“Machine learning model to predict precipitation levels based on climatic conditions in LAMAR - Junín, Peru”

Autores: Alvites, L. (20221943), Holguin R. (20221466).

Abstract - In this project, two datasets were used, collected at the Atmospheric Microphysics and Radiation Laboratory in Junín. This laboratory is close to agricultural areas, so the question arises whether it is possible to predict nearby rainfall using these data. Preprocessing techniques such as cyclic coding, null elimination, and selection of the most relevant columns were used. A classification model was then trained to differentiate the "Rain" and "No Rain" classes. A regression model was then trained to estimate precipitation in millimeters in the data with the "Rain" class. Problems of imbalance in classification and regression were encountered, so upsampling, downsampling, and the Synthetic Minority Over Sampling Technique for Regression were used to obtain better results.

1. Introduction

According to the Junín Regional Agriculture Directorate (2022), Huancayo has nearly 23,000 hectares of crops that produce more than 222,000 tons of food annually. This translates into a gross income of approximately 200 million soles. However, due to its geographic location, Huancayo and the entire Junín region are constantly affected by both a lack of rainfall and excessive rainfall, which harms agricultural production. For example, in 2022, the lack of rainfall affected part of the region's crops. Furthermore, excessive rainfall can cause pests and flooding of rivers that flow through the agricultural area. Again, these disasters mean considerable economic losses for farmers in the region.

In parallel, the Atmospheric Microphysics and Radiation Laboratory (LAMAR) has been in operation at the Huancayo Observatory in Junín, Peru, since 1947. This laboratory has been responsible for constantly collecting atmospheric data using traditional instruments such as thermometers and barometers, as well as other automatic instruments such as rain gauges or pressure sensors [2]. The data published for the most recent years (2018 to 2023) contain relevant information, ranging from humidity and temperature at different altitudes, as well as wind direction and atmospheric pressure.

Considering the above, the hypothesis is that it is possible to predict daily precipitation levels in agricultural areas based on atmospheric data collected in these same areas.

Thus, the main objective of the project is to build, adjust, and evaluate a cascade regression model based on machine learning techniques to predict precipitation levels, with the aim of serving as a support tool in the prevention of climatic phenomena affecting the Junín region. Although the research will initially focus on this region, it is hoped that the developed model can be replicated in other agricultural areas of the country facing similar conditions, considering that the data used mostly come from traditional measuring instruments.

2. Related Works

Development of a machine learning model for meteorological precipitation forecasting at a subseasonal scale in Peru

This investigation explains how using the PCA method with backward elimination can improve the performance of linear regression ML models. This method involves selecting the components that represent the greatest variance in the data (PCA) and then comparing trained models by removing some of the components, thereby discarding the component combinations that yield the worst results. [1]

Impact of SMOGN on Regression Models for Crop Yield Prediction in Mizoram Agriculture

This post explains how a synthetic technique can be used to generate more minority data for Gaussian noise regression. For example, this post uses this technique to balance a disproportionate variable as part of data preprocessing using SMOGN. [3]

Modeling and forecasting rainfall patterns in India: a time series analysis with XGBoost algorithm

This study compares traditional statistical models and machine learning models, such as Random Forest or XGBoost, for predicting rainfall patterns in India. It is shown that XGBoost outperforms the other methods, especially when capturing complex and nonlinear rainfall patterns. [4]

3. Methodology

1. Obtaining the datasets

Through the National Open Data Platform, two datasets from different stations in LAMAR, Huancayo, have been obtained. Each one contains measurements of atmospheric variables.
The data dictionaries and variables provided by each station are explained below.

- **Automatic Weather Station (AWS)**: Captures data on atmospheric variables such as Temperature, Humidity, Precipitation (Target), Atmospheric Pressure, Wind Speed and Wind Direction.
- **Torre Gradiente Station**: Similar to AWS, it collects atmospheric data but at different altitudes. This station collects data such as temperature, wind speed, humidity, and wind direction at different altitudes. It also includes the soil heat flux variable.

It's important to note that these datasets share the variables Year, Month, Day, and Time, ranging from 2018 to 2024. For this reason, the datasets will be combined according to these time variables to obtain a single dataset. Finally, the following variables can be observed.

#	Variable	#	Variable
1	Date	15	Humidity (6 m)
2	Temperature	16	Humidity (12 m)
3	Humidity	17	Humidity (18 m)
4	Precipitation	18	Humidity (24 m)
5	Atmospheric Pressure	19	Humidity (29 m)
6	Wind Speed	20	Wind speed (2 m)
7	Wind direction	21	Wind speed (6 m)
8	Temperature(2 m)	22	Wind speed (12 m)
9	Temperature(6 m)	23	Wind speed (18 m)
10	Temperature(12 m)	24	Wind speed (24 m)
11	Temperature(18 m)	25	Wind speed (29 m)
12	Temperature(24 m)	26	Wind speed (24 m)
13	Temperature(29 m)	27	Wind direction (29 m)
14	Humidity (2 m)	28	Heat flow in the earth

Table 1: Variables of the initial dataset. Source: Prepared by the authors.

2. Data preprocessing and identification

Once the new data table was obtained, a deep cleanup was performed by removing rows with nulls, so that those rows with a non-null target could be preserved. First, redundant columns were eliminated since, according to the data dictionary, rows such as Temperature and Temperature (2 m) had the same data, similarly occurring with Humidity, Wind Speed, and Wind Direction. Also, during this stage, many rows were identified that had the target variable set to 0.

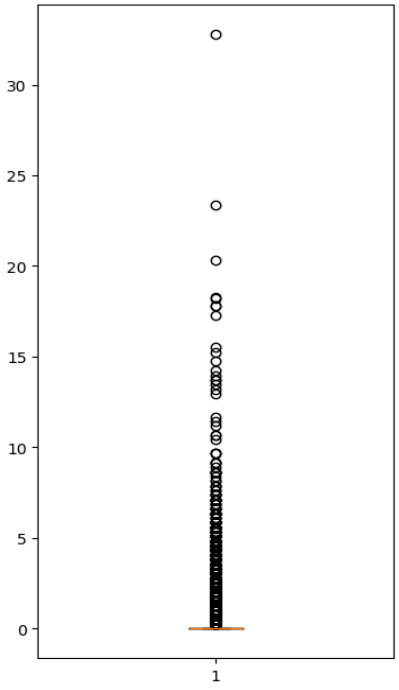

Figure 1: Box plot of the Precipitation variable. Source: Prepared by the authors.

This abnormal number of zeros (and data points greater than 0 as outliers) assumes that on an average day at LAMAR, it does not rain, so the data point is recorded as 0. Therefore, due to the significant imbalance in the data, the SMOGN technique will be used at this stage to enrich the data with precipitation points greater than 0.

Below is a graph by date of the Precipitation variable with a PP threshold > 1 mm.

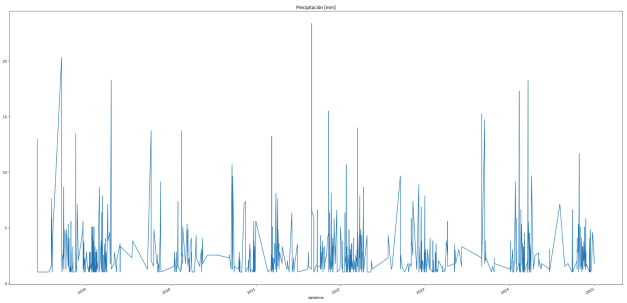

Image 2: Graph of precipitation > 1 mm by date. Source: Prepared by the authors.

The histogram for the Precipitation variable is also presented using the same threshold. We can see a high density at values close to 0.

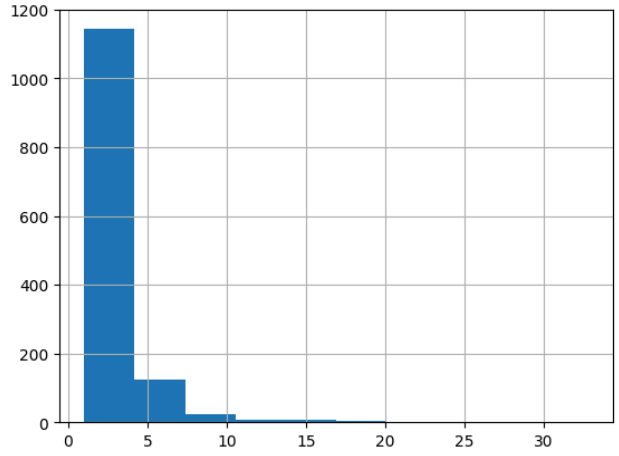

Image 4: Histogram of precipitation > 1 mm. Source: Prepared by the authors.

3. Training the models

For training purposes, the data will be divided into 80% for active training use and 20% for exclusive testing. The largest division will be called the training data set, and the smallest will be called the test data set.

During training, the K-Fold Cross Validation technique (dividing the test dataset into K samples to perform validations during training) will be used to obtain an average performance of each algorithm based on different partitions of the test data. This is done to avoid bias and potential overfitting during training. It will also help select the best-performing model based on the RMSE, RMA, or R2 statistics.

During this process, different pipelines were also used for training with linear regression algorithms, which will allow us to transform the data without affecting the training during the execution of the K-Fold CV technique. This step is important because many data can be found at different scales, and a change in one variable could be highly significant for that variable. However, a change of the same magnitude in another variable might not alter it or overshadow the change in the other variable. For example, the variables Temperature, Humidity, and Atmospheric Pressure use different units, which could lead to their existence at different scales. The pipelines used were Min-Max Scaler, Standardizer, L1 Normalizer, and L2 Normalizer.

Therefore, to implement the classification and regression models, the problem was initially transformed into a binary classification task to determine the presence or absence of precipitation. To do this, a classification model was trained using a dataset divided into training and test data, and its performance was evaluated using appropriate metrics. From this model, the probability of precipitation was calculated for all records. Subsequently, only cases with positive precipitation were selected to train a regression model to estimate the amount of rainfall. This model was also evaluated using test data. The final prediction of expected precipitation was obtained by combining the probability of rainfall with the estimated amount, multiplying both values. Finally, the accuracy of this composite prediction was evaluated considering only cases with actual rainfall, which allowed us to validate the effectiveness of the approach in estimating precipitation levels.

The following image (Thihlum et al.) shows the flow used as a reference for preprocessing, training, and selecting the most optimal models for this project.

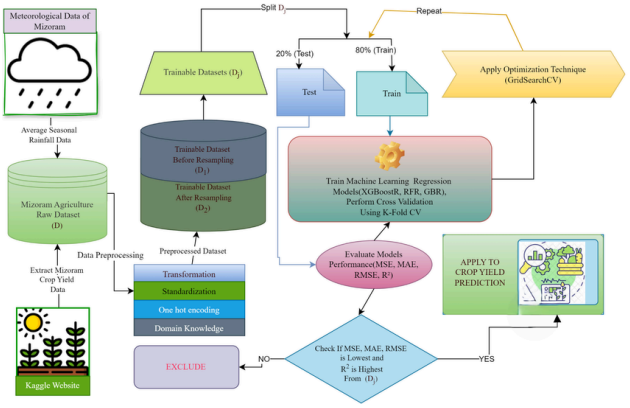

Figure 3: Model construction flow for precipitation prediction. Source: Thihlum, Z. et al. (2025). Impact

of SMOGN on Regression Models for Crop Yield Prediction in Mizoram Agriculture.

The flow shown was subsequently adapted with the methods and methodology described in this section.

4. Experimentation and Results

■ Experimental setup:

As mentioned previously, the data was obtained through Peru's open data repository. According to the website of the laboratory that owns the data, the data was collected using traditional instruments mentioned in the previous section.

These data were then further processed, selecting the most important variables (data collected at a height of 2 meters) and discarding some, such as "Ground Heat Flux." Furthermore, due to the nature of some variables, it was decided to use cyclic coding, for example, for the variable "Wind Direction (degrees)" or for the variable "Month." Additionally, variables such as "It rained the day before" were added to the dataset to enrich the data.

Because a cascade model was used, a Confusion Matrix was used for the classification phase, using three key metrics to evaluate model performance: precision, recall, and f1-score. Precision indicates how accurate the model is at predicting rainy days, that is, how many of those predictions were actually correct. Recall, on the other hand, measures the model's ability to identify all the days on which it actually rained. Finally, the f1-score represents a balance between both metrics, being especially useful when classes are unbalanced, as in this case, where rainless days are much more frequent.

Furthermore, to evaluate the performance of the regression model responsible for estimating the amount of precipitation, four main metrics were used. The Mean Squared Error (MSE) and the Mean Absolute Error (MAE) measure the difference between actual and predicted values, with the MSE being more sensitive to large errors and the MAE being easier to interpret because it expresses the average error directly. On the other hand, the Explained Variance Score indicates how much of the behavior of the target variable is explained by the model, while the R^2 Score measures the degree of fit of the model to the data, with a value close to 1 representing a highly accurate prediction. Together, these metrics reflect whether the regression model performed well in estimating rainfall amounts. It is important to note that the SMOGN technique will be used to enrich the data and reduce sample imbalance.

During the experiments conducted for the classification stage, the Random Forest Classifier algorithm was used due to its ability to handle data with nonlinear relationships and its good performance with noisy data. Different precipitation thresholds were explored to define the boundary between cases labeled as "Raining" and "Not Raining." Initially, any value greater than zero was considered rain, but this generated unrealistic results, as very small values, such as 0.01 mm, could be due to instrumental error or minimal condensation. Therefore, different cutoff values were tested, and it was finally decided to consider records with precipitation greater than 1 mm as "rain," which improved model consistency and reduced classification noise.

In the regression stage, various machine learning algorithms were experimented with to identify which offered the best performance in estimating precipitation amounts. The models evaluated included simple linear regression, Ridge, Lasso, ElasticNet, K-Nearest Neighbors (with different neighbor values), regression trees, and more complex models such as Random Forest Regressor and XGBoost Regressor. Each model was trained under the same conditions, using cross-validation to ensure consistent results. Comparing the error metrics obtained, it was observed that the tree-based models, especially XGBoost Regressor, achieved significantly better fit by better capturing the nonlinear relationships present in the atmospheric data.

Additionally, as part of the experimental process, the impact of different combinations of atmospheric variables on model performance was evaluated. Tests were conducted using both a reduced set of variables, such as temperature, humidity, and atmospheric pressure, and a larger set that included data at different altitudes, wind speed and direction, and soil heat flux. The objective was to determine whether including more variables contributed to improving the model's predictive capacity or, on the contrary, introduced noise or redundancy. These tests made it possible to identify which variables were most relevant for classification and thus optimize the feature selection of the final model.

■ Results and Discussion:

Various thresholds were experimented with to determine what is rain and what is not, and the following confusion matrices were obtained for the thresholds 0 mm, 8 mm and 1 mm.

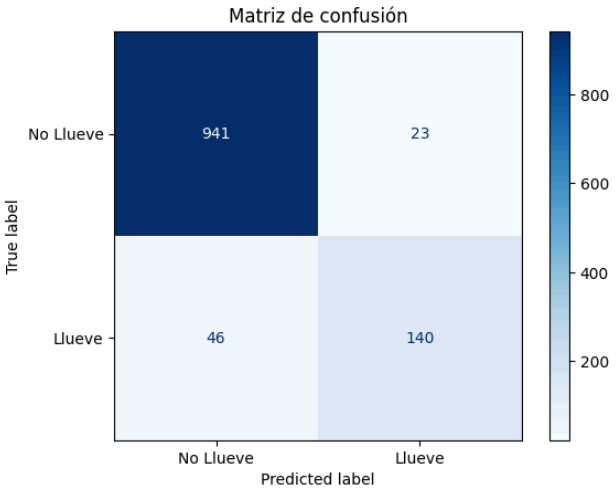

Table 3: Confusion matrix using the threshold Precipitation > 0 mm. Source: Prepared by the authors.

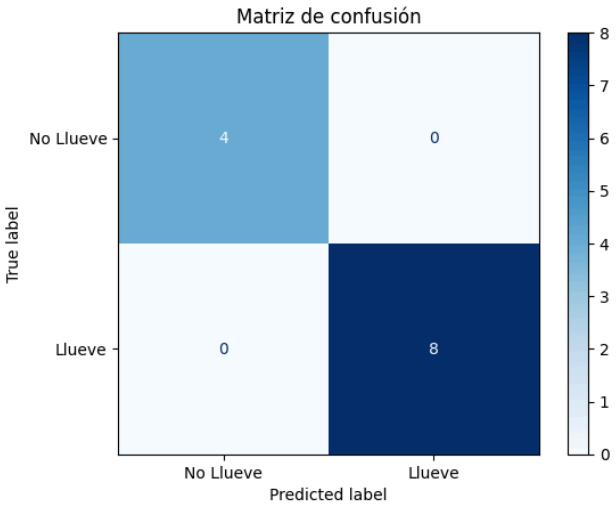

Table 4: Confusion matrix using the Precipitation > 8 mm threshold. Source: Prepared by the authors.

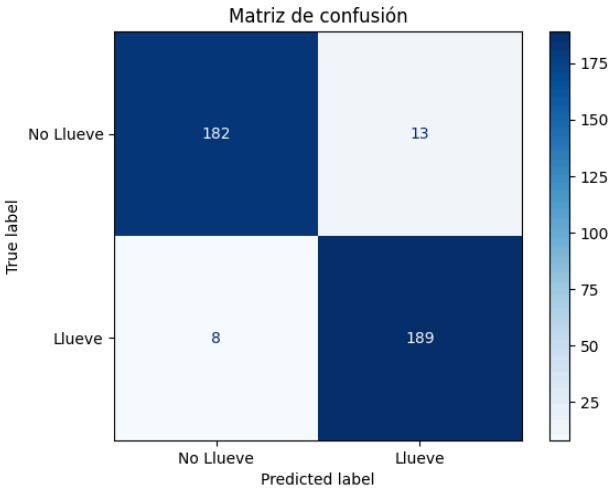

Table 5: Confusion matrix using the Precipitation > 1 mm threshold. Source: Prepared by the authors.

It is demonstrated that it is extremely important to define an appropriate threshold to determine how much precipitation in millimeters should be classified as rain. For example, setting the rainfall threshold to any data greater than 0 mm is not suitable for the model. Similarly, setting a threshold too high, such as 8 mm, is counterproductive, as very few data exceed this threshold, resulting in poor model training due to potential overfitting.

After the initial classification, to train the classification model, we use the data classified as rain and apply the SMOGN technique to generate noisy samples. With this, we can obtain the following histogram for the Precipitation variable and improve the balancing.

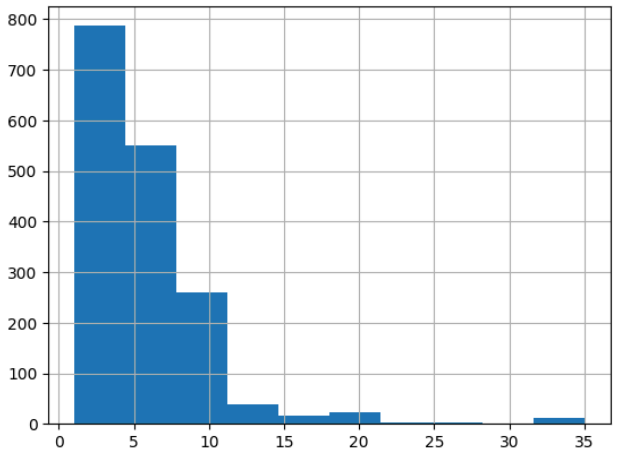

Image 5: New histogram with data generated by SMOGN for Precipitation > 1 mm. Source: Prepared by the authors.

With this newly balanced data, we test the aforementioned algorithms and obtain the following results. For this project, we will use XGBoost Regressor.

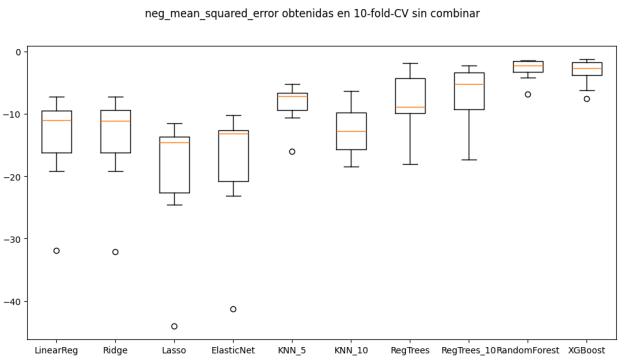

Figure 6. Boxplot of the algorithms used for regression (XGBoost on the right). Source: Prepared by the authors.

For comparison, the training image without SMOGN is shown, followed by the results of the model using SMOGN. For the first model, we obtain an R^2 metric of approximately 0.45, while for the model using SMOGN, better results are achieved with an R^2 of approximately 0.80..

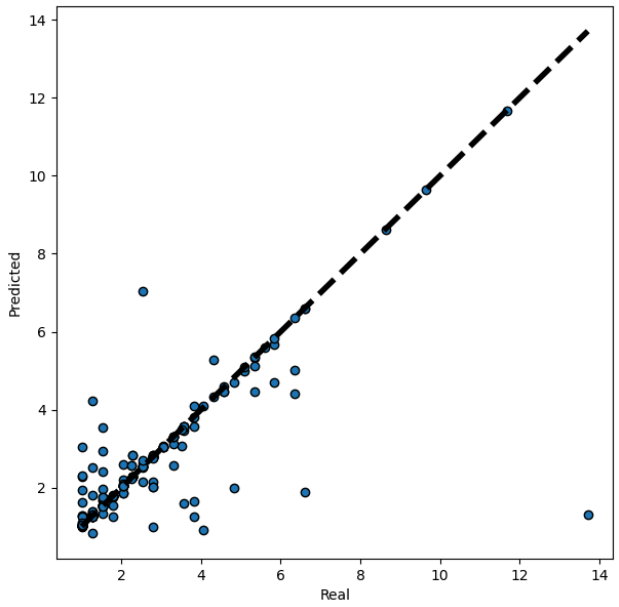

Image 7: Actual result vs. model prediction with XGBoost without SMOGN. Source: Prepared by the authors.

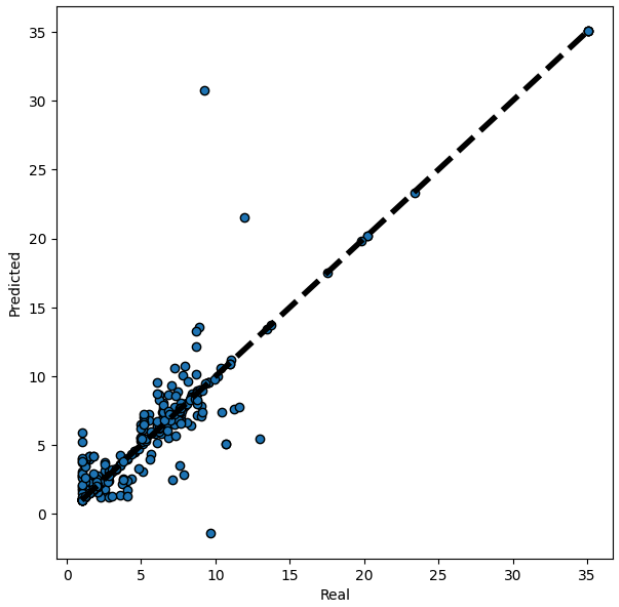

Image 8: Actual result vs. model prediction with XGBoost using SMOGN. Source: Prepared by the authors.

We can notice how XGBoost and RandomForest are the ones with the best results in the boxplot, while KNN_10 and RegTrees_10 are the ones with the worst results, this may be because they are prone to overfitting and remember that even though we have data only for rain, these are still predominant in values close to 0. The latter can be observed in Image 5, where there seems to be a predominance of the model in resulting in values close to 0, which coincides with the distribution of the rain data for that threshold.

5. Conclussion

The importance of downsampling to the "No Rain" category is highlighted, as it is the most predominant category in the dataset, and this generates deficiencies in both the classification and regression models. It can also be concluded that the model for classifying the "Rain" and "No Rain" classes is effective due to its high accuracy, which partially supports the initial hypothesis of being able to determine whether or not precipitation will occur on a given day using local meteorological data.

Furthermore, it can be concluded that, despite the high variability and limited information available on other meteorological conditions that influence the area, the results obtained are encouraging. The XGBoost-based regression model, chosen for its ability to handle nonlinear relationships, its robustness to noisy data, and its good performance in prediction tasks, achieved a coefficient of determination (R^2) of 0.8, indicating a good fit to the data. However, it is recognized that the incorporation of more relevant and representative variables of the local climatic context could further improve the model's predictive capacity. Better class balancing and more appropriate data clustering would also contribute to more accurate and generalizable results, especially in regions where rainfall is rare or of low intensity.

Finally, it should be noted that the classification model works best with predetermined thresholds and that, with data knowledge, better decisions can be made to avoid overfitting or a poorly trained model.

6. Suggestions for future work

A possible improvement to the current approach would be to more systematically optimize the classification threshold that defines whether a day is considered rainy or not. Instead of setting an arbitrary value, such as 1 mm, a more rigorous analysis could be applied that considers the practical impact of different precipitation levels. Furthermore, a direct regression approach could be explored, avoiding the separation into two stages (classification and regression), which would allow the model to jointly learn both the occurrence and

amount of rainfall, thus avoiding the loss of information between phases.

Another possible improvement in data preprocessing would be to perform daily data clustering. For example, instead of having many records with data from a specific hour, a mean or median per day for each variable could be obtained and the mm of rainfall summed over a day. This could reduce data imbalance and contribute to training a model more suited to the project's objectives.

Furthermore, the developed model could be adapted to address extreme events, such as heavy rainfall or prolonged droughts, which have a significant impact on local agriculture. Furthermore, because the variables used are common to many meteorological stations, this approach has the potential to be replicated in other agricultural regions of the country, thus extending its benefits beyond the local context of Junín.

7. Link from work repository

https://github.com/rodholguin/TA_IA_GRUPO7

8. References

[1]. Urbina, B., & Takahashi, K. (2023). Desarrollo de un modelo de Machine Learning para el pronóstico meteorológico de precipitación a escala subestacional en Perú. Boletín científico El Niño, Instituto Geofísico del Perú, vol. 10 n.o 09, págs. 13-17. http://hdl.handle.net/20.500.12816/5540

[2] Instituto Geofísico del Perú (s. f.). *Laboratorio de Microfísica Atmosférica y Radiación (LAMAR) Observatorio de Huancayo*. Instituto Geofísico del Perú. http://met.igp.gob.pe/huayao/

[3] Thihlum, Z., Ambeth Kumar, V.D., Chawngsangpuii (2025). Impact of SMOGN on Regression Models for Crop Yield Prediction in Mizoram Agriculture. In: Patel, K.K., Santosh, K., Gomes de Oliveira, G., Patel, A., Ghosh, A. (eds) Soft Computing and Its Engineering Applications. icSoftComp 2024. Communications in Computer and Information Science, vol 2430. Springer, Cham. https://doi.org/10.1007/978-3-031-88039-1_14

[4] Mishra, P., Al Khatib, A. M. G., Yadav, S., Ray, S., Lama, A., Kumari, B., ... & Yadav, R. (2024). Modeling and forecasting rainfall patterns in India: a time series analysis with XGBoost algorithm. Environmental Earth Sciences, 83(6), 163. http://dx.doi.org/10.1007/s12665-024-11481-w