stage, exploratory data analysis was conducted. For this purpose, a Jupiter Notebook was created and named “2023195\_Integrated\_CA2.ipynb”.

## Data Preparation

There are various abbreviations in the data set, and necessary arrangements have been made for clarity. To understand the data set, first, the data was presented graphically. The months with the most rainfall was displayed to understand the distribution of rainfall.

A graph of blue bars

Description automatically generated with medium confidence

1. Highest Monthly Precipitation Rainfall

Since the rainfall prediction for future years will be made with a retrospective analysis, it is necessary to understand the previous rainfall values first. Before building the model, training the model and making future predictions, the model was established to visualize and examine the rainfall graph of the past years.

A graph with blue dots

Description automatically generated

1. Ireland Annually Precipitation (1941-2023)

## Preparation for Modelling Phase

* Data was imported into the database using Apache Spark running in a Jupiter notebook. Spark was chosen because it is "an open-source data processing engine for large data sets" that will allow for the processing and storage of large data.
* Correlation test was performed to ensure compatibility of existing data with each other. Since there is no specific correlation between the relevant data, it was concluded that we can make predictions using year data. The dependent variable 'Amount of Precipitation' column was chosen because there was no correlation between the data and in order to make more accurate predictions. Histogram image was chosen to visualize the data.

A comparison of a graph

Description automatically generated

1. Histogram of Precipitation Amount and Kernel Density Estimate of Rainy Days

* After the relevant visualization and analysis, some developmental techniques began to be used to make the data set suitable for deep learning. Checked for null values in the data frame. Null values were detected and eliminated. Since annual precipitation values were tried to be estimated, the required precipitation estimation column for annual precipitation was developed by entering the details of the month with the most rain and the day with the most precipitation.

A screenshot of a computer

Description automatically generated

CCT College Dublin

.

1. .head() Function

* The 'new\_date' column is converted from a string to a datetime column and a list of missing dates is created based on the first and last date in the dataset. Missing dates are added, and the corresponding sentiment scores are set to NA. The data is then grouped by the date column, which is aggregated by the average value.

## After the Preparation Phase

After EDA processes, impressions were visualized with the data set containing all columns in order to make comments about the current and past situation. In the second part, to examine the planned predicted rainfall, a separate data frame was created with the columns created by combining months and years and the predicted rainfall data. The planned skin learning methods were carried out with this data set.

To store the resulting 'df1' using Apache Hive, it is first necessary to start the Hadoop Distributed File System (HDFS) and its YARN monitor by running the 'start-dfs.sh' and 'start-yarn.sh' commands in it which is then necessary to restart the Hive data warehouse system by calling 'Hive' in the terminal. Once all systems are running, a Spark session is created on the laptop with Hive connection settings. Pandas df is converted to Spark data frame and data is written to a new table in the database using the '.write.saveAsTable()' command. Selecting the first five rows from the resulting table shows that the data is stored there correctly.

A graph of blue lines

Description automatically generated

1. Time Series Data Visualization

## Time Series Analyses

* Before creating a time series analysis, certain tests were performed to check for trend and seasonality as well as whether the data is stationary. “A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time” (Duke University, 2019).
* A Dickey-Fuller test is applied to the data with the results indicating that the data is non-stationary. The data is made stationary by applying differentiation, which can “help stabilize the mean of a time series by removing changes in the level of a time series, and therefore eliminating (or reducing) trend and seasonality” (Hyndman and Athanasopoulos, 2018). The NA value created by this transformation is filled using the column mean, interpolation is not possible in this case as there is no prior value.

A screenshot of a computer

Description automatically generated

1. Results of Dickey-Fuller Test

* Seasonality, trends and residuals are plotted, with a seasonally decomposed column being created containing the sentiment scores. The created columns are tested to find which has the lowest p-value, and is therefore most stationary.
* According to the Dickey-Fuller unit root test results, it was concluded that the data set was stationary. This means that the time series exhibits statistically stable behavior and there is no particular trend or fluctuation over time. These results will be used in future modeling and forecasting studies to ensure that the stationarity assumption is met.
* Stationarity test; It was carried out with another test, KPSS. According to the KPSS test results, it can be said that the time series is stationary and does not contain any statistically significant change. This means that the time series remains constant around a certain mean and variance over time.

A white background with black text

Description automatically generated

1. Results of KPSS Test

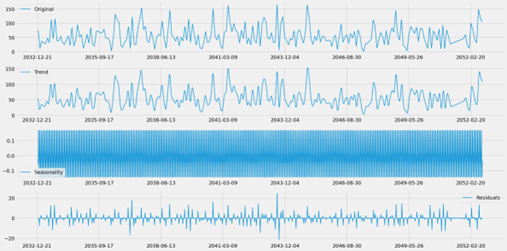
* Since the time series is stationary, differentiating or any other method will not be used to make it stationary.
* Autocorrelation and Partial Autocorrelation analysis was performed. The diagram with the results is depicted below,

A graph showing autocorrelation and autocorrelation

Description automatically generated

1. Autocorrelation and Partial Autocorrelation

* Multiplicative and Additive decomposition were calculated to understand the seasonality and patterns in the time series.



1. The Trend, The seasonality and The Residuals

* In this scenario, our KPSS statistics were obtained as 0.3250760. This value is smaller than any calculated critical value. The P value is 0.1 and is large. In this case, the null hypothesis cannot be rejected at any level of significance, and we can conclude that the time series is stationary around a linear trend. Another important test regarding the predictability of time series was carried out. Sample Entropy performed. The method is a modification of Approximate Entropy. This method has the advantages of being independent of data length and easy to apply. This test measures the complexity of a time series and is based on the probability that two parts of the series are similar. The probability is lower, and the mode of the time series is complex or unpredictable. This means that the time series has a moderate level of disorder or complexity.

# arima model

The Autoregressive Integrated Moving Average (ARIMA) model is a popular time series forecasting model, able to represent several different types of time series and particularly suitable for “data [which has been] differenced in order to achieve weak stationarity” (Lim, 2019)

## Arima Model Overview

Since the rainfall forecasts did not follow a certain pattern and a consistent forecast was tried to be obtained for the scattered data, the ARIMA model was first applied to the entire data set and then applied for the last 20 years.

### The data is normalised using a MinMaxScaler() function, to attempt to improve the results of the model. The data is then split into training and testing sets, with 7184 days in the train set and 1437 days in the test, roughly an 80/20 split.

### This part of the model will not attempt to make the time series stationary because it is stationary. A stationary time series is a series whose statistical properties, such as mean and variance, do not change over time. If the time series is non-stationary, differencing can be used to extract trends or seasonal components. It will not be used in this model.

## Hyperparameter Tuning

There are a number of different approaches and strategies for hyperparameter tuning, there is no standard set for time series analysis in particular. Some approaches include a grid search, “a brute force approach constrained by a predefined set of hyperparameters combinations, i.e., grid points” (Bakhashwain and Sagheer, 2021), and random search, where the predefined hyperparameter sets are sampled uniformly at random, typically leading to better-learned models through less computational time.

A function is created to find the best parameters for the ARIMA model, by testing a range of p, d and q values, and returning the combination which provides the best result. This is run on the training set, and the results are applied as parameters in the ARIMA model.

A screenshot of a computer

Description automatically generated

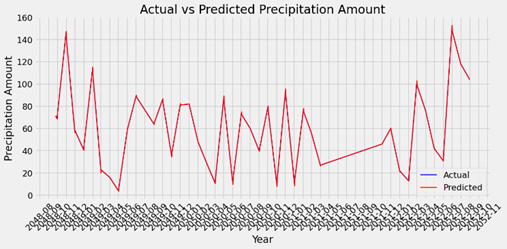
1. ARIMA Test Result

The model is then created and fitted using the training data. A report is printed and the residuals are plotted. Predictions are made against the test set, the result is plotted and the Root Mean Square Error (RMSE) is calculated. Predications are also plotted against the entire dataset to visualise the variance.

## Model

The model is then created and fitted using the training data. A report is printed, and the residuals are plotted. Predictions are made against the test set; the result is plotted, and the Root Mean Square Error (RMSE) is calculated.

#### Positioning Figures and Tables: Predications are also plotted against the entire dataset to visualise the variance.



1. Actual and Predicted Precipitation Amount

#### Predications: Using the model created, predictions are made at 7 day, 31 day and 90 day periods into the future, representing a week, a month, and three months respectively.

|  |  |
| --- | --- |
| new\_date | Predicted\_Precipitation |
| 02/09/2023 | 103.570635 |
| 03/09/2023 | 103.162942 |
| 04/09/2023 | 102.775827 |
| 05/09/2023 | 102.408253 |
| 06/09/2023 | 102.059232 |
| ... | … |
| 28/08/2024 | 95.493595 |
| 29/08/2024 | 95.493595 |
| 30/08/2024 | 95.493595 |
| 31/08/2024 | 95.493595 |
| 01/09/2024 | 95.493595 |
|  |  |
| [366 rows x 1 columns] |  |

# lstm

LSTM is a type of artificial neural network designed to identify patterns in successive data over time, such as time series. In particular, it can capture long-term dependencies through memory cells, making it particularly powerful in time series analysis. This model learns patterns in historical data and tries to predict future values based on these patterns. LSTM is ideal for identifying long-term patterns in time series, such as rainfall amounts in previous years. This allows seasonal changes, annual trends and other long-term factors to be taken into account in forecasts. Flexible models such as LSTM can quickly adapt to such dynamic changes and increase the accuracy of forecasts.

## LSTM Model Overview

Since it was decided which columns in the dataset would be selected for the LSTM model, it was decided that using which years to make predictions would be healthier for the future. Due to the need to improve libraries that allow representations such as Keras on the Linux system, the LSTM model was implemented without using the Jupiter interface but Ubuntu. Due to the size of the data set, future predictions were made using data from the last 20 years, as the model was required to work properly.

|  |
| --- |
| Forecasted Precipitation Amounts for Future Years: |
| 2023-09-02 42.800000 |
| 2023-09-03 46.733333 |
| 2023-09-04 50.666667 |
| 2023-09-05 54.600000 |
| 2023-09-06 58.533333 |
| ... |
| 2028-08-26 136.000000 |
| 2028-08-27 135.000000 |
| 2028-08-28 134.000000 |
| 2028-08-29 133.000000 |
| 2028-08-30 132.000000 |
| Freq: D, Name: Forecast, Length: 1825, dtype: float64 |

## Model

In the LSTM model, Epochs are used to organize the training process of the model. Each epoch means that all samples in the training data set are seen once by the model. This allows the model to see the entire dataset and learn general patterns across the dataset.

The MSE-Epochs chart, which shows the change in Mean Squared Error (MSE) values throughout the epochs, is very important. This graph is used to evaluate the performance of the ANN model in the training process. The lower the MSE value, the higher the success of the model. Additionally, the Gradient - Epoch graph shows the change in the gradient values of the ANN during training. Gradient is an optimization algorithm used to adapt the network to training data. The values in this graph show how accurately the gradient value is calculated.

A screenshot of a computer program

Description automatically generated

1. Epochs

By determining the number of epochs, how long the model would continue the training process was determined and the model was established. Epochs are also used to evaluate the performance of the model. During training, at the end of each epoch the model can be tested on a separate “validation” data set. This validation dataset consists of data that the model has not seen before and is often used to measure the model's ability to generalize. By examining the validation results between epochs, it was checked whether the model had problems such as overfitting or underfitting.

## Predictions

#### Pre-Prediction Test: In LSTM and other deep learning models, periods are the basic building blocks of the education process. They are important for the model to learn patterns in the dataset, evaluate its performance, and optimize the training process. Correctly determining the number of periods can increase the generalization ability of the model and prevent undesirable situations (such as overfitting). Therefore, the LSTM model was chosen for this prediction.

#### Prediction: Since careful selection and monitoring of periods in the training of the model will be critical for the success of the model, looking at the number of observations and periods in the last year of this model, we can say that the past data are compatible with the model. It was concluded that LSTM is the most successful model of the future and the two models used.

A graph showing a number of blue and red lines

Description automatically generated

1. LSTM Forecated and Precipitation Amount

# Discussion and conclusion

The initial dataset went through a series of transformations where it was stored entirely in an oracle database before being processed and reduced using an ARIMA and LSTM analysis models programmed from a Jupiter notebook using Apache Spark and the PySpark package. This data was then stored using Apache Hive and accessed again to perform a time series analysis on the extracted results. When the scores were initially plotted over time, there were no significant differences in predicted precipitation. The ARIMA model performed well in predicting future sentiment scores.

The LSTM estimator was more adept at predicting future values and was decided to be the model used to create the predicted values to be used in the final dashboard. Although the values predicted by the LSTM model appeared to be more accurate than the ARIMA models, over longer time periods the predictions had less variance and approached zero. It is possible that there is insufficient data and explanatory variables unavailable in time, making it difficult to make predictions for the imminent future.

In terms of data processing, Apache Spark was chosen as the data processing tool as it can be easily integrated with Python using PySpark. in the Python API for Apache spark, enables the performance of “real-time, large-scale data processing in a distributed environment using Python.” (Apache, 2023). The only disadvantage of the operations performed via Oracle format Ubuntu was that libraries such as Keras were not successful in visualization. Relevant visualizations were made on Windows Jupiter and also recorded.

There may be a variety of factors to consider when choosing between machine learning models. The interpretability of a model is how explainable the result is. ARIMA models provide an explanation of how past values and error terms affect the current value.

In LSTM and other deep learning models, periods are the basic building blocks of the educational process. They are important for the model to learn patterns in the dataset, evaluate its performance, and optimize the training process. Correctly determining the number of periods can increase the generalization ability of the model and prevent undesirable situations such as overlearning. It was concluded that LSTM is the most successful model of the future and the two models used. The reason why the method is so successful is that it is fast and does not require a large amount of data.

##### Acknowledgment *(Heading 5)*

This study was created for the Advanced Data Analysis and Big Data Storage & Processing Integrated CA assignment of the CCT College Dublin MSc Data analytics section.

##### References

1. Lim, Y.H. (2019). SAAS Berkeley. [online] saas.berkeley.edu. Available at: https://saas.berkeley.edu/rp/arima [Accessed 5 Nov. 2023].
2. Greg Schulz, 2012, Cloud and Virtual Storage Networking, CRC Press.
3. Big Data Systems: A 360-degree Approach, Jawad Ahmed Shamsi, Muhammad Ali Khojaye, CRC Press, 2021.
4. Spark in Action, Second Edition, Jean-Georges Perrin, Rob Thomas, Manning Publishers, 2020
5. https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html
6. Duke University (2019). Stationarity and differencing of time series data. [online] Duke.edu. Available at: https://people.duke.edu/~rnau/411diff.htm [Accessed 5 Nov. 2023].
7. Bakhashwain, N. and Sagheer, A. (2021). Online Tuning of Hyperparameters in Deep LSTM for Time Series Applications. International Journal of Intelligent Engineering and Systems, 14(1), pp.212–220. doi:https://doi.org/10.22266/ijies2021.0228.21.
8. Győrödi, C.A., Dumşe-Burescu, D.V., Zmaranda, D.R. and Győrödi, R.Ş. (2022). A Comparative Study of MongoDB and Document-Based MySQL for Big Data Application Data Management. Big Data and Cognitive Computing, 6(2), p.49. doi:https://doi.org/10.3390/bdcc6020049.
9. Rashid, A. (2013). Learning MongoDB. Packt Publishing.
10. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts. Retrieved from <https://otexts.com/fpp2/>
11. Internet Archive: About IA (no date). Available at: https://archive.org/about/ (Accessed: 29 April 2023).
12. Sanh, V. et al. (2020) ‘DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter’. arXiv. Available at: http://arxiv.org/abs/1910.01108 (Accessed: 18 May 2023).
13. Timeseries — darts documentation (no date). Available at: https://unit8co.github.io/darts/generated\_api/darts.timeseries.html (Accessed: 23 May 2023).
14. Verma, Y. (2021) Complete Guide To Dickey-Fuller Test In Time-Series Analysis, Analytics India Magazine. Available at: https://analyticsindiamag.com/complete-guide-to-dickey-fuller-test-in-time-series-analysis/ (Accessed: 24 May 2023).
15. Prasad, A., 2021. Importance of Machine Learning in Making Investment Decision in Stock Market. s.l.:s.n.
16. P. Lee, Z. H. Y. T., 2022. Trend Prediction Model of Asian Stock Market Volatility Dynamic Relationship Based on Machine Learning. s.l.:s.n.
17. McKinney W. 2012, Python for Data Analysis, O'Reilly Media [ISBN: 13: 978-14493]