SUMMER TRAINING REPORT

On

Forecasting of Water Discharge in Dams

Submitted to Guru Gobind Singh Indraprastha University, Delhi (India)
in partial fulfillment of the requirement for the award of the degree of

B.TECH

in

INFORMATION TECHNOLOGY

Submitted By
Sunaina Rustagi
Roll. No. 08915003120

DEPTT. OF INFORMATION TECHNOLOGY

MAHARAJA SURAJMAL INSTITUTE OF TECHNOLOGY ,
NEW DELHI-110058

DECEMBER 2022

ACKNOWLEDGEMENT

A research work owes its success from commencement to completion, to the people in love with researchers at various stages. Let me in this page express my gratitude to all those who helped us in various stage of this study. First, I would like to express my sincere gratitude indebtedness to **Dr. Tripti Sharma** (HOD, Department of Information Technology, Maharaja Surajmal Institute of Technology,New delhi) for allowing me to undergo the summer training of 30 days at **National Hydroelectric Power Corporation**.

I am grateful to our guide **Mr. Sunil Vishwakara (SM IT) sir and Mr. Rajan Trivedi sir (GM IT)**, for the help provided in completion of the project, which was assigned to me. Without his friendly help and guidance it was difficult to develop this project.

I am also thankful to **Dr. Suman Mann** for her true help, inspiration and for helping me to preparation of the final report and presentation.

Last but not least, I pay my sincere thanks and gratitude to all the Staff Members of **T&HR Department** of the company for their support and for making our training valuable and fruitful.

Submitted By:
Sunaina Rustagi
08915003120

CERTIFICATE

This is to certify that Ms. Sunaina Rustagi of Bachelor of Information Technology ,has completed Summer Training on the topic **Forcasting of Water Discharge** from National Hydroelectric Power Corporation organization as partial fulfillment of Bachelor of Engineering CSE. The summer Training report and presentation by him is genuine work done by him and the same is being submitted for evaluation.

Signature

एनएचपीसी लिमिटेड
(भारत सरकार का उद्यम)
NHPC Limited
(A Govt. of India Enterprise)

आज़ादी का अमृत महोत्सव

प्रशिक्षण एवं मानव संसाधन विकास विभाग
T&HRD Division
एनएचपीसी ऑफिस कॉम्प्लेक्स, सैक्टर-33, फ़रीदाबाद
(हरियाणा) – 121003
NHPC Office Complex, Sector-33,
Faridabad (Haryana)-121003
फोन/Phone: 0129-2278695, 2256564

एचआरडी/वीटी/2022-23/२२१६

दिनांक: 26/09/2022

जो कोई भी इससे संबंधित है उसके लिए

TO WHOMSOEVER IT MAY CONCERN

यह प्रमाणित किया जाता है कि महाराजा सुरजमल इंस्टीट्यूट ऑफ टेक्नोलोजी, जनकपुरी, नई दिल्ली की छात्रा सुश्री सुनैना रुस्तगी, बी.टेक (इन्फॉर्मेशन टेक्नोलोजी) ने सफलतापूर्वक एनएचपीसी लिमिटेड के सूचना प्रोद्योगिकी एवं संचार विभाग में दिनांक 28.07.2022 से 11.09.2022 तक अपना ओद्योगिक/ ग्रीष्मकालीन प्रशिक्षण "Prediction/Forecasting of Water Discharge" (Machine learning Model in Python using ARIMA Model) विषय पर पूरा कर लिया है।

प्रशिक्षण के दौरान उनका आचरण उत्कृष्ट था। हम उनके उज्ज्वल और सफल भविष्य की कामना करते हैं।

This is to certify that Sunaina Rustagi, a student of B.Tech (Information Technology) from Maharaja Surajmal Institute of Technology, Janakpuri, New Delhi, has successfully completed her Industrial / Summer training from 28.07.2022 to 11.09.2022 in IT & C Division of NHPC Limited on the Subject "Prediction/Forecasting of Water Discharge" (Machine learning Model in Python using ARIMA Model).

Her conduct during the training was excellent. We wish her a bright and successful future ahead.

26/09/22

(अनिरुद्ध सिंह)
व.प्रबंधक (एचआरडी)

CANDIDATE'S DECLARATION

I, **Sunaina Rustagi**, Roll No **08915003120**, B.Tech (Semester- 5th) of the Maharaja Surajmal Institute of Technology, New Delhi hereby declare that the Training Report entitled **"Forcasting of Water Discharge"** is an original work and data provided in the study is authentic to the best of my knowledge. This report has not been submitted to any other Institute for the award of any other degree.

Sunaina Rustagi
(Roll No. 08915003120**)**

Place: New, Delhi
Date: 24th December, 2022

ABOUT THE ORGANIZATION

NHPC Limited (erstwhile National Hydroelectric Power Corporation) is an Indian government hydropower board under the ownership of Ministry of Power, Government of India that was incorporated in the year 1975 with an authorised capital of ₹2,000 million and with an objective to plan, promote and organise an integrated and efficient development of hydroelectric power in all aspects.

At present, NHPC is a Mini Ratna Category-I Enterprise of the Govt. of India with an authorised share capital of ₹150,000 Million . With an investment base of over ₹387,180 Million Approx., NHPC is among the top ten companies in the country in terms of investment. Baira Suil Power station in [Salooni] Tehsil of Chamba district was the first project undertaken by NHPC.

Presently NHPC is engaged in the construction of 3 projects aggregating to a total capacity of 3130 MW. NHPC has planned to add 1702 MW during 12th Plan period of which 1372 MW has been completed. 5 projects of 4995 MW are awaiting clearances/Govt. approval for their implementation. Detailed Projects reports are being prepared for 3 projects of 1130 MW. Besides, 3 projects of 1230 MW are under development through its JV, Chenab Valley Power Projects Pvt. Ltd. in J&K. Here are some working projects by NHPC:

S.no.	Power Plant	State	Commissioned Capacity (MW)	year of commission
1	Baira Siul	Himachal Pradesh	180[4]	1981
2	Loktak	Manipur	105	1983
3	Salal	Jammu and Kashmir	690	1987
4	Tanakpur	Uttarakhand	120	1992
5	Chamera-I	Himachal Pradesh	540	1994
6	Uri-I	Jammu and Kashmir	480	1997
7	Rangit Dam	Sikkim	60	1999
8	Chamera II Hydroelectric Plant	Himachal Pradesh	300	2004
9	Indira Sagar*	Madhya Pradesh	1000	2005
10	Dhauliganga-I	Uttarakhand	280	2005
11	Dul Hasti	Jammu and Kashmir	390	2007
12	Omkareshwar*	Madhya Pradesh	520	2007
13	Teesta-V	Sikkim	510	2008
14	Sewa-II	Jammu and Kashmir	120	2010
15	Chamera-III	Himachal Pradesh	231	2012
16	Teesta Low Dam - III Hydropower Plant	West Bengal	132	2013
17	Nimmo Bazgo	Ladakh	45	2013
18	Chutak	Ladakh	44	2012–13
19	Uri-II	Jammu and Kashmir	240	2013
20	Parbati-III	Himachal Pradesh	520	2014
21	Jaisalmer Wind Farm	Rajasthan	50	2016
22	Teesta Low Dam - IV Hydropower Plant	West Bengal	160	2016
23	Kishenganga	Jammu and Kashmir	330	2018
24	Theni Solar farm	Tamil Nadu	50	2018

Table 1: Projects created by NHPC that are generating power

LIST OF TABLES

LIST OF FIGURES

INDEX

CHAPTER 1

INTRODUCTION

1.1 Problem Statement:

Every dam that is producing power and is fully operational uses the hydroelectric concept. The water from the rivers is collected on the downstream side of the dam. To generate power, the water is allowed to proceed once it reaches a certain level. The water turbine then rotates, converting the kinetic energy of the falling water into mechanical energy at the turbine shaft. Simply stated, falling water drives the water turbine. Through the use of an associated alternator, the turbine converts mechanical energy into electrical energy. The following describes the "basic operating concept of a hydroelectric power plant." The water runs through the downstream section of the river.

Water level monitoring at dams is a basic measurement. The main requirements for such a measurement are reliability and very high accuracy. Water Level Sensors are deployed in the upstream region, 15-20 km away, to monitor the water level and safety of the upstream. The necessary data is regularly updated by these sensors. The water level (m) and discharge (m3/sec) are directly proportional, meaning that as the water level upstream rises, so will the discharge.

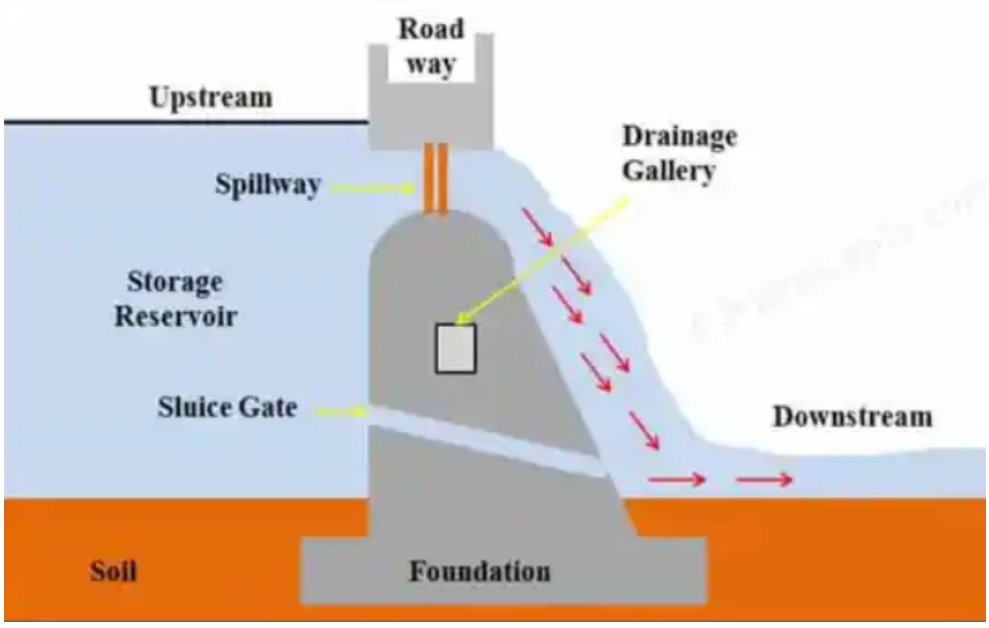

Fig 1: Structure of dam and upstream-downstream

After investigating several machine learning techniques, develop a prediction model that can forecast discharge level based on water level for monsoon and non-monsoon seasons. Also, design a user interface that displays accuracy and the finished product.

1.2 Need:

Water level Sensors are installed in a dam for safety measures. They are employed to gauge the ocean's depth, keep an eye on the dam's bottom water pressure, figure out how much flow is being produced, and control water levels. By giving users advance notice of any potential flooding, remote water level monitoring contributes to the development of an early warning system by giving them important time to move valuables, preserve assets, and secure other property.

1.3 Objective:

The goal is to build a machine learning model that can learn from the provided dataset on its own. The model need to be able to forecast and track the trend of water discharge and dam level over the course of many seasons. The problem deals with reinforcement learning area of machine learning.

1.4 Workflow:

The office has divided the internship into 6 weeks/45 days in accordance with the college curriculum. Under the mentor's supervision, the internship divided the problem implementation into six pieces. The workflow goes as follows:

- Week 1 -
    - Getting the understanding of the problem
    - Understanding dataset
    - Studying features and trend of the given dataset

- Week 2 -
    - Getting to know about different models that can be used
    - Learning about different machine learning models

- Week 3 -
    - Dataset preprocessing
    - Feature extraction

- Week 4 -
    - Implementation of different models on the dataset
    - Comparing accuracy of models

- Week 5 -
    - Final model selection after comparison
    - Prediction of water discharge

- Week 6 -
    - Conclusion/Result
    - Final submission of project
    - Final submission of report

CHAPTER 2

UNDERSTANDING DATASET AND PROBLEM STATEMENT

2.1 Getting the understanding of the problem:

Understanding the issue at hand primarily entails realizing the necessity of monitoring the water level and water discharge in dams in order to assure safety and accurate power generation calculations. We must be certain of the following tasks in order to accurately comprehend the issue:

- State the goal for the product you are refactoring.
- Determine whether the goal is best solved using ML.
- Verify you have the data required to train a model.

In this case all the above factors are satisfied.

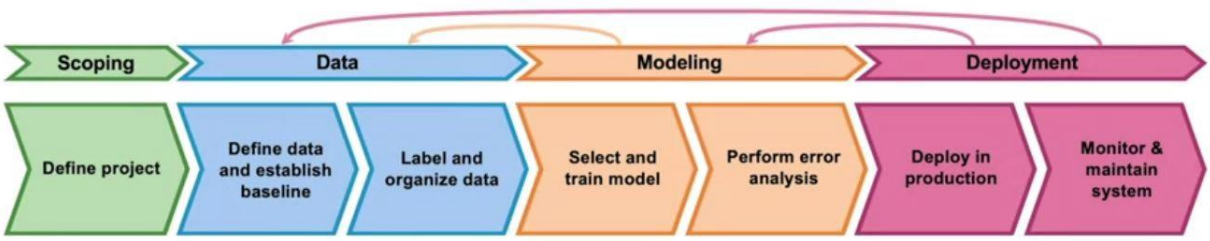

Fig 2: ML Project lifecycle for implementation

State the goal: Begin by stating your goal in non-ML terms. The goal is the answer to the question, "What are we trying to accomplish?". Here, our goal is to train the model and predict the discharge values for the given water level values.

Identify whether machine learning will be the most effective method for achieving the goal: Some people think of machine learning (ML) as a tool that can solve any problem. In truth, machine learning is a specialised technology meant to solve specific issues. Given the requirement to train the data and then predict the outcome following the training, the reinforcement learning field of machine learning will be utilised for this challenge.

Data: Data is the driving force of ML. To make good predictions, you need data that contains features with predictive power. Your data should have the following characteristics:

- Abundant. The more relevant and useful examples in your dataset, the better your model will be.
- Consistent and reliable. Having data that's consistently and reliably collected will produce a better model. For example, an ML-based weather model will benefit from data gathered over many years from the same reliable instruments.
- Trusted. Understand where your data will come from. Will the data be from trusted sources you control, like logs from your product, or will it be from sources you don't have much insight into, like the output from another ML system?
- Available. Make sure all inputs are available at prediction time in the correct format. If it will be difficult to obtain certain feature values at prediction time, omit those features from your datasets.
- Correct. In large datasets, it's inevitable that some labels will have incorrect values, but if more than a small percentage of labels are incorrect, the model will produce poor predictions.
- Representative. The datasets should be as representative of the real world as possible. In other words, the datasets should accurately reflect the events, user behaviors, and/or the phenomena of the real world being modeled. Training on unrepresentative datasets can cause poor performance when the model is asked to make real-world predictions.

For this problem statement our data fulfill all the required characteristics mentioned above.

2.2 Understanding dataset:

	A	B	C	D	E	F
1	Date	Time	Water_Level(m)	Discharge_Through_Rating_Curve(cumecs)		
2	13.01.2022	14:00:00	2842.6	11.4921		
3	13.01.2022	15:00:00	2842.5	7.9225		
4	13.01.2022	16:00:00	2842.5	7.9225		
5	13.01.2022	17:00:00	2842.5	7.9225		
6	13.01.2022	18:00:00	2842.5	7.9225		
7	13.01.2022	19:00:00	2842.5	7.9225		
8	13.01.2022	20:00:00	2842.5	7.9225		
9	13.01.2022	21:00:00	2842.5	7.9225		
10	13.01.2022	22:00:00	2842.5	7.9225		
11	13.01.2022	23:00:00	2842.5	7.9225		
12	14.01.2022	0:00:00	2842.5	7.9225		
13	14.01.2022	1:00:00	2842.5	7.9225		
14	14.01.2022	2:00:00	2842.5	7.9225		
15	14.01.2022	3:00:00	2842.5	7.9225		

Fig 3: Dataset Screenshot

The company gave away the dataset, which was compiled from data collected for the Kiru project in Kishtwar, Jammu and Kashmir, India. In the Kishtwar area of Jammu & Kashmir, the Kiru Hydroelectric Project (624 MW) is being considered for construction on the Chenab River. The plan calls for a Run of River Scheme. It is pertinent to note that the 1000 MW Pakal Dul Project, 624 MW Kiru Project, 540 MW Kwar Project, 930 MW Kirthai Project, and the 850 MW Ratle Project—which has been restored as a joint venture between the Center and the UT—are all situated close to one another.

Dataset provided consists of 4 fields:

- Date
- Time
- Water level in meters
- Water discharge through rating curve in cubic meter/sec

The dataset consist of 33128 rows of data and 4 columns. The date column in the dataset contains date in format dd/mm/yyyy. Second column comprises of the time of the day at which the data was recorded for the the 2 main values, ie, water level and discharge, according to 24 hour clock. Values of water level and discharge are measured in meters and cubicmeter/second respectively. These two column contains values in float format.

Fig 4: Kiru project, Kishtwar, Jammu & Kashmir, India

2.3 Studying features and trend of the given dataset:

Features are the basic building blocks of datasets. The quality of the features in your dataset has a major impact on the quality of the insights you will gain when you use that dataset for machine learning. Additionally, different business problems within the same industry do not necessarily require the same features, which is why it is important to have a strong understanding of the business goals of your data science project.

Located at an elevation of None meters (0 feet) above sea level, Kishtwar has a Temperate highland tropical climate with dry winters climate (Classification: Cwb). The district's yearly

temperature is 27.05ºC (80.69ºF) and it is 1.08% higher than India's averages. Kishtwar typically receives about 29.93 millimeters (1.18 inches) of precipitation and has 31.1 rainy days (8.52% of the time) annually.[1]

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Nov	Oct	Dec	Year
Record high °C (°F)	27.0 (80.6)	30.0 (86.0)	36.0 (96.8)	41.0 (105.8)	45.0 (113.0)	48.0 (118.4)	50.0 (122.0)	44.0 (111.2)	39.0 (102.2)	39.0 (102.2)	32.0 (89.6)	27.0 (80.6)	50.0 (122.0)
Average high °C (°F)	19.44 (66.99)	22.49 (72.48)	26.31 (79.36)	33.91 (93.04)	38.88 (101.98)	41.97 (107.55)	39.02 (102.24)	35.41 (95.74)	34.96 (94.93)	33.12 (91.62)	26.21 (79.18)	21.2 (70.16)	31.08 (87.94)
Daily mean °C (°F)	14.96 (58.93)	17.85 (64.13)	22.06 (71.71)	29.87 (85.77)	35.33 (95.59)	38.54 (101.37)	35.73 (96.31)	32.43 (90.37)	31.27 (88.29)	28.26 (82.87)	21.66 (70.99)	16.6 (61.88)	27.05 (80.69)
Average low °C (°F)	8.53 (47.35)	10.52 (50.94)	13.52 (56.34)	20.64 (69.15)	25.88 (78.58)	29.94 (85.89)	28.29 (82.92)	25.65 (78.17)	23.41 (74.14)	18.96 (66.13)	14.7 (58.46)	10.08 (50.14)	19.18 (66.52)
Record low °C (°F)	4.0 (39.2)	5.0 (41.0)	7.0 (44.6)	13.0 (55.4)	19.0 (66.2)	24.0 (75.2)	22.0 (71.6)	21.0 (69.8)	17.0 (62.6)	14.0 (57.2)	10.0 (50.0)	4.0 (39.2)	4.0 (39.2)
Average precipitation mm (inches)	24.97 (0.98)	42.67 (1.68)	34.22 (1.35)	15.65 (0.62)	13.39 (0.53)	25.16 (0.99)	60.92 (2.4)	72.58 (2.86)	29.1 (1.15)	1.71 (0.07)	19.23 (0.76)	19.6 (0.77)	29.93 (1.18)
Average precipitation days (≥ 1.0 mm)	1.91	1.82	2.91	2.27	2.91	2.64	4.82	5.27	3.09	0.64	1.82	1.0	2.59
Average relative humidity (%)	41.52	46.56	45.72	32.03	23.23	25.92	47.17	64.47	55.39	31.95	35.27	35.72	40.41
Mean monthly sunshine hours	7.47	10.26	10.52	12.48	13.78	13.99	13.79	12.68	11.37	9.62	8.13	8.27	11.03

Table 2: Weather data, Kishwar, Jammu & Kashmir

A wet day is one with at least 0.04 inches of liquid or liquid-equivalent precipitation. The chance of wet days in Kishtwār varies significantly throughout the year. The wetter season lasts 5.9 months, from March 8 to September 6, with a greater than 21% chance of a given day being a wet day. The month with the most wet days in Kishtwār is July, with an average of 10.4 days with at least 0.04 inches of precipitation. The drier season lasts 6.1 months, from September 6 to March 8. The month with the fewest wet days in Kishtwār is November, with an average of 1.7 days with at least 0.04 inches of precipitation.[2]

To show variation within the months and not just the monthly totals, the graph below show the rainfall accumulated over a sliding 31-day period centered around each day of the year. Kishtwār experiences significant seasonal variation in monthly rainfall. Rain falls throughout the year in Kishtwār.[2] The month with the most rain in Kishtwār is July, with an average rainfall of 3.5 inches. The month with the least rain in Kishtwār is November, with an average rainfall of 0.5 inches.

If we look at the precipitation trend of Kishtwar, we can clearly see that the peak rainfall is received in the month of July. Therefore, in the dataset provided we should have a trend similar to this.

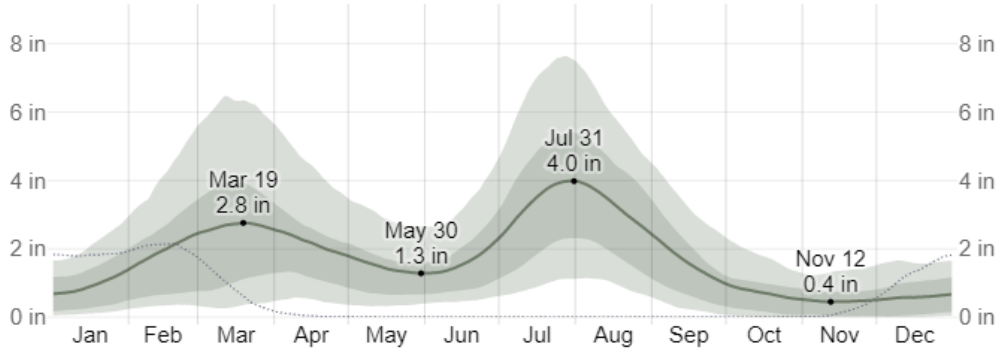

Fig 5: Average monthly rainfall in Kishtwar[1]

A "trend" is an upwards or downwards shift in a data set over time. As shown in the graph below, the tendency is consistent for the early dates, but for a brief time at the end, the discharge increases. The tendency is as stated in the issue statement, i.e., the relationship between water level (m) and discharge (m3/sec) is one of direct proportionality, which means that as the water level upstream rises, so will the discharge. The graph is drawn using raw data that has not been preprocessed, or filtered.

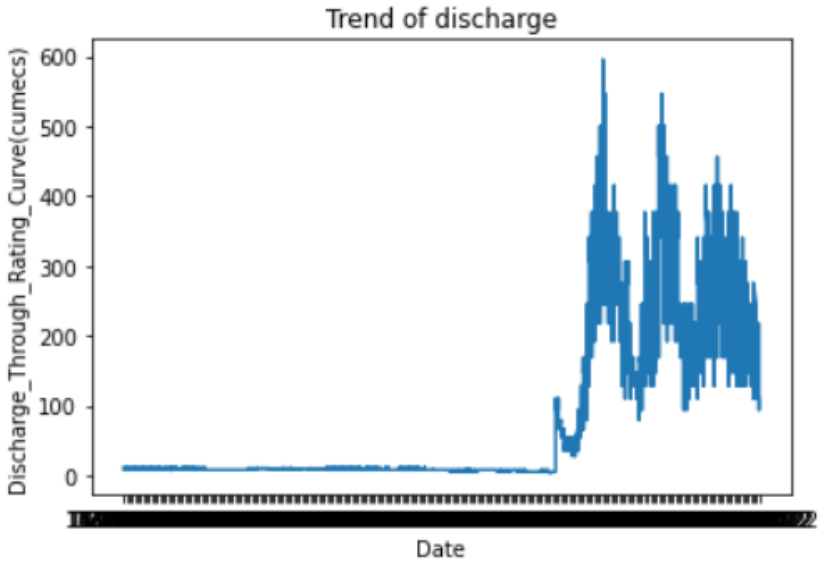

Fig 6: Trend of water discharge according to dataset

The monsoon season in Kiru, Kishtwar arrives around June end and ends around October. During the monsoon season the water level rises due the access rainfall in the hills which cause the river water overflow. As the water quantity in the river rises the water discharge for the same also increases. In the dataset, the water level rise is shown from 17th June to 28th August, as the data is given till 28th August, 2022.

CHAPTER 3

RESEARCHING MACHINE LEARNING MODELS

3.1 Getting to know about different models that can be used:

Machine Learning is a part of Data Science, an area that deals with statistics, algorithmics, and similar scientific methods used for knowledge extraction. Engineers can use ML models to replace complex, explicitly-coded decision-making processes by providing equivalent or similar procedures learned in an automated manner from data. ML offers smart solutions for organizations that want to implement decision processes that are just too complex to be manually coded. Statistical techniques (including machine learning, predictive modeling, and data mining). Predictive analytics helps us to understand possible future occurrences by analyzing the past. Machine-learning algorithms can learn from and make predictions on data, data-driven decisions.

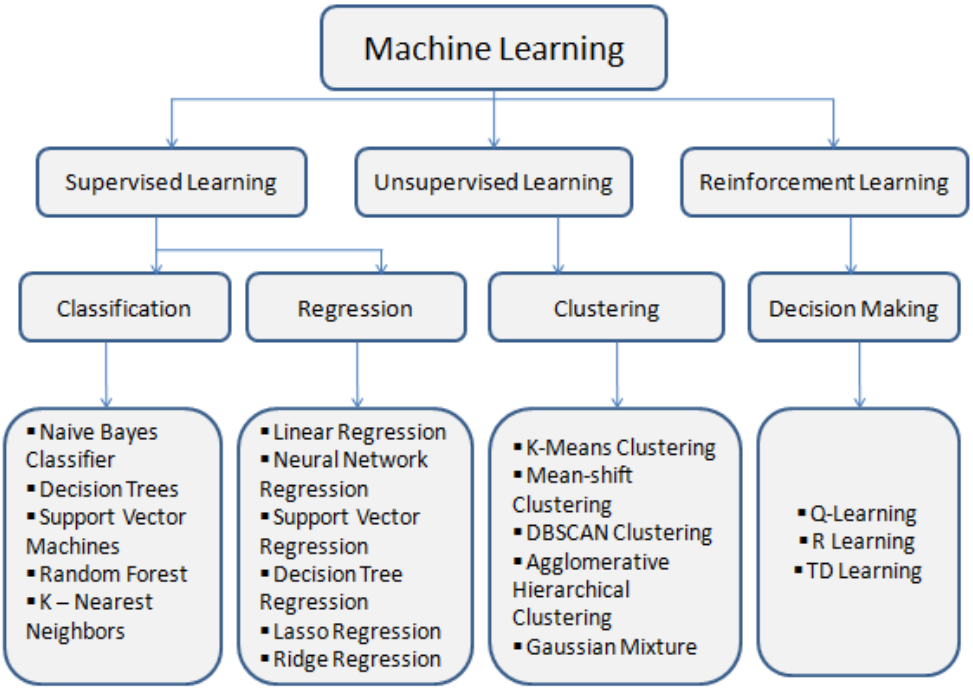

Fig 7: Machine Learning techniques and models[3]

The most widely used predictive models are:

- **Decision trees:**
  Decision trees are a simple, but powerful form of multiple variable analysis. They are produced by algorithms that identify various ways of splitting data into branch-like segments. Decision trees partition data into subsets based on categories of input variables, helping you to understand someone's path of decisions.

- **Regression (linear and logistic):**
  Regression is one of the most popular methods in statistics. Regression analysis estimates relationships among variables, finding key patterns in large and diverse data sets, and how they relate to each other.

- **Neural networks:**
  Patterned after the operation of neurons in the human brain, neural networks are a variety of deep learning technologies. They're typically used to solve complex pattern recognition problems – and are incredibly useful for analyzing large data sets. They are great at handling nonlinear relationships in data – and work well when certain variables are unknown

Other classifiers:

- **Time Series Algorithms:** Time series algorithms sequentially plot data and are useful for forecasting continuous values over time.

- **Clustering Algorithms:** Clustering algorithms organize data into groups whose members are similar.

- **Outlier Detection Algorithms:** Outlier detection algorithms focus on anomaly detection, identifying items, events, or observations that do not conform to an expected pattern or standard within a data set.

- **Ensemble Models:** Ensemble models use multiple machine learning algorithms to obtain better predictive performance than what could be obtained from one algorithm alone.

- **Factor Analysis:** Factor analysis is a method used to describe variability and aims to find independent latent variables.

- **Naïve Bayes:** The Naïve Bayes classifier allows us to predict a class/category based on a given set of features, using probability.

- **Support vector machines:** Support vector machines are supervised machine learning techniques that use associated learning algorithms to analyze data and recognize patterns.

These machine learning methods are typically employed in predictive analysis. Despite the differences between each technique, the basic idea of training never changes. The model is designed in a way that enables it to recognise trends in the dataset and forecast future values in accordance with those trends.

3.2 Learning about different machine learning models:

The problem statement is solely dependent on the time. The water level varies drastically in the monsoon months (July - Sept), hence it can be seen that the water level varies with date and time. Also for we can see that there is a formation of clusters of some some values as the time changes.

With this conclusion we come to a learning that we can use either a clustering algorithm or a time series algorithm.

In this project I will be implementing 3 algorithms on the given dataset and finally will analysis them on the basis of their accuracy. The algorithms are as follows:

- Linear Regression
- K means Clustering
- Arima model

Linear Regression - A variable's value can be predicted using linear regression analysis based on the value of another variable. The dependent variable is the one you want to be able to forecast. The independent variable is the one you're using to make a prediction about the value of the other variable.

With the help of one or more independent variables that can most accurately predict the value of the dependent variable, this type of analysis calculates the coefficients of the linear equation. The differences between expected and actual output values are minimised by linear regression by fitting a straight line or surface. The best-fit line for a set of paired data can be found using straightforward linear regression calculators that employ the "least squares" technique. Then, using Y(independent variable), you estimate the value of X (the dependent variable) .

In the provided dataset, dependent variable is the Water discharge which is dependent on the independent variable which is Water level which changes with date. Therefore, X will be Water discharge and Y will be Water Level for us.

K means Clustering - K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids.

In out dataset, the data creates clusters for monsoon seasons and for non-monsoon seasons. In the monsoon seasons the data gives very low values whereas for monsoon months it gives very large values. This can form clusters through the k - means algorithm.

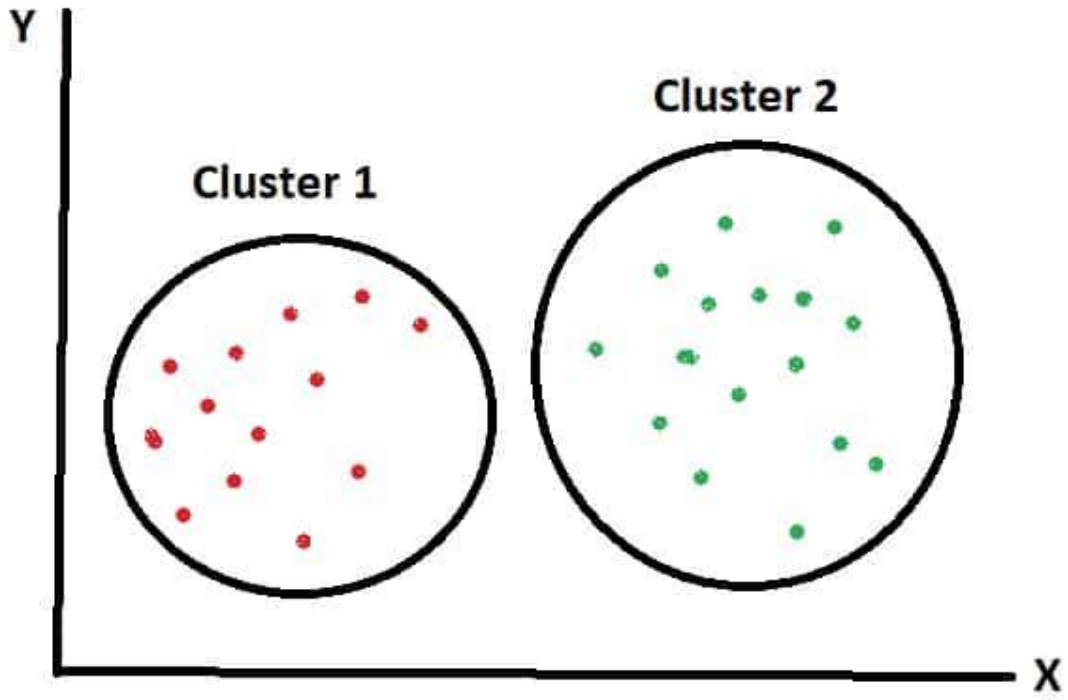

Fig 8: Example of k means clustering

ARIMA model - Autoregressive Integrated Moving Average (ARIMA) models have many uses in many industries. It is widely used in demand forecasting, such as in determining future demand in food manufacturing. That is because the model provides managers with reliable guidelines in making decisions related to supply chains. ARIMA models can also be used to predict the future price of your stocks based on the past prices.[4]

ARIMA models are a general class of models used for forecasting time series data. ARIMA models are generally denoted as ARIMA (p,d,q) where p is the order of autoregressive model, d

is the degree of differencing, and q is the order of moving-average model. ARIMA models use differencing to convert a non-stationary time series into a stationary one, and then predict future values from historical data. These models use "auto" correlations and moving averages over residual errors in the data to forecast future values.

- **AutoRegressive - AR(p)** is a regression model with lagged values of y, until p-th time in the past, as predictors. Here, p = the number of lagged observations in the model, ε is white noise at time t, c is a constant and φs are parameters.

$$\hat{y}_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \epsilon_t$$

- **Integrated I(d) -** The difference is taken d times until the original series becomes stationary. A stationary time series is one whose properties do not depend on the time at which the series is observed.

$By_t = y_{t-1}$ where B is called a backshift operator

Thus, a first order difference is written as

$$y'_t = y_t - y_{t-1} = (1 - B)y_t$$

In general, a *d* th-order difference can be written as

$$y'_t = (1 - B)^d y_t$$

- **Moving average MA(q)** - A moving average model uses a regression-like model on past forecast errors. Here, ε is white noise at time t, c is a constant, and θs are parameters.

$$\hat{y}_t = c + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}$$

Combining all of the three types of models above gives the resulting ARIMA(p,d,q) model.

CHAPTER 4

DATASET PREPROCESSING

4.1 Dataset Preprocessing:

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.[5]

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

It involves below steps:

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data

For this project, Google Colab is being used. Colaboratory, or “Colab” for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education.

Getting the dataset - Dataset was provided by the organization from Kiru project. The dataset consisted of 4 columns, viz, Date, Time, Water level, Discharge.

Importing the libraries - Python libraries are collections of modules that contain useful codes and functions, eliminating the need to write them from scratch. Initially for preprocessing, following libraries are imported in Google collab:

- Numpy - NumPy is a popular Python library for multi-dimensional array and matrix processing because it can be used to perform a great variety of mathematical operations. Its capability to handle linear algebra, Fourier transform, and more, makes NumPy ideal for machine learning and artificial intelligence (AI) projects, allowing users to manipulate the matrix to easily improve machine learning performance.

- Pandas - Pandas is another Python library that is built on top of NumPy, responsible for preparing high-level data sets for machine learning and training. It relies on two types of data structures, one-dimensional (series) and two-dimensional (DataFrame).

- Sklearn - Scikit-learn is a very popular machine learning library that is built on NumPy and SciPy. It supports most of the classic supervised and unsupervised learning algorithms, and it can also be used for data mining, modeling, and analysis.

Importing dataset - After importing the python libraries, pandas library is used to import the dataset into the collab file. The dataset should either be in csv format (.csv) or in excel format (.xlsx).

Finding missing data - The problem of missing value is quite common in many real-life datasets. Missing value can bias the results of the machine learning models and/or reduce the accuracy of the model. After getting the data description from fig 8, we can conclude that Discharge column contains missing values as the total values in the dataset are 33128 and discharge column contains 33103 rows of data only.

```
data.describe()
```

	Water_Level(m)	Discharge_Through_Rating_Curve(cumecs)
count	33128.000000	33103.000000
mean	2840.759705	75.286291
std	79.517391	116.500745
min	0.000000	3.103300
25%	2842.500000	7.922500
50%	2842.500000	7.922500
75%	2843.700000	127.883900
max	2847.300000	595.212500

Fig 9: Dataset description

Also from the data description we can get information about the max, min, standard deviation, and 25, 50 and 75th percentile of the data. Here, for finding total missing values, I simpply used the code “data.isna().sum()” which gave me the following results.

```
Date                                  0
Time                                  0
Water_Level(m)                        0
Discharge_Through_Rating_Curve(cumecs)    25
dtype: int64
```

Fig 10: Missing value count

The missing values were dropped by using “data.dropna()” as there were very less number of missing values present as compared to the dataset points.

Encoding categorical data - Encoding categorical data is a process of converting categorical data into integer format so that the data with converted categorical values can be provided to the

different models. All the data in this dataset is already in numerical form which can be understood by the models for training and prediction purposes.

For smooth prediction I preprocessed the data by calculation the average of data given on different time of a particular date. By doing this we can have one value for a day instead of having multiple values at multiple periods.

```
d = []
for i in dates:
  d.append(data['Water_Level(m)'].where(data['Date'] == i).mean())

d
```

```
 2844.5507142857145,
 2844.6283687943264,
 2844.543356643357,
 2844.384285714286
```

```
e = []
for i in dates:
  e.append(data['Discharge_Through_Rating_Curve(cumecs)'].where(data['Date'] == i).mean())

e
```

```
 364.54197142857146,
 396.25895390070923,
 361.07700139860134,
 304.2248085714286,
 276.79468680555556
```

Fig 11: Finding avg and creating atomic entrys for each day

After the above implementation, the dataset received was as follows:

	Date	Water Level Avg	Discharge Avg
0	13.01.2022	2842.510000	8.279460
1	14.01.2022	2842.504167	8.071233
2	15.01.2022	2842.537500	9.261100
3	16.01.2022	2842.508333	8.219967
4	17.01.2022	2842.508333	8.219967
...	...	...	...

Fig 12: Final dataset after preprocessing

Final dataset contains data for 229 days (rows) and 3 columns as date, water level avg, discharge avg.

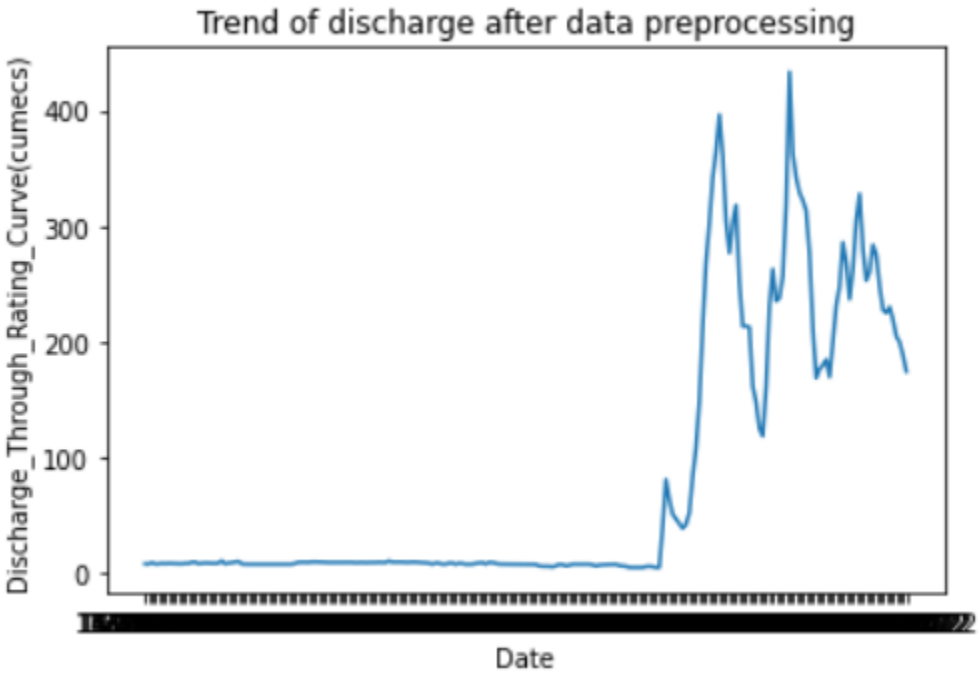

Fig 13 : Graph of trend of water discharge after data preprocessing

The trend can now be seen clearly after the data preprocessing.

After the completion of data cleaning and pre processing, final step of data splitting arrives. Data splitting is when data is divided into two or more subsets. Typically, with a two-part split, one part is used to evaluate or test the data and the other to train the model. Data splitting is an important aspect of data science, particularly for creating models based on data. This technique helps ensure the creation of data models and processes that use data models -- such as machine learning -- are accurate.

In a basic two-part data split, the training data set is used to train and develop models. Training sets are commonly used to estimate different parameters or to compare different model performance.

The testing data set is used after the training is done. The training and test data are compared to check that the final model works correctly. With machine learning, data is commonly split into three or more sets. With three sets, the additional set is the dev set, which is used to change learning process parameters.

The data for this dataset is split into a 70:30 ratio, ie, 70% data will be used for training the model and the rest 30% will be used for testing.

CHAPTER 5

IMPLEMENTATION

5.1 Implementation of different models on the dataset

Time Series Plot is used to observe various trends in the dataset over a period of time. In such problems, the data is ordered by time and can fluctuate by the unit of time considered in the dataset (day, month, seconds, hours, etc.). When plotting the time series data, these fluctuations may prevent us to clearly gain insights about the peaks and troughs in the plot. So to clearly get value from the data, we use the rolling average concept to make the time series plot.

The rolling average or moving average is the simple mean of the last ‘n’ values. It can help us in finding trends that would be otherwise hard to detect. Also, they can be used to determine long-term trends. You can simply calculate the rolling average by summing up the previous ‘n’ values and dividing them by ‘n’ itself. But for this, the first (n-1) values of the rolling average would be Nan.

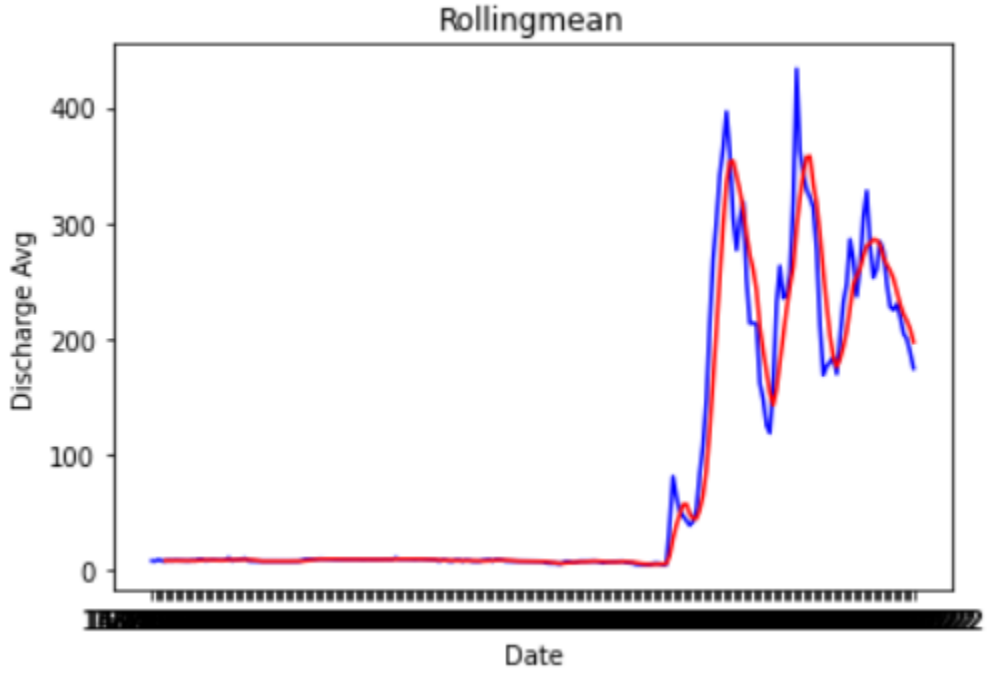

Fig 14: Rolling mean graph

In this graph it can be clearly seen that there exist a trend. Initially the graph is quite smooth and have low values but as it approaches the monsoon months the discharge grows.

As mentioned above, 3 algorithms will be tested in this section on the given dataset:

- Linear Regression
- K means clustering
- ARIMA model

Linear Regression:

Regression is a technique for investigating the relationship between independent variables or features and a dependent variable or outcome. It's used as a method for predictive modeling in machine learning, in which an algorithm is used to predict continuous outcomes.

Solving regression problems is one of the most common applications for machine learning models, especially in supervised machine learning. Algorithms are trained to understand the relationship between independent variables and an outcome or dependent variable. The model can then be leveraged to predict the outcome of new and unseen input data, or to fill a gap in missing data.

Regression analysis is an integral part of any forecasting or predictive model, so is a common method found in machine learning powered predictive analytics. Alongside classification, regression is a common use for supervised machine learning models. This approach to training models required labeled input and output training data. Machine learning regression models need to understand the relationship between features and outcome variables, so accurately labeled training data is vital.[6]

To check if regression can be implemented, we need to check the relation between the dependent and independent variables. Correlation can be used for the same The most commonly used techniques for investigating the relationship between two quantitative variables are correlation and linear regression.

Correlation quantifies the strength of the linear relationship between a pair of variables, whereas regression expresses the relationship in the form of an equation.Correlation determines if two variables have a linear relationship while regression describes the cause and effect between the two. As the correlation is very low for this data regression can't be used.

```
data['Water Level(m)'].corr(data['Discharge Through Rating Curve(cumecs)'])
```

```
0.04236314775632064
```

Fig 15: Correlation

K Means Clustering:

K-Means clustering is one of the unsupervised algorithms where the available input data does not have a labeled response. Clustering is a type of learning wherein data points are grouped into different sets based on their degree of similarity. There is no labeled data for this clustering, unlike in supervised learning. K-Means performs the division of objects into clusters that share similarities and are dissimilar to the objects belonging to another cluster.

The term ‘K’ is a number. You need to tell the system how many clusters you need to create. For example, K = 2 refers to two clusters. There is a way of finding out what is the best or optimum value of K for a given data.

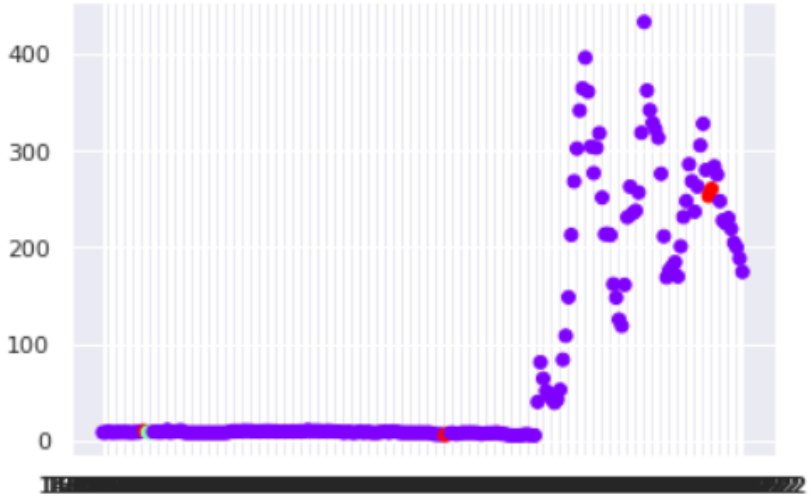

Fig 16: Discharge Avg to date scatter plot of clustering

```
model.fit(x_scaled)
```

```
KMeans(n_clusters=25, random_state=32)
```

```
from sklearn.metrics import silhouette_samples, silhouette_score
silhouette_score_average = silhouette_score(x_scaled, model.predict(x_scaled))
print(silhouette_score_average)
```

```
0.6067297369026806
```

Fig 17: Silhouette score avg

The silhouette coefficient or silhouette score kmeans is a measure of how similar a data point is within-cluster (cohesion) compared to other clusters (separation). The score here represents that 60% resemblance is there in one point to other but that is not enough in our case as we have taken the cluster value as 25 as well.

ARIMA Model:

Here, we will be using ARIMA model to predict the discharge wrt water level. ARIMA is a form of regression analysis that indicates the strength of a dependent variable relative to other changing variables. The final objective of the model is to predict future time series movement by examining the differences between values in the series instead of through actual values. It is generally used for short term prediction of stock market.

It's a model used in statistics and econometrics to measure events that happen over a period of time. The model is used to understand past data or predict future data in a series.

```
print('Test RMSE: %f' % rmse)
```

```
Test RMSE: 34.708550
```

Fig 18: Test RMSE

As the score of ARIMA Model is highest, I will be using this model as my final model, I have explained the model building and its outcomes in the next section.

5.2 Comparing accuracy of models

The comparison of accuracy of the models can be seen below:

Algorithm/Model	**Accuracy Score**
Linear Regression	0.0426
K means clustering	0.6067
ARIMA model	34.7085 (RMSE)

Table 3: Accuracy comparison table

According to the comparison ARIMA model will be the best suited model for the prediction of water discharge.

CHAPTER 6

FINAL SELECTION

6.1 Final model selection after comparison

ARIMA, short for 'Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

Figure given below is the summary of the ARIMA model for the given dataset.

SARIMAX Results

Dep. Variable:	Discharge Avg	No. Observations:	229
Model:	ARIMA(1, 0, 0)	Log Likelihood	-1005.286
Date:	Sun, 18 Dec 2022	AIC	2016.571
Time:	04:49:55	BIC	2026.872
Sample:	0	HQIC	2020.727
	- 229		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	77.6938	138.116	0.563	0.574	-193.009	348.397
ar.L1	0.9838	0.012	82.702	0.000	0.960	1.007
sigma2	374.9443	16.375	22.898	0.000	342.850	407.038

Ljung-Box (L1) (Q):	37.11	Jarque-Bera (JB):	755.73
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	312.88	Skew:	1.26
Prob(H) (two-sided):	0.00	Kurtosis:	11.54

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Fig 19: ARIMA model summary

We are using the order of (1,0,0) for the model. ARIMA(1,0,0) = first-order autoregressive model: if the series is stationary and autocorrelated, perhaps it can be predicted as a multiple of its own previous value, plus a constant.

6.2 Prediction of water discharge

In the fit() method, where the required formula is used and perform the calculation on the feature values of input data and fit this calculation to the transformer. For applying the fit() method (fit transform in python) .fit() is used in front of the transformer object. For changing the data, transformation is done, in the transform() method, where the calculations are applied that were calculated in fit() to every data point in feature F. The .transform() method is used in front of a fit object because we transform the fit calculations.

After fitting and transforming the model, we get the prediction and expected results as shown in the figure given below:

```
predicted=43.580071, expected=43.149316
predicted=39.994238, expected=38.936662
predicted=36.130157, expected=42.273621
predicted=39.515315, expected=52.411177
predicted=49.762386, expected=83.600653
predicted=82.273052, expected=108.341855
predicted=107.681940, expected=148.200524
predicted=147.739460, expected=212.762415
predicted=212.316640, expected=268.240647
predicted=267.825839, expected=302.231219
predicted=301.854783, expected=341.543342
predicted=341.191786, expected=364.541971
predicted=364.211283, expected=396.258954
predicted=395.945890, expected=361.077001
predicted=360.704308, expected=304.224809
predicted=303.639861, expected=276.794687
predicted=276.104139, expected=302.967034
predicted=302.347066, expected=318.058511
```

Fig 20: ARIMA model predictions

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

In our case, RMSE turns out to be 34. The model as shown above give acceptable predictions. Following graph shows us the plot between the test values and the redicted values. The graph almost fits the line and hence its output can be accepted.

Red line shows the predicted values plot and the blue line shows the plot for test values.

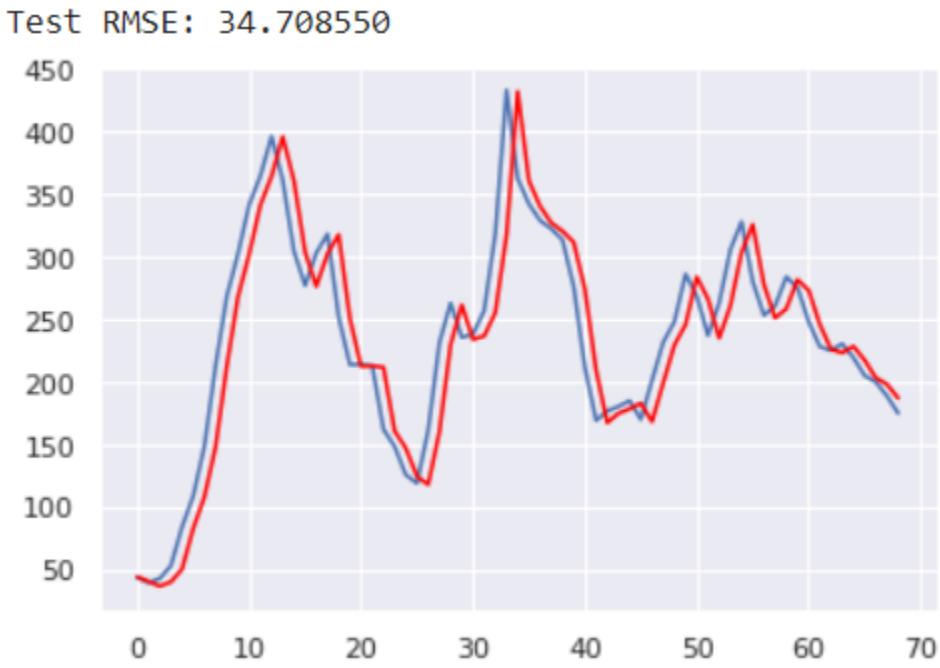

Fig 21: Final graph after prediction of test and prediction values

CHAPTER 7

RESULT

7.1 Result:

As a result it can be stated that the ARIMA model will be best for the prediction of water discharge. The final graph formed from the original and predicted data is very similar. The ARIMA model is a good implementation of the time series problem. For further prediction, the model can predict values if there is an input given and the user will receive output accordingly. The output will depend on the date and water level of the dam at that particular day.

7.2 Final Submission of project:

Final project was presented and accepted by the mentor. The project was build on ARIMA model, and required data cleaning and preprocessing was done the dataset provided on Kiru project of Kishtwar, Jammu and Kashmir. The project was submitted in the organization and was approved by the GM IT and SM IT.

Fig 22: Kiru project, NHPC

7.3 Final submission of report:

The final report describes the activities that have been carried out and the results achieved over the entire project period. You are also to state whether the project has achieved the results set out in the grant proposal.

My sincere efforts made me accomplish the task of completing the internship. I have taken efforts in this project however, it would not have been possible without the kind support and help of many individuals.

I submitted the report and want to specifically thank the Training and Human Resources Department for providing me with the wonderful opportunity to do an internship at the Information Technology Department of NHPC Ltd. Additionally, I want to thank the IT Department for providing me with the opportunity to work on the fantastic Prediction of Water Discharge project and for mentoring me throughout the internship.

REFERENCES

1. Rainfall yearly record of Kishtwar
   https://tcktcktck.org/india/jammu-and-kashmir/kishtwar

2. Weather, Kishtwar
   https://weatherspark.com/y/108425/Average-Weather-in-Kishtw%C4%81r-India-Year-Round

3. ML Models
   https://www.analyticsvidhya.com/blog/2021/03/everything-you-need-to-know-about-machine-learning/

4. ARIMA model
   https://www.capitalone.com/tech/machine-learning/understanding-arima-models/

5. Data Preprocessing
   https://www.javatpoint.com/data-preprocessing-machine-learning

6. Regression
   https://www.seldon.io/machine-learning-regression-explained#:~:text=Regression%20is%20a%20technique%20for,used%20to%20predict%20continuous%20outcomes.