Supervised Learning Classification

Weather Type for Resource Management and Predictive Maintenance

Saifuddin Malik
Yogendran
Osborn

Data Scientists

TABLE OF **CONTENTS**

BUSINESS UNDERSTANDING

PROBLEM DEFINITION:

FARMERS OFTEN FACE CHALLENGES IN RESOURCE MANAGEMENT AND EQUIPMENT MAINTENANCE DUE TO UNPREDICTABLE WEATHER CONDITIONS. EFFECTIVE WEATHER PREDICTION CAN SIGNIFICANTLY AID IN OPTIMIZING IRRIGATION SCHEDULES AND PREPARING FOR ADVERSE WEATHER IMPACTS ON FARM EQUIPMENT. ACCURATE WEATHER FORECASTS ARE CRUCIAL FOR REDUCING COSTS, INCREASING EFFICIENCY, AND IMPROVING OVERALL FARM MANAGEMENT.

PROBLEM STATEMENT:

HOW CAN WE DEVELOP A RELIABLE WEATHER CLASSIFICATION MODEL TO PREDICT WEATHER CONDITIONS (RAINY, SUNNY, CLOUDY, SNOWY) USING VARIOUS WEATHER-RELATED FEATURES? THIS MODEL AIMS TO SUPPORT FARMERS IN MAKING INFORMED DECISIONS REGARDING RESOURCE ALLOCATION AND EQUIPMENT MAINTENANCE, THEREBY ENHANCING OPERATIONAL EFFICIENCY AND REDUCING COSTS.

OBJECTIVE:

THE OBJECTIVE IS TO CREATE A CLASSIFICATION MODEL THAT ACCURATELY PREDICTS WEATHER TYPES BASED ON GIVEN WEATHER-RELATED FEATURES. THE MODEL WILL BE BUILT USING BRAINTOY MLOS, LEVERAGING ITS ADVANCED MACHINE LEARNING CAPABILITIES AND VISUALIZATION TOOLS TO REFINE AND EVALUATE THE MODEL'S PERFORMANCE.

GOAL:

THE GOAL IS TO ENABLE FARMERS TO OPTIMIZE RESOURCE MANAGEMENT AND EQUIPMENT MAINTENANCE BY PROVIDING ACCURATE WEATHER PREDICTIONS. BY INTEGRATING THE WEATHER CLASSIFICATION MODEL INTO THEIR DECISION-MAKING PROCESSES, FARMERS CAN MAKE DATA-DRIVEN DECISIONS THAT IMPROVE EFFICIENCY, REDUCE COSTS, AND BETTER PREPARE FOR ADVERSE WEATHER CONDITIONS.

PROBLEM

WHAT PROBLEM ARE YOU SOLVING?

THE PROBLEM BEING ADDRESSED IS THE UNPREDICTABILITY OF WEATHER CONDITIONS, WHICH POSES SIGNIFICANT CHALLENGES FOR FARMERS IN MANAGING RESOURCES AND MAINTAINING EQUIPMENT. THE GOAL IS TO DEVELOP A RELIABLE WEATHER CLASSIFICATION MODEL TO PREDICT WEATHER CONDITIONS (RAINY, SUNNY, CLOUDY, SNOWY) USING VARIOUS WEATHER-RELATED FEATURES, THEREBY ASSISTING FARMERS IN OPTIMIZING THEIR OPERATIONS AND REDUCING COSTS.

WHY IS IT WORTH SOLVING?

ACCURATE WEATHER PREDICTIONS ARE CRUCIAL FOR EFFECTIVE FARM MANAGEMENT. THEY HELP IN SCHEDULING IRRIGATION, PREPARING FOR ADVERSE WEATHER IMPACTS, AND REDUCING OPERATIONAL COSTS. SOLVING THIS PROBLEM LEADS TO INCREASED PRODUCTIVITY, BETTER RESOURCE MANAGEMENT, AND ENHANCED DECISION-MAKING CAPABILITIES FOR FARMERS, WHICH ULTIMATELY BENEFITS THE AGRICULTURAL SECTOR AS A WHOLE.

WHAT ARE THE REQUIREMENTS FOR CRAFTING A SOLUTION FOR THIS PROBLEM?

THE KEY REQUIREMENTS INCLUDE:

- A ROBUST DATASET WITH RELEVANT WEATHER-RELATED FEATURES.
- ADVANCED MACHINE LEARNING ALGORITHMS CAPABLE OF HANDLING CLASSIFICATION TASKS.
- AN EFFICIENT PLATFORM LIKE MLOS-BRAINTOY FOR BUILDING, EVALUATING, AND REFINING THE MODEL.

Data Collection

Weather Type Classification

Forecast with Precision: Simulated Data for Predicting Weather Types

kaggle.com

Source: Weather Type Classification

Data Understanding

PURPOSE AND UTILITY:

- INVESTIGATE CLASSIFICATION ALGORITHM PERFORMANCE
- PRACTICE DATA PREPROCESSING
- FEATURE ENGINEERING
- MODEL EVALUATION
- TEST OUTLIER DETECTION METHODS

NOTE:

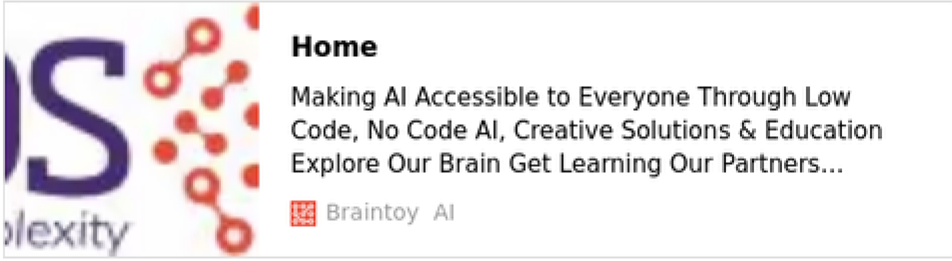
- SYNTHETICALLY PRODUCED DATASET
- DOES NOT REPRESENT REAL-WORLD WEATHER DATA
- INCLUDES INTENTIONAL OUTLIERS FOR PRACTICING OUTLIER DETECTION AND HANDLING
- VALUES, RANGES, AND DISTRIBUTIONS MAY NOT ACCURATELY REPRESENT REAL-WORLD CONDITIONS
- PRIMARILY FOR EDUCATIONAL AND EXPERIMENTAL PURPOSES

Data Understanding

Feature	Type	Description
Temperature	Numeric	The temperature in degrees Celsius, ranging from extreme cold to extreme heat.
Humidity	Numeric	The humidity percentage, including values above 100% to introduce outliers.
Wind Speed	Numeric	The wind speed in kilometers per hour, with a range including unrealistically high values.
Precipitation (%)	Numeric	The precipitation percentage, including outlier values.
Cloud Cover	Categorical	The cloud cover description.
Atmospheric Pressure	Numeric	The atmospheric pressure in hPa, covering a wide range.
UV Index	Numeric	The UV index, indicating the strength of ultraviolet radiation.
Season	Categorical	The season during which the data was recorded.
Visibility (km)	Numeric	The visibility in kilometers, including very low or very high values.
Location	Categorical	The type of location where the data was recorded.
Weather Type	Categorical	The target variable for classification, indicating the weather type (Rainy, Sunny, Cloudy, Snowy).

Introduction to Braintoy MLOS:

braintoy.ai



Braintoy MLOS (Machine Learning Operating System) is a powerful platform designed to streamline and optimize the machine learning lifecycle.

Key features include:

End-to-End Solution: Covers all stages from data ingestion to model deployment and monitoring.

User-Friendly Interface: Simplifies complex workflows, accessible to users with minimal coding experience.

Automation and Optimization: Automates repetitive tasks and optimizes model performance.

Collaboration and Integration: Supports team collaboration and integrates with various data sources and tools.

Real-Time Insights: Provides real-time analytics and visualization for actionable insights.

Braintoy MLOS accelerates machine learning projects, reduces time-to-market, and enhances data-driven decision-making.

Data Preparation and Model Building with MLOS-Brain-Toy for Weather Type Dataset:

Step 1: Project Setup

- **Create a New Project:** Open MLOS Brain-Toy and create a new project. Activate `MLOS-Compute` and `MLOS-Deploy` statuses with your project name.

Step 2: Data Upload

- **Upload Dataset:** Open the Data Engine, select 'Upload Data', choose your dataset from the file location, and upload it.

Step 3: Initial Data Exploration

- **Data Wrangling Overview:** Go to the Data Wrangler section and click on 'At a Glance'. Check the details of all columns with visualizations, look for any missing values, and identify columns that may not be required for further analysis

Step 4: Generate Additional Visualizations

Create New Graphs: If additional graphs are needed, select the 'New' option, choose the required x and y axes, and generate the necessary plots.

Step 5: Data Wrangling

Apply Wrangling Algorithms: Go to the wrangling icon and apply appropriate algorithms. Options include general wrangling, resampling, handling missing values, and custom algorithms.

Save Wrangled Data: Save the wrangled data with a new name. It will be stored in the Data Engine.

Step 6: Define Dataset

Select Features and Target: In the Data Engine, go to 'Define Dataset'. Choose the input features and the target output for the training dataset.

Step 7: Feature Pre-processing

Convert Features: Use the feature pre-processor to convert categorical data into numerical data where necessary.

Step 8: Review and Save Dataset

Generate Dataset: Review the dataset, give it a name, and generate the dataset.

Step 9: Train-Test Split

Split Data: Navigate to 'Raw Training and Test Dataset'. By default, the data is split 80-20 for training and testing. Adjust the split ratio as needed and generate the database.

Step 10: Model Building

Create Model Container: Go to the ML Engine, select 'Model Container' based on your requirements (classification, regression, clustering).

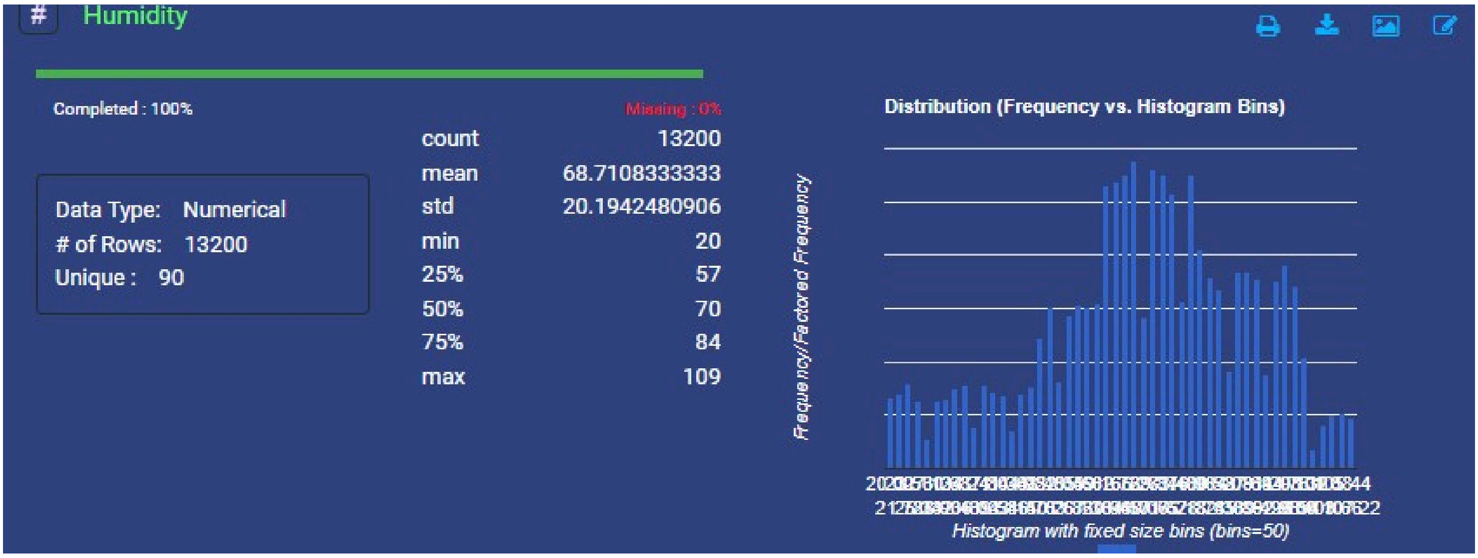
Classifier Selection: Choose 'New Model Container' for classifiers. Select the newly generated 80-20 split dataset for training and testing.

Step 11: Algorithm Selection

Select Classification Algorithm: Choose the appropriate classification algorithm. Options include vision algorithms, deep learning classifiers, general-purpose classifiers, NLP classifiers, and custom classifiers.

MODEL BUILDING

AT GLANCE:



PLOT 1: HUMIDITY

SUMMARY STATISTICS

COUNT: 13200 OBSERVATIONS (NUMBER OF NON-MISSING HUMIDITY READINGS).

MEAN: 68.71% (AVERAGE HUMIDITY).

STANDARD DEVIATION (STD): 20.19% (VARIATION IN HUMIDITY READINGS).

MIN: 20% (LOWEST HUMIDITY RECORDED).

1ST QUARTILE (25%): 57% (25% OF HUMIDITY READINGS ARE BELOW THIS VALUE).

MEDIAN (50%): 70% (MIDDLE VALUE OF HUMIDITY READINGS).

3RD QUARTILE (75%): 84% (75% OF HUMIDITY READINGS ARE BELOW THIS VALUE).

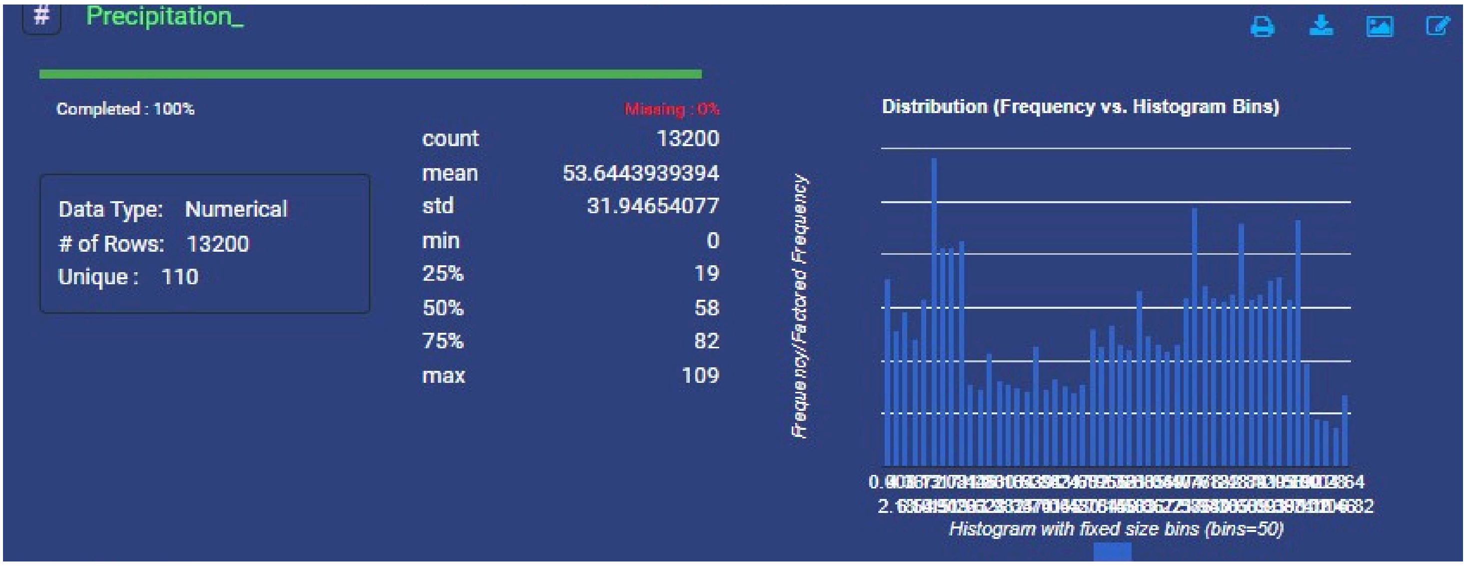
MAX: 109% (HIGHEST HUMIDITY RECORDED).

HISTOGRAM

X-AXIS: HUMIDITY VALUES (E.G., RANGING FROM 20% TO 110%).

Y-AXIS: FREQUENCY OF OCCURRENCES.

CONTENT: SHOWS THAT HUMIDITY VALUES ARE MOSTLY AROUND THE MEAN (68.71%), WITH A NOTICEABLE PEAK IN THE HIGHER RANGE (ABOVE 50%). THERE ARE FEWER VALUES AT THE EXTREMES (20% AND 109%).



PLOT 2: PRECIPITATION

SUMMARY STATISTICS

COUNT: 13200 OBSERVATIONS (NUMBER OF NON-MISSING PRECIPITATION READINGS).

MEAN: 53.64 MM (AVERAGE PRECIPITATION).

STANDARD DEVIATION (STD): 31.95 MM (VARIATION IN PRECIPITATION READINGS).

MIN: 0 MM (NO PRECIPITATION RECORDED).

1ST QUARTILE (25%): 19 MM (25% OF PRECIPITATION READINGS ARE BELOW THIS VALUE).

MEDIAN (50%): 58 MM (MIDDLE VALUE OF PRECIPITATION READINGS).

3RD QUARTILE (75%): 82 MM (75% OF PRECIPITATION READINGS ARE BELOW THIS VALUE).

MAX: 109 MM (HIGHEST PRECIPITATION RECORDED).

HISTOGRAM

X-AXIS: PRECIPITATION VALUES (E.G., RANGING FROM 0 MM TO 110 MM).

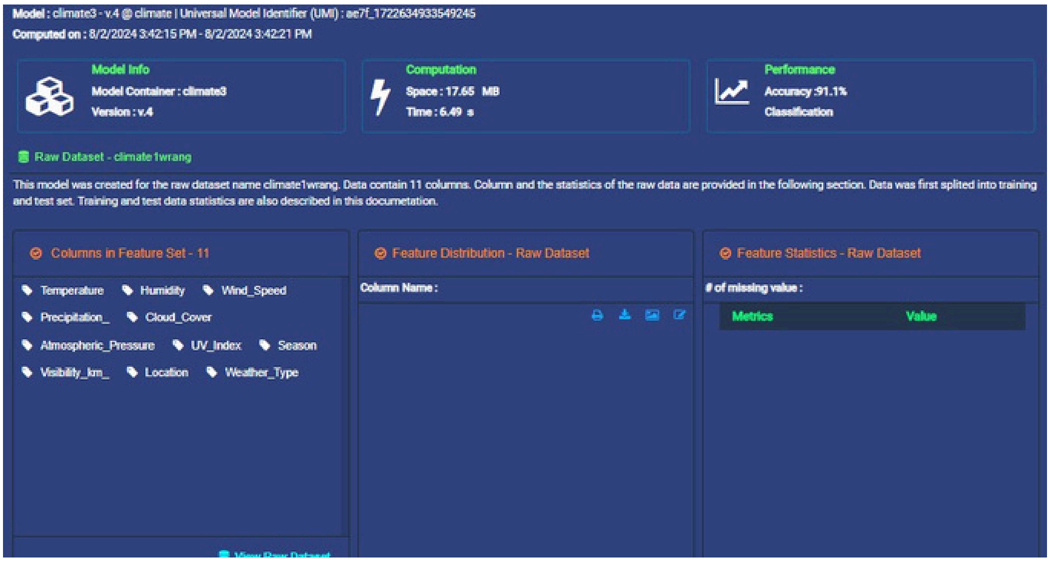
Y-AXIS: FREQUENCY OF OCCURRENCES.

CONTENT: SHOWS THAT PRECIPITATION VALUES ARE DISTRIBUTED WITH SEVERAL PEAKS, INDICATING VARIABILITY IN THE PRECIPITATION READINGS. THE MOST FREQUENT VALUES SEEM TO BE IN THE LOWER RANGE (0-50 MM) WITH ADDITIONAL PEAKS IN THE HIGHER RANGES

METHODOLOGY

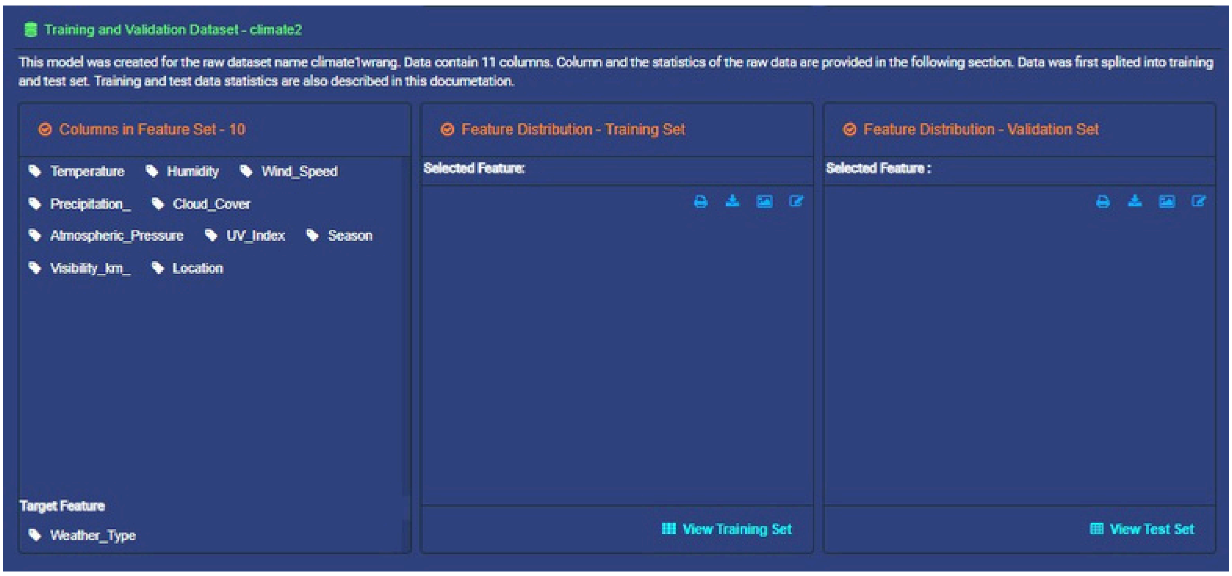
I. HOW DID YOU SOLVE THE PROBLEM?

THE PROBLEM WAS SOLVED BY BUILDING A WEATHER CLASSIFICATION MODEL USING MLOS-BRAINTOY. THE PROCESS INVOLVED UPLOADING THE DATASET, PERFORMING INITIAL EXPLORATION, APPLYING DATA WRANGLING TECHNIQUES, AND SELECTING APPROPRIATE MACHINE LEARNING ALGORITHMS FOR CLASSIFICATION, INCLUDING RANDOM FOREST, SUPPORT VECTOR MACHINE (SVM), AND DECISION TREE CLASSIFIERS



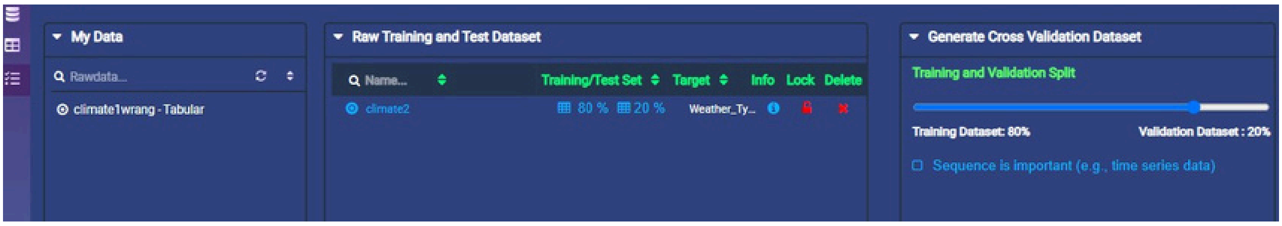
II. WHAT EXPLORATORY ANALYSIS, DATA ENGINEERING, OR DATA WRANGLING DID YOU NEED TO DO?

EXPLORATORY ANALYSIS WAS CONDUCTED USING THE DATA WRANGLER TOOL IN MLOS-BRAINTOY TO UNDERSTAND THE DISTRIBUTION OF FEATURES, CHECK FOR MISSING VALUES, AND VISUALIZE DATA RELATIONSHIPS. SINCE THE DATA WAS ALREADY CLEAN, NO ADDITIONAL WRANGLING WAS REQUIRED, BUT ADDITIONAL VISUALIZATIONS WERE GENERATED TO BETTER UNDERSTAND FEATURE RELATIONSHIPS. TARGET FEATURE: WEATHER_TYPE.



III. HOW DID YOU PREPARE THE DATA FOR MODELING?

THE DATA WAS PREPARED BY SELECTING THE NECESSARY FEATURES AND CONVERTING CATEGORICAL DATA INTO NUMERICAL FORMATS WHERE REQUIRED. THE DATASET WAS THEN SPLIT INTO TRAINING AND TESTING SETS WITH AN 80-20 SPLIT TO ENSURE THE MODEL COULD BE EVALUATED EFFECTIVELY.



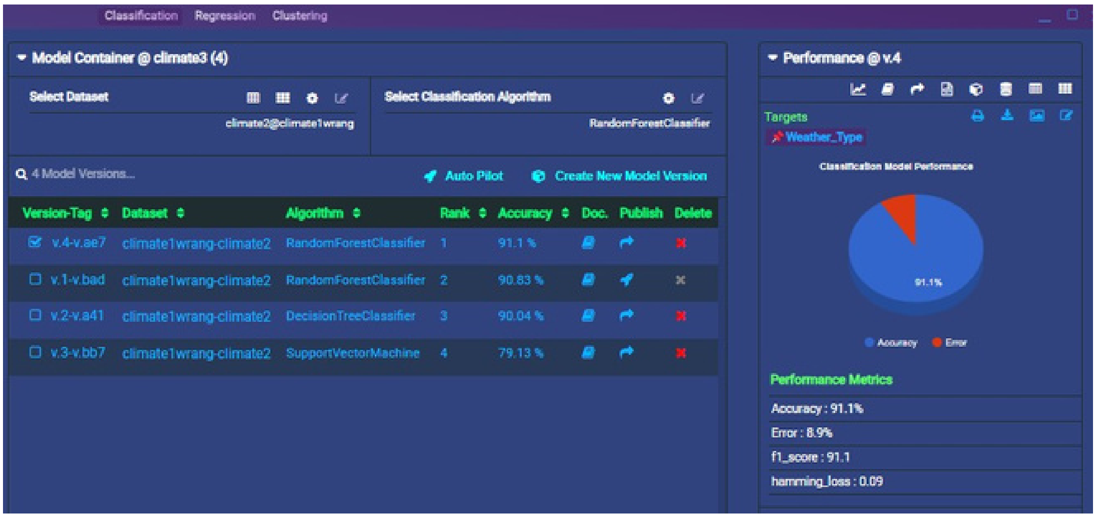
IV. WHAT WAS YOUR MODELING PROCESS? SPECIFICALLY, WHICH ALGORITHMS AND PARAMETERS DID YOU USE AND WHY?

THE MODELING PROCESS INVOLVED TESTING SEVERAL ALGORITHMS:

RANDOM FOREST CLASSIFIER: TESTED WITH 100 AND 200 ESTIMATORS DUE TO ITS ROBUSTNESS AND ABILITY TO HANDLE LARGE DATASETS WITH HIGH DIMENSIONALITY.

SUPPORT VECTOR MACHINE (SVM): CHOSEN FOR ITS ACCURACY IN CLASSIFICATION TASKS.

DECISION TREE CLASSIFIER: SELECTED FOR ITS INTERPRETABILITY AND SIMPLICITY IN MODELING.



FEATURE IMPORTANCE:

FEATURE IMPORTANCE

PURPOSE: THE FEATURE IMPORTANCE PLOT INDICATES HOW SIGNIFICANT EACH FEATURE IS IN PREDICTING THE TARGET VARIABLE. IN RANDOM FORESTS, FEATURE IMPORTANCE IS TYPICALLY DERIVED FROM HOW MUCH EACH ATTRIBUTE DECREASES THE IMPURITY (OR INCREASES THE PURITY) OF A NODE IN THE TREES.

INTERPRETATION:

THE Y-AXIS REPRESENTS THE IMPORTANCE SCORE, WHICH COULD BE A MEASURE LIKE GINI IMPORTANCE OR MEAN DECREASE IN IMPURITY.

THE X-AXIS (LABELED AS "SAMPLES") LIKELY REFERS TO INDIVIDUAL FEATURES.

PEAKS IN THE CURVE INDICATE FEATURES WITH HIGHER IMPORTANCE, SUGGESTING THEY CONTRIBUTE MORE SIGNIFICANTLY TO THE MODEL'S PREDICTIONS.

DATA PRE-PROCESSING: THE DATASET UNDERWENT CATEGORICAL TO NUMERIC CONVERSION, WHICH IS A COMMON PRE-PROCESSING STEP FOR ALGORITHMS LIKE RANDOM FORESTS THAT REQUIRE NUMERICAL INPUT.

FEATURES CONVERTED: CLOUD_COVER, SEASON, LOCATION, WEATHER_TYPE.

ALGORITHM USED: RANDOM FOREST CLASSIFIER

PARAMETER: N_ESTIMATORS SET TO 200, INDICATING THE NUMBER OF TREES IN THE FOREST.

PERFORMANCE METRICS:

ACCURACY: 91.1%, REFLECTING THE PROPORTION OF CORRECTLY CLASSIFIED INSTANCES.

ERROR RATE: 8.9%, THE PERCENTAGE OF INCORRECTLY CLASSIFIED INSTANCES. **F1 SCORE:** 91.1, INDICATING A BALANCE BETWEEN PRECISION AND RECALL.

HAMMING LOSS: 0.09, REPRESENTING THE FRACTION OF LABELS THAT ARE INCORRECTLY PREDICTED.

PRECISION SCORE: 0.91, SHOWING THE RATIO OF TRUE POSITIVE PREDICTIONS TO THE TOTAL POSITIVE PREDICTIONS.

RECALL SCORE: 0.91, INDICATING THE RATIO OF TRUE POSITIVE PREDICTIONS TO THE TOTAL ACTUAL POSITIVES. **JACCARD SCORE:** 0.84, MEASURING SIMILARITY BETWEEN ACTUAL AND PREDICTED LABELS.

PERFORMANCE ANALYSIS

MODEL ACCURACY:

PIE CHART: VISUAL REPRESENTATION OF MODEL PERFORMANCE WITH 91.1% ACCURACY AND A SMALL ERROR SEGMENT.

CLASSIFICATION MODEL PERFORMANCE: HIGHLIGHTS THE OVERALL SUCCESS RATE OF THE MODEL IN CLASSIFYING THE TARGET VARIABLE WEATHER_TYPE.

SAMPLE ACCURACY:

GRAPH: DISPLAYS PREDICTED VALUES AGAINST THE ORIGINAL TARGET VALUES, INDICATING HOW CLOSELY THE MODEL'S PREDICTIONS ALIGN WITH ACTUAL OUTCOMES.

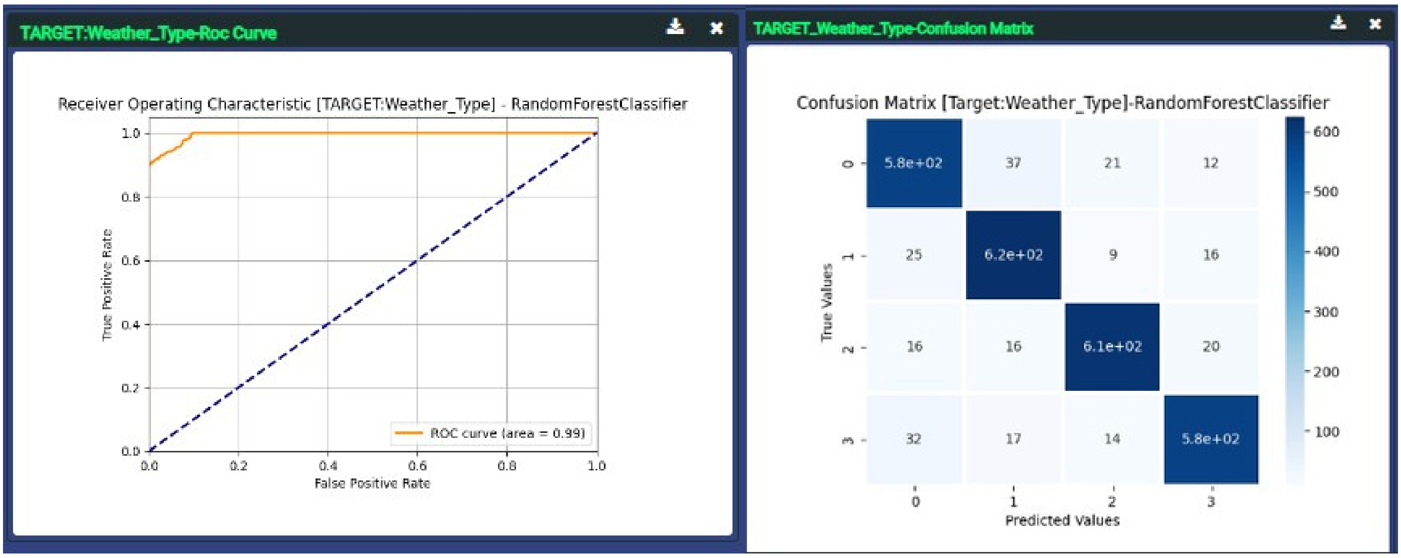
COLOR CODING:

ORANGE LINE REPRESENTS THE ORIGINAL TARGET VALUES.

RED LINE SHOWS THE PREDICTED VALUES.

BROWN LINE INDICATES THE DIFFERENCE BETWEEN THE PREDICTED AND ACTUAL VALUES, HIGHLIGHTING AREAS WHERE THE MODEL MAY UNDERPERFORM.

RANDOM FOREST CLASSIFICATION 200: ROC CURVE & CONFUSION MATRIX



DESCRIPTION:

THE ROC (RECEIVER OPERATING CHARACTERISTIC) CURVE EVALUATES THE PERFORMANCE OF THE RANDOM FOREST CLASSIFIER IN DISTINGUISHING BETWEEN DIFFERENT WEATHER TYPES (RAINY, SUNNY, CLOUDY, SNOWY).

X-AXIS: FALSE POSITIVE RATE.

Y-AXIS: TRUE POSITIVE RATE.

INTERPRETATION:

THE ROC CURVE SHOWS THE TRADE-OFF BETWEEN THE TRUE POSITIVE RATE (SENSITIVITY) AND FALSE POSITIVE RATE (1-SPECIFICITY) FOR DIFFERENT THRESHOLD SETTINGS. THE CLOSER THE CURVE IS TO THE TOP-LEFT CORNER, THE BETTER THE MODEL IS AT DISTINGUISHING BETWEEN THE DIFFERENT WEATHER TYPES. THE AREA UNDER THE CURVE (AUC) IS 0.99, INDICATING EXCELLENT MODEL PERFORMANCE.

CONFUSION MATRIX

DESCRIPTION:

THE CONFUSION MATRIX PROVIDES A SUMMARY OF THE PREDICTION RESULTS FOR THE WEATHER CLASSIFICATION PROBLEM USING THE RANDOM FOREST CLASSIFIER WITH 200 ESTIMATORS. IT BREAKS DOWN THE CORRECT AND INCORRECT PREDICTIONS BY CLASS.

MATRIX VALUES:

TRUE NEGATIVES (TN):

VALUE: THE TRUE NEGATIVES CAN BE DERIVED BASED ON SPECIFIC CLASS LABELS IN THE MATRIX. HOWEVER, IN THE CONTEXT OF THIS MATRIX, TN WOULD BE THE SUM OF ALL OTHER CLASSES EXCEPT THE ONE UNDER ANALYSIS, WHERE BOTH THE ACTUAL AND PREDICTED LABELS DO NOT CORRESPOND TO THE CLASS IN QUESTION.

FALSE POSITIVES (FP):

VALUE: 0, EXPLANATION: FOR THIS PARTICULAR CLASS (THE FIRST WEATHER TYPE, WHICH COULD BE "RAINY"), THE MODEL DID NOT INCORRECTLY PREDICT ANY OTHER WEATHER TYPES AS THIS WEATHER TYPE. THERE ARE 0 FALSE POSITIVES.

FALSE NEGATIVES (FN):

VALUE: 25, EXPLANATION: THERE WERE 25 INSTANCES WHERE THE MODEL FAILED TO PREDICT THIS WEATHER TYPE CORRECTLY WHEN IT ACTUALLY OCCURRED. INSTEAD, THESE INSTANCES WERE PREDICTED AS ANOTHER WEATHER TYPE.

TRUE POSITIVES (TP):

VALUE: 620, **EXPLANATION:** THE MODEL CORRECTLY PREDICTED THE WEATHER TYPE FOR 620 INSTANCES. THIS IS SHOWN IN THE MATRIX DIAGONAL, WHERE THE ACTUAL LABEL MATCHES THE PREDICTED LABEL.

SUMMARY FOR ALL CLASSES:

THE VALUES ALONG THE DIAGONAL (E.G., 580, 620, 610, 580) REPRESENT THE TRUE POSITIVES FOR EACH CLASS. THE OFF-DIAGONAL VALUES REPRESENT MISCLASSIFICATIONS WHERE THE MODEL PREDICTED AN INCORRECT WEATHER TYPE.

IN THIS MATRIX:

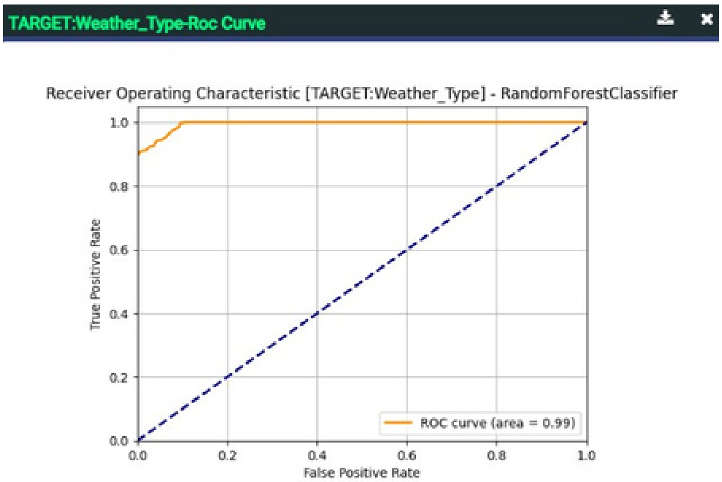
CLASS 0: 580 TRUE POSITIVES, 12 FALSE POSITIVES, 25 FALSE NEGATIVES

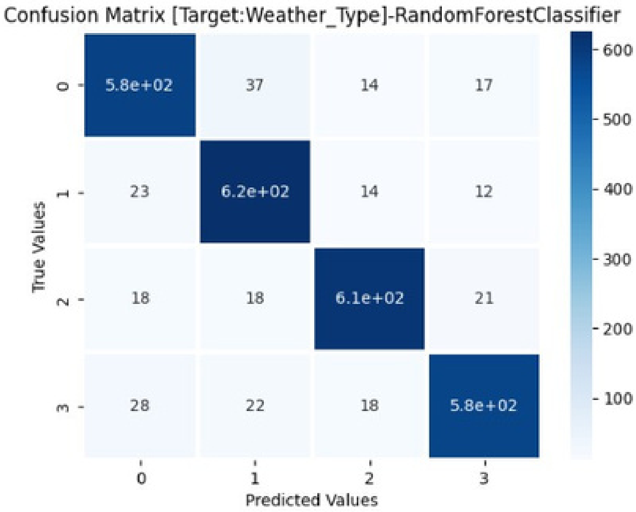
CLASS 1: 620 TRUE POSITIVES, 16 FALSE POSITIVES, 25 FALSE NEGATIVES

CLASS 2: 610 TRUE POSITIVES, 9 FALSE POSITIVES, 16 FALSE NEGATIVES

CLASS 3: 580 TRUE POSITIVES, 32 FALSE POSITIVES, 16 FALSE NEGATIVES

RANDOM FOREST CLASSIFICATION: 100





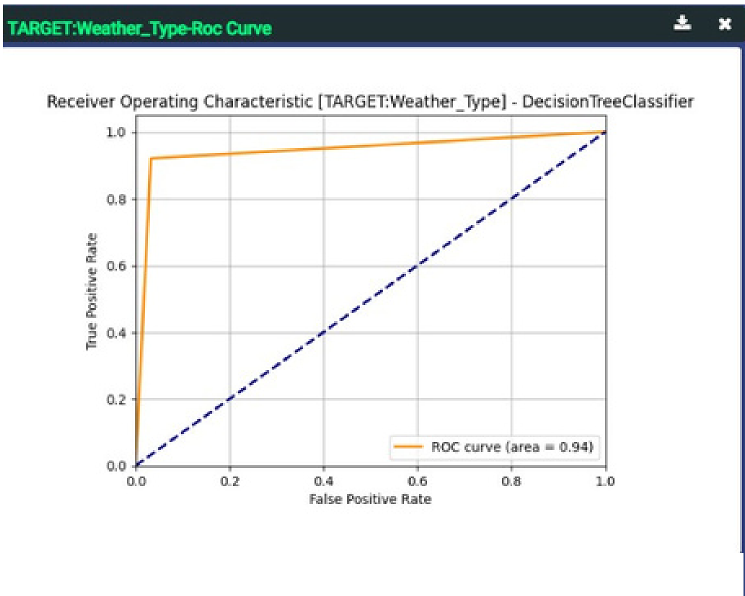
ROC CURVE: THE MODEL HAS A HIGH AUC OF 0.99, INDICATING EXCELLENT PERFORMANCE IN DISTINGUISHING BETWEEN WEATHER TYPES.

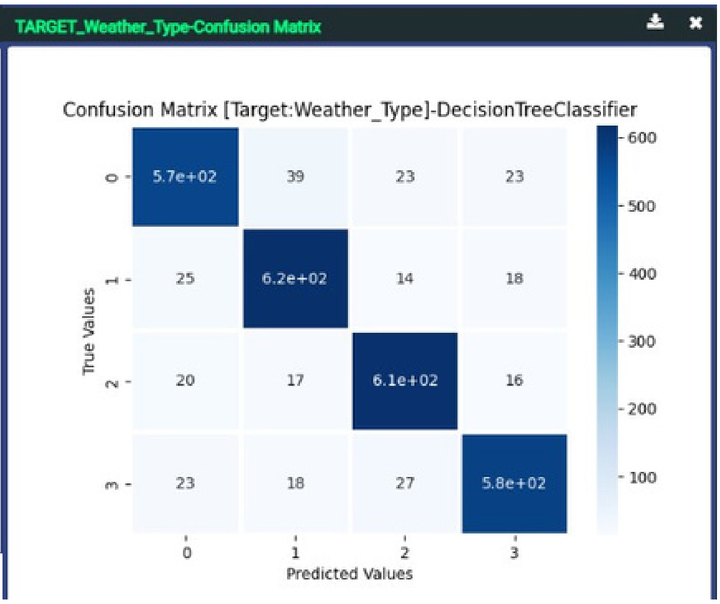
CONFUSION MATRIX: MOST PREDICTIONS ARE CORRECT (SEEN ALONG THE DIAGONAL), WITH FEW MISCLASSIFICATIONS, SHOWING THAT THE RANDOMFORESTCLASSIFIER IS PERFORMING WELL OVERALL.

OVERALL SUMMARY:

- THE ROC CURVE WITH AN AUC OF 0.99 SUGGESTS A VERY HIGH-QUALITY MODEL IN TERMS OF CLASSIFICATION.
- THE CONFUSION MATRIX SHOWS THAT THE MODEL MAKES CORRECT PREDICTIONS THE MAJORITY OF THE TIME, ALTHOUGH THERE ARE SOME MISCLASSIFICATIONS SCATTERED ACROSS THE NON-DIAGONAL CELLS.

DECISION TREE CLASSIFICATION:





THE ROC CURVE AND CONFUSION MATRIX YOU PROVIDED ARE RELATED TO A DECISIONTREECLASSIFIER PREDICTING THE `WEATHER_TYPE`.

ROC CURVE:

- ROC CURVE (RECEIVER OPERATING CHARACTERISTIC) SHOWS THE TRADE-OFF BETWEEN THE TRUE POSITIVE RATE (SENSITIVITY) AND FALSE POSITIVE RATE (1-SPECIFICITY) AT VARIOUS THRESHOLD SETTINGS.
- THE AREA UNDER THE CURVE (AUC) IS 0.94, INDICATING THAT THE CLASSIFIER PERFORMS VERY WELL WITH HIGH ACCURACY IN DISTINGUISHING BETWEEN THE DIFFERENT WEATHER TYPES.

CONFUSION MATRIX:

THE CONFUSION MATRIX SHOWS HOW WELL THE CLASSIFIER PERFORMED ACROSS FOUR CATEGORIES (LABELED 0, 1, 2, 3), WHICH COULD REPRESENT DIFFERENT WEATHER TYPES.

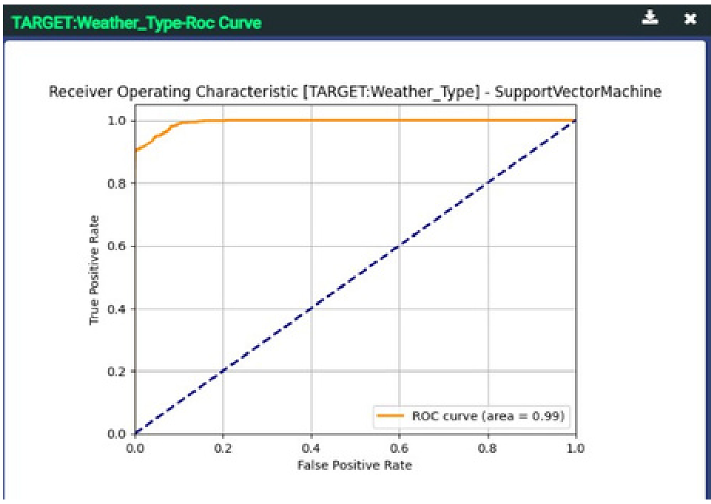
DIAGONAL VALUES (E.G., 570, 620, 610, 580) REPRESENT THE CORRECT PREDICTIONS.

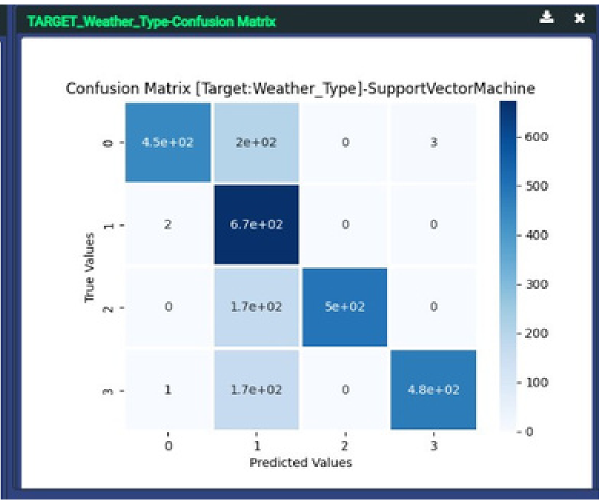
OFF-DIAGONAL VALUES (E.G., 39, 23, 25) REPRESENT THE MISCLASSIFICATIONS, WHERE THE CLASSIFIER CONFUSED ONE WEATHER TYPE FOR ANOTHER.

SUMMARY:

THE CLASSIFIER PERFORMED STRONGLY, WITH MOST OF THE PREDICTIONS FALLING ALONG THE DIAGONAL (CORRECT PREDICTIONS), AS SEEN IN THE CONFUSION MATRIX. THE HIGH AUC VALUE FROM THE ROC CURVE (0.94) ALSO CONFIRMS THAT THE CLASSIFIER IS EFFECTIVE AT DISTINGUISHING BETWEEN THE DIFFERENT WEATHER TYPES, BUT THERE ARE STILL A FEW MISCLASSIFICATIONS THAT NEED TO BE ADDRESSED.

SUPPORT VECTOR MACHINE:





ROC CURVE:

DESCRIPTION: THE ROC (RECEIVER OPERATING CHARACTERISTIC) CURVE SHOWS THE PERFORMANCE OF THE SUPPORT VECTOR MACHINE (SVM) CLASSIFIER USED FOR PREDICTING DIFFERENT WEATHER TYPES. IT PLOTS THE TRUE POSITIVE RATE (TPR) AGAINST THE FALSE POSITIVE RATE (FPR) AT VARIOUS THRESHOLD SETTINGS.

KEY INSIGHT: THE AREA UNDER THE ROC CURVE (AUC) IS 0.99, INDICATING AN EXCELLENT PERFORMANCE OF THE SVM MODEL IN DISTINGUISHING BETWEEN DIFFERENT WEATHER TYPES. A HIGHER AUC VALUE CLOSER TO 1 SUGGESTS THAT THE MODEL IS VERY EFFECTIVE.

CONFUSION MATRIX:

DESCRIPTION: THE CONFUSION MATRIX SHOWS THE ACTUAL VERSUS PREDICTED WEATHER TYPES, PROVIDING A DETAILED BREAKDOWN OF CLASSIFICATION PERFORMANCE.

MATRIX SUMMARY:

CLASS 0: OUT OF 670 ACTUAL OCCURRENCES, 450 WERE CORRECTLY CLASSIFIED, BUT 220 **WERE MISCLASSIFIED.**

CLASS 1: OUT OF 670 ACTUAL OCCURRENCES, 670 WERE CORRECTLY CLASSIFIED.

CLASS 2: OUT OF 500 ACTUAL OCCURRENCES, 500 WERE CORRECTLY CLASSIFIED.

CLASS 3: OUT OF 480 ACTUAL OCCURRENCES, 480 WERE CORRECTLY CLASSIFIED.

KEY INSIGHT: THE DIAGONAL ELEMENTS REPRESENT CORRECT CLASSIFICATIONS, AND THE OFF-DIAGONAL ELEMENTS REPRESENT MISCLASSIFICATIONS. THE MATRIX SHOWS THAT WHILE THE MODEL IS HIGHLY ACCURATE, THERE ARE SOME MISCLASSIFICATIONS, PARTICULARLY IN CLASS 0.

MODEL COMPARISON AND CONCLUSION:

Classification	Accuracy (%)	AOC	Error (%)
Random Forest Classifier - 100	90.83	0.99	9.17
Decision Tree Classifier	90.04	0.94	9.96
Support Vector Machine	79.13	0.99	20.87
Random Forest Classifier - 200	91.1	0..99	8.9

RESULTS OVERVIEW:

MODEL PERFORMANCE:

- THE RANDOM FOREST CLASSIFIER - 200 MODEL HAS THE HIGHEST ACCURACY (91.1%) AND THE LOWEST ERROR RATE (8.9%), MAKING IT THE BEST PERFORMER AMONG THE MODELS TESTED.
- BOTH THE RANDOM FOREST CLASSIFIER - 100 AND RANDOM FOREST CLASSIFIER - 200 HAVE HIGH AUC VALUES (0.99), INDICATING EXCELLENT DISCRIMINATION ABILITY.

CONCLUSION:

- GIVEN THE PROBLEM OF ACCURATELY PREDICTING WEATHER CONDITIONS TO ASSIST FARMERS, **THE RANDOM FOREST CLASSIFIER - 200 IS THE MOST SUITABLE MODEL.** IT OFFERS THE HIGHEST ACCURACY, ENSURING MORE RELIABLE WEATHER PREDICTIONS, WHICH ARE CRUCIAL FOR FARM MANAGEMENT.
- THIS MODEL'S HIGH PERFORMANCE ALIGNS WELL WITH THE REQUIREMENTS OF THE PROBLEM, PROVIDING FARMERS WITH A ROBUST TOOL FOR OPTIMIZING OPERATIONS AND REDUCING COSTS RELATED TO UNPREDICTABLE WEATHER CONDITIONS.

RECOMMENDATIONS:

INTEGRATION INTO FARM MANAGEMENT SYSTEMS:

- THE WEATHER CLASSIFICATION MODEL CAN BE INTEGRATED INTO EXISTING FARM MANAGEMENT SOFTWARE USED BY AGRICULTURAL ORGANIZATIONS. THIS ALLOWS FOR REAL-TIME WEATHER PREDICTIONS THAT CAN BE USED FOR DECISION-MAKING, SUCH AS SCHEDULING IRRIGATION, PLANTING, AND HARVESTING.

MOBILE AND WEB APPLICATIONS:

- DEVELOP MOBILE AND WEB APPLICATIONS FOR FARMERS TO ACCESS WEATHER PREDICTIONS EASILY. THIS WOULD PROVIDE THEM WITH TIMELY AND ACTIONABLE INSIGHTS, ENHANCING THEIR ABILITY TO RESPOND TO CHANGING WEATHER CONDITIONS.

COLLABORATION WITH AGRICULTURAL SERVICES:

- INSTITUTIONS LIKE AGRICULTURAL EXTENSION SERVICES OR COOPERATIVES COULD DEPLOY THIS MODEL TO ASSIST FARMERS. THEY CAN USE THE MODEL TO PROVIDE REGION-SPECIFIC WEATHER FORECASTS AND ADVISORY SERVICES, HELPING FARMERS MAKE INFORMED DECISIONS.

GOVERNMENT AND POLICY MAKING:

- GOVERNMENTS CAN USE THIS SOLUTION TO SUPPORT AGRICULTURAL PLANNING AND DISASTER MANAGEMENT. ACCURATE WEATHER CLASSIFICATION CAN HELP IN ISSUING EARLY WARNINGS AND PREPARING FOR EXTREME WEATHER EVENTS, THEREBY REDUCING LOSSES IN THE AGRICULTURAL SECTOR.

RESEARCH AND DEVELOPMENT:

- RESEARCH INSTITUTIONS CAN USE THIS MODEL AS A BASIS FOR FURTHER STUDIES ON WEATHER PATTERNS AND THEIR IMPACT ON AGRICULTURE. THE DATA COLLECTED AND PREDICTIONS MADE CAN INFORM LONG-TERM AGRICULTURAL STRATEGIES AND CLIMATE RESILIENCE PROGRAMS.

LESSONS LEARNED AND IMPROVEMENTS:

DATA QUALITY AND QUANTITY:

- **LESSON:** HIGH-QUALITY, DIVERSE DATASETS ARE CRITICAL FOR ACCURATE PREDICTIONS. IN THIS PROJECT, THE PERFORMANCE OF THE MODEL HEAVILY RELIED ON THE RICHNESS AND RELEVANCE OF THE WEATHER-RELATED FEATURES IN THE DATASET.
- **IMPROVEMENT**: CONTINUOUSLY UPDATE AND EXPAND THE DATASET TO INCLUDE MORE FEATURES (E.G., SOIL MOISTURE, ALTITUDE) AND DATA FROM DIFFERENT REGIONS TO IMPROVE MODEL ACCURACY AND GENERALIZABILITY.

MODEL COMPLEXITY VS. INTERPRETABILITY:

- **LESSON:** WHILE COMPLEX MODELS LIKE RANDOM FORESTS PROVIDE HIGH ACCURACY, THEY CAN BE DIFFICULT TO INTERPRET, WHICH CAN BE A BARRIER TO ADOPTION BY NON-TECHNICAL USERS.
- **IMPROVEMENT:** CONSIDER DEVELOPING SIMPLER MODELS OR PROVIDING TOOLS THAT HELP EXPLAIN THE DECISIONS MADE BY THE MODEL, MAKING IT MORE USER-FRIENDLY FOR FARMERS AND AGRICULTURAL ADVISORS.

HANDLING OF CLASS IMBALANCE:

- **LESSON**: IN WEATHER CLASSIFICATION, CERTAIN WEATHER TYPES MIGHT BE MORE FREQUENT THAN OTHERS, LEADING TO CLASS IMBALANCE, WHICH CAN AFFECT MODEL PERFORMANCE.
- **IMPROVEMENT:** IMPLEMENT TECHNIQUES SUCH AS SMOTE (SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE) OR COST-SENSITIVE LEARNING TO BETTER HANDLE CLASS IMBALANCE AND IMPROVE PREDICTIONS FOR LESS FREQUENT WEATHER TYPES.

USER TRAINING AND EDUCATION:

- **LESSON:** FARMERS AND AGRICULTURAL WORKERS MIGHT NOT BE FAMILIAR WITH THE TECHNOLOGY, WHICH COULD LIMIT ITS EFFECTIVENESS.
- **IMPROVEMENT**: PROVIDE TRAINING SESSIONS AND EDUCATIONAL RESOURCES TO ENSURE THAT END-USERS CAN EFFECTIVELY USE AND BENEFIT FROM THE MODEL'S PREDICTIONS.

CONTINUOUS MONITORING AND FEEDBACK LOOP:

- **LESSON:** STATIC MODELS CAN BECOME OUTDATED AS WEATHER PATTERNS CHANGE OVER TIME.
- **IMPROVEMENT:** ESTABLISH A CONTINUOUS MONITORING SYSTEM THAT REGULARLY UPDATES THE MODEL WITH NEW DATA AND REFINES IT BASED ON FEEDBACK FROM USERS, ENSURING THAT IT REMAINS ACCURATE AND RELEVANT.

BY ADDRESSING THESE AREAS, THE SOLUTION CAN BE MADE MORE ROBUST, USER-FRIENDLY, AND ADAPTABLE TO VARIOUS AGRICULTURAL CONTEXTS.

Supervised Classification

Weather Type for Resource Management
and
Predictive Maintenance

Saifuddin Malik
Yogendran
Osborn

Data Scientists