**RAINFALL PREDICTION IN INDIA USING LINEAR REGRESSION**

School

Department

Course

Student Name

Project supervisor

# **Declaration**

Name:

Registration No:

Signature…………………….

Date………………………….

**Supervisor details**

Name……………………………….

Signature: ………………………………

Date……………………………….

# **Dedication**

I dedicate my project to everyone in my family who has helped me succeed academically by providing encouragement and financial assistance. Moreover, I would want to convey my heartfelt gratitude to my friends and instructors for their generosity when I was in my academic years. I'm grateful for their unwavering support and prayers for my university studies.

Thank you for your support.

May the almighty God bless them.

# **Acknowledgement**

First and foremost is to thank God for his mercies and grace during the academic journey at the university.

My acknowledgements go to my parent for their continued great support, both morally and financially, during my learning period at the university. I would also like to thank my friends and group mates who have helped me compile this project report. I also acknowledge the support from my supervisor, who has been there for my consultation and given me guidelines on how to start and finish this project.

I thank the university fraternity for ensuring that I have acquired adequate skills and expertise in my course. The school has provided all the facilities required to complete my course.

GOD BLESS YOU.

# **Abstract**

As a means of predicting rain in India, this study provides the most modern machine-learning-based models. Rainfall is always a huge problem around the world since it impacts all of the primary factors on which humans rely. Predicting rain in the current climate is difficult because it is both unpredictable and imprecise. SVMs, Navie Bayes, Logistic Regression, Random Forests, and Multilayer Perceptrons are some of the machine learning methods we use to analyze rainfall data in India (MLP). Because we're interested in getting the best possible result and a more accurate forecast of rainfall, this is why we're doing it.

Table of Contents

[**Declaration** 2](#_Toc99713319)

[**Dedication** 3](#_Toc99713320)

[**Acknowledgement** 4](#_Toc99713321)

[**Abstract** 5](#_Toc99713322)

[**CHAPTER ONE - Introduction** 8](#_Toc99713323)

[Problem statement 10](#_Toc99713324)

[**CHAPTER TWO - Literature Review** 16](#_Toc99713325)

[**CHAPTER THREE - RESEARCH METHODOLOGY** 30](#_Toc99713326)

[Introduction 30](#_Toc99713327)

[Development Methodology 30](#_Toc99713328)

[Requirement 31](#_Toc99713329)

[Software Design 31](#_Toc99713330)

[Development and Coding 32](#_Toc99713331)

[Integration and Testing 33](#_Toc99713332)

[Implementation and Deployment 33](#_Toc99713333)

[Review 33](#_Toc99713334)

[Data collection methods 33](#_Toc99713335)

[Summary 34](#_Toc99713336)

[**CHAPTER FOUR - MATERIALS AND METHODS** 35](#_Toc99713337)

[Proposed solution 35](#_Toc99713338)

[**Machine Learning** 35](#_Toc99713339)

[**SVM** 36](#_Toc99713340)

[**Navie Bayes** 37](#_Toc99713341)

[**Random Forest** 38](#_Toc99713342)

[**MLP (MULTILAYER PRECEPTRON)** 39](#_Toc99713343)

[**CHAPTER FIVE - Result and discussion** 43](#_Toc99713344)

[Train dataset: 44](#_Toc99713345)

[Summary and Conclusion 44](#_Toc99713346)

[RECOMMENDATIONS 45](#_Toc99713347)

[Future work 45](#_Toc99713348)

[Appendix: Tools 45](#_Toc99713349)

[REFERENCE 46](#_Toc99713350)

**CHAPTER ONE - Introduction**

With the rise of global warming, climate change has been accelerated and has a profound effect on humanity. As a result, sea levels rise, floods become more frequent, and other climate-related issues arise. Climate change has a negative impact on rainfall, and this is one of the more serious consequences. Predicting rainfall has become increasingly difficult in recent years, according to the world's finest meteorologists. Rainfall has a direct impact on a wide range of human activities, including agriculture, construction, power generation, and tourism (Luo et al., 2018). There is a need for better rainfall forecasting because of this. As a result of climate change, predicting rainfall becomes increasingly challenging. As a result of its arbitrary features, rainfall series are typically categorized as stochastic. Floods and droughts are becoming more common, as evidenced by Uttarakhand, India, suffering the worst natural calamity in June of 2013. Around 400 percent more rain fell than usual during the monsoon season. More than 100,000 pilgrims and visitors were stranded during their "Char Dham Yaatra" as a result of the heavy rain. Prior to the disaster, the government, major industries, risk management authorities, and the scientific community had no idea what was about to happen, and no one predicted it. Land slides, another important geohazard, have resulted in human deaths and considerable property damage around the world as a result of all of these factors (Raimi et al., 2021).

Researchers and engineers have been developing models for decades that can be used to predict various outcomes with high accuracy. Prediction and categorization can both be accomplished through the use of machine learning. If you're interested in learning more about the many methods that can be used, you can check out the following: Self-adaptive approaches are used by ANNs for metrology projections, where instances and functional ties between data, even if their relationships are unknown or difficult to articulate, are learned from and captured by the system.

As an ANN technique for dealing with enormous amounts of data and difficult issues, deep learning has just lately emerged as a viable option [5]. It is essentially a series of structures that are taught. The weight and layer learning rate are the two most significant changes applied to this model. Bioinformatics, computer vision, and language processing all benefit greatly from the widespread use of deep learning (Hariri et al., 2019).

Our experiment relies on data from the Indian government's official website. All over India, rainfall data has been collected for more than a decade. Accurate forecasting of rainfall is becoming increasingly important as the world's attention turns to the pressing issue of water scarcity. As a result, our research aims to develop a model that can accurately forecast rainfall in India.

Our research relies on rainfall data obtained from the Indian government's official website. The rainfall observations in Asia have been taking place for the better part of a decade. Rain forecasting is critical in Asian countries, where water scarcity is a major concern (Greve et al., 2018). Our goal here is to maximize the findings and select the best model for rain forecasting in Asia only.

Every person on the planet is affected by climate change, which is a global issue. People all throughout the world are suffering as a result of this climate change. There was a movement from regional to global climate change from the middle to the end of the 20th century, which is referred to as "climate change". CO2 emissions from burning fossil fuels are a key contributor to the pollution that is contributing to the rise in global temperatures (Nordhaus, 2019). Known as global warming, climate change is going place. Earth's average surface temperature has risen as a result of global warming. As a result, people all across the world are experiencing shifts.

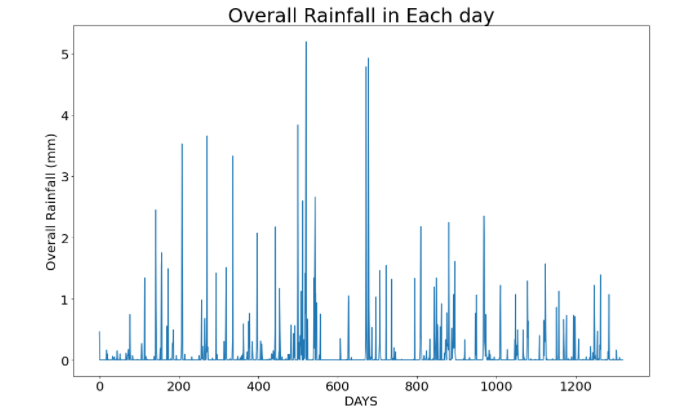
## Problem statement

The weather plays a vital role in our daily lives. As a result, the Prediction must be as precise as feasible. Rainfall forecasting is an important part of human life since it provides the most important resource needed for human survival, fresh water. Every human being needs fresh water for everything from drinking to farming to cleaning, and it's constantly critical.

It is increasingly difficult to anticipate the future due of climate change.

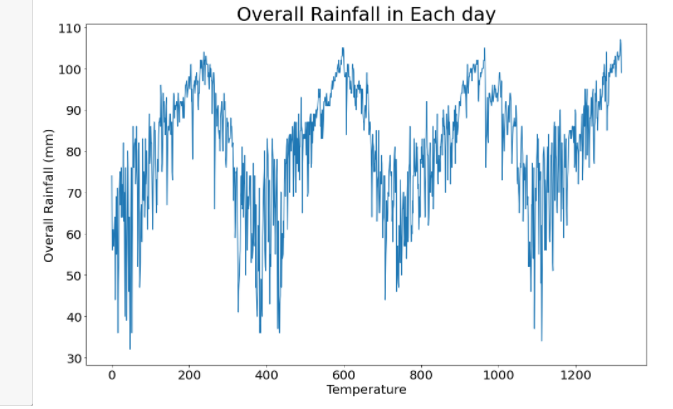
Global warming has risen to the top of the global agenda. People are trying to figure out how climate change affects the economy, from production to infrastructure, and how to detect such patterns (Alves et al., 2020). It's just as difficult to predict rain as it is to forecast rain, yet the accuracy rate is high. Scientists are employing machine learning and deep learning to anticipate rainfall patterns because the old method of predicting rainfall is ineffective.

Farmers, especially frames, are particularly vulnerable to adverse weather forecasts because their entire crop depends on rain, and agriculture is an important element of every economy. As a result, there is a reasonable chance of making an accurate rainfall estimate. There are a variety of machine learning approaches that may be used to predict rainfall, but accuracy is always a challenge (Agamile et al., 2021). Drought, floods, and extreme heat in the summer are just a few of the problems that can be brought on by rain. As a result, water supplies around the planet will be affected. As we can see in figure below there has been a significant decrease in annual rainfall in millimeters. This is our primary concern.



We go into greater depth about India's rainfall on a daily basis. Also, look at the rainfall on a monthly basis. We can see from the graph above that day 500 and day 700 are the wettest days of the year.

Climate change is a major concern for humanity in the modern era. Climate change is having a devastating effect on the lives of people around the world. The term "climate change" refers to the shift from the middle to the end of the 20th century in the focus from regional climate change to global climate change. Attributable to the high pollution level in the atmosphere, which is mostly owing to the increased CO2 concentration in the atmosphere due to fossil fuel use, this is all occurring. Global warming is another name for climate change. The entire planet is altering as a result of the rise in average global surface temperature.



The temperature has been slowly rising over the last few decades, as the graph illustrates. Unless all of them are regulated, the climate will get worse, as stated by NASA. NASA keeps a close check on any change in temperature.

This climate change or global warming is largely regarded to be caused primarily by human greenhouse gas emissions.

People's lives are being affected by this climate change, and it is becoming more and more pronounced. As a result, this is affecting all of the areas in which humans rely. In order for a human to survive, he or she needs water, food, and air. However, global warming has impacted all three of these sectors.

Changes in the climate have made it increasingly difficult to plant crops, raise cattle or fish in the same locations as in the past. Agricultural productivity is also affected. Farmers are the primary source of food for humans, yet they are also negatively affected as a result of this. A large number of toxic gases are present in the air, making it dangerous for humans to breathe. In order for humans to survive, they need to be able to inhale and exhale (Cronin et al., 2018).

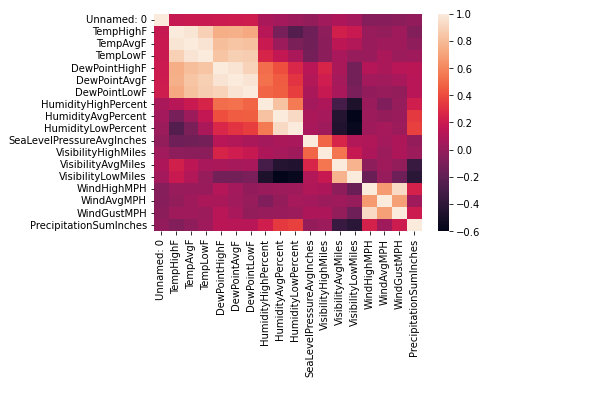
Also, water, which is essential to human survival. However, the availability of fresh water is quickly reducing as a result of climate change. There is no shortage of fresh drinking water in any country. It's not simply the temperature that's shifting. In addition, the water cycle as a whole is harmed as well. As the planet warms, the atmosphere's moisture-holding capacity increases. That's why water vapor concentrations fluctuate as much as rainfall and evaporation rates.

Every human being needs fresh water for everything from drinking to farming to cleaning, and it's constantly critical. Climate change is partly to blame for the decline in its availability in the future.

Only 2% of the water on our planet is fresh enough to drink, with the remaining 99% being salty. Nearly 70% of these 25 originate from snow and ice, 30% from groundwater, and fewer than 5% come from surface water, such as lakes, rivers, and the like, with the remaining 0.05 percent coming from the air we breathe. Rainfall is also a key supply of fresh water in this region, and it is this that fills the lake or river. Because of the rain, the groundwater level rises as well. However, rainfall patterns are changing as a result of global warming. As a result of severe drought, many people over the world have been compelled to leave their homes and countries or regions in order to find other areas to live where they can access clean water in order to continue their lives. Water cycles are also impacted by this pollution. The more moisture the atmosphere can hold as the planet warms. In other words, the amount of water vapour, rainfall, and air currents vary.

Climate migration has emerged as a new problem in the world as a result of this, in which people are compelled to relocate to find water.

Most of India is experiencing a drought. As a result, farmers were unable to carry out their work, as water is an essential aspect of farming. On this page you can see a graph of recent rainfall in India.



Each day`s rainfall in India is depicted as a heat map in the graph above. As can be seen, the amount of rain that falls in India varies greatly. Rainfall forecasts can help improve resource management.

This study covers a wide range of topics. People are currently dealing with serious issues as a result of climate change. If we are able to accurately predict the weather, it will be extremely beneficial to the entire human race.

# **CHAPTER TWO - Literature Review**

A variety of Machine Learning techniques were proposed by Steve Oberlin for Big Data processing. Complex and powerful data sets were analyzed using Machine Learning and other AI approaches. Machine Learning is employed in Netflix's recommendation algorithms, which analyze the audience's ratings and preferences. Human language is analyzed and portrayed by IBM which uses a variety of Machine Learning methods. He use linear regression, data manipulation, Perception, and k-means to reveal links and patterns in data. The nature of the prediction is the primary consideration when selecting a Machine Learning algorithm. Some predictions are estimates, while others are categorical. The complexity of the algorithm and the consequent rise in computational demands were also discussed (Noor et al., 2020).

Singh, Jainender, developed a machine learning technique that could provide promising outcomes for addressing security vulnerabilities in software, its technology, and theories. Using machine learning methods like Support Vector Machine (SVM), Nave Bays classifiers, and clustering approaches, he stressed the need of mining sparse, partial, and ambiguous data to produce optimal results(Deka et al., 2019). It would help people in a variety of professions, including health care, education, and trade.

Some of the most recent improvements in Machine Learning have been proposed by Junfei Qiu. New advanced learning method called "Representation Learning," which makes use of useful information extracted from data when building classifiers and predictions. By gathering as much data as possible, it seeks to improve computing and statistical efficiency. Representation learning includes feature selection, feature extraction, and metric learning. Biological DNA identification and image categorization are two examples of big data applications that use active learning, a more advanced form of machine learning. Users are tasked with retrieving data from a limited number of key labelled instances, therefore reducing costs and enhancing accuracy while delivering optimal outcomes by using semi-supervised machine learning techniques. Machine learning for Big Data processing was also mentioned by him. Big Data raises a slew of issues, including the unpredictability and incompleteness of the data it generates, as well as its sheer size. On top of that, he offered cures. Parallel and distributed big data processing using ADMM (alternating direction methods of multipliers) looks promising. It efficiently divides the many variables, making it easier to discover a solution to a big amount of data. High-speed data has necessitated the development of the Extreme Learning Method, providing a faster learning speed, outstanding performance, and minimal human intervention.

If you're looking for something in an enormous collection of unstructured text, k-means clustering can help. Yasir Safeer presented this algorithm. He came up with a method for depicting papers that would enhance the clustering results. The K-means method was used to build a more accurate representation of document content for the algorithm. In addition, he suggested using a domain dictionary to get more accurate document similarity results.

Robotics and Big Data Processing play an important role in Roheet Bhatnagar's presentation on Machine Learning (BDA). Machine learning and big data analytics are mutually supportive fields of study. He spoke about the future of Machine Learning for Big Data, and how it will evolve. 'Data Meaning' refers to methods for improving the intelligence of machine learning so that it can better understand text or data. Technique integration is a growing trend in data processing and integration. Machine learning techniques such as classification, regression, and cluster analysis can be used to undertake analyses and make predictions about the future based on current trends. determine if there is any relationship between the given sets of data

Using Machine Learning Algorithms, Alexandra L'Heureux showed new techniques of handling Big Data. Traditional tools are no longer equipped to handle the storage, transfer, or efficiency of Big Data. An important part of Data Analytics is Machine Learning, which has the ability to learn from data and deliver insights, predictions, and decisions that are driven by that information. Computational efficiency would be impossible because of the massive expansion in size, space, and time complexity of Support Vector Machines (SVM). For example, the Curse of Modularity occurs when an algorithm's borders collapse as data size increases. Map Reduce is a solution to this. Parallel processing of big data volumes is made possible by this customizable and scalable approach. It employs an iterative strategy. The Curse of modularity can also be solved using K-means. Machine Learning paradigms are used to address the efficiency gaps created by Big Data. It makes it easier to deal with vast amounts of data. It can deal with unclean and noisy data because of its adaptability.

In order to link Type 2 Diabetes Mellitus to a specific genotype and phenotype, the proposed study relies on Zheng and Tao's findings (T2DM). Feature engineering and machine learning are employed in suggested framework for the accurate diagnosis, detection, and therapy of T2DM patients. In order to make the comparisons, a variety of learning models, including SVM, K-Nearest Neighbors (KNN), Decision Tree-J48, and Random Forest, were used (RF). The WEKA package is used on the various models' engineered features, as previously indicated. Every learning model is cross-validated four times and their average and standard deviations are presented. Performance indexes for the three levels of feature selection (J48 SVM and RF) are all over 95% for J48 and SVM respectively. While SVM and RF have a 98% and 98% accuracy respectively, LR is the most accurate with a 99%. KNN and NB are not affected by sparsity or noise in the records database because they are not affected by the three levels of feature selection as are KNN and LR.

With six distinct methods, including Linear Discriminant Analysis, Generalized Linear Model, Recursive Trees, SVM, and Nave Bayes, the SVM is a powerful tool for analyzing data sets, Madeeh Nayer Algedway created a model for diagnosing diabetes in patients. A GLM wholistic classifier generates a confusion matrix, which is used to construct four performance metrics such as accuracy (accuracy), precision (precision), recall (recall), and f-measure (f-measure). Stacking ensembles can be used to achieve a stronger prediction model for greater accuracy and f-measure. Each of the six algorithms should be run separately for a different performance model measure, such as f-measure, accuracy, precision or recall. All six algorithms do cross-validation using 10-cross-validation. A stack of algorithms, including LDA, KNN and the Recursive tree, yields 96.4 percent accuracy and an f-measure of 0.956%. [45] When used in conjunction with other feature selection strategies, the KNN's shortcomings have been shown to be overcome. Stacking ensemble with three algorithms is superior to other models, according to the study's findings, models including support vector machines, generalized linear models, and naive bayes can accurately predict diabetes in a patient.

In the suggested study, Huang, Guan-Mau, demonstrated that diabetic nephropathy can be detected in type 2 diabetes patients. Diabetes, excessive cholesterol, and hypertension all contribute to diabetic nephropathy, a kidney disease. Huang, Guan-Mau et alproposed .'s work claims that a decision tree-based model that incorporates genotype and clinical data from 345 type 2 diabetic patients can detect renal impairment at an early stage. Data are first categorised by gender, and then comparison techniques such as Naive Bayes (which uses a five-fold cross-validation strategy) are tested against each other for accuracy. The comparison between WEKA and LibSVM yielded the results presented below. Comparing performance between individuals and combinations of clinical variables is one of the four categories of results cited in the paper. Individual and genetic traits can be compared to see how each performs in comparison. integrating clinical and genomic data to assess performance. Diabetes nephropathy training dataset gender-based performance in classification decisions trees. 5 fold cross-validation method using decision tree classification algorithm yields accuracy, specificity and sensitivity of 85.27 percent, 83.32% and 85.24% respectively. Finally, whereas using genetic and clinical factors alone yields poor results, combining the two yields a markedly improved outcome.

Following a K-Means clustering process for outlier detection, a Genetic Algorithm (GA) was used to pick features, and a Support Vector Machine (SVM) was employed to categorize the Pima Indian Diabetes dataset from the UCI repository. SVM's 10-fold cross validation technique uses the Genetic Algorithm's selection tools and output to classify data. It is possible to obtain 98.82% accuracy with SVM and K-Mean clustering, and a Genetic Algorithm model. The following conclusion is drawn from the proposed research method: For the SVM, the lowest and highest classification accuracy are 98.43 and 99.21 percent, while the average degree of accuracy is 98%. (a) K-means outlier identification percentage is 33.46 percent, out of 768 occurrences, 511 samples were picked and 257 samples were considered outliers. 3 is the minimum and 6 is the maximum number of attributes that can be specified. Data on pregnancies, PG concentration, and age are all important considerations for putting together this dataset. In comparison to the modified K-means algorithm, the SVM classification model shows a 2.08 percent improvement in accuracy.

Based on her findings, Arianna Dagliati suggests that machine learning algorithms be integrated into the data mining pipelines. MOSAIC, an EU-funded research to discover diabetes influence factors, collected around 1,000 patients' electronic health records between 2012 and 2016. This sort of data mining profiling includes the building and validation of predictive models. SVM, LR, NB, and RF are among the classification models that are employed, along with Nave Bayes and NB (RF). The study makes use of a four-stage data mining pipeline. The center profiling method is used to bias selection and specify performance criteria, such as demographics and care practices. According to specialists, diabetes duration and BMI are the main risk factors for retinopathy and neuropathy, whereas high blood pressure is the main risk factor for retinopathy and kidney disease. These findings suggest a potential for combining data mining and machine learning in order to improve disease prediction and treatment.

By employing machine learning ensembles, Santillana, Mauricio, the fundamental contribution of the proposed system is to gather and combine influenza-like disease activities. SVM, SLR, and AdaBoost regression with a decision tree are all employed in the proposed research, as well as a number of other common machine learning techniques. In addition to CDC data, near real-time hospital visit records from Athena Health and Google Trends, as well as influenza-related Twitter microblogging postings, FluNearYou and Google Flu Trends have all been created as weak ILI predictors. There are several methods for determining the weak predictors of performance, including the use of Pearson correlations, RMSEs, MAEs, RMSPEs, and hit rates. "This week," "next week," and "in two weeks" are the three time horizons for which AdaBoost has the best accuracy, lowest RMSE, and lowest MAPE. In order to effectively anticipate influenza flu within one, two, or three weeks, ensemble models are essential, according to this conclusion.

In this proposed investigation, the SVM-FFA algorithm (SVM-FFA) provides the greatest performance for forecasting the transmission of malaria. In Rajasthan, the malaria epidemic is prevalent, resulting in death and disease, and the lack of access to basic health care exacerbates the problem, as suggested by the proposed research study. SVM-FFA, ARMA, ANN, and SVM are four of the model systems that have been developed for the two regions. With the help of MATLAB's LibSVM package and IBM's SPSS program, the proposed models are created. Use of R2 statistics and parameters such as NMSE delivers an accurate prediction of malaria incidences using SVM-FFA for training and testing. When it comes to predicting malarial incidences before they occur, we found that the unique strategy of FFA and SVM is better than existing models. This means that authorities can take better measures for their specific communities and locations.

Using machine learning methods and data mining approaches, 2536 cases of type 2 diabetes were discovered in the Tabriz, Iran, population. SVM, ANN, decision trees, and the k-nearest neighbor algorithm are just some of the classification methods incorporated into the system, and uses 10-fold cross validation in a Bayesian network for disease diagnosis. With 97.5 percent sensitivity and 96.6 percent specificity, an ANN with eight hidden neuron layers achieves 97.18 percent accuracy in terms of performance.

If you are over the age of 55, you are at high risk for cardiovascular and cerebrovascular events (strokes, syncopal events) due to the use of heart rate variability (HRV) analysis. Techniques including Nave Bayes classifier, decision trees, Random Forests, boosting meta-learning (AdaboostM1), SVM, and artificial neural networks (MLP) are employed in the proposed study to perform HRV feature-based prediction (MLP). In the debate, the best results were achieved by C4.5 and AB using chi-squared feature selection algorithms, while the lowest results were achieved by MLP and NB using co-relation based methods. Radio frequency outperformed with an accuracy rate of 85.7%, a sensitivity rate of 71.4%, and a specificity rate of 87.8%. In the suggested system, which also contains SVM as a feature selection method, Random Forest is selected as the best approach.

Several data mining and machine learning approaches combined into the proposed study can uncover biomarkers associated with depression. During the years 2009 and 2010, researchers at the National Health and Nutrition Examination Survey (NHANES) obtained data. Biomarkers in the study are discovered using multiple imputations and machine learning boosted regression and logistic regression. Details on the specifics of the two methods that were employed in this paper are provided in the following sections (MCAR). Since it can handle a wide range of variables and a wide range of likelihood loss functions, the use of boosted regression is useful. This unique hybrid method approach, which combines boosted regression with statistical regression, may considerably benefit large datasets of biomarkers. The multi-step hybrid technique may now be used to deal with missing data and complex survey populations.

An anti-diabetic drug failure prediction is included in Kang, Seokho's suggested work, which takes into account the exponential rise in type 2 diabetes in humans worldwide. SVM, one of the finest approaches for training huge medical datasets, is the approach described in this paper. The new E3-SVM method E3-SVM's proposed research effort is separated into two phases in order to improve the efficacy and ensemble effectiveness of SVM. Data selection parameters, or k values, have been identified as one of the system's most significant flaws. K values have an impact on the selection of data points at various levels, and this impacts the prediction and performance of data The suggested system has found that Bagging performs better when the bootstrap size is minimal. Performance can be improved with even a few minutes of training.

A genuine concern for the risk of diabetic neuropathy, which might result in cardiac arrest when diabetes spreads throughout the body, is expressed in the proposed system by H. F Jelinek. Allan vectors are employed to measure HRV, and ECG recordings are used for machine learning and automatic detection. Diabetes neuropathy can be accurately diagnosed with the Graph-Based Machine Learning Technology (GBMLS), which is part of the system under consideration in this proposal. For testing the results, the Scikit learns were utilized, along with other algorithms such as SVM, CBGF and others that come with the Scikit learning package. Multi-Scale Allen Vector (MAV) and GBLMS are the top classifiers for the GBLMS approach, with high specificity and sensitivity of 98 and 89 percent, respectively, which surpass other classifiers like Random Forest.

As proposed by Khan and Saranjam, the proposed method for diagnosing the Dengue Virus efficiently in 84 patients collected from the Holy Family Hospital in Rawalpindi, Pakistan in the autumn of 2015 incorporates Raman spectroscopy and Support Vector Machine. By measuring the wavelength shift caused by inelastically scattered radiation, Raman Spectroscopy provides information on chemical and structural properties. SVM results and Raman spectra are used to generate a confusion matrix, which is then evaluated using the 10-fold cross-validation procedure. In terms of precision, sensitivity, and specificity, each of the three polynomial kernel functions of order 1 has a better outcome when these factors are taken into account. SVM and Raman Spectroscopy were used in conjunction with various kernels to effectively filter the data and accurately classify all of the features, as demonstrated in the proposed study.

By applying the Support Vector Machine, Kesorn and Kraisak suggest in their paper the implementation of a surveillance system to monitor how dengue and Aedes aegypti mosquito transmission rates compare across similar climates and regions (SVM). Temperature, humidity, wind speed, Ae. aegypti larva infection rate, male mosquito infection rate, and female mosquito infection rate are all examined in the proposed research effort for the dengue epidemic rate between 2007 and 2013. Also taken into account are the population density and mortality rates. It is necessary to complete this method in three steps. There are a number of different types of support vector machines, however the RBF kernel outperforms the other two, including KNN, Decision Tree, and Neural Networks (NN). The accuracy, sensitivity, and specificity of the SVM, as well as its overall performance metrics for prediction and result evaluation, are verified using a 10-fold cross-validation technique. In a 10-fold cross-validation model with two parameters C (the regularization parameter) and 2, the SVM RBF kernel produces values of 8 and 0.1250.

The SVM-RBF kernel outperforms all others with an accuracy of 96.296 percent, a sensitivity of 87.47 percent, and a specificity of 87.47 percent.

Twitter microblogs are used to detect early epidemics of influenza-like illness (ILI) in Guido Zuccon's suggested method. Nave Bayes and SVM are used in the proposed study to show that, for low threshold values, Logistics, like SMO, performs in a similar way, but that, for high threshold values, it loses its effectiveness when using 10-fold cross validation. [47] Even with this massive unstructured data, Map Reduce can be used to extract useful information from the tweets. Twitter messages are analyzed using Medtex text analysis software, which uses a number of standard features such as word tokens, stems, and n-grames; as well as the existence of Twitter username, hashtags, URLS and emoticons. Comparatively to the unbalanced sheet of unknown data validation, classifiers using balanced sheet F-measures perform better in cross-validation.

Extraction of tunable Q-wavelet transform (TQWT) characteristics from raw ECG heart rate signals is employed in the proposed method to detect coronary artery disease (CAD), which is a primary cause of heart failure, arrhythmias, and sudden death in patients (Electrocardiogram). For Q between 24 and 30, the Morlet Kernel function with 3-fold cross validation yields 99.7 percent accuracy, 99.6 percent sensitivity, and 99.8 percent specificity, with a Matthews correlation coefficient ranging from 0.9956 to 0.9956 for Q ranging from 24 to 30. Detection of heart disease, diabetes, eye disease, and neurological disease has become clear as a result of the intended research.

Using the BagMOOV new ensemble algorithms proposed by Bashir Saba and Usman Qamar, cardiac disease is predicted. The foundation of this framework is a multi-objective weighted voting system based on an improved bagging approach for forecasting and analysis. Tests based on ANOVA and 10-fold cross-validation were conducted to see if the predicted work was accurate (Analysis of Variance). Additionally, a new ensemble method is utilized to create a DSS for the diagnosis of heart illness. For this proposed study, the datasets have been given class labels of class 0 and 1, which designate different aspects of the datasets. The BagMOOV ensemble generates accurate and efficient solutions for all datasets. It has an 84.78 percent accuracy, 73.47 percent sensitivity, 91.1 percent specificity, and an f-measure of 81.30 percent with the proposed work. This study. It is most accurate to employ the DSS as a classifier in this investigation.

As part of this proposed study, Ives Cavalcante Passos will use machine learning algorithms to analyze mood disorders as a human behavioral attribute. MATLAB is used to create a system that can predict whether or not a person would try suicide by employing three machine learning algorithms: SVM, RVM, and Least Average Shrinkage and Selection Operator (LASSO) are the three most commonly used machine learning techniques (RVM) The Leave-one-out cross-validation approach is often used for analyzing training and testing datasets. The RVM performed the best out of the three algorithms, correctly predicting whether or not 103 out of 144 patients will attempt suicide, with an accuracy of 72%, a sensitivity of 72.1%, and a specificity of 71.3 percent (chi-squared equals 0.00001). The RVM's accuracy remained at 71.4 percent even when the confusion matrix was used.

It's possible to predict flu gripe, dengue, malaria, cholera, leptospirosis and Chikungunya outbreaks by examining common symptoms using Rane's approach. For example To carry out this study, researchers combed through the medical data of 316 patients at a hospital in Nasik, Maharashtra (India). It is possible to use open source software and tenfold cross-validation to test methods such as Decision Trees, Artificial Neural Network, SMO, K-Nearest Neighbor, and Naive Bayes using these open source software and SMO. Starting with this work, which has an exceptionally low acceptance result, ANN surpasses SVM in terms of parameter values.

Maternal health risk is defined as preeclampsia in the proposed study by Moreira, Mário WL, which is a risk of hypertension in pregnant women. Preeclampsia is difficult to diagnose manually, thus a probabilistic method called a Bayesian Network was developed and tested on more than 20 pregnant women with hypertension to help speed up the diagnosis process. The Bayesian network studies each symptom a pregnant woman experiences, such as headache, giddiness, nausea/vomiting, and edema, in great detail in order to construct a decision support system for making a diagnosis of preeclampsia.

When we had some labeled knowledge with an enormous amount of unlabeled knowledge, Sally Nihilist predicted the most likely learning scenarios and gave us a "curtaining" approach for using the unlabeled knowledge to improve supervised learning algorithms. According to her, there are two distinct supervised learning algorithms, each of which produces a hypothesis that specifies how an instance's area will be divided, The area of an instance can be divided into a single category for each tree of the call tree. This led her to the conclusion that 2supervised learning approaches can be used to label information in a range of different settings and circumstances.

From the perspective of applied mathematics modeling, Zoubin Ghahramani provided a brief overview of unsupervised learning. According to him, unsupervised learning could be driven by concepts of data abstraction and reasoning. Additional models in unsupervised learning were also examined by him in his review. Statistical models like the Graphical model play an important role in learning systems for many kinds of completely different kinds of knowledge, and he stressed their importance for reasoning under uncertainty. Statistics gives an unifying framework for learning from information. As a parting shot, he said, "statistics."

An extensive empirical evaluation of 10 different types of supervised learning procedures has been provided by Rich Caruana et al. in the previous decade. A wide range of techniques are available to help you make sense of your data, from SVMs to neural networks to supply regression to naive Bayes to random forests. In addition, they investigated and examined the impact on model performance of Platt Scaling and Isotonic Regression calibration. the instructional practices had been evaluated according to several performance-based criteria

Man Bayu Kanigoro, Galih Salman, Yaya Heryadi, and Bayu Kanigor Foretelling is the subject of extensive research. Deep learning is being used by the author to make predictions. "Deep" in the context of AI means that the neural networks used in deep learning have more layers than the "shadow" ones often used in machine learning. Using the meteorological data, it generated an image of a probe frame. The findings have been used to investigate heuristic NN exploitation methods for rain prediction supported weather datasets in an ongoing manner.

# **CHAPTER THREE - RESEARCH METHODOLOGY**

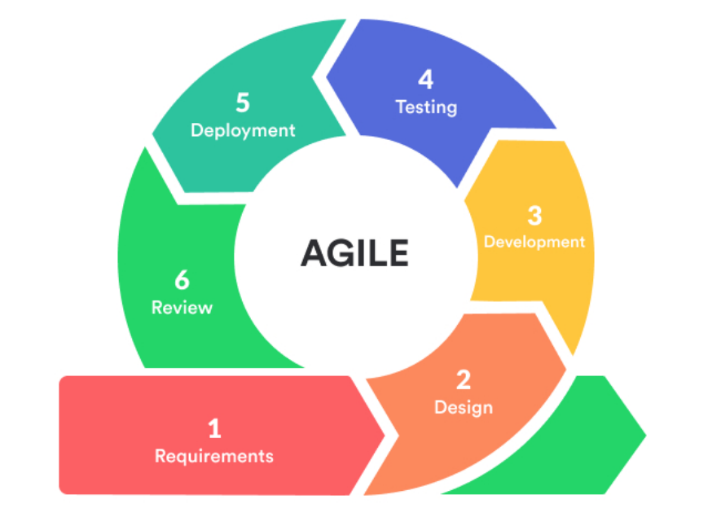
## Introduction

This chapter explains the research instruments that were used in the study. It outlines the methodology used to develop the system by the developer. The chapter also contains the methods used to gather data for the system and justification as to why the developer opted for this model.

## Development Methodology

System methodology simply divides the software development process into distinct phases or stages containing activities to manage and plan better. The Agile Methodology used in this study emphasizes people, relationships, and interactions rather than tools. There is a strong emphasis on working software rather than documentation in this approach, and customers are involved in every step of the development process. Agile is well-equipped to deal with the complexities and variability inherent in the development process. ' Each sprint or iteration of an Agile project has a set duration and deliverables, but there is no priority given to the order in which they are completed. Working software or any other tangible, tested product focuses on sprints.

Agile emphasizes teamwork, efficiency, and input from multiple departments and clients and a focus on team strengths and efficiency. With the Agile strategy, providing a working, tested, and prioritized features to customers is the primary goal (Chamakiotis, 2020).



***Figure 3.1: Agile Methodology process.***

## Requirement

Before beginning any design work, the product owner must develop the preliminary paperwork that outlines the project's needs. They represent the project's intended conclusion. As an illustration, consider the functionalities supported by a text editor. Several features will not be initially supported, such as the ability to incorporate video or multiple font sizes.

It is often advisable to reduce these initial requirements as much as possible, adding just those necessary features and ignoring those that are rarely utilized. Once the program has been released, and the fundamental functionality is working properly, developers can begin working on these features. Developers who opt to disregard this stage risk a condition known as "feature creep," The project is regularly expanded to include new, non-essential features, thus diverting their attention away from the most vital ones.

Both the client and the product owner iterate on their needs, refining them as necessary.

## Software Design

Before beginning development, the Product Owner brings together their team and goes over what has to be done. Once these needs have been outlined, the team begins brainstorming how best to meet them. By way of illustration, they specify which languages, frameworks, and libraries will be used in the project.

Iterations two and three focus on implementing new features and the system's internal structure, respectively. For the User Interface (UI), the designers develop a preliminary mock-up as part of the System Development Life Cycle (SDLC) (UI). A product's interface and overall user experience are critical if it is to be sold to the general public. When looking at potential competitors, it's always a good idea to see what they're doing properly and their flaws (Sunnen, Arend, & Maquil, 2017).

The basic design is refined or reworked in subsequent iterations to accommodate the new functionality.

## Development and Coding

This phase of software development is all about writing code and translating design documents into working software. This SDLC stage typically takes the most time because it is the foundation of the entire process. The Unified Modelling Language is one of the design tools that should be utilized (UML). A system's boundaries, structure, and behaviour can be described using this analysis method. In addition, a flowchart can be used to depict all of the logic steps in your program visually. The third tool is a case diagram, which shows how different iterations interact with one another. Thirdly, the Data Movement Diagram (DFD) is used to depict the flow of information in the system. It is possible to model interactions between items using a sequential diagram as the fifth tool. Activity diagrams can be used to represent the sequence of events and the control flow in a system.

## Integration and Testing

All bugs and compatibility issues are addressed at this part of the development process and any other issues that may arise. QA ensures the code is error-free and that the business goals of the solution are met by running a set of tests.

## Implementation and Deployment

A demo version of the application is installed on a server and made available to consumers for use. Updates to the software that has already been installed include new features and fixes for bugs.

## Review

During the last stages of development, the product owner meets with the Development Team to assess how far along they are in fulfilling the requirements. The product owner listens to the team's suggestions for resolving the issues that developed during the previous stages and incorporates their suggestions. Following a new iteration or scaling Agile, the Agile software development lifecycle stages start anew either in the current stage or in the following stage.

## Data collection methods

The primary data collection method is from sleep sensor iOS app data and self-reported sleep data and detection of social isolation only using the sleep quality data using machine learning models and giving the high efficiency.

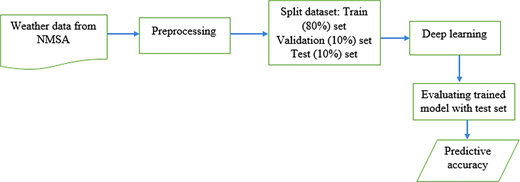
## Summary

Learning software development life cycle is essential in developing (SDLC) computer system projects. Researchers have the will to choose the development methodology they want to follow in line with the type of system they are developing to meet their needs effectively.

# **CHAPTER FOUR - MATERIALS AND METHODS**

## Proposed solution

In order to build a rain prediction model, meteorological data must be collected. Weather data from the NMSA is used in this context. Preprocessing is the process of removing empty entries, resolving missing values, and normalizing the data after it has been collected and gathered. The deep-learning module uses the preprocessed data to learn from it and make predictions about future rainfall based on data that has not yet been observed. As depicted in this diagram, the suggested rainfall forecast model includes the following modules:



### **Machine Learning**

We're heading toward a future in which everything is managed and handled by machines. Using artificial intelligence, we can educate machines to learn on their own, which is referred to as machine learning. In machine learning, we educate the machine to learn from its prior data and aim to better its results in the future by learning from its mistakes. The use of various tools, methodologies, and strategies is an important part of machine learning. Using these approaches and techniques, machines and humans alike can get new insights into previously unknown phenomena by combining existing data with previously unexplored sources of information. Occasionally, we'll attempt to capture an action and then replicate that action in a new situation. This in turn helps people gain a better grasp of the problem. There is a history of statistics in machine learning methods. The true message hidden in a vast amount of data can be extracted to aid in the exploration of more complex learning models.

Data analysis can use both machine learning and classical statistics techniques, but there are significant differences in the underlying principles and characteristics of each. For example, machine learning has the ability to process large data sets and real-time streams with mixed values types, the ability to select from a variety of learning models and controlling parameters to capture the non-linear or high-order structure in data, and the ability to visualize the information for making a decision.

Even though learning is a never-ending process, we aim to stimulate it in machines in the same way we do in humans. Machine learning is divided into three main categories, each with its own distinct advantages. It's important to consider all of these factors while developing a learning model. For example, supervised learning requires that a system be given a list of labels and a desired result. To "learn," the algorithm compares its output to the outputs it was trained on and adjusts the model accordingly, thereby allowing it to "learn." To put it another way, because the data used for unsupervised learning is unlabeled, it is broken up according to how similar the input data are. The algorithm's learning is left to detect similarities in the data it receives. Machine learning algorithms that promote unsupervised learning are very useful when dealing with large amounts of unlabeled data.

### **SVM**

It is well recognized as the hyperplane that SVMs use to deal with classification challenges. Your forecast confidence can be gauged based on how far away you are from the hyperplane. Since the hyperplane is a distance away from the nearest data point, we must widen the gap between the two. SVM employs a kernel technique to transform low-dimensional input data into high-dimensional feature space in order to find the best fit line or decision boundaries. The optimal option is the one that leaves the largest space between the two classes' edges, but no hyperplane connects them.

The SVM is a A Support Vector Machine (SVM) optimizer uses it to handle the optimization problem that emerges during training. In most cases, it's used to break down a larger difficulty into smaller ones [8]. Predictions of rainfall were made using SVM-based models in Lin et al. [17], both with and without typhoon features.

SVM class concepts are separated using a separating hyperplane, which is a n1-dimensional plane that is an analagous to at n-dimensional space. Maximum margin hyperplane: The hyperplane is selected so that both classes are separated by a maximum distance from the plane. Below is the equation for the separating hyperplane:



There are four terms in this equation: |b|/ ||w||, which is the perpendicular distance between the hyperplane and origin, and |[w||2], which is a Euclidean norm of w. There are two types of features that need to be separated in Xi, which is a two-dimensional feature matrix

### **Navie Bayes**

The Bayes theorem, which is the foundation of the Navie Bayes classifier, is used to train this machine. As a predictive modeler, it's an excellent algorithm. Navie bayes is a popular classifier for training datasets with large dimensionality. Classifier Navie employs probability theory to classify data and assumes that one characteristic is independent of other attributes, which is what makes it probabilistic. This is why it's called a probabilistic classifier.

Using the Bayes law, we can calculate an event's conditional probability based on our prior knowledge, which is the likelihood of an event occurring if we already know about it. When using the naive Bayes method, with the prior probability P, we may calculate the posterior probability P(C|A) (C), prior probability P(A), and prior probability (C|C) using Bayes' theorem.

Because event C is more likely to happen if A is true than A is less likely to happen, we can calculate the conditional probability, or likelihood, of C occurring based on A being true. We can then multiply this likelihood by the chance of C occurring based on A's probability.

### **Random Forest**

A classification tree is built using the supervised learning method of random forest. Each tree in the forest has an input data vector that is used to categorize a new object based on feature vectors in the forest. To classify a forest tree, a "vote" is given to each tree, and the trees with the most "votes" are taken into account. pseudo-code From the total number of "m" features, randomly select "k" features. Using the best split, break the element down into smaller components.

It is possible for algorithms to perform classification and regression tasks at the same time. As the number of trees in a forest grows, so does its accuracy and robustness.

### **MLP (MULTILAYER PRECEPTRON)**

MLPs are a subset of feed forward ANNs. At least three layers of nodes are required for an MLP. A nonlinear activation function is used for each node except for the input nodes, which have linear activation. For the purpose of training, MLP employs a supervised learning approach known as back propagation [11].

In a directed graph, a multilayer perceptron is a neural network that links multiple layers in a single direction. There are no linear activation functions in this system, save for the input nodes. MLPs use a supervised learning approach called backpropagation. MLP is a deep learning technique since it has numerous layers of neurons.

Its goal is to be able to mimic the functioning of the human brain. An artificial neural network made out of hardware and GPUs is encouraged to use multiple layers. An input layer receives its output from the previous layer, and vice versa. When it comes to deep learning, algorithms can either be supervised or unsupervised, depending on the specific task at hand.

We drew on rainfall statistics from the data.gov.in website run by the Indian government. Using machine learning techniques, we are trying to determine the optimum method for predicting rainfall. The following are the steps that follow this one:

* **Data from an open repository, such as data.gov.in, can be used in the first step.**
* **Step 2: Cleansing and preprocessing of data, as well as the choice of features.**
* **Output will be an algorithm with the optimum outcome in Step 3.**

The dataset comes from data.gov.in, the Indian government's official website.

Data pre-processing is used to prepare the data for analysis. We rely on information gleaned from the Indian government's official website. We may use this data to compute the amount of rainfall in millimeters over India. Rainfall totals for each day of the specified year, from January through December, are included in this set.

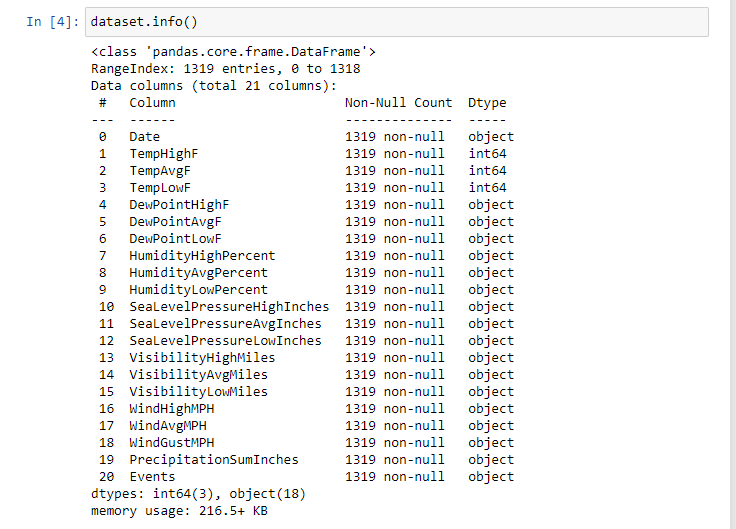
There is a lot of noise in the data, and certain pieces are missing. Feature selection is necessary for us to get better data for processing and analysis. In the paper, we used daily rainfall volume in millimeters for the experiment.

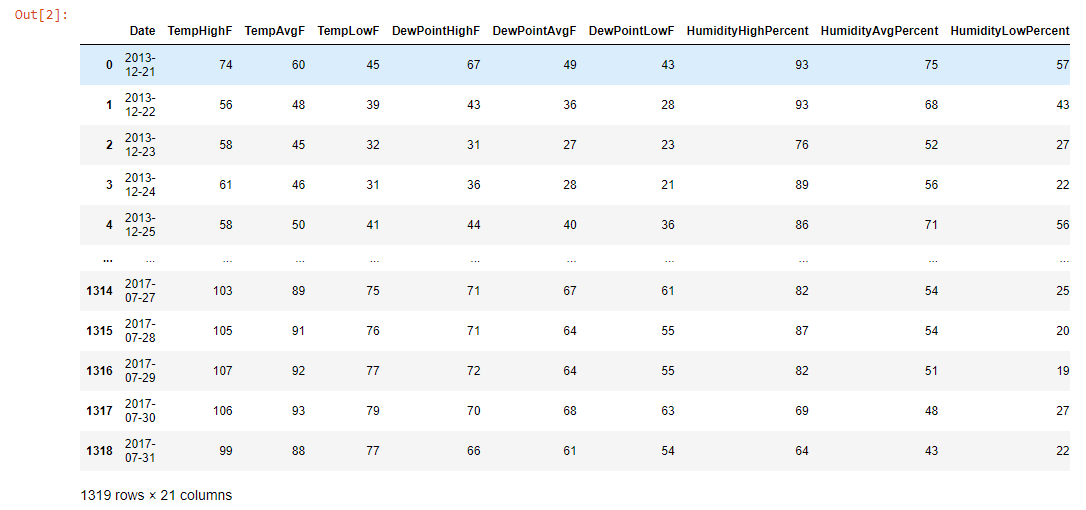
Feature scaling, or data normalization, is used to standardize the independent variables' range. We used the conventional scalar formula as follows to rescale data between [0,1]:

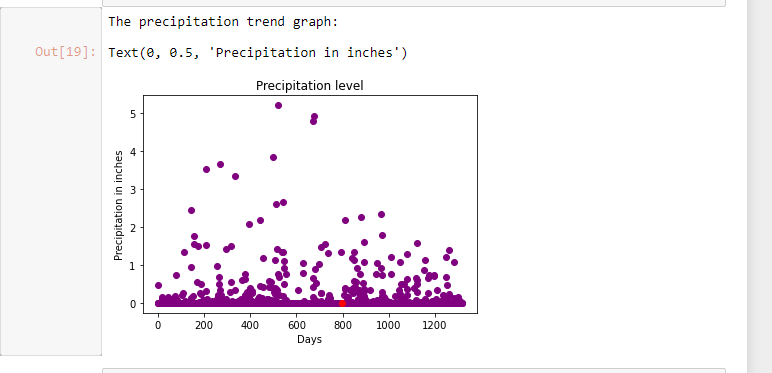


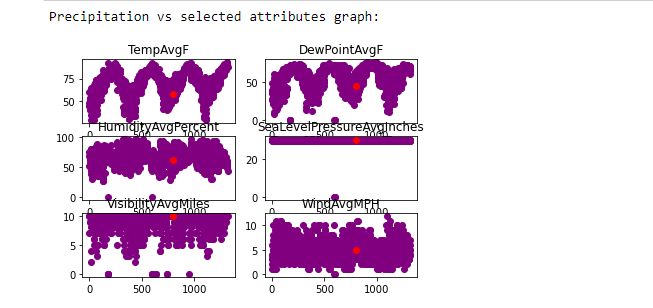
With the help of several machine learning algorithms, we can forecast the next day`s rainfall by using train data from prior days.

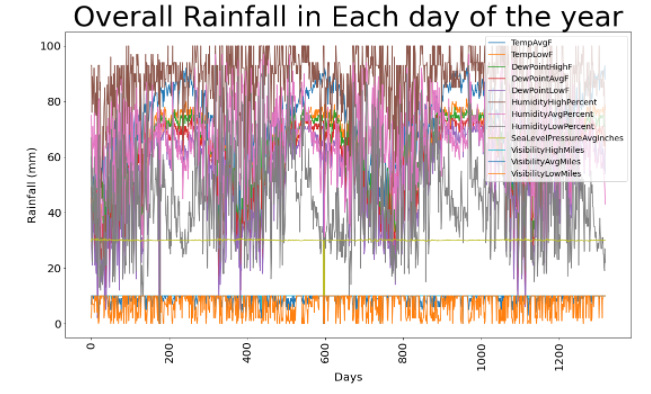
We used the first few months of the year to train the data and anticipate the next month's rainfall.











# **CHAPTER FIVE - Result and discussion**

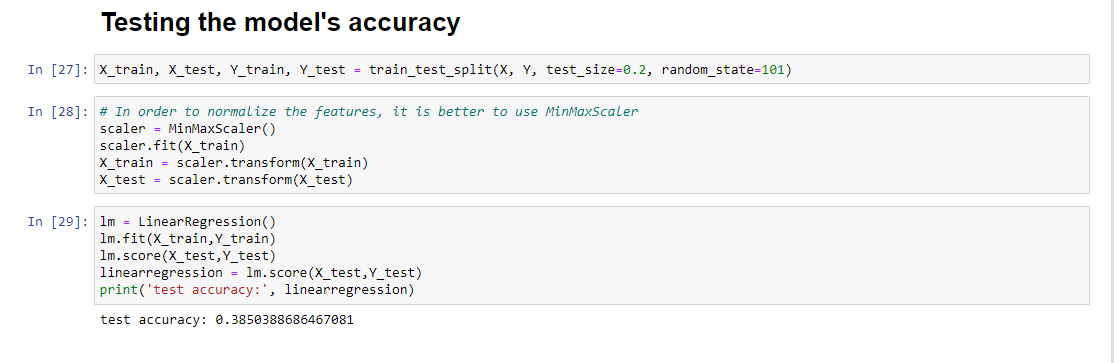
Predicting the effects of climate change is difficult and unexpected these days, but it has always been a big concern for the world as a whole. Human activity is contributing to the current trend of global warming. This causes a rise in sea level, increased floods, and a variety of other climate-related problems. Rainfall is one of the more harmful repercussions of climate change. Despite the difficulty of rain forecasting, several of the world's leading authorities are taking it into account.

The primary goal of this research is to identify an algorithm that is capable of accurately predicting rainfall. We obtained the most recent rainfall data for India from the Indian government's official website. The accuracy of the algorithm is shown in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | Linear regression | 0.38 | 38% |

Using the linear regression approach, we can observe that rainfall prediction is the best.

## Train dataset:



## Summary and Conclusion

Data on rainfall was classified using machine learning techniques in this study's supervised model. Machine learning techniques were employed to investigate the accuracy of rainfall forecasts. Classifiers like the SVM, Random Forest, Naive Bayes, and MLP have been compared in various studies (Multilayer Perceptron). The most accurate method of predicting rainfall in India has been found to be through the application of machine learning.

Industry-specific applications of machine learning are becoming more and more commonplace. We're turning to technology to assist us deal with the ever-increasing volume and variety of data. In addition, rainfall forecasts might be fairly precise due to the high accuracy score in weather predictions It is our goal to increase our efforts in the areas of storm forecasting and crop forecasting in conjunction with rainfall predictions in the future.

## RECOMMENDATIONS

In order to help agro farmers prepare their farms for harvest, I would recommend this system to weather forecasting stations.

Flying during the rainy season can be hazardous, especially during periods of heavy hail. I would recommend this model to airports so that flight plans can be created to prevent doing so.

## Future work

When it comes to software, there's always space for improvement. As a result of the short amount of time available for its creation, this is the case. It is possible to add a user interface and different types of users to the software if it has more Time. This is the initial step toward making the software network-ready and more resilient. Aside from this original afterthought, there are many ways to improve the software, like enhancing its ability to perform calculations and adding additional flexibility.

## Appendix: Tools

The tools that will be used are Jupyter notebook using Python language. There are a series of steps in developing the technical solution, and we use different libraries such as NumPy, pandas, and for visualization, we use matplotlib, seaborn. In addition, the sklearn library is used for training the data sets.

## REFERENCE

Chamakiotis, P. (2020). Virtual teams as creative and Agile work environments. Agile Working and Well-Being in the Digital Age, 133-142.

Raimi, M. O., Vivien, O. T., & Oluwatoyin, O. A. (2021). Creating the healthiest nation: Climate change and environmental health impacts in Nigeria: A narrative review. Morufu Olalekan Raimi, Tonye Vivien Odubo & Adedoyin Oluwatoyin Omidiji (2021) Creating the Healthiest Nation: Climate Change and Environmental Health Impacts in Nigeria: A Narrative Review. Scholink Sustainability in Environment. ISSN.

Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. Journal of Big Data, 6(1), 1-16.

Nordhaus, W. (2019). Climate change: The ultimate challenge for economics. American Economic Review, 109(6), 1991-2014.

Rainfall has a direct impact on a wide range of human activities, including agriculture, construction, power generation, and tourism.

Greve, P., Kahil, T., Mochizuki, J., Schinko, T., Satoh, Y., Burek, P., ... & Wada, Y. (2018). Global assessment of water challenges under uncertainty in water scarcity projections. Nature Sustainability, 1(9), 486-494.

Alves, F., Leal Filho, W., Casaleiro, P., Nagy, G. J., Diaz, H., Al-Amin, A. Q., ... & Azeiteiro, U. M. (2020). Climate change policies and agendas: Facing implementation challenges and guiding responses. Environmental Science & Policy, 104, 190-198.

Agamile, P., Dimova, R., & Golan, J. (2021). Crop Choice, Drought and Gender: New Insights from Smallholders’ Response to Weather Shocks in Rural Uganda. Journal of Agricultural Economics, 72(3), 829-856.

Cronin, J., Anandarajah, G., & Dessens, O. (2018). Climate change impacts on the energy system: a review of trends and gaps. Climatic change, 151(2), 79-93.

Noor, M. N., & Haneef, F. (2020). A Review on Big Data and Social Network Analytics Techniques. Researchpedia Journal of Computing, 1(1), 39-49.

Deka, B., Maji, P., Mitra, S., Bhattacharyya, D. K., Bora, P. K., & Pal, S. K. (Eds.). (2019). Pattern Recognition and Machine Intelligence: 8th International Conference, PReMI 2019, Tezpur, India, December 17-20, 2019, Proceedings, Part I (Vol. 11941). Springer Nature.

Sunnen, P., Arend, B., & Maquil, V. (2017). "Okay, yes, it's true" – doing discovering work in a tangible-user-interface-mediated joint problem-solving physics activity. EDULEARN17 Proceedings.