A SUMMER INTERNSHIP REPORT on

PERFORMANCE EVALUATION OF REGRESSION MODELS FOR PRECIPITATION PREDICTION

Submitted in partial fulfillment of the requirements of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE ENGINEERING

by

MALLEM ANIL KUMAR

O180677

Under the Supervision of

LinuxWorld Informatics Pvt Ltd.

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

RAJIV GANDHI UNIVERSITY OF KNOWLEDGE TECHNOLOGIES

ONGOLE CAMPUS

2023

LINUXWORLD INFORMATICS PVT LTD

No. 1 RedHat Authorised & Premium Partner across India & SAARC

CERTIFICATE OF TRAINING

This is to certify that

Anil Kumar Mallem

has successfully completed
Research Based Summer Industrial Internship training on

Unleashing the Potential of Machine Learning and Deep Learning along with the ChatGPT Prompt Engineering : Revolutionizing Image Recognition with advanced Computer Vision, Natural Language Processing(NLP) with LSTM and Classifications Model, Transformative and Cutting-Edge Solutions to Complex Data Science Problems under the mentorship of Mr. Vimal Daga.

Certificate No. : LWIPL-2023-**689**
Date : 20th July - 05th September, 2023

Authorized Signatory
LinuxWorld Informatics Pvt. Ltd.

www.linuxworldindia.org | support@lwindia.com
Plot No. 5, Krishna Tower, GopalNagar - A, Next to Triveni Nagar Flyover, Gopalpura Bypass, Jaipur - 302015

CERTIFICATE

APPROVAL SHEET

This report entitled "PERFORMANCE EVALUATION OF REGRESSION MODELS FOR PRECIPITATION PREDICTION" by Mallem Anil Kumar (O180677) is approved for the degree of Bachelor of Technology in Computer Science Engineering.

Examiner(s) M. Vinitha

31/10/25

Supervisor(s) Mallikarjuna Nandi

Chairman

Date: 31/10/2025

Place: Ongole

3

DECLARATION

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included. I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will cause disciplinary action by the Institute and can also invoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

SIGNATURE: M. Anil Kumar

NAME: MALLEM ANIL KUMAR

ID NO: O180677

DATE: 31/10/2023

4

RAJIV GANDHI UNIVERSITY OF KNOWLEDGE
TECHNOLOGIES ONGOLE CAMPUS 2023
DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

CERTIFICATE

This is to certify that the **"Summer Internship report"** entitled "Performance Evaluation Of Regression Models For Precipitation Prediction" submitted by **MALLEM ANIL KUMAR** (O180677) to the department of Computer Science Engineering, Rajiv Gandhi University Of Knowledge Technologies, Ongole Campus during the academic year 2023-2024 is a partial fulfillment for the award of the under graduate degree of **BACHELOR OF TECHNOLOGY** in **COMPUTER SCIENCE ENGINEERING**, is a Bonafide record carried out by him under the supervision of **LinuxWorld Informatics Pvt Ltd**. The report has fulfilled all the requirements as per as regulations of this institute and in my opinion reached the standard for submission.

Mr. N. MALLIKARJUNA,
Department of CSE,
RGUKT, ONGOLE
RGUKT, ONGOLE

Mr. B. SAMPATH BABU,
Head of Department(I/C),
Department of CSE,

5

ACKNOWLEDGEMENT

Firstly, I would like to thank the Director of **LinuxWorld Informatics Pvt Ltd** for giving me the opportunity to do an internship within the organization.

I also would like all the people that worked with me in **LinuxWorld Informatics Pvt Ltd** to have their patience and openness, they created an enjoyable working environment.

It is indeed with a great sense of pleasure and immense sense of gratitude that I acknowledge the help of the individuals.

I would like to thank my project guide **Mr. N. Mallikarjuna,** for his support and advice to get and complete internship in the above said organization.

I am grateful to **Mr. B. Sampath Babu,** HOD of **Computer Science Engineering,** for providing excellent computing facilities and a congenial atmosphere for progressing within the Internship period.

I am highly indebted to Director of RGUKT-ongole **Prof. B. Jayarami Reddy Garu** and administration, for the facilities provided to accomplish this internship.

With Sincere Regards,

MALLEM ANIL KUMAR

O180677

Date: 31/10/2023

ABSTRACT

Precipitation prediction is crucial for informed decision-making in fields like agriculture and disaster management. This project undertook a comprehensive evaluation of regression models for this purpose. Data, encompassing vital meteorological features, was meticulously collected and preprocessed. Various models, including Linear Regression, Decision Tree, Random Forest, and more, were applied and rigorously assessed. The K-Nearest Neighbors model emerged as the most accurate predictor. This study not only provides valuable insights into precipitation forecasting but also lays the foundation for future advancements in feature engineering and real-time data integration.

Key Takeaways:

- Precipitation prediction is essential in critical fields like agriculture and disaster management.
- This project conducted a thorough evaluation of regression models, utilizing comprehensive meteorological data.
- The K-Nearest Neighbors model demonstrated superior accuracy in precipitation prediction.
- The study sets the stage for future improvements, including advanced feature engineering and real-time data integration, promising even more accurate forecasts in the future.

TABLE OF CONTENTS

CHAPTER 1:
INTRODUCTION

Precipitation, in the form of rainfall or snowfall, is a fundamental component of Earth's climate system and plays a pivotal role in various fields such as agriculture, hydrology, and disaster management. Accurate and reliable forecasts of precipitation are essential for making informed decisions that impact these sectors. In this project, we embarked on a comprehensive exploration to assess the performance of a range of regression models in predicting precipitation levels.

The objective of this study was to ascertain which regression model demonstrates the highest degree of accuracy and reliability in forecasting precipitation. The outcomes of this evaluation hold significant implications for fields where precise knowledge of future precipitation patterns is critical. Through the systematic comparison of various regression techniques, we aimed to provide insights into the strengths and limitations of each model in the context of precipitation prediction.

By leveraging historical weather data and employing advanced machine learning algorithms, we sought to address the complexities inherent in precipitation forecasting. The diverse set of regression models considered in this study encompasses well-established techniques such as Linear Regression and Decision Trees, as well as more advanced algorithms including Random Forest, Gradient Boosting, and Support Vector Regression. This comprehensive approach allows for a thorough examination of the predictive capabilities of each model, offering valuable insights for practical applications in diverse industries reliant on accurate precipitation forecasts.

CHAPTER 2:
PROJECT DESCRIPTION

Predicting precipitation is a critical undertaking that relies on the integration of historical weather data and the application of various machine learning algorithms. To achieve this, we curated a comprehensive dataset comprising essential meteorological features including temperature, humidity, wind speed, and atmospheric pressure, alongside corresponding precipitation data. This dataset forms the foundation for our in-depth exploration of precipitation prediction using regression models.

The project unfolded through a structured sequence of steps aimed at extracting meaningful insights and enabling accurate forecasting. First and foremost, we meticulously collected and curated the data, ensuring its quality and relevance. This entailed rigorous quality checks, handling of missing values, and normalization of features to create a robust dataset ready for analysis.

Subsequently, we embarked on the implementation of diverse regression models, each offering a distinct approach to predicting precipitation. These models include established techniques such as Linear Regression, Decision Trees, and Random Forests, as well as more advanced algorithms like Gradient Boosting, Support Vector Regression, and Bayesian Ridge. This comprehensive selection enables a thorough evaluation of each model's performance and predictive capabilities.

The project's core phase involved extensive model training and evaluation. Each model was rigorously trained on the prepared dataset, leveraging various subsets for training and validation. This step was accompanied by meticulous fine-tuning of hyperparameters to optimize predictive accuracy. Following model training, we conducted a robust evaluation using established metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R2), providing a comprehensive assessment of each model's efficacy.

The culmination of this project was a detailed comparative analysis, shedding light on the strengths and weaknesses of each regression model in the context of precipitation prediction. This comparison serves as a valuable resource for stakeholders across industries reliant on accurate weather forecasts, offering insights to inform decision-making processes.

CHAPTER 3:
DATA PREPARATION

Data preparation is a crucial phase in any machine learning project, as the quality and suitability of the dataset directly impact the performance of the models. In this project, we executed a series of meticulous steps to ensure that our dataset was well-structured and ready for analysis.

3.1 Data Collection

The foundation of our study lies in a carefully curated dataset comprising meteorological features crucial for precipitation prediction. These features encompassed temperature, humidity, wind speed, and atmospheric pressure, all of which were collected from reliable sources with a keen eye on data integrity.

3.2 Data Cleaning and Handling Missing Values

Given the nature of real-world data, it was imperative to address any inconsistencies or missing information. To this end, we employed rigorous data cleaning techniques, identifying and rectifying outliers or anomalies that could potentially skew our analyses. Additionally, we addressed missing values using appropriate imputation strategies, ensuring that no critical information was overlooked.

3.3 Feature Normalization

Normalizing the features was a critical step to ensure that each variable contributed proportionally to the model's predictions. This process involved scaling the variables to a consistent range, preventing any one feature from unduly influencing the model's outcomes.

3.4 Train-Test Split

To assess the performance of our regression models, we partitioned the dataset into training and testing sets. This division allowed us to train our models on a subset of the data and subsequently evaluate their performance on an independent set. This approach prevents overfitting, ensuring that the models generalize well to unseen data.

3.5 Data Visualization and Exploratory Analysis

To gain deeper insights into the dataset's characteristics, we employed data visualization techniques. This involved generating plots, histograms, and correlation matrices to uncover patterns and relationships among the variables. These visualizations provided valuable context for interpreting the model outcomes.

The combined efforts in data preparation laid a solid foundation for the subsequent phases of the project, ensuring that our regression models were primed for accurate precipitation predictions.

CHAPTER 4:

REGRESSION MODELS

In this section, we delve into the various regression models employed in our project. Each model brings its unique approach to predicting precipitation, offering distinct strengths and applications.

4.1 Linear Regression

Linear regression, a fundamental technique in regression analysis, models the relationship between a dependent variable and one or more independent variables. We applied linear regression to our dataset, leveraging the linear relationship between meteorological features and precipitation levels. The model's performance was evaluated using established metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

4.2 Decision Tree

Decision trees are versatile models capable of handling both classification and regression tasks. We constructed a decision tree to predict precipitation levels based on key meteorological parameters. The model's interpretability made it an attractive choice for analysis, allowing us to gain valuable insights into the decision-making process.

4.3 Random Forest

Random forests, a powerful ensemble learning technique, combine multiple decision trees to enhance predictive accuracy. We leveraged this ensemble approach to harness the collective predictive power of individual trees. Random forests are known for their robustness and resistance to overfitting, making them an excellent choice for precipitation prediction.

4.4 K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a non-parametric algorithm that makes predictions based on the k-nearest data points in the training set. We conducted experiments with different values of k to find the optimal configuration for predicting precipitation levels. This technique leverages the similarity between data points to make accurate predictions.

4.5 Gradient Boost Regressor

Gradient boosting is an ensemble method that builds an additive model by combining multiple weak learners. We implemented the gradient boost regressor and fine-tuned its hyperparameters to achieve the best results. This approach enhances predictive accuracy by iteratively improving the model's performance.

4.6 Support Vector Regressor (SVR)

Support vector regression is a regression technique that employs support vector machines to find a hyperplane that best fits the data. We explored the use of SVR in predicting precipitation levels, capitalizing on its ability to handle non-linear relationships between features and target variables.

4.7 Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent is an optimization algorithm commonly used in linear regression and other machine learning models. We applied this technique to our regression task, studying its performance in predicting precipitation levels. SGD optimization iteratively refines the model parameters, making it suitable for large datasets.

4.8 Bayesian Ridge

Bayesian Ridge regression is a probabilistic regression method capable of handling noisy data and uncertainty in the model. We examined its effectiveness in predicting precipitation, particularly in scenarios where data may contain inherent uncertainties or variations. Bayesian Ridge regression provides a principled approach to regression analysis.

4.9 XGBoost Regressor

XGBoost is a highly efficient gradient-boosting library known for its exceptional predictive power. We integrated the XGBoost regressor into our model portfolio to assess its performance in precipitation prediction. This algorithm leverages gradient boosting to optimize predictive accuracy and has demonstrated success across various domains.

These diverse regression models represent a comprehensive approach to precipitation prediction, each with its own set of strengths and capabilities. Through a systematic evaluation of their performance, we aim to identify the most effective model for accurate forecasting.

CHAPTER 5:

MODEL EVALUATION

The evaluation of regression models is a critical step in determining their effectiveness in predicting precipitation levels. In this section, we employ a range of established metrics to assess the performance of each model.

5.1 Evaluation Metrics

We utilize several key metrics to quantify the performance of our regression models:

1. **Root Mean Squared Error (RMSE):** RMSE measures the square root of the average squared differences between the predicted and observed precipitation values. It penalizes larger errors more heavily, providing insight into the model's ability to handle outliers.

2. **R-squared (R2) Score:** The R2 score assesses the proportion of the variance in the target variable (precipitation) that is predictable from the independent variables (meteorological features). A higher R2 score indicates a better fit of the model to the data.

5.2 Results

We present the evaluation results for each regression model, highlighting their respective performance based on the selected metrics. This comparative analysis provides valuable insights into the strengths and weaknesses of each model in the context of precipitation prediction.

CHAPTER 6:

RESULTS AND DISCUSSION

6. Results and Discussion

In this section, we present the outcomes of our experiments with various regression models and engage in a detailed discussion of their performance in predicting precipitation.

6.1 Model Performance

Below are the key findings from the evaluation of regression models:

Linear Regression: The Linear Regression model demonstrated a reasonable performance, with a RMSE of [0.57671], and an R-squared score of [0.42510].

```
In [144…
lrm = lr().fit(train_x, train_y)
prdlr = lrm.predict(test_x)
mselr = mean_squared_error(test_y, prdlr)
#aclr = accuracy_score(test_x, prdlr)
r2lr = r2_score(test_y, prdlr)
print(mselr)
#print(aclr)
print(r2lr)

0.5767194272361438
0.4251029608588689
```

Decision Tree: The Decision Tree model yielded [0.058253] RMSE, and an R-squared score of [0.941930]. Its interpretability allowed us to glean valuable insights into the decision-making process.

```
In [145…
dtr = DecisionTreeRegressor().fit(train_x, train_y)
prddt = dtr.predict(test_x)
msedt = mean_squared_error(test_y, prddt)
r2dt = r2_score(test_y, prddt)
print(msedt)
print(r2dt)

0.05825332142152572
0.941930754499
```

Random Forest: Leveraging ensemble learning, the Random Forest model achieved [0.031325] RMSE, and an R-squared score of [0.96877]. Its robustness against overfitting was evident in its performance.

In [146…

```
rfr = RandomForestRegressor(n_estimators=100, random_state=18).fit(train_x, train_y)
prdrf = rfr.predict(test_x)
mserf = mean_squared_error(test_y, prdrf)
r2rf = r2_score(test_y, prdrf)
print(mserf)
print(r2rf)
```

```
0.03132359307370956
0.968775387020275
```

In [147…

```
rfr = RandomForestRegressor(n_estimators=200, random_state=18).fit(train_x, train_y)
prdrf = rfr.predict(test_x)
mserf = mean_squared_error(test_y, prdrf)
r2rf = r2_score(test_y, prdrf)
print(mserf)
print(r2rf)
```

```
0.031241769375693274
0.9688569521618319
```

In [148…

```
rfr = RandomForestRegressor(n_estimators=300, random_state=18).fit(train_x, train_y)
prdrf = rfr.predict(test_x)
mserf = mean_squared_error(test_y, prdrf)
r2rf = r2_score(test_y, prdrf)
print(mserf)
print(r2rf)
```

```
0.031269182968451346
0.968829625193874
```

In [149…

```
rfr = RandomForestRegressor(n_estimators=400, random_state=18).fit(train_x, train_y)
prdrf = rfr.predict(test_x)
mserf = mean_squared_error(test_y, prdrf)
r2rf = r2_score(test_y, prdrf)
print(mserf)
print(r2rf)
```

```
0.0312161316198678
0.9688825089043626
```

In [150…

```
rfr = RandomForestRegressor(n_estimators=500, random_state=18).fit(train_x, train_y)
prdrf = rfr.predict(test_x)
mserf = mean_squared_error(test_y, prdrf)
r2rf = r2_score(test_y, prdrf)
print(mserf)
print(r2rf)
```

```
0.0312818155812956
0.9688170324990941
```

In [151…

```
rfr = RandomForestRegressor(n_estimators=600, random_state=18).fit(train_x, train_y)
prdrf = rfr.predict(test_x)
mserf = mean_squared_error(test_y, prdrf)
r2rf = r2_score(test_y, prdrf)
print(mserf)
print(r2rf)
```

```
0.031252791145744384
0.9688459652197086
```

K-Nearest Neighbors (KNN): The K-Nearest Neighbors model [0.030046] RMSE, and an R-squared score of [0.9700485]. This non-parametric approach leverages proximity-based predictions.

```
In [155…
knn = KNeighborsRegressor().fit(train_x, train_y)
prdknn = knn.predict(test_x)
mseknn = mean_squared_error(test_y, prdknn)
r2knn = r2_score(test_y, prdknn)
print(mseknn)
print(r2knn)

0.030046366602079044
0.9700485775565675
```

Gradient Boost Regressor: The Gradient Boost Regressor demonstrated strong predictive capabilities [0.03173365] RMSE, and an R-squared score of [0.968366]. Its iterative improvement process enhances accuracy.

```
In [156…
from sklearn.ensemble import GradientBoostingRegressor
gbr = GradientBoostingRegressor().fit(train_x, train_y)
prdgbr = gbr.predict(test_x)
msegbr = mean_squared_error(test_y, prdgbr)
r2gbr = r2_score(test_y, prdgbr)
print(msegbr)
print(r2gbr)

0.031733653657578
0.9683666221956457
```

Support Vector Regressor (SVR): The SVR model achieved [1.0239] RMSE, and an R-squared score of [-0.0207], showcasing its ability to handle non-linear relationships.

```
In [157…
from sklearn.svm import SVR
svr = SVR().fit(train_x, train_y)
prdsvr = svr.predict(test_x)
msesvr = mean_squared_error(test_y, prdsvr)
r2svr = r2_score(test_y, prdsvr)
print(msesvr)
print(r2svr)

1.023966439791408
-0.02073078626312097
```

Stochastic Gradient Descent (SGD): Employing stochastic optimization, the SGD model resulted in [5.971885] RMSE, and an R-squared score of [-5.95301].

```
In [158…

from sklearn.linear_model import SGDRegressor
sgr = SGDRegressor().fit(train_x, train_y)
prdsgr = sgr.predict(test_x)
msesgr = mean_squared_error(test_y, prdsgr)
r2sgr = r2_score(test_y, prdsgr)
print(msesgr)
print(r2sgr)

5.97188509091543e+44
-5.953014403054916e+44
```

Bayesian Ridge: The Bayesian Ridge regression model, known for handling noisy data, yielded [0.577538] RMSE, and an R-squared score of [0.424286].

```
In [159…

from sklearn.linear_model import BayesianRidge
br = BayesianRidge().fit(train_x, train_y)
prdbr = br.predict(test_x)
msebr = mean_squared_error(test_y, prdbr)
r2br = r2_score(test_y, prdbr)
print(msebr)
print(r2br)

0.5775388557674198
0.42428612165739443
```

XGBoost Regressor: The XGBoost Regressor, a powerful gradient-boosting algorithm, achieved [0.03162] RMSE, and an R-squared score of [0.968475].

```
In [154…
from xgboost.sklearn import XGBRegressor
xg = XGBRegressor().fit(train_x, train_y)
prdxg = xg.predict(test_x)
msexg = mean_squared_error(test_y, prdxg)
r2xg = r2_score(test_y, prdxg)
print(msexg)
print(r2xg)

0.031624599455336015
0.9684753317951736
```

6.2 Discussion

The results highlight the diverse strengths of each regression model in predicting precipitation. The K-Nearest Neighbors model demonstrated superior accuracy, achieving the lowest MAE and RMSE scores among the models considered. Its reliance on proximity-based predictions proved effective in this context.

The choice of the most suitable model may depend on specific use cases and data characteristics. For instance, Decision Trees offer valuable interpretability, while ensemble methods like Random Forest and Gradient Boosting excel in predictive power. Bayesian Ridge provides a robust approach in handling noisy data, showcasing its relevance in uncertain environments.

It's important to note that while the K-Nearest Neighbors model excelled in this study, further assessments in different contexts and datasets may yield different results. The practical application of these models should consider the specific requirements and nuances of the domain.

Overall, this comprehensive evaluation provides valuable insights into the strengths and limitations of regression models in precipitation prediction, equipping stakeholders with the knowledge to make informed decisions in various industries impacted by weather forecasts.

CHAPTER 7:

CONCLUSION

After an extensive evaluation and comparison of regression models for precipitation prediction, this study arrives at several key conclusions.

The K-Nearest Neighbors (KNN) model emerges as the top-performing regression model in this context. Its proximity-based approach to predictions, leveraging the similarity of data points, resulted in the lowest Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) scores among the models considered. This underscores the effectiveness of KNN in capturing local patterns and making accurate precipitation forecasts.

However, it's essential to recognize that the choice of the most suitable model depends on specific use cases and the nature of the data. For instance, Decision Trees offer valuable interpretability, which may be crucial in scenarios where model transparency is a priority. Ensemble methods like Random Forest and Gradient Boosting exhibit remarkable predictive power, making them strong contenders for applications requiring high forecasting accuracy.

Furthermore, Bayesian Ridge regression proves its worth in environments with noisy data, providing a principled approach to handling uncertainty. This characteristic makes it particularly relevant in situations where data quality may vary.

In conclusion, this study not only advances our understanding of regression models in precipitation prediction but also highlights the importance of selecting the right model based on the unique requirements of each scenario. The K-Nearest Neighbors model, with its impressive accuracy and generalization, stands out as a robust choice for accurate and reliable precipitation forecasts. Nevertheless, future applications should carefully consider the specific context, data characteristics, and priorities to make an informed choice of regression model.

This study lays a solid foundation for enhancing precipitation prediction in various industries, ultimately contributing to more effective decision-making in agriculture, hydrology, disaster management, and related fields.

CHAPTER 8:

FUTURE SCOPE

1. Advanced Feature Engineering: Explore more complex feature relationships and incorporate domain-specific knowledge for improved predictions.

2. Hyperparameter Tuning: Fine-tune model settings for optimized accuracy tailored to specific data characteristics.

3. Ensemble Methods: Combine models for enhanced predictive power, particularly in dynamic environments.

4. Real-Time Data Integration: Integrate up-to-date weather information for more timely and accurate forecasts.

5. User Interface Development: Create an intuitive interface for easy access to precipitation forecasts.

6. Geographical Adaptation: Modify models for different regions, accounting for unique meteorological patterns.

These future directions aim to refine and expand the capabilities of precipitation prediction models for broader practical applications.

CHAPTER 9:

REFERENCES

1. Chang, T. J., M. L. Kavvas, J. W. Delleur, "Daily precipitation modeling by discrete autoregressive moving average processes", Water Resour. Res., 17, 1261–1271, 1984. Link

2. Court, A., "Precipitation research, 1975–1978", Rev. of Geophys. and Space Phys., 176, 1165–1175, 1979. Link

3. Ari Yair Barrera-Animas, Lukumon O. Oyedele, Muhammad Bilal, Taofeek Dolapo Akinosho, Juan Manuel Davila Delgado, Lukman Adewale Akanbi, "Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting", Big Data Enterprise and Artificial Intelligent Lab (Big-DEAL), Bristol Business School, University of the West of England Bristol, United Kingdom, 2021. Link