**Cloudburst Prediction Practicum Final Report**

**1. Problem Statement**

Cloudbursts are an extreme weather condition characterized by sudden and intense precipitation exceeding 100mm/hr or 3.94 inches/hr. Meteorologically, these events result from rapid atmospheric instability and moisture convergence, often associated with severe weather patterns like thunderstorms or monsoon dynamics. Understanding and predicting cloudbursts is critically important due to their substantial socioeconomic impacts, including infrastructure damage, loss of life, disruptions to economic activities, and long-term psychological effects on affected communities. Accurate early prediction allows authorities to implement timely disaster mitigation strategies, potentially saving lives and reducing economic losses.

However, there remain tremendous technical challenges today that prevent high-accuracy and timely cloudburst predictions. Cloudbursts typically occur in small, localized areas, complicating detection and prediction efforts. Technical issues also arise from limitations in satellite, radar, and sensor technologies. For example, ultrasonic sensors can be easily obstructed by physical objects; geostationary satellites have limited resolution and temporal coverage; multiple Doppler weather radars can monitor moving cloud droplets and help to provide a forecast for the next three hours, but they are expensive and installing them widely may not be feasible. Moreover, the lack of globally shared weather data exacerbates predictive challenges, often compounded by geopolitical issues that restrict data access and sharing. Communication gaps have further impeded disaster preparedness, with instances where successfully predicted cloudburst events failed to reach local authorities and communities in time for effective action.

The nature of cloudbursts and the technical challenges we face impose difficulties during the prediction modelling processes. Firstly, the rarity of cloudbursts creates a class imbalance issue, complicating machine learning model training and reducing prediction reliability. Secondly, cloudburst prediction is hindered by data sparsity due to limited weather station coverage, especially in less developed or geographically challenging regions. Another significant challenge is ensuring the generalizability of models trained on datasets from specific geographical areas, as weather dynamics and climate patterns vary widely across regions, particularly between temperate and tropical climates.

The primary objective of this cloudburst prediction project, as part of the INDRA initiative, is to develop a machine learning model capable of detecting cloudburst occurrences within a 3-hour prediction window, achieving accuracy above existing baselines (currently around 50%). The model aims to enhance early warning systems and disaster preparedness, supporting practical real-world applications through clearly documented methodologies and actionable insights.

**2. Literature Review**

This review synthesizes findings from three recent studies on cloudburst and extreme precipitation prediction:

**Murakami et al. (2022)**

Murakami et al. leveraged a deep learning autoencoder model to detect precipitation anomalies in Japan, correlating rising extreme events with climate change. Their method employed unsupervised anomaly detection, identifying unusual rainfall patterns through reconstruction errors (mean squared error thresholds). The approach is beneficial for broad anomaly detection but lacks precision and recall metrics, making it less suited for short-term, event-specific predictions. Despite this limitation, the methodology provides insight into anomaly detection frameworks and emphasizes the importance of climate trend analyses in predicting large-scale weather anomalies.

**Patil & Kulkarni (2023)**

Patil and Kulkarni combined Global Forecast System (GFS) forecasts with machine learning algorithms including XGBoost and neural networks. Their binary classification model achieved impressive metrics: accuracy (88%), F1 score (0.84), precision (0.85), and recall (0.83). They also developed a multiclass model that maintained robust performance. This work highlights the effectiveness of integrating meteorological forecast data with advanced supervised machine learning techniques for short-term forecasting. The high F1 score and balanced precision and recall demonstrate strong predictive capabilities for operational early-warning systems, making this approach highly relevant to our project’s objectives.

**Shani & Nagappan (2024)**

Shani and Nagappan utilized Gramian Angular Fields (GAF) transformation combined with Convolutional Neural Networks (CNN) to classify cloudburst events. Their model achieved an accuracy of 86.4%, though the recall was moderate at 0.58, indicating a substantial proportion of missed cloudburst events. This approach underscores the potential of CNN models to effectively capture complex temporal-spatial patterns inherent in precipitation data but also highlights critical limitations regarding recall. Such CNN-based architectures provide a valuable foundation for our INDRA project, suggesting potential improvements in data handling or feature engineering to increase recall.

**Comparative Summary and Bridging to INDRA**

Murakami et al.'s study offers foundational knowledge on anomaly detection for large-scale climate trends but lacks specificity for real-time operational use. In contrast, Patil and Kulkarni demonstrate how supervised ML combined with forecast models significantly improves short-term predictive accuracy, making their approach directly applicable to our 3-hour prediction window. Shani and Nagappan’s CNN architecture presents valuable insights into sophisticated pattern recognition, despite its lower recall, suggesting possible areas for refinement.

Overall, these studies collectively inform our project's direction: leveraging high-resolution datasets, incorporating supervised learning with forecast model data, and exploring CNN methodologies with enhanced feature engineering to improve recall. This combined methodological insight guides our approach toward developing a robust and operationally useful cloudburst prediction model for the INDRA project.

**3. Data Processing and Feature Engineering**

The NOAA Local Climatological Data (LCD) dataset was selected due to its comprehensive hourly meteorological observations. The dataset used here comes from the Miami International Airport station (Station ID: 72202012839), spanning from January 2015 to December 2024. For cloudburst prediction, the most relevant variables included DryBulb temperature, WetBulb temperature, DewPoint, wind speeds, precipitation, and pressure readings.

Data cleaning was critical due to extensive missing values and reporting inconsistencies. Initial preprocessing steps involved loading data, selecting relevant hourly variables, filtering by report priority (FM-15 > FM-12 > FM-16 > SOD > SOM), and converting timestamps into datetime formats. Missing value patterns were visualized, and duplicates or clear outliers were systematically removed. These steps ensured a cleaner and more consistent dataset, improving the reliability and feasibility of building effective machine learning models.

Feature engineering was essential to capture meteorological dynamics relevant to cloudburst prediction. Time-based features were created (e.g., hour, month, weekend flags, sinusoidal transformations to represent cyclical patterns). Weather-derived variables included vapor pressure, wind U/V components, and precipitation. Lagged features were introduced (1-hour, 3-hour, and 6-hour intervals) to capture temporal dependencies, improving the predictive capability of the model. These terms are created because the project objective is to predict whether this event will occur within the defined window based on current and past meteorological conditions. We would like to investigate whether weather conditions preceding cloudbursts have a significant influence on their occurrence. Interaction terms were also considered to capture potential nonlinear relationships and combined effects between meteorological variables, as interactions between variables such as humidity and temperature might critically affect the likelihood of cloudbursts. That is, the use of single variables without considering their interaction effects might lead to an underestimation of their combined influence on cloudburst occurrence, thereby limiting the model's ability to detect complex atmospheric conditions associated with extreme rainfall.

A screenshot of a computer

Description automatically generated

Example code snippet illustrating key feature engineering steps:A screen shot of a computer code

Description automatically generated

A screen shot of a computer

Description automatically generatedA screen shot of a computer code

Description automatically generated

Additionally, domain knowledge informed by a 2022 article from The Hindu (<https://www.climate.rocksea.org/images/climate/Cloudbursts_TheHindu_11Sep2022.pdf>) guided the selection of more nuanced predictors. Two examples implemented include: (1) the use of the Clausius-Clapeyron relationship to model temperature–moisture dynamics, which calculates potential moisture capacity based on dry bulb and dew point temperature, and (2) rapid temperature change indicators to capture sudden heating or cooling events that may precede cloudbursts. These physics-informed transformations help align meteorological theory with data-driven approaches, enhancing the selection, interpretability, and relevance of engineered features.

A screenshot of a computer program

Description automatically generated

A screen shot of a computer code

Description automatically generated

Two significant challenges emerged during data processing:

1. **Class Imbalance**: Cloudbursts, as rare events, created severe imbalance in the dataset. To address this, I first relaxed the threshold defined for a cloudburst. Then, synthetic minority oversampling (SMOTE) was applied, ensuring balanced class representation for effective model training.
2. **Missing Values**: Extensive missing data required strategic imputation and filtering techniques. Linear interpolation was applied to weather variables with moderate missingness, while variables with substantial gaps or inconsistent reporting were excluded from modeling.

These careful preprocessing steps and feature engineering methodologies were foundational in creating a robust dataset, suitable for training reliable predictive models.

1. **Model Implementation and Evaluation**

I created both a random forest model and an autoencoder model. The Random Forest model was chosen for its robust handling of noisy, imbalanced, and high-dimensional meteorological data. Key hyperparameters used were:

* Number of Trees (n\_estimators): 100
* Maximum Depth (max\_depth): 10
* Minimum Samples per Leaf (min\_samples\_split): 5

Model training involved splitting the dataset into training and testing sets based on weeks to maintain temporal integrity. Feature importance analysis revealed that "HourlyPrecipitation" was the most influential variable (makes sense since cloudburst is basically precipitation), followed by "HourlyPressureChange," "HourlyWindSpeed," "HourlyDryBulbTemperature," and "HourlyRelativeHumidity."

The evaluation metrics obtained were:

* Precision: 0.5
* Recall: 0.005
* F1-score: 0.0099

It indicates that only half of the predicted cloudbursts were actually cloudbursts. The 0.5% low recall rate suggests nearly all actual cloudbursts (heavy rainfall events) were missed.

The evaluation metrics obtained for the model that contains interactive features were:

* Precision: 0.2381
* Recall: 0.0084
* F1-score: 0.0162

These results indicate medium precision but very low recall, reflecting a critical challenge in predicting rare events. Comparing models, introducing second-order interaction terms lowered precision without substantially improving recall. Both confusion matrices for the basic and interacted models highlighted a significant imbalance: the model accurately identified all true negatives (no heavy rain), but only correctly captured a small fraction of actual cloudburst events (true positives). That is, in the base model, only 3 out of 597 actual cloudburst events were correctly identified. In the interaction model, 5 were detected, but at the cost of many more false positives.

An additional experiment introduced a "Dry Period Duration" feature, which measures the time elapsed since the last recorded precipitation. This variable was expected to offer insights into atmospheric recharging—long dry spells followed by sudden saturation are sometimes linked with heavy rainfall. However, the model evaluation showed that this feature did not significantly boost prediction outcomes. The model with the dry period feature achieved a precision of 1.0000 and a recall of 0.0304, with an F1-score of 0.0600—nearly identical to the model without this feature. The confusion matrix confirmed that although all true negatives were correctly classified (no false alarms), only 34 of 1119 actual cloudbursts were captured, resulting in a very low recall.

Visual inspection of the confusion matrices and feature importance plots supported this finding: the addition of the dry period feature did not shift the model's reliance away from "HourlyPrecipitation," nor did it provide substantial discriminative power on its own. While precision remained perfect, indicating no false positives, this high precision actually reflects an overly conservative model that rarely predicts a cloudburst unless it is extremely certain—resulting in the vast majority of actual events being missed. This behavior suggests that the model is biased toward minimizing false positives at the expense of recall, which is problematic in early-warning systems where detecting as many real events as possible is critical. This suggests that the feature, although theoretically justifiable, lacks standalone predictive strength in this application, requires improvement in data class balance, or may require more complex integration.

The model's limited recall emphasizes a trade-off: extremely high precision to avoid false positives (important in disaster preparedness to prevent unnecessary alarm and resource strain), versus higher recall needed to ensure timely warnings. Further experimentation with more advanced modeling frameworks or hybrid features may be necessary to fully leverage the dry period duration concept.

**5. Ethical Considerations**

Deploying predictive models for cloudburst events involves crucial ethical implications. Foremost is balancing the model's precision and recall. The low recall identified poses severe ethical consequences; false negatives, or missed predictions, could result in loss of life, property damage, and inadequate community preparedness. Conversely, false positives may cause unnecessary evacuation, panic, economic disruption, and reduced trust in future predictions. This necessitates transparent communication of uncertainty and clear articulation of risk levels to affected communities and stakeholders.

Privacy and data security are additional concerns. Although meteorological data typically contains limited personally identifiable information, robust security practices are essential to prevent data breaches or unauthorized data manipulation that might mislead forecasts or hinder timely responses.

Transparency and interpretability of the model are also critical. Given the potential impact of cloudburst predictions on public safety, ensuring that stakeholders fully understand model limitations, decision rationales, and confidence intervals is paramount. Transparent models help maintain public trust, enabling more effective communication during disaster events.

To mitigate these concerns, the following approaches are recommended:

* Regular performance evaluation and updates to the model using the most recent and comprehensive data.
* Clear risk communication strategies, explicitly conveying prediction confidence and uncertainty levels to decision-makers and communities.
* Implementation of robust cybersecurity measures to ensure data integrity and reliability.
* Prioritizing transparent, explainable models or providing supplementary interpretability methods to clarify predictions, thus reinforcing public trust and informed decision-making.

**6. AI Assistants Use and Process Reflection**

Throughout this project, I incorporated AI tools like Claude and ChatGPT into my workflow to support technical, analytical, and proofreading tasks. I relied on ChatGPT particularly during the model development and evaluation stages. It served as a conversational reference to clarify unfamiliar statistical concepts (such as precision, recall, and F1-score) and to help explore trade-offs between different modeling strategies. Claude was most helpful in the implementation phase—it walked me through Python coding blocks and helped troubleshoot syntax issues, offering alternative ways to structure code logic more clearly.

However, the AI-generated content was not used blindly. For instance, when I initially asked ChatGPT to analyze results from my Random Forest model, it incorrectly interpreted the table by referencing the wrong metrics. I quickly recognized the inconsistency, so instead of relying on its conclusions, I asked it to define the metric formulas and wrote the interpretation myself. This process showed me that AI tools are valuable for structure, guidance, and concept explanation, but not always reliable for autonomous statistical interpretation.

One of the most valuable takeaways from using AI in this project was learning how to balance automation with judgment. While AI suggestions improved my efficiency, especially when switching between writing and debugging, I consistently verified their applicability to my specific dataset, which was highly imbalanced and difficult to model. I learned to adapt AI-generated ideas rather than adopt them wholesale. Still, the AI helped reduce the cognitive load and improve work efficiency when switching between coding, interpreting results, and writing up findings.

Ultimately, this project taught me how to effectively incorporate AI support into technical workflows and utilize AI as a collaborative partner. I learned to prompt AI tools more precisely, verify suggestions against trusted sources, and iterate independently when solutions didn’t generalize. These skills, especially debugging with AI and translating results into formal writing, will transfer well to future data science challenges where time efficiency and technical communication are critical.

**7. Conclusion and Future directions**

In conclusion, this practicum—working extensively with data, Python, machine learning models, and meteorological content—provided me with a strong foundation in end-to-end applied forecasting and rare event prediction. I gained hands-on experience designing and implementing a pipeline for predicting cloudbursts using NOAA hourly weather data, where I carried out data processing, feature engineering, iterative model refinement, and careful evaluation of prediction quality.

The Random Forest model emerged as the most interpretable and robust classifier for this task, successfully identifying top predictors such as hourly precipitation, pressure change, and wind speed. Feature engineering steps like incorporating lagged weather variables, interaction terms, and dry period duration were grounded in meteorological reasoning and literature. While these additions slightly improved model structure, they did not yield large performance gains, largely due to the challenge of detecting rare cloudburst events.

The main limitation across models was low recall, even when precision remained high. This suggests that our models were overly conservative, detecting few true cloudbursts while avoiding false alarms. From an operational standpoint, this is a critical weakness, since timely identification of cloudbursts is central to disaster prevention.

Future directions include adopting more advanced imbalance handling methods like cost-sensitive learning or SMOTE-ENN, exploring hybrid models that combine classification with anomaly detection, and broadening the dataset to include multiple geographies for generalizability. These refinements would make the model more scalable and better suited for deployment under the INDRA framework, ultimately contributing to more reliable early warning systems in high-risk regions.