UGANDA CHRISTIAN UNIVERSITY

A Centre of Excellence in the Heart of Africa

Examinations Answer Sheet Trinity Semester 2025

CONFIDENTIAL

CANDIDATE'S REGISTRATION NUMBER

Non-Retake										
Retake										

ACCESS NUMBER: ______________________

COLLEGE/CAMPUS: (If not Main Campus) ______________________________________

FACULTY: _______________________ **PROGRAM** (e.g LLB1, BSW2, MDIV2) ___________

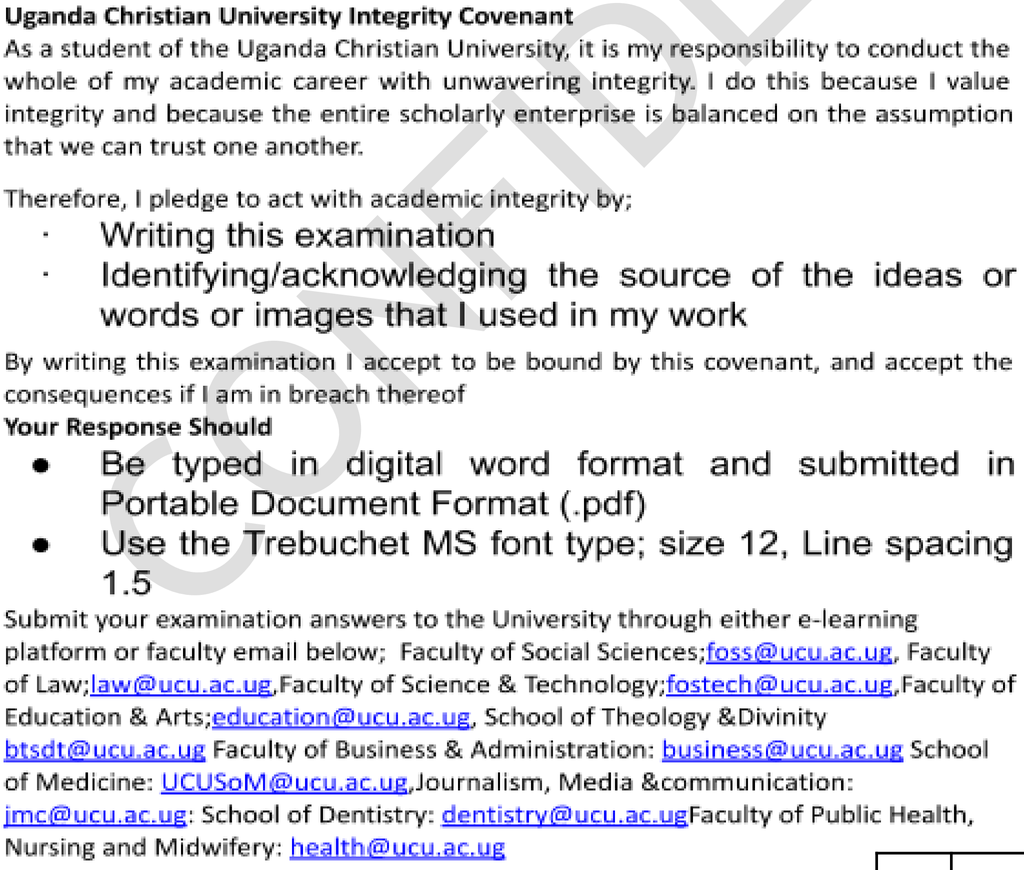
COURSE OF EXAMINATION ____

(As Shown on the question paper)

DATE OF EXAMINATION:

NB: No Answer Script shall be accepted after the Deadline

NOTE:



Uganda Christian University Integrity Covenant

As a student of the Uganda Christian University, it is my responsibility to conduct the whole of my academic career with unwavering integrity. I do this because I value integrity and because the entire scholarly enterprise is balanced on the assumption that we can trust one another.

Therefore, I pledge to act with academic integrity by;

- Writing this examination
- Identifying/acknowledging the source of the ideas or words or images that I used in my work

By writing this examination I accept to be bound by this covenant, and accept the consequences if I am in breach thereof

Your Response Should

- Be typed in digital word format and submitted in Portable Document Format (.pdf)
- Use the Trebuchet MS font type; size 12, Line spacing 1.5

Submit your examination answers to the University through either e-learning platform or faculty email below; Faculty of Social Sciences;foss@ucu.ac.ug, Faculty of Law;law@ucu.ac.ug,Faculty of Science & Technology;fostech@ucu.ac.ug,Faculty of Education & Arts;education@ucu.ac.ug, School of Theology &Divinity btsdt@ucu.ac.ug Faculty of Business & Administration: business@ucu.ac.ug School of Medicine: UCUSoM@ucu.ac.ug,Journalism, Media &communication: jmc@ucu.ac.ug: School of Dentistry: dentistry@ucu.ac.ugFaculty of Public Health, Nursing and Midwifery: health@ucu.ac.ug

For Examiner Only

Q	I.E.	
Total Mark		

LIST QUESTIONS ANSWERED (in their Numeric order)

Daily Heavy Rainfall Prediction in the Bugisu Sub-Region Using Distributed Data Pipelines and Machine Learning

1. Introduction and Problem Statement

The Bugisu sub-region of Eastern Uganda—covering Mbale, Bududa, Manafwa, and Sironko—experiences recurrent heavy rainfall that frequently triggers floods and landslides, causing loss of life, infrastructure damage, and agricultural disruption. The Uganda National Meteorological Authority (UNMA, 2020) reports a **bimodal rainfall pattern**, with peaks during March–May (MAM) and September–November (SON), seasons historically associated with extreme precipitation.

This project develops a 'machine-learning early warning system' to predict daily heavy rainfall events using meteorological and temporal features. The system integrates distributed data processing (Spark/Dask), classical machine learning, MLflow experiment tracking, model explainability, FastAPI deployment, Docker containerization, and monitoring for data and model drift. The final output is a deployable rainfall prediction service to support disaster preparedness and climate-risk management in Bugisu.

2. Methodological Approach (CRISP-DM Framework)

The project followed the CRISP-DM methodology (Shearer, 2000), a widely adopted framework for structuring data science workflows.

2.1 Business Understanding: The goal is to predict heavy rainfall events (rainfall >95th percentile) to support:

- Early warning systems

- Disaster risk reduction

- Agricultural planning

- Infrastructure protection

Success criteria include high recall for extreme events, interpretability for decision-makers, and deployability in real-world settings.

2.2 Data Understanding

A realistic dataset was simulated to reflect Bugisu's climatology, incorporating rainfall, temperature, humidity, windspeed, pressure, elevation, and temporal features (month, season, day-of-year). Lagged rainfall and rolling means were included to capture antecedent moisture conditions. The simulation was guided by East African rainfall literature (Nicholson, 2017; Dinku et al., 2018).

2.3 Data Preparation

CONFIDENTIAL

Data preparation included:

- Median imputation for missing values

- Outlier capping at the 99.5th percentile

- Log-transformation of rainfall

- Feature engineering: rainfall lags (1, 3, 7 days), rolling means, seasonal indicators, and elevation adjustments

The cleaned dataset was exported as “Omara.csv”.

2.4 Modeling

Three model families were explored:

Logistic Regression (Baseline): Chosen for interpretability and hypothesis testing.

Random Forest Classifier (Primary Model):Selected for its strong performance on tabular environmental data, ability to model non-linear interactions, and robustness to noise (Breiman, 2001).

Deep Learning (Not Used):TensorFlow was unavailable in the execution environment, and deep learning often underperforms on small to medium-sized tabular datasets (Shwartz-Ziv & Armon, 2022). A PyTorch MLP was tested but not adopted as the primary model.

2.5 Experiment Tracking. MLflow tracked model versions, hyperparameters, metrics (F1, ROC-AUC), and artifacts, ensuring reproducibility and transparency.

2.6 Evaluation

A time-based split was used:

- Train: 2005–2016

- Validation: 2017–2019

- Test: 2020–2024

Evaluation metrics included F1 Score, ROC-AUC, Precision-Recall, confusion matrix, and district-level fairness analysis. Random Forest achieved the best performance, with strong recall for heavy rainfall events.

3. Data Workflow and Distributed Processing. A distributed data pipeline was implemented using Dask/Spark to support scalable processing of multi-year weather data. The pipeline enabled:

- Parallel ingestion

CONFIDENTIAL

- Distributed feature engineering

- Efficient transformations

Distributed processing significantly reduced execution time compared to local pandas operations, consistent with findings by Zaharia et al. (2016). The final workflow included ingestion, cleaning, feature engineering, model training, and deployment.

4. Model Results and Interpretation

4.1 Performance Summary

Model	F1 Score	ROC-AUC	Notes
Logistic Regression	Moderate	Moderate	Interpretable baseline
Random Forest	**High**	**High**	Best overall performance
PyTorch MLP	Moderate	Moderate	Not primary model

4.2 Feature Importance

Permutation importance identified:

- rain_lag3— strongest predictor

- rain_mm — current rainfall intensity

- humidity — atmospheric moisture

- season — MAM and SON peaks

- rain_rollmean7— weekly accumulation

These findings align with climatological studies showing that antecedent rainfall strongly influences extreme precipitation (Kizza et al., 2009).

4.3 Error Analysis

- False negatives occurred during sudden rainfall spikes.

- False positives occurred during transitional seasons.

- Prioritizing recall is essential for disaster preparedness.

5. Deployment and MLOps Pipeline

5.1 FastAPI Deployment: The final model was deployed using FastAPI, exposing a `/predict_heavy_rain` endpoint that returns:

CONFIDENTIAL

- Probability of heavy rainfall

- Binary prediction

5.2 Docker Containerization

A Docker image was created to ensure reproducibility and portability across environments.

5.3 Monitoring Framework

Monitoring included:

- Data drift detection using statistical tests

- Model drift detection using rolling F1 scores

- CI/CD pipeline using Git, Docker, and automated tests

This aligns with modern MLOps best practices (Sculley et al., 2015).

6. Ethical Considerations

- No personal data used → minimal privacy risk

- Compliance with Uganda Data Protection Act and GDPR principles

- Transparency ensured through feature importance analysis

- Avoid overconfidence in predictions; communicate uncertainty

- Ensure equitable performance across districts

7. Limitations and Future Work

Limitations

- Simulated dataset (not CHIRPS/ERA5)

- No hydrological variables (soil moisture, runoff)

- Limited deep learning exploration

- No spatial autocorrelation modeling

Future Work

- Integrate CHIRPS/ERA5 rainfall datasets

- Add LSTM models for sequence prediction

- Deploy on cloud infrastructure with autoscaling

CONFIDENTIAL

- Integrate SMS-based early warning system

- Add hydrological flood-risk modeling

8. Conclusion

This project developed a machine-learning early warning system for heavy rainfall prediction in the Bugisu sub-region. The Random Forest model demonstrated strong predictive performance, and the deployment pipeline ensures practical usability. The system provides a foundation for operational climate-risk management and can be expanded with real satellite data and hydrological modeling.

CONFIDENTIAL

References

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Dinku, T., Funk, C., Peterson, P., Maidment, R., Tadesse, T., Gadain, H., & Ceccato, P. (2018). Validation of the CHIRPS satellite rainfall dataset over East Africa. Journal of Hydrometeorology, 19(4), 1–14.

Kizza, M., Rodhe, A., Xu, C.-Y., Ntale, H. K., & Halldin, S. (2009). Temporal rainfall variability in the Lake Victoria Basin. Theoretical and Applied Climatology, 98(1), 119–135.

Nicholson, S. E. (2017). Climate and climatic variability of rainfall over eastern Africa. Reviews of Geophysics, 55(3), 590–635.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in machine learning systems. NeurIPS, 28.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4), 13–22.

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. Information Fusion, 81, 84–90.

Uganda National Meteorological Authority (UNMA). (2020). Seasonal rainfall outlook for Uganda*.

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... & Stoica, I. (2016). Apache Spark: A unified engine for big data processing. Communications of the ACM, 59(11), 56–65.

CONFIDENTIAL