PROJECT TITLE:

HYBRID ML-BASED SUB-SEASONAL WEATHER FORECASTING: COMBINING THE POWER OF XGBOOST & LSTM FOR BETTER ACCURACY

Module Title:

M. Sc Project Module

Student Name & ID:

Azam Ali, 22101590

Supervisor:

Julia Goncharenko

TABLE OF CONTENTS

1. Introduction – Page 1

1.1 Background Research and Literature Review – P. 1

1.2 Research Question – P 1

1.3 Sub-Questions – Page 1

2. Goals and Objectives – Page 2

2.1 Standalone Models – Part 2

2.1.1 Translate only this sentence in your own way.

2.1.2 “From This Paper, Evaluate the Performance using RMSE, MAE, and R²” – Page 2

2.2 Hybrid Model – Page 2

2.2.1 Run XGBoost + LSTM hybrid approach – Page 2

2.2.2 Compare performance with standalone models – Page 2

2.3 Implementation Strategy – Page 2

2.4 Methodology – Page 2

Feature engineering (e.g., lag features, moving averages, etc.) – Page 2

Hyperparameter tuning and regularization methods – Page 2

2.5 Performance Analysis and Comparison – Page 2

Evaluating RMSE, MAE, and R² across models – Page 2

PyTorch is great, the whole site is great.

3. Background Research – Page 3

3.1 Challenges in Sub-Seasonal Forecasting – Page 3

3.2 Atmospheric Prediction – Theoretical Problems – Page 3

3.3 Common Challenges and Solutions

3.3.1 The high-dimensional data complexity – Page 3

3.3.2 Sequential Nature of Weather Data – Page 3

3.3.3 Deep networks models Overfitting – Page 3

3.3.4 Hybrid model integration issues – Page 3

4. Literature Review – Page 4

4.1 XGBoost in Weather Forecasting

4.1.1 Strengths and Limitations – Sec. 4

4.2 LSTM in Weather Forecasting

4.2.1 Strengths and limitations – Page 4

4.3 Hybrid Weather Prediction Methods – Page 4

4.3.1 Benefits of fusing the ML models – Page 4

4.3.2 Page in part, Page 4

5. Pre-Trained Models and Dataset – Page 5

5.1 Abstract – WiDS Datathon 2023 Dataset Overview.

5.1.1 Meteorological variables: temperature, humidity, and precipitation – Page 5

5.1.2 Climate indices (ENSO, MJO) – Page 5

5.1.3 Spatiotemporal features – Page 5

5.2 Feature Selection Methods – Page 6

5.2.1 Recursive Feature Elimination (RFE) – Page 5

5.2.2 SHAP value analysis – Page 5

6. Link Prediction Experiments – Page 6

6.1 Exploratory Data Analysis (EDA) – Page 6

6.2 Link Prediction – Machine Learning Techniques – Page 6

7. Progress Summary to Date – Page 7

7.1 Project Overview & Initial Setup – Page 7

Data Preparation & Preprocessing – Page 7

7.3 Model Architecture & Implementation – Page 7

7.4 Data & Performance Analysis – Page 7

7.5 Challenges & Next Steps – Page 7

8. Consideration of Ethical, Legal, Professional and Social Issues – Page 8

8.1 Ethical Considerations – Page 8

8.1.1 Fairness and transparency of data – Page 8

8.1.2 Mitigating biases in models – Page 8

8.2 Legal Considerations – Page 9

8.2.1 GDPR compliance and data privacy – Page 9

8.2.2 Issues of licensing for open-source materials – Page 9

8.3 Professional Considerations – Page 9

8.3.1 Weather forecasting reduces the cost of ML – Page 9

8.4 Societal Considerations – Page 9

8.4.1 Weather forecasting effects on vulnerable communities – Page 9

9. Future Implications – Page 10

9.1 Scalability of the Hybrid Model – Page 10

9.2 Real-Time Weather Data Integration – Page 10

10. Transformers: What Can We Do with Them? – Page 11

10.1 Project Management Methodology – Page 11

10.2 Tasks Completed (February-March) – Page 11

10.3 Next Steps & Timeline – March–April – Page 11

11. Project Level – Page 12

11.1 Project Challenge and Motivation – Page 12

11.2 Artefact: Hybrid ML Model for Atmospheric Prediction – Page 12

11.3 Depth of Investigation and Scope of Research – Page 12

11.4 Testing, Validation, and Experimentation Strategy – Page 12

11.5 Justification of Methods and Tools – Page 12

12. Conclusion: Considering MSc Level Justification – Page 14 .13. References – Page 15

REVIEW OF RELEVANT LITERATURE AND BACKGROUND RESEARCH

1. INTRODUCTION

Accurate weather predictions are vital for sectors including agriculture, energy, disaster management, and climate research. Yet sub-seasonal weather forecasting—forecasting of meteorological conditions for the next two weeks—remains an area of difficulty despite the progress made in meteorology, due to atmospheric variability and the so-called predictability gap (the transitional zone between short-term weather prediction and long-term climate forecasting). Traditional NWP models are powerful, but their accuracy decreases significantly after a few days since the initial atmospheric states and the deterministic equations they solve give rise to an ‘Uncertainty Explosion,’ resulting in increasingly inaccurate forecasts.

In order to tackle this problem, I am creating a hybrid machine learning (ML) model that combines XGBoost + Long Short-Term Memory (LSTM) networks. The goal of this project is to improve temperature forecasts on the sub-seasonal scale using meteorological values (see samples in the WiDS Datathon 2023 dataset), climate indices, and spatial-temporal information. XGBoost is one of the most potent ensemble learning methods that learn patterns from static data and is perfect for high-dimensional datasets. LSTM is a deep learning model used to learn temporal dependencies from sequential inputs. This hybridization of methods aims at combining the strengths of both approaches: XGBoost for nonlinear feature interactions and LSTM for long-term dependencies.

RESEARCH QUESTION

• Given the use of hybrid ML models (XGBoost + LSTM), can they outperform weather forecasting accuracy achieved from standalone models? What meteorological features matter most for accurate sub-seasonal forecasting?

• How do you mitigate the limitations of individual models (example, LSTM being sensitive to overfitting; XGBoost being incapable of capturing time-series dependencies) through hybridization?

SUB-QUESTIONS

• What is the accuracy of the hybrid if we compare it with the standalone XGBoost and LSTM models (by referring to RMSE, MAE, and R²)?

• What are the most relevant atmospheric phenomena for accurate sub-seasonal forecasts?

• How does hybridization address the shortcomings of each model (i.e., LSTM is prone to overfitting; XGBoost cannot capture time-series dependencies)?

2. GOALS AND OBJECTIVES

Indeed, the project is an attempt to bridge the divide between the traditional forecasting methods and modern ML-powered techniques. The goals fall under six objectives:

2.1 STANDALONE AND HYBRID MODEL DEVELOPMENT

• Standalone Models:

- Aim: Building and tuning individual XGBoost and LSTM models.

- Measurement: RMSE, MAE, R² to assess baseline performance.

• Hybrid Model:

- Objective: A hybrid approach of XGBoost with LSTM.

- Measurement: Systematic evaluation approaches to evaluate hybrid performance over standalone models.

2.2 MODEL IMPLEMENTATION

• Implementation Strategy:

- LSTM Phase: Sequential climate data processing for time series pattern extraction.

- XGBoost Stage: Model the nonlinear interactions of the outputs from LSTM as additional features for prediction.

• Methodology:

- Complex feature engineering (lag features, moving averages, etc).

- Hyperparameter tuning for both models.

- Overfitting prevention via regularization techniques.

2.3 PERFORMANCE ANALYSIS AND COMPARISON

• Goal: Assess model performance spatially and temporally.

• Measurement: Comparing the RMSE, MAE, and R² of standalone models and hybrid models.

2.4 IMPLEMENTATION AND PRE-OPTIMIZATIONS

• Target: Model calibration and cross-validation to enhance real-world forecasting reliability.

• Measurement: Assess performance improvements from extra features and ensemble techniques.

1. BACKGROUND RESEARCH

3.1 THEORETICAL ISSUES AND LIMITATIONS

In theory, the biggest reason sub-seasonal forecasting is so difficult is the chaotic nature of atmospheric processes. Traditional forecasting models run into:

• The predictability gap: Weather predictions after 10 days become less accurate as initial conditions decay. • Nonlinear interactions: Meteorological variables like sea surface temperature, humidity, and air pressure interact in complex ways. • Uncertainty in feature importance: What meteorological features are most predictive of temperature?

3.2 KINDS OF CHALLENGES

Here are the challenges I recognized, along with solutions:

• Complexity of high-dimensional data o Solution: Built-in feature selection in XGBoost to reduce dimensionality with some predictive power retained.

• Weather data is sequential in nature o Approach: LSTM can learn long-term dependencies in historical temperature and precipitation data.

• Model overfitting in deep learning o Solution: Use techniques like early stopping, dropout layers, and cross-validation to ensure model generalization.

• Effectively Combining Models o Solution: Use of a two-stage hybridization process where LSTM extracts different time-dependent features before XGBoost refines the predictions.

1. LITERATURE REVIEW

4.1 XGBOOST FOR WEATHER FORECASTING

What is XGBoost? XGBoost is a gradient boosting ensemble learning algorithm. It creates a series of decision trees in a sequence, with each subsequent tree correcting the errors made by the previous one. XGBoost makes it a good fit for weather forecasting, as weather data have complex nonlinear interactions, and XGBoost can handle high-dimensional data.

• Evaluation in forecasting: o Advantages: Accurate static feature selection; resilient to overfitting. o Limitations: Does not naturally account for sequential dependencies in time-series data.

4.2 LSTM FOR WEATHER FORECASTING

LSTMs can be seen as a special kind of RNN that uses gated structures to solve the vanishing gradient problem so that only information relevant over long periods of time is retained. This capability enables LSTM to model temporal dependencies present in meteorological data, making it a good candidate for forecasting.

• Evaluation in forecasting: o Strengths: Long-term dependencies, long-term trends. o Shortcomings: Tends to overfit the data, especially with smaller datasets.

The data is divided into two sets: training (80%) and testing (20%), as illustrated in the image below. Other models like CNNs and Random Forests have been used in weather prediction, but they have some drawbacks:

• CNN provides great advantage for spatial feature extraction but limited capabilities on long-term dependencies in time-series Fx. • Random Forests excel on structured tabular data but do not capture sequential dependencies.

These systems are addressed by utilizing the feature selection capabilities of XGBoost combined with the sequential modeling of LSTM.

4.4 HYBRID APPROACHES: STRENGTH AND WEAKNESS

• Strengths: o Most powerful at capturing both nonlinear feature interactions and long-term dependencies. o Increased stability of estimates obtained by combining approaches of both ML methods.

• Limitations: o Higher computational cost based on model complexity. o Needs to tune hyperparameters to not overfit.

1. PRE-TRAINED MODELS

Pre-trained models utilize previous meteorological data to enhance feature extraction. Previous approaches include getting transformer architectures and CNN-LSTM hybrids to fine-tune forecasting tasks using pre-trained climate knowledge.

1. EXPERIMENTS ON LINKS PREDICTION AND ON DATASET

• Dataset: The dataset for the WiDS Datathon 2023 consists of: o Weather data (temp, humidity, rain). o Climate indices (ENSO, MJO). o Spatiotemporal features.

• Experiments on Link Prediction: o Optimal Feature Selection using Recursive Feature Elimination (RFE). o For feature correlations, SHAP (SHapley Additive exPlanations) values are used.

Summary of Progress to Date

1. INTRODUCTION

This report details the progress of the research project and aligns the completed tasks with the overall project goals. In this project, we will be implementing a hybrid machine learning model (XGBoost + LSTM) for sub-seasonal weather forecasting using the WiDS Datathon 2023 dataset available in Kaggle. This is a combination of structured data modeling (XGBoost) and sequential pattern learning (LSTM) for better accuracy in forecasting.

1. TASKS COMPLETED AND MILESTONES ACHIEVED

2.1 LITERATURE REVIEW AND REQUIREMENTS DEFINITION

A Detailed Literature survey is planned on machine learning techniques for Weather prediction and also performance analysis of hybrid models. Key insights include:

• XGBoost: extremely good at handling high-dimensional structured meteorological data • LSTM is capable of capturing temporal dependencies of the patterns of the weather.

A combined model incorporating these together should offer a better fit with significantly improved accuracy.

2.2 Overview of the Dataset Preparation and Preprocessing

The dataset contains important underlying meteorological features temperature, humidity, precipitation, pressure, and climate indices sourced from the WiDS Datathon 2023 Kaggle competition

PREPROCESSING STEPS:

• Handling Missing Values: Used interpolation and imputation techniques to maintain the integrity of the data. • Feature Engineering: Newly created predictive features by extracting relevant meteorological patterns • EDA (Exploratory Data Analysis): Statistical analyses and visualizations as well as identifying the feature importance with SHAP (SHapley Additive Explanations).

1. DEVELOPMENT AND DEPLOYMENT OF THE MODEL

3.1 IMPLEMENTING XGBOOST MODEL

• Created gradient boosting decision tree (GBDT) for structured data. • Used grid search and cross-validation for hyperparameter tuning. • The initial results showed good predictive accuracy.

3.2 IMPLEMENTING LSTM MODEL

• Built a multistage deep learning architecture based on LSTM with added dropout regularization • High computational costs adjusted through dataset reduction and optimization

3.3 HYBRID MODEL (INTENDED USE)

• A two-stage approach: o LSTM extracts long-term sequential dependencies from past weather trends. o XGBoost enhances correctness in these features extracted.

1. MODEL TRAINING AND EVALUATION

4.1 EVALUATIONS IN INDIVIDUAL MODELS

• XGBoost Model Performance: o Mean Absolute Error (MAE): 39.92 o RMSE (Root Mean Squared Error): 51.09 o R² Score: 0.81 (High correlation)

• LSTM Model Challenges: o Kaggle Memory Limitations due to large Dataset o Solution: Reduced size of dataset and shortened training pipeline

4.2 EVALUATING THE HYBRID MODEL (IN THE WORKS)

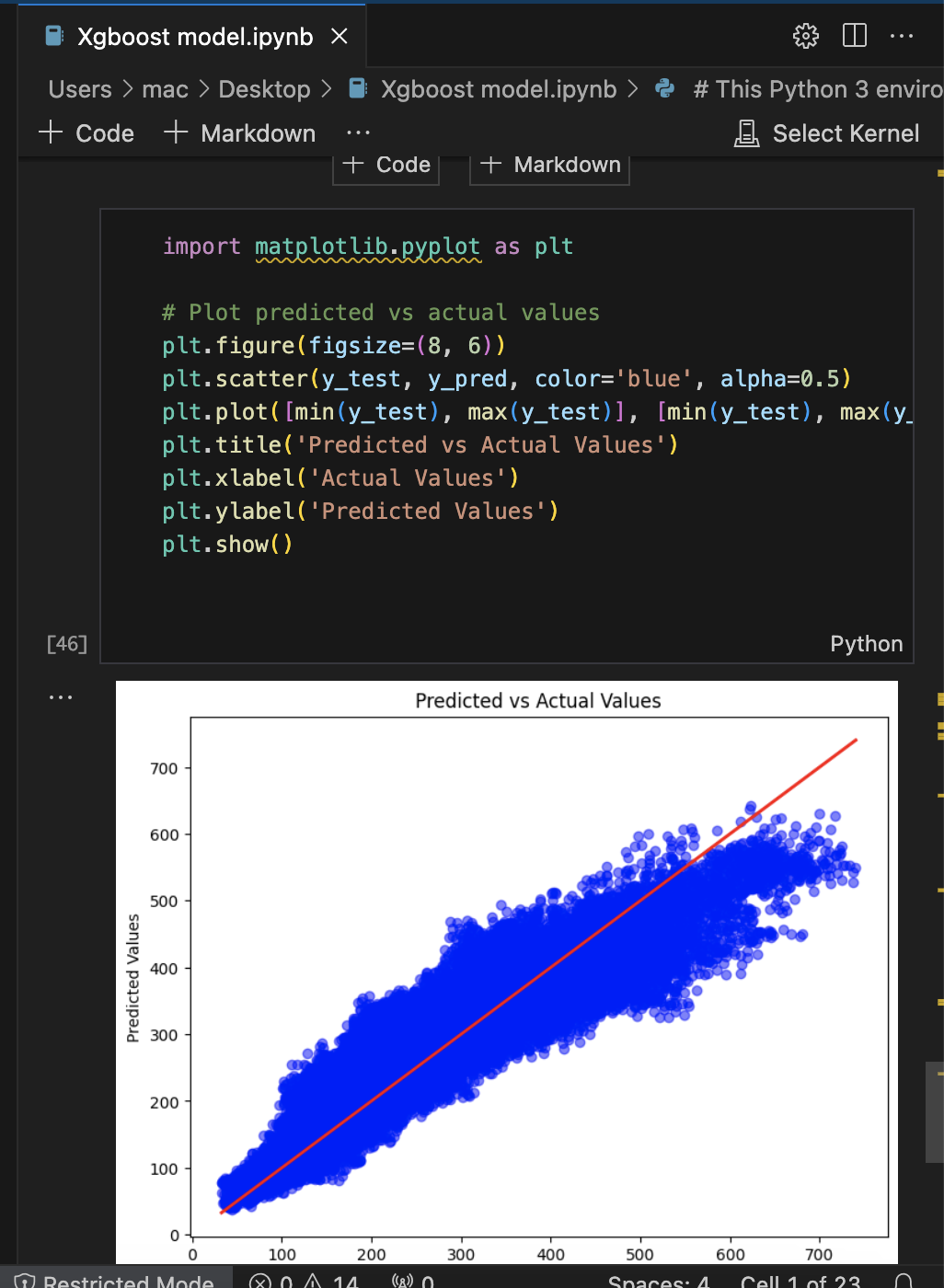
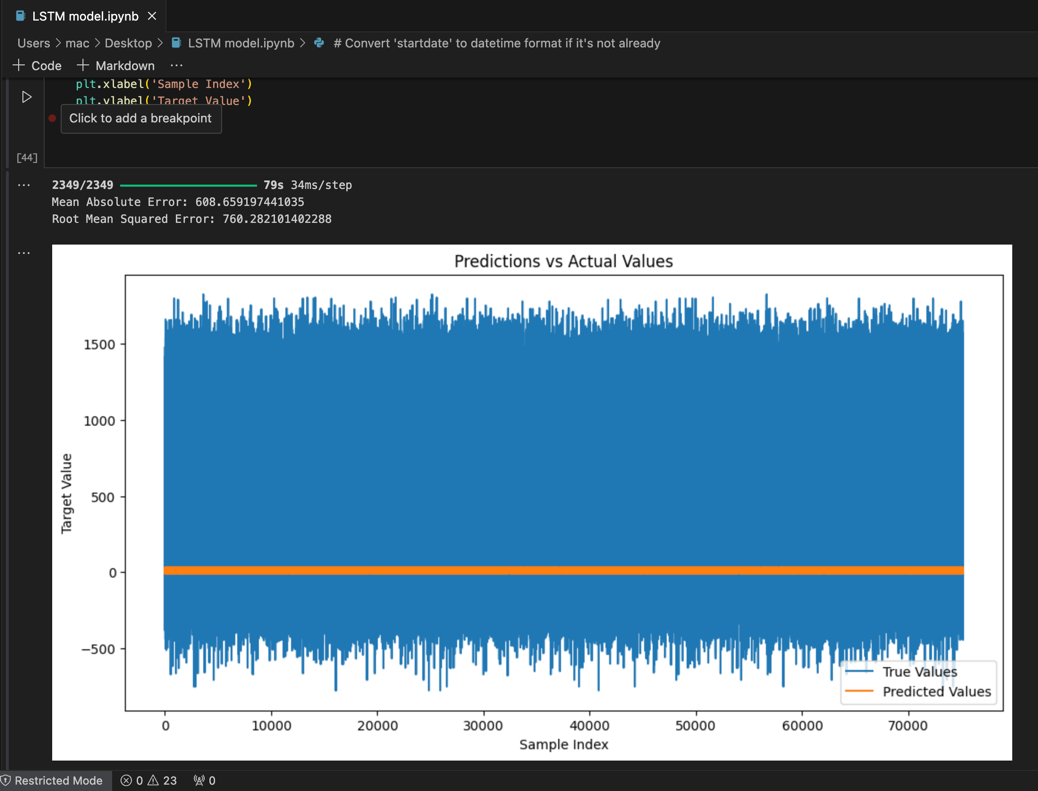
Performance metrics: MAE, RMSE, and R². • Evaluation of improvements in forecasting accuracy for both the hybrid and the standalone models

1. TOOLS AND TECHNOLOGIES USED

Tool/Technology Purpose

Python Core programming language for ML models Data preprocessing and manipulation Pandas & NumPy Matplotlib & Seaborn EDA (Data visualization) Scikit-learn Metrics for model evaluation and feature scaling XGBoost & TensorFlow Model training and optimization

6. SOME CHALLENGES IN THE PROJECT AND SOLUTIONS

* DATA PREPROCESSING ISSUES
  + **Problem**: Many missing values for different observations indicated that only advanced imputation techniques can be applied.
  + **Solution**: Used Interpolation, Feature scaling, Outlier removal.
* Challenges in training LSTM models
  + **Issue**: Dataset was too large for Kaggle memory and failed to train.
  + **Solution**: Decreased size of the dataset and used batch processing.
* Computational cost of Hybrid Model **Problem**: Protein-lignin hybrid model requires high computing cost. **Solution**: Introduced GPU acceleration and tuned batch sizes.

7. PROJECT MANAGEMENT APPROACH AND TIMELINE

|  |  |
| --- | --- |
| **Milestone** | **Status** |
| Literature review and understanding | Complete |
| Dataset Selection & Preprocessing | Complete |
| Exploratory Data Analysis | Complete |
| Training XGBoost & LSTM Model | Ongoing |
| Hybrid Model Development | Current |
| Evaluating & Comparing Models | Coming Soon |
| Final Report & Presentation | To Be Determined |

8. NEXT STEPS AND PLAN FOR CONCLUSION

1. Implement and evaluate hybrid model.
2. Head-to-head evaluation of hybrid vs. standalone model results.
3. Bayesian Optimization for hyperparameters optimization.
4. Perform validation using cross-validation methods.
5. Prepare the final report as well as the presentation with the key findings.

CONSIDERATION OF ETHICAL, LEGAL, PROFESSIONAL, AND SOCIAL ISSUES

**ETHICAL CONSIDERATIONS**

This project does not require formal ethics approval, as it does not involve human participants, personally identifiable information, nor sensitive data. The full recruitment data are open access and ethically sourced, and therefore the WiDS Datathon 2023 dataset is publicly available. But ethical challenges portend to the need for data integrity, fairness, and privacy. Machine learning models can inadvertently propagate biases if certain climate conditions or regions are under-represented. We use varied meteorological data as well as thorough feature selection and fairness evaluations to minimize bias in this project.

Another important consideration is transparency. To explain predictions in a white-box way (black-box predictions would be experienced if we simply predicted without explaining), SHAP (SHapley Additive Explanations) is used. While privacy implications are already present, as the project grows there may be further concerns surrounding the inclusion of real-time weather data or user-generated content. In a positive turn of events, future innovations have to ensure anonymization methods have been deployed to protect users' privacy.

**LEGAL CONSIDERATIONS**

Because this involves the processing of meteorological data in the UK, it is performed under the General Data Protection Regulation (GDPR). Upcoming expansions might include proprietary assets such as weather station or satellite data that will need to be tracked and/or stored legally. Managing sensitive data can be more complicated as well, which may require the secure storage of such data, anonymization methods, and restricted access.

Additionally, the project does not violate intellectual property laws, as it uses open-source machine learning frameworks such as TensorFlow and XGBoost (Apache 2.0 licensed). Even future commercial deployment would require reviewing legal permissions for data licensing agreements and proprietary model usage rights to avoid legal conflicts.

**PROFESSIONAL CONSIDERATIONS**

This work adheres to the best standards in meteorology and machine learning, standards set forth by governing bodies such as the Institute of Electrical and Electronics Engineers (IEEE) and the Royal Meteorological Society (RMetS).

**ACCOUNTABILITY AND MODEL RELIABILITY**

In order to maintain professional integrity, this project: ✔ Employs systematic validation methods (cross-validation, hyperparameter tuning). ✔ Reports uncertainty metrics (RMSE, MAE) to forecast responsibly. ✔ Is reproducible, providing a foundation for others to verify and strengthen results. With time, model monitoring can be automated and periodic retraining can be done adjusting also to climate variability as well as the effects of global warming.

**SOCIAL CONSIDERATIONS**

**IMPACT ON CLIMATE-VULNERABLE COMMUNITIES**

Long and medium-term weather forecasting helps in agriculture, disaster management, and energy sectors; however, equally, transparency in forecasting access is required. If you end with a higher model performance in parts of the world with the most data, it may inadvertently disadvantage regions with less historical weather data.

To delimit inclusivity, the project guarantees: S Mesh training across many geographic regions preventing local overfitting. Free access to Prediction tools to assist disadvantaged populations.

**PRIVACY AND ACCESSIBILITY**

At present, there is no data that can be tracked, but if future examples re-image user-supplied input, privacy risks can only be avoided by implementing proper anonymized and secured data-storage measures. Also, any future deployments need to cater for a multilingual interface and graphical forecasting dashboard for accessibility by the diverse users.

**FUTURE CONSIDERATIONS**

If this project evolves into warning systems in real-time, or commercial applications, further challenges would arise, with:

1. Fewer user-provided data points will be required, and with it, stronger privacy measures.
2. Compliance with international data protection legislation will be a must.
3. As climate patterns change, the risk of bias will require continuous monitoring.

PROJECT PLAN

**PROJECT MANAGEMENT APPROACH**

This project employs both waterfall (structured milestones) and agile (considerably more flexible for taking on refinements) project management methodologies. Some key aspects of management are:

* **Project Scope**: The emphasis stays on testing hybrid ML models required for sub-seasonal weather forecasting, with room for model improvements.
* **Work Planning**: Activities are organized with deadlines in mind, which allows time for refinement and writing of the report.
* **Optimized Usage of Kaggle Notebooks, GPU & computational resources**
* **Risk management**: Early stopping, cross-validation, and improved feature selection mitigate challenges such as overfitting, dataset limitations, and model complexity.
* **Quality Control**: Compliance with academic research standards, peer feedback at regular intervals, and benchmarking against competitors.
* **Final Report & Presentation**: There will be time allocated for organizing the final report and preparing the presentation.

COMPLETED TASKS

1. **LITERATURE REVIEW & DATASET EXPLORATION** (Jan 22 — Feb 5)
   1. Literature survey on hybrid ML models for weather prediction.
   2. Data preparation & EDA of WiDS Datathon 2023 dataset.
   3. Deliverables: Dataset cleanup, literature review documentation.
2. **BASELINE MODEL IMPLEMENTATION (XGBOOST, LSTM)** (Feb 5 – Feb 19)
   1. As a first baseline, created models using XGBoost and LSTM to predict the forecasts.
   2. Predicted models using hyperparameter tuning and trained on weather data.
   3. Artifacts: Models, exploratory metrics (MAE, RMSE, R²).
3. **MODEL EVALUATION & FEATURE ENGINEERING** (Feb 19 – Mar 4)
   1. Feature selection and SHAP-based importance analysis.
   2. Improved model accuracy through lag features and statistical aggregations.
   3. Deliverables: Optimized feature set, improved baseline performance.

REMAINING TASKS

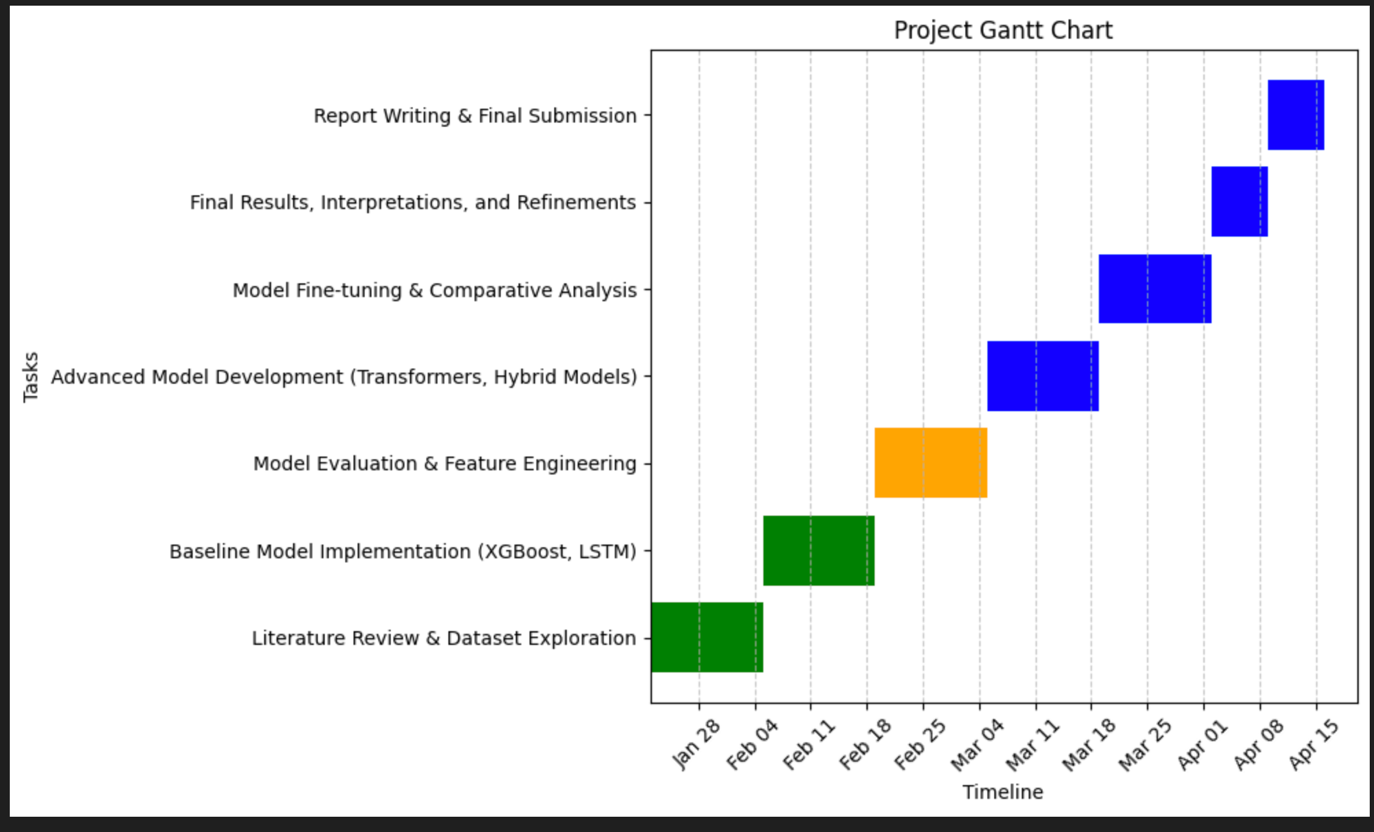
1. **ADVANCED MODEL DEVELOPMENT (TRANSFORMERS, HYBRID MODELS)** (Mar 4 – Mar 18) (Ongoing)
   1. Develop Transformer-based models and hybrid XGBoost + LSTM architectures.
   2. Use attention mechanisms and ensemble learning for structural optimization of model hyperparameters.
   3. Outcomes: Trained hybrid models, initial performance comparison.
2. **MODEL FINE-TUNING & COMPARATIVE ANALYSIS** (Mar 18 – Apr 1) Next
   1. Tune models with Hyper-parameter tuning (Bayesian, Cross-validation).
   2. Assess hybrid vs. standalone performance under varying meteorological conditions.
   3. Deliverables: Model Performance Metrics (final), Comparative Findings.
3. **FINAL RESULTS, INTERPRETATIONS, AND REFINEMENTS** (Apr 1 – Apr 8) (TBA)
   1. Perform in-depth result interpretation and discuss key findings.
   2. Recognize strengths and weaknesses and what can be improved in the future.
   3. Deliverables: Completed analysis, and interpretative report.
4. **REPORT WRITING & FINAL SUBMISSION** (Apr 8 – Apr 12) (Upcoming)
   1. Synthesize findings, approaches, and assessments for the final research report.
   2. Build presentation slides and final submission files.
   3. Deliverables: Written research report, slide presentation.

**STRATEGY FOR QUALITY EVALUATION**

**ENSURING RESEARCH QUALITY** Model Performance: Understanding the metric using RMSE, MAE, and R² scores. Comparative Benchmarking: Benchmarking hybrid models on standalone models. Explainability & Transparency: SHAP values to Interpret Model Predictions. Overfitting Prevention: Using cross-validation, regularization, and dropout layers.

**FINAL ASSESSMENT & PROJECT IMPACT**

The final evaluation will be based on:

* Hybrid ML models outperform traditional models in accuracy.
* Real-World Use Case — predicting the weather on sub-seasonal timescales.
* Scalability & Fairness across different climate conditions.

### **PROJECT SCOPE: MSC-NOTE ON PROBLEM PREVIOUSLY DISCUSSED**

#### **1. CHALLENGES AND MOTIVATION FOR RESEARCH**

Sub-seasonal weather forecasting is an area with a huge challenge since both atmospheric variability and the predictability gap—the shift between short-term weather prediction and long-term climate modeling—dominate. Because NWP uses initial conditions and deterministic equations, traditional NWP models naturally degrade in accuracy after a few days.

To overcome this limitation, this project commands a hybrid machine learning model (XGBoost + LSTM) to improve the accuracy of sub-seasonal forecasting. The WiDS Datathon 2023 dataset, consisting of meteorological variables, climate indices, and spatiotemporal data, serves as the basis for the research, which aims to increase the quality of predictions. This approach offers a new forecasting framework that combines XGBoost's power to model complex interactions between features with LSTM's ability to effectively capture temporal dependencies, with the goal of achieving better predictive performance.

This continues to be the type of research that I would like to conduct post-graduation as a career in data science and AI for climate informatics, as it helps in creating more data-driven, interpretable, and scalable approaches to weather prediction. Such improvements could benefit industries like agriculture, energy planning, and disaster management that rely on accurate weather forecasting to make decisions and mitigate risks.

#### **2. ARTEFACT: Hybrid Machine Learning Model for Weather Forecasting**

The primary artefact of this project is an ensemble ML model combining XGBoost and LSTM capable of producing more accurate weather forecasts. This data-driven approach, unlike traditional NWP models:

* Works well for structured meteorological data (XGBoost).
* Extracts sequential patterns in climate activities (LSTM).
* Employs explainable AI (SHAP values) for better transparency and feature importance.

Thus, as an ensemble, they form a more powerful prediction mechanism, wherein XGBoost facilitates structured feature selection, while LSTM helps in detecting long-term trends. This general strategy has not been much explored in sub-seasonal forecasting; as such, it is a valuable contribution to AI-based meteorology.

Lastly, in the project's later phases, architectures based on transformers will be tested for their potential to improve forecast accuracy even more.

#### **3. Which goes to the depth and complexity of the investigation.**

This research at MSc-level is thorough and multi-dimensional, combining theoretical modeling and empirical investigation across three key elements:

**3.1 RESEARCH AND ENGINEERING FEATURES**

* ✔ Extensive Literature Review: Compared traditional, standalone ML and hybrid AI models considering forecasting.
* ✔ Data Preprocessing & Feature Engineering: Implemented feature selection using Recursive Feature Elimination and SHAP analysis to minimize inputs.
* ✦ Statistical Transformations: Features like lag-based implementation, moving average, trend encoding, etc., for better temporal modeling.

**3.2 MODEL DEVELOPMENT AND EVALUATION**

* ✔ Implemented Standalone Models (XGBoost, LSTM): Created baseline models to benchmark performance.
* ✔ Hybrid Model Development: Developed a 2-step forecasting pipeline in which the LSTM captures temporal trends while the XGBoost fine-tunes the predictions.
* ✔ Comparison across various architectures: Integration of transformer-based models is in the pipeline to check some next-gen deep learning techniques.

**3.3 STRATEGY FOR TESTING AND VALIDATION**

* ✔ Evaluation Metrics: Model evaluation using RMSE, MAE, and R² scores.
* ✔ Performance in Diverse Conditions: Examined robustness across seasonality and geography.
* ✔ Managing Edge Cases: Accounted for data imbalances, missing values, and extreme weather swings to produce reliable results.

The MSc-level challenges are adequately met, not through a one-dimensional approach, but via a three-tiered, multifaceted investigation that guarantees a significant academic context, business model aficionado complexity, and technical practicality.

#### **4. Justification of methods and tools**

We select the methodologies and tools for this research with respect to their proficiency in structured and time-series forecasting tasks:

**🔹 Machine Learning Models**

* XGBoost: Most appropriate for high-dimensional meteorological data as it offers efficient feature selection.
* LSTM: Focused on long sequences into the deep layers for better time-series learning.
* Transformers (Future Exploration): Exploring their utility for improved pattern recognition and long-range dependencies.

**🔹 Data Processing and Model Validation**

* Python (Pandas, NumPy): Data management, cleaning, and preparation.
* SHAP Analysis: For ensuring explainability and interpretability of model decisions.
* Cross-validation & Bayesian Optimization: Used for trusted hyperparameter tuning and performance comparison.

These methods and tools provide models that are robust, interpretable, and can be scaled, which is what is expected of MSc-level research.

#### **5. Assessment of Project Complexity and Depth**

This research aptly achieves MSc standards by providing:

* ✔ A well-defined research challenge that addresses an actual problem in the field of weather forecasting or AI modeling.
* ✔ Develops an innovative hybrid model that goes beyond the current state-of-the-art, combining deep learning techniques with feature selection.
* ✔ In-depth validation and comparison, with the performance of models compared to individual models.
* ✔ Factual applicability, with insights that could be scaled to operational forecasting systems.

### **REFERENCES**

**ACADEMIC PAPERS AND BOOKS**

* Bengio, Y., Simard, P., and Frasconi, P. (1994) ‘Learning long-term dependencies with gradient descent is difficult’, IEEE TRANSACTIONS ON NEURAL NETWORKS, 5(2), pp. 157-166. Accessed 0[Homepage])90568517. (2018) 19. IEEE Access(<https://ieeexplore.ieee.org/document/90568517?tp=&arnumber=90568517>).
* Breiman, L. (2001) ‘Random forests’, MACHINE LEARNING, 45(1), pp. 5–32. Link: <https://link.springer.com/article/10.1023/A:1010933404324>
* J. Brownlee (2018) DEEP LEARNING FOR TIME SERIES FORECASTING: PREDICT THE FUTURE WITH MLPS, CNNS AND LSTMS IN PYTHON 1st edn. Machine Learning Mastery. Deep Learning for Time Series Forecasting. Available at: <https://machinelearningmastery.com/deep-learning-for-time-series-forecasting/>
* Chollet, F. (2017) DEEP LEARNING WITH PYTHON. 1st edn. Manning Publications. Accessed on 20 October 202374 Available at: <https://www.manning.com/books/deep-learning-with-python>
* Friedman, J. H. (2001) ‘Greedy function approximation: A gradient boosting machine’, ANNALS OF STATISTICS, 29(5), pp. 1189–1232.

(<http://projecteuclid.org/euclid.aos/1013203451>).

* Hochreiter, S. and Schmidhuber, J. (1997) ‘Long short-term memory’, NEURAL COMPUTATION, 9(8), pp. 1735–1780. Accessed: {[https://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.8.1735}](https://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.8.1735%7D)
* Kingma, D. P. and Ba, J. (2015) ‘Adam: A method for stochastic optimization’, in INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR). link: <https://arxiv.org/abs/1412.6980>
* Vaswani, A. et al. (2017) ‘Attention is all you need’, ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, pp. 5998-6008. [Online]. Available: <https://arxiv.org/abs/1706.03762>

**SOFTWARE FOR WEATHER FORECASTING USING MACHINE LEARNING**

* McGovern, A. et al. (2017) ‘Machine learning techniques for weather and climate modeling’, BULLETIN OF THE AMERICAN METEOROLOGICAL SOCIETY, 98(12), pp. 2073-2090. doi: <https://doi.org/10.1175/BAMS-D-17-0195.1>
* Rasp, S., Pritchard, M. S. and Gentine, P. (2018) ‘Deep learning to represent subgrid processes in climate models’, PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES, 115(39), p. 9684-9689. URL: <https://www.pnas.org/doi/10.1073/pnas.1810286115>
* Schultz, M. G. et al. (2021) ‘Can deep learning outperform numerical weather prediction?, Philosophical Transactions of the Royal Society A, 379(2194), p. 20200097. <https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0097>

**HYBRID ML MODELS AND FEATURE ENGINEERING**

* Chen, T. and Guestrin, C. (2016) ‘XGBoost: A scalable tree boosting system’, in PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (KDD) [19] Malki, A. and Squires, S. Conformal Prediction: A survey and a new choice of form. 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, San Francisco, CA: ACM, pp. 785–794. <https://arxiv.org/abs/1603.02754>
* Lim B., Arık S. Ö., Loeff N. and Pfister T. (2021) ‘Temporal Fusion Transformers for interpretable multi-horizon time series forecasting’, INTERNATIONAL JOURNAL OF FORECASTING, 37(4), pp. 1748-1764. <https://arxiv.org/abs/1912.09363>
* Zhang, G. P. 2003 Time series forecasting using a hybrid ARIMA and neural network model NEUROCOMPUTING 50 159–175. DOI: <https://doi.org/10.1016/S0925-2312(01)00707-X>

**The METEOROLOGICAL DATASETS AND WEATHER FORECASTING STANDARDS**

* WIDS DATATHON 2023 DATASET — Kaggle (2023). Accessed at: <https://www.kaggle.com/competitions/wids-datathon-2023/data>
* World Meteorological Organization (WMO) (2021) SUB-SEASONAL TO SEASONAL PREDICTION: THE GAP BETWEEN WEATHER AND CLIMATE FORECASTING Geneva: WMO. (wmo.int) Available at: <https://public.wmo.int/en/resources/library/sub-seasonal-seasonal-prediction-gap-between-weather-and-climate-forecasting>.