CS 422 – Introduction to Machine Learning
Fall 2023

Alex Akoopie
Henry nan

Assignment 6

The purpose of this assignment is to use a *multilayer perceptron (MLP)* machine learning model. This dataset should have at least 3 continuous, real-valued input features and 1 continuous, real-valued output label or 1 categorical output label. There should be three instances of the MLP machine learning model with different combinations of hyperparameters. 80% of the data will be used for training and 20% of the data will be used for testing.

Dataset: https://www.kaggle.com/datasets/ananthr1/weather-prediction/data

Columns:

1. date : The date the sample was taken.
2. precipitation : The amount of precipitation in cm.
3. temp_max : The maximum temperature recorded for the day in ℃.
4. temp_min : The minimum temperature recorded for the day in ℃.
5. wind : The average wind speed recorded for the day in mph.
6. weather : The observed weather {rain, sun, fog, drizzle, snow}.

Data analysis

Structure:

	date	precipitation	temp_max	temp_min	wind	weather
0	2012-01-01	0.0	12.8	5.0	4.7	drizzle
1	2012-01-02	10.9	10.6	2.8	4.5	rain
2	2012-01-03	0.8	11.7	7.2	2.3	rain
3	2012-01-04	20.3	12.2	5.6	4.7	rain
4	2012-01-05	1.3	8.9	2.8	6.1	rain

	date	precipitation	temp_max	temp_min	wind	weather
516	2013-05-31	0.0	19.4	11.1	2.5	sun
1265	2015-06-19	0.5	23.9	13.3	3.2	rain
1234	2015-05-19	0.0	21.7	11.7	2.6	sun
720	2013-12-21	5.6	8.9	5.6	2.3	rain
1121	2015-01-26	0.0	16.1	6.1	2.2	fog

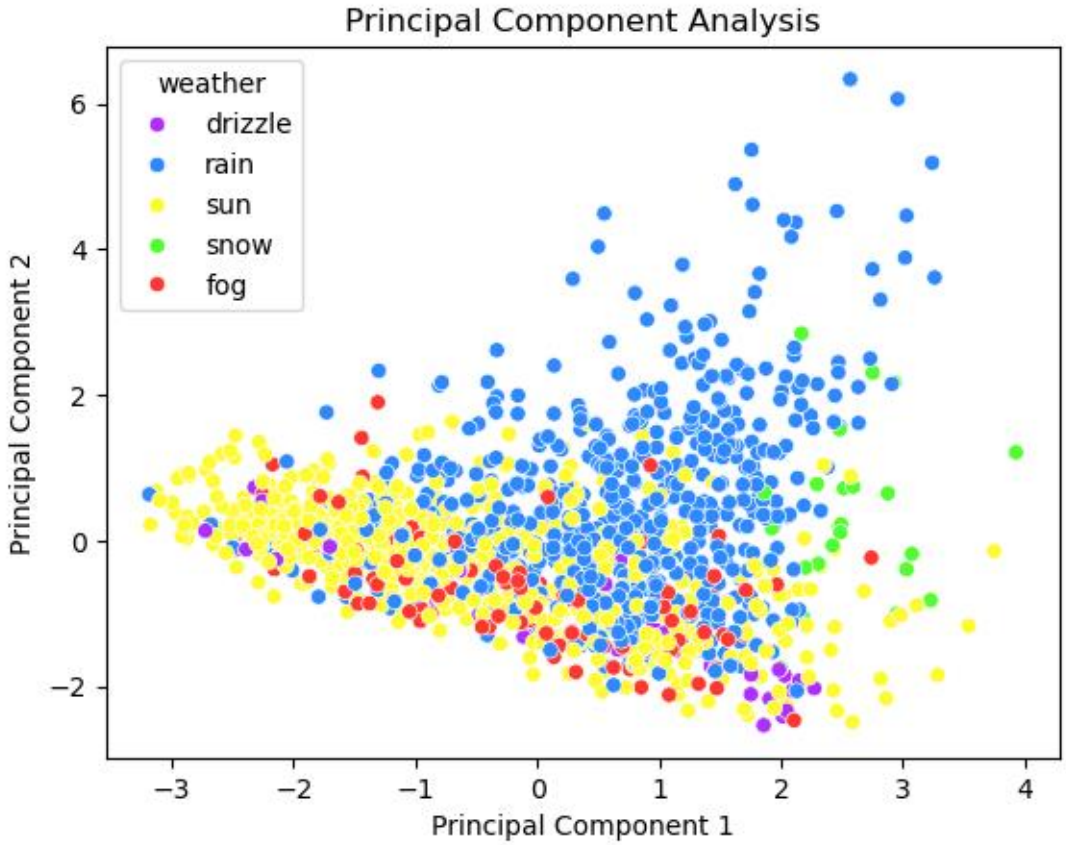
	date	precipitation	temp_max	temp_min	wind	weather
1456	2015-12-27	8.6	4.4	1.7	2.9	rain
1457	2015-12-28	1.5	5.0	1.7	1.3	rain
1458	2015-12-29	0.0	7.2	0.6	2.6	fog
1459	2015-12-30	0.0	5.6	-1.0	3.4	sun
1460	2015-12-31	0.0	5.6	-2.1	3.5	sun

Description:

	precipitation	temp_max	temp_min	wind
count	1461.000000	1461.000000	1461.000000	1461.000000
mean	3.029432	16.439083	8.234771	3.241136
std	6.680194	7.349758	5.023004	1.437825
min	0.000000	-1.600000	-7.100000	0.400000
25%	0.000000	10.600000	4.400000	2.200000
50%	0.000000	15.600000	8.300000	3.000000
75%	2.800000	22.200000	12.200000	4.000000
max	55.900000	35.600000	18.300000	9.500000

The dataframe itself has 1461 instances. Of which, if we discriminate by column 6, then 641 are *rain*, 640 are *sun*, 101 are *fog*, 53 are *drizzle*, and 26 are *snow*. If we choose to use column 6 as our label, then the data is going to be highly imbalanced. From the description, we can see the continuous, real-valued columns have varying centers and spreads. Let us examine how stringent each of these real-valued columns are to determining the class of column 6. We will reduce the dimensionality to 2 components to make it easier to visualize on a 2-dimensional graph. But before that, we need to scale the data to make the relationships between individual instances easier to recognize.

Principal Component Analysis:



We can see that there are barely any instances of *drizzle*, *snow*, and *fog*. This is going to make it more difficult to train the neural network. Based off the principal component analysis, it seems that there are some clear distinctions between what makes a data instance *sun* and *rain*. The majority influence of *sun* comes from principal component 1 while the majority influence of *rain* comes from principal component 2.

Intuition

We obviously want columns 2-5 to act as our features (input) while column 6 will act as our label (output). The features are all continuous and real-valued while the label is categorical and has 5 classes. Column 1 is going to be useless for our purposes. We need to create more data instances for the minority class labels to fix the imbalanced issue. Finally, the data should be scaled before training the model. We will use 3 instances of the MLPClassifer machine learning model with different activations. One will use identity, another will use tanh, and the last will use relu.

Data preprocessing

1) Remove column 1
2) Apply Synthetic Minority Oversampling Technique (SMOTE) to fix imbalanced data
3) Standardize the data using the standard scaler

Data preprocessing analysis

Structure:

	precipitation	temp_max	temp_min	wind	weather
0	-0.487872	-0.191751	-0.252996	1.082515	drizzle
1	1.312598	-0.468099	-0.642196	0.940622	rain
2	-0.355727	-0.329925	0.136204	-0.620200	rain
3	2.865296	-0.267119	-0.146850	1.082515	rain
4	-0.273137	-0.681640	-0.642196	2.075766	rain

	precipitation	temp_max	temp_min	wind	weather
2971	2.835692	-0.535607	-0.402483	1.862926	snow
192	-0.487872	1.692436	1.215349	-0.194521	fog
1666	-0.487872	1.396821	1.109336	-0.710351	drizzle
558	-0.487872	0.637291	1.215349	-0.691147	sun
2872	-0.314018	-1.185484	-0.951875	2.592468	snow

	precipitation	temp_max	temp_min	wind	weather
3200	0.623649	-0.585403	-0.814832	-0.345543	snow
3201	1.844059	-1.055144	-0.612787	0.795177	snow
3202	0.672951	-1.183382	-1.031395	1.426090	snow
3203	0.265712	-1.121506	-1.123533	1.089175	snow
3204	-0.487872	1.387292	1.475538	-0.910343	sun

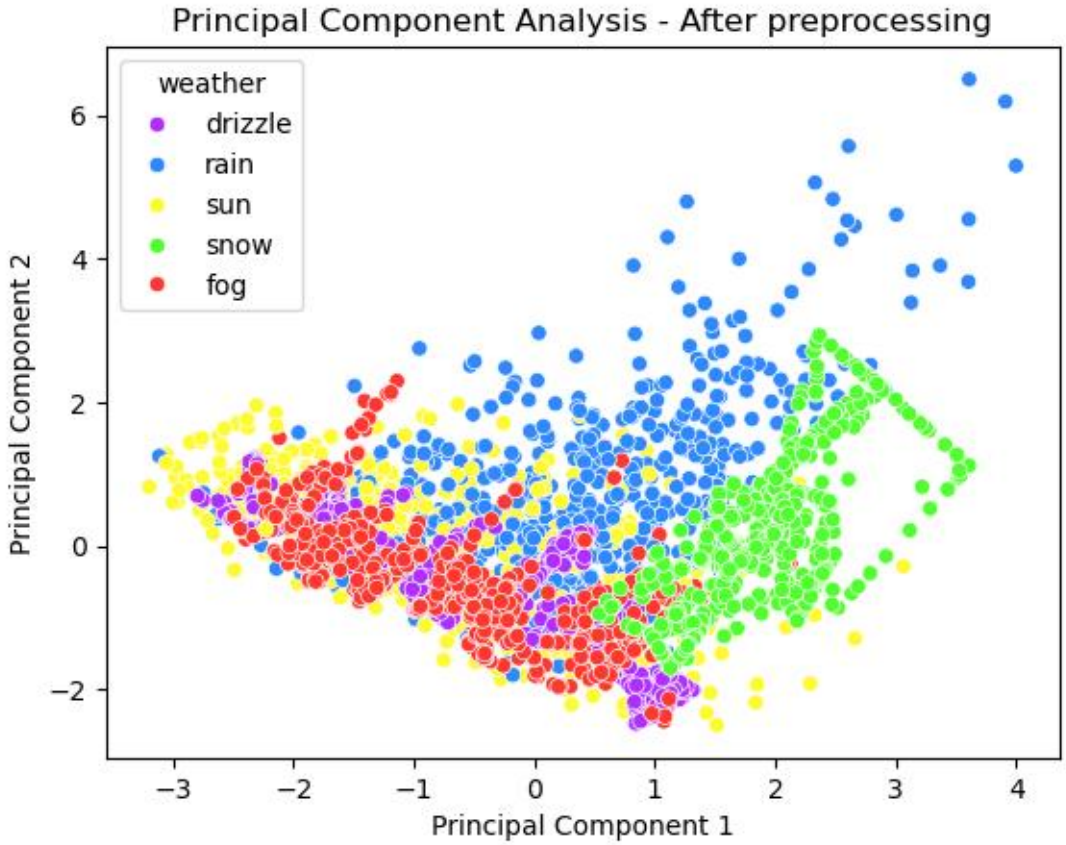
Description:

	precipitation	temp_max	temp_min	wind
count	3.205000e+03	3.205000e+03	3.205000e+03	3.205000e+03
mean	1.773586e-17	1.773586e-17	1.773586e-17	-2.660378e-17
std	1.000156e+00	1.000156e+00	1.000156e+00	1.000156e+00
min	-4.878716e-01	-2.000571e+00	-2.393595e+00	-1.968183e+00
25%	-4.878716e-01	-8.350532e-01	-9.429409e-01	-7.620932e-01
50%	-4.878716e-01	-1.289450e-01	-5.839593e-02	-2.654679e-01
75%	5.722461e-02	8.170183e-01	8.775211e-01	6.568363e-01
max	8.745727e+00	2.672213e+00	2.099894e+00	4.487946e+00

Now that the data preprocessed, there are more data instances to accommodate the minority columns. In fact, there are 641 instances across all 5 class labels. That makes 3205 data instances

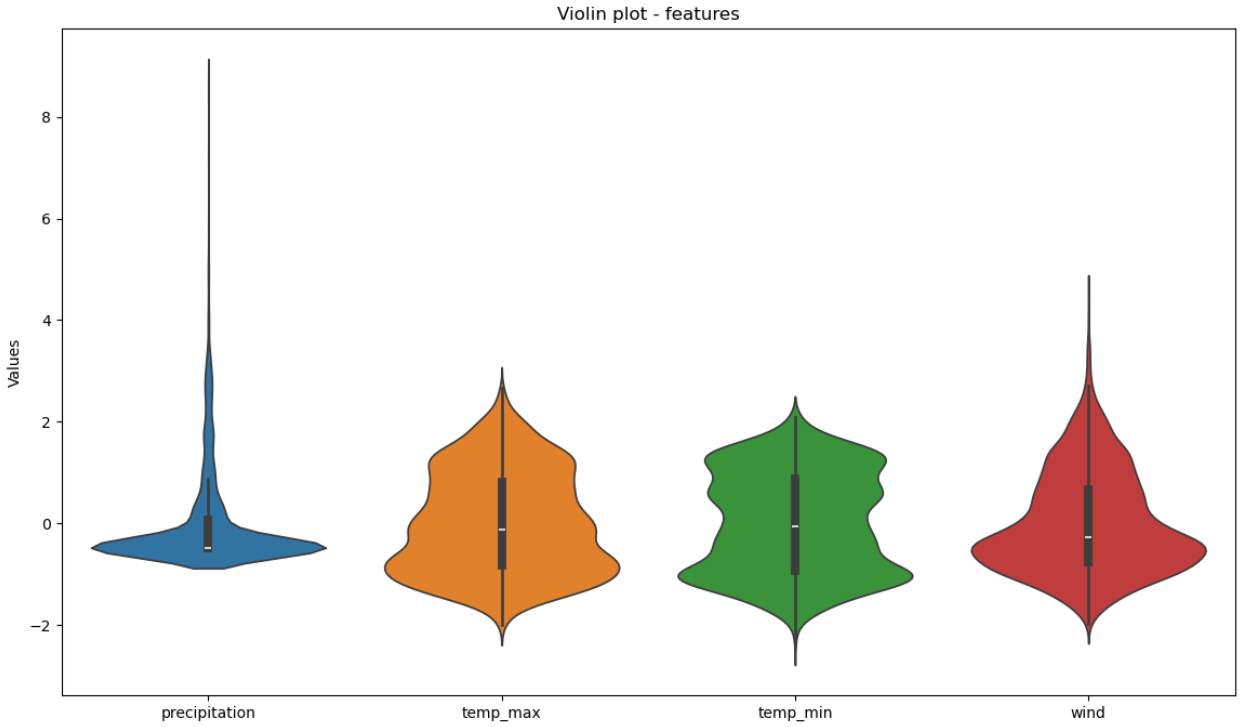
total. Scaling the data made all the means close to 0 and all the standard deviations close to 1 for each feature.

Principal Component Analysis:



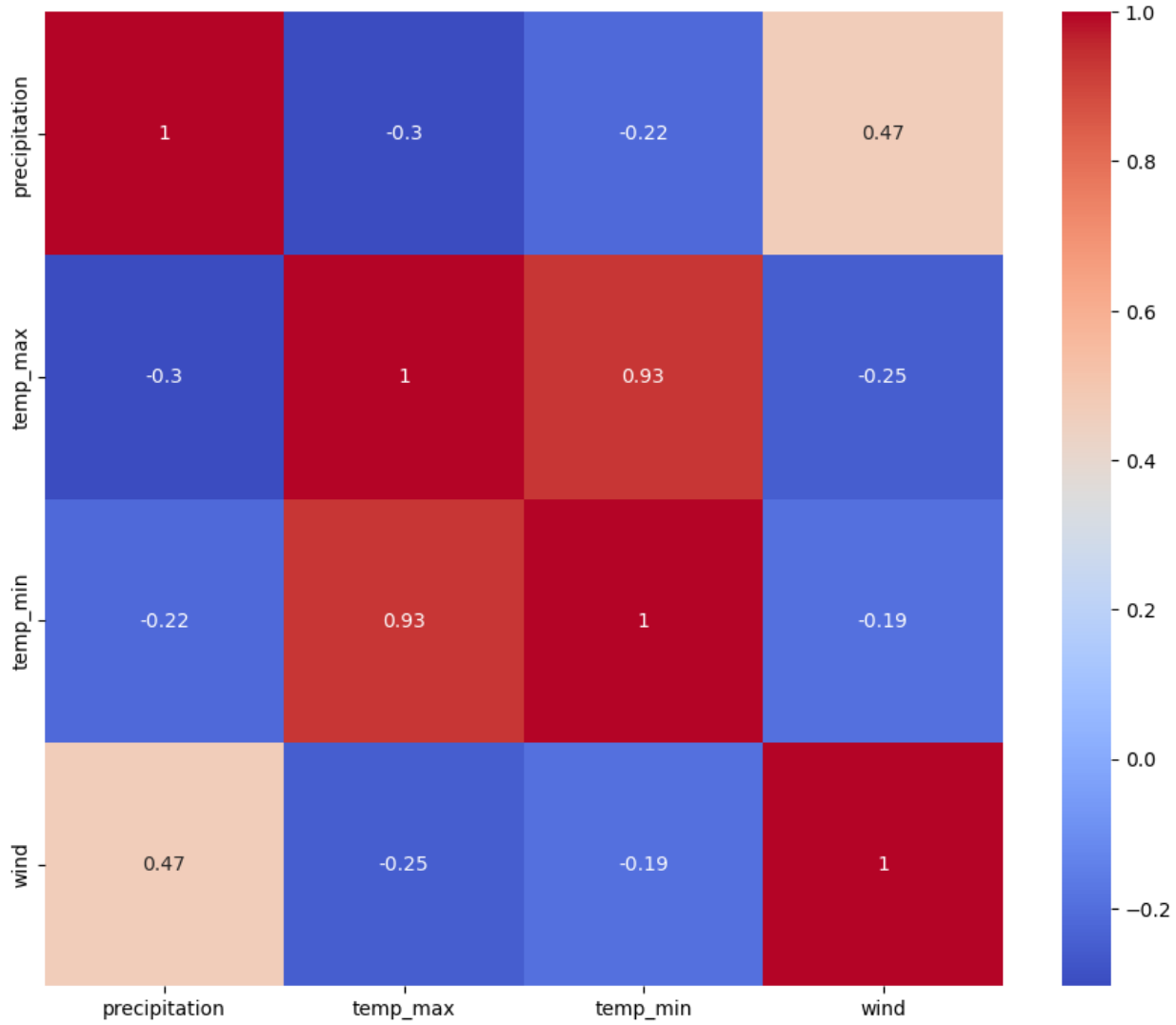
Now that we created more data instances, it makes it easier to visualize how the 2 principal components affect which class label a data instance belongs to. There is a weird artifact happening with the *snow* class to the right and we believe that is a side effect from using SMOTE on a very tiny (used to be 26 instances!) class count.

Violin plot:



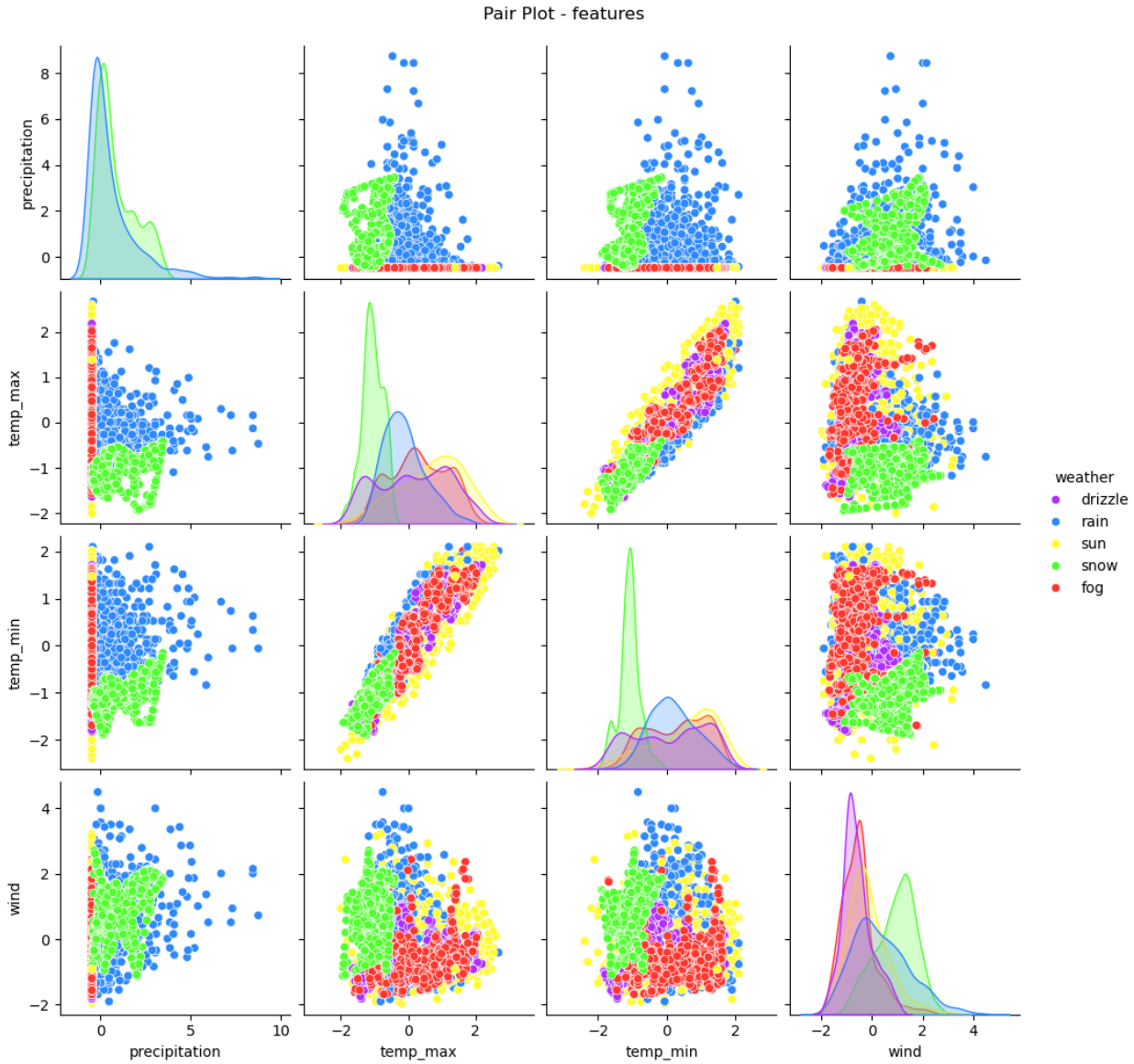
This is how the distributions look after scaling the data. We didn't bother printing a histogram or box plot because the violin plot does both, and it does it rather well. The *precipitation* feature has a lot of outliers while everything else is more uniformly distributed.

Correlation heatmap:



There are a lot of promising correlations here, particularly with *wind* and *precipitation*. When we first used this dataset on Assignment 4, the correlations were not as strong.

Pair plot:



This shows us how the features are correlated with each other based off the class labels.

Machine learning analysis

Identity Activation - Training Scores:

Accuracy: 0.6193447737909517

Sensitivities:

- drizzle: 0.4187380497131931
- rain: 0.3110236220472441
- sun: 0.826
- snow: 0.962890625
- fog: 0.5854126679462572

Specificities:

- drizzle: 0.8153662894580107
- rain: 0.8552631578947368
- sun: 0.9824414715719063
- snow: 0.9563318777292577
- fog: 0.7842298288508558

Precision: 0.6237652075361504

Recall: 0.6193447737909517

F1-score: 0.618097015844325

Log Loss: 0.7914983329523178

Identity Activation - Testing Scores:

Accuracy: 0.625585023400936

Sensitivities:

- drizzle: 0.4576271186440678
- rain: 0.2556390977443609
- sun: 0.851063829787234
- snow: 0.9302325581395349
- fog: 0.6083333333333333

Specificities:

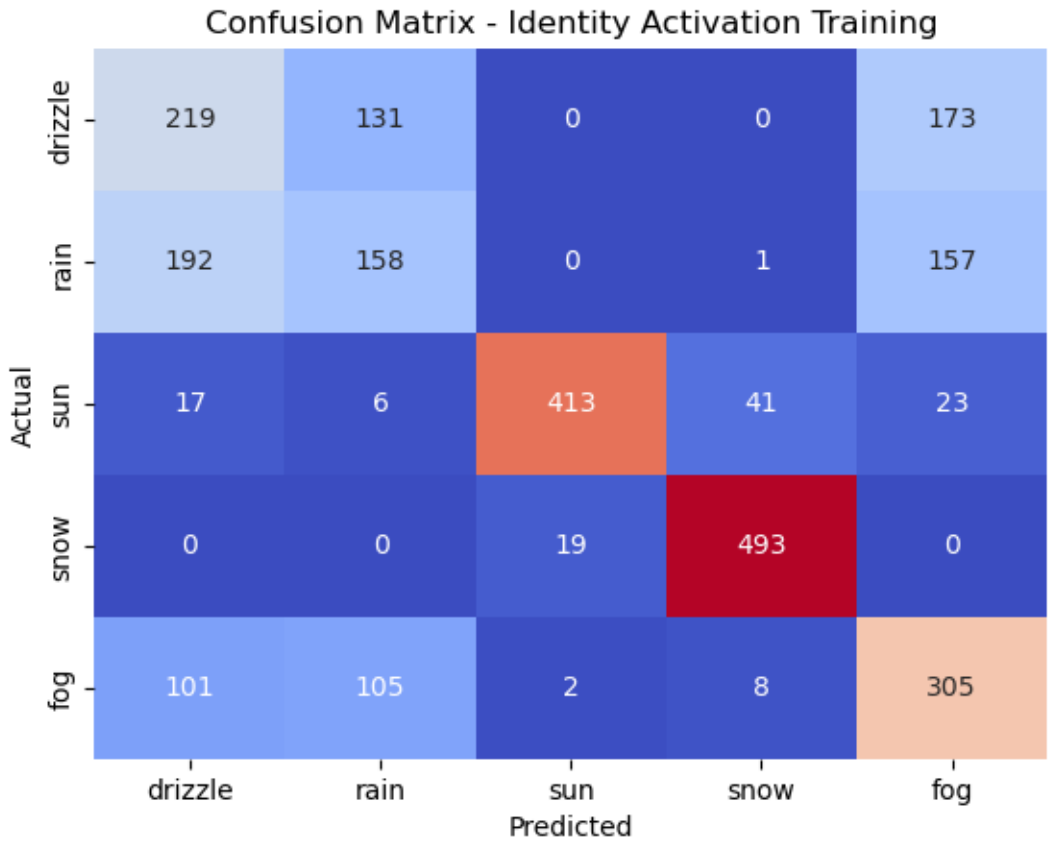
- drizzle: 0.7940503432494279
- rain: 0.8843373493975903
- sun: 0.9689655172413794
- snow: 0.9493243243243243
- fog: 0.8078817733990148

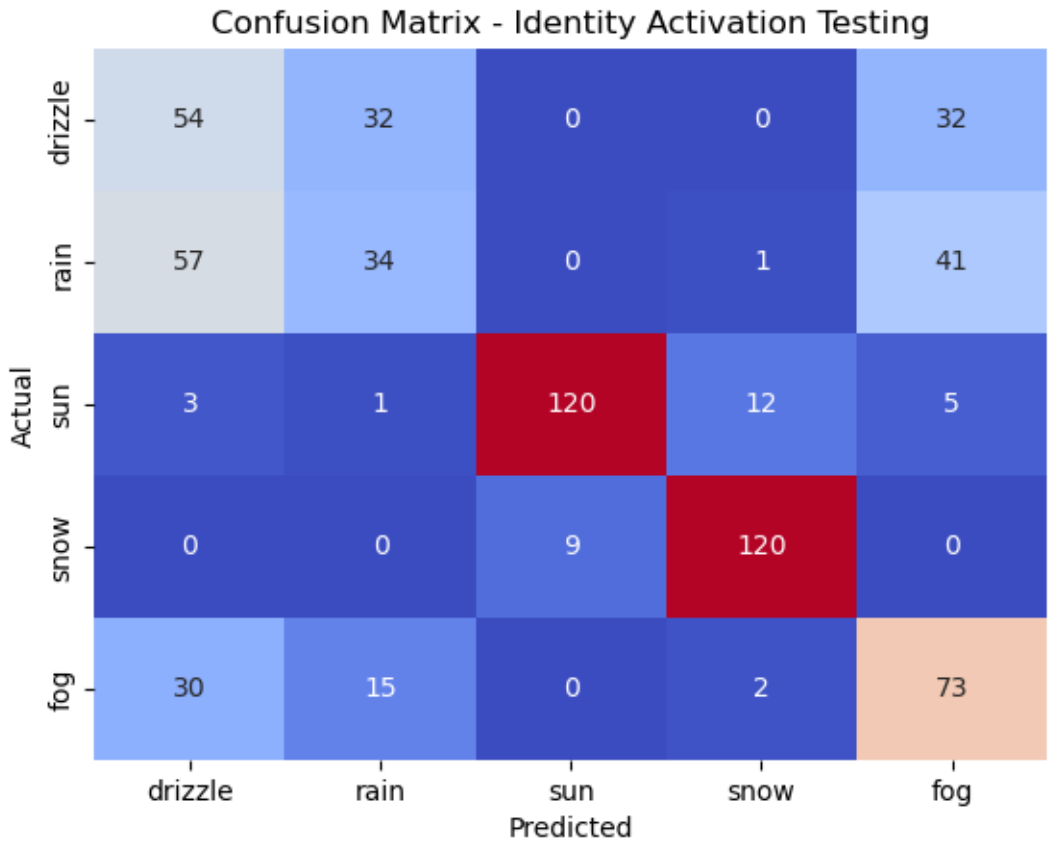
Precision: 0.6290780715458164

Recall: 0.625585023400936

F1-score: 0.6208453116281453

Log Loss: 0.7647719552057832




Confusion Matrix - Identity Activation Testing
drizzle
rain
sun
snow
fog
Actual
54
32
0
0
32
57
34
0
1
41
3
1
120
12
5
0
0
9
120
0
30
15
0
2
73
drizzle
rain
sun
snow
fog
Predicted

Tanh Activation - Training Scores:

Accuracy: 0.6942277691107644

Sensitivities:

- drizzle: 0.4588910133843212
- rain: 0.5984251968503937
- sun: 0.86
- snow: 0.974609375
- fog: 0.5892514395393474

Specificities:

- drizzle: 0.8974358974358975
- rain: 0.8353140916808149
- sun: 0.9904622157006603
- snow: 0.9756283320639756
- fog: 0.8441260744985674

Precision: 0.7035880069407743

Recall: 0.6942277691107644

F1-score: 0.6958206399064745

Log Loss: 0.6748225502956261

Tanh Activation - Testing Scores:

Accuracy: 0.7098283931357254

Sensitivities:

- drizzle: 0.4661016949152542
- rain: 0.5338345864661654
- sun: 0.8865248226950354
- snow: 0.9767441860465116
- fog: 0.65

Specificities:

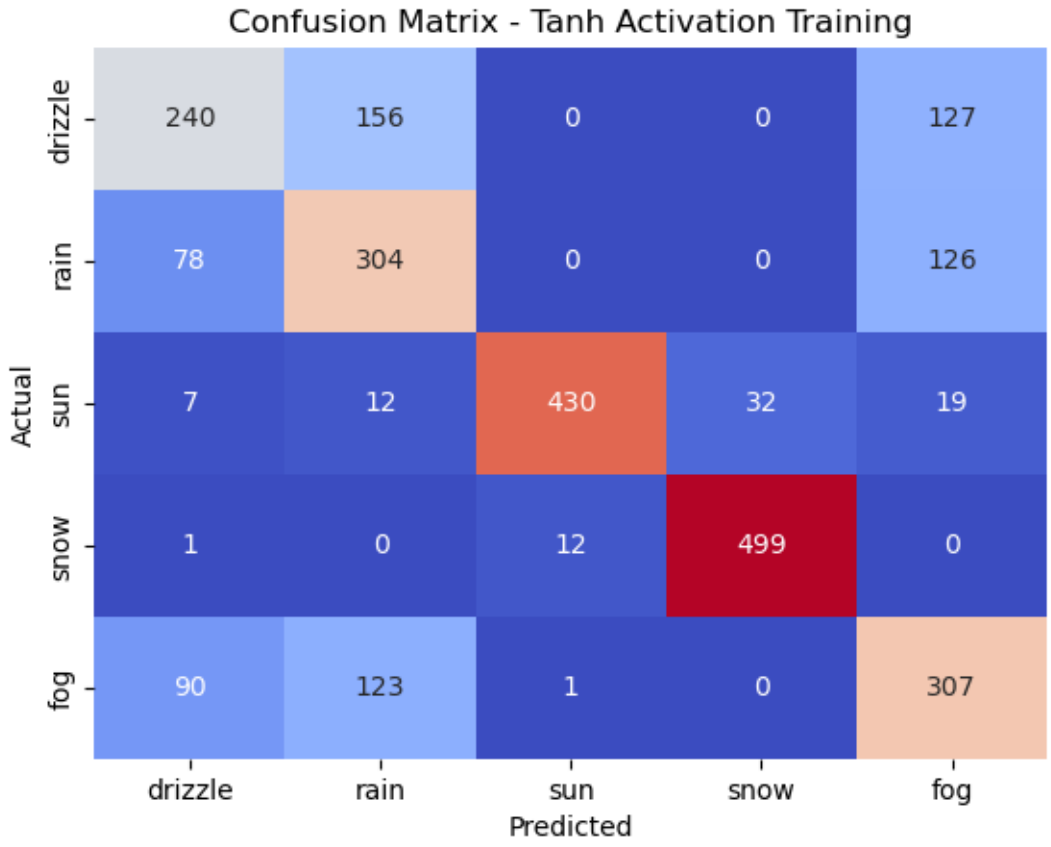
- drizzle: 0.9029345372460497
- rain: 0.8458149779735683
- sun: 0.9880239520958084
- snow: 0.9791666666666666
- fog: 0.8587699316628702

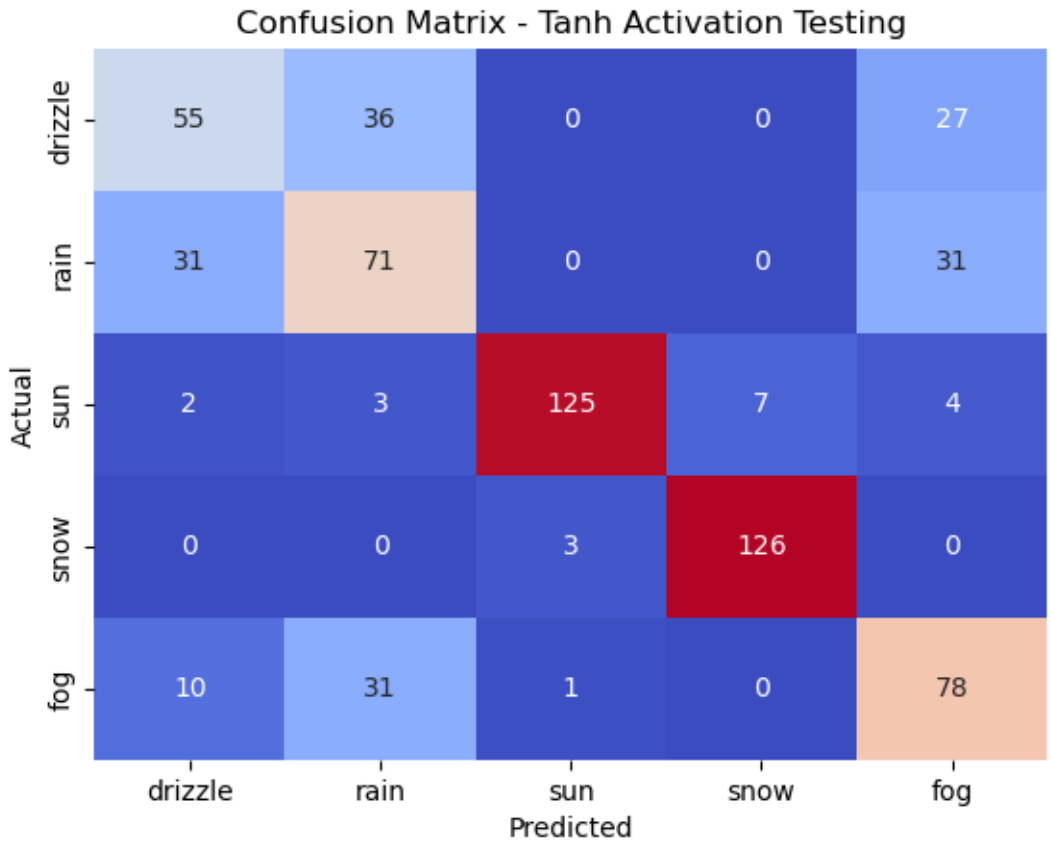
Precision: 0.7158996835433208

Recall: 0.7098283931357254

F1-score: 0.7108447602578483

Log Loss: 0.6492236702597707




Confusion Matrix - Tanh Activation Testing
drizzle
rain
sun
snow
fog
Actual
55
36
0
0
27
31
71
0
0
31
2
3
125
7
4
0
0
3
126
0
10
31
1
0
78
drizzle
rain
sun
snow
fog
Predicted

Relu Activation - Training Scores:

Accuracy: 0.7445397815912637

Sensitivities:

- drizzle: 0.4588910133843212
- rain: 0.5984251968503937
- sun: 0.86
- snow: 0.974609375
- fog: 0.5892514395393474

Specificities:

- drizzle: 0.8974358974358975
- rain: 0.8353140916808149
- sun: 0.9904622157006603
- snow: 0.9756283320639756
- fog: 0.8441260744985674

Precision: 0.7511126311688794

Recall: 0.7445397815912637

F1-score: 0.7466867823470288

Log Loss: 0.599368815484719

Relu Activation - Testing Scores:

Accuracy: 0.7581903276131046

Sensitivities:

- drizzle: 0.4661016949152542
- rain: 0.5338345864661654
- sun: 0.8865248226950354
- snow: 0.9767441860465116
- fog: 0.65

Specificities:

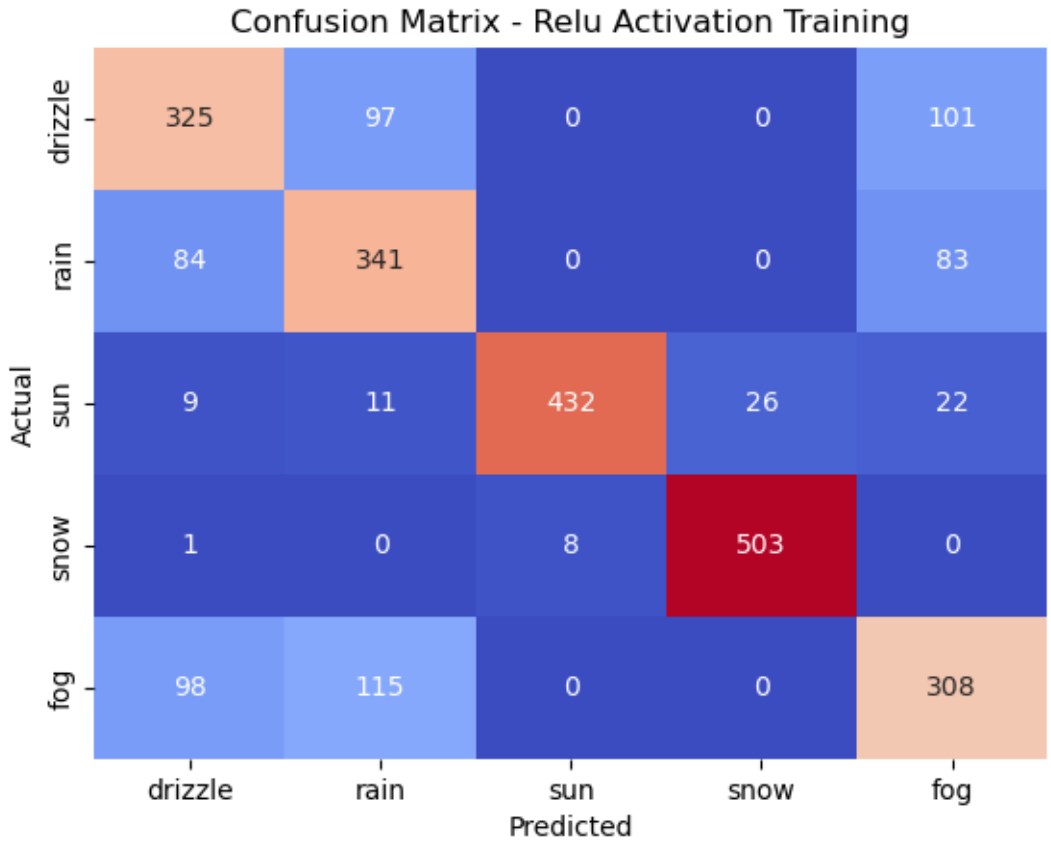
- drizzle: 0.9029345372460497
- rain: 0.8458149779735683
- sun: 0.9880239520958084
- snow: 0.9791666666666666
- fog: 0.8587699316628702

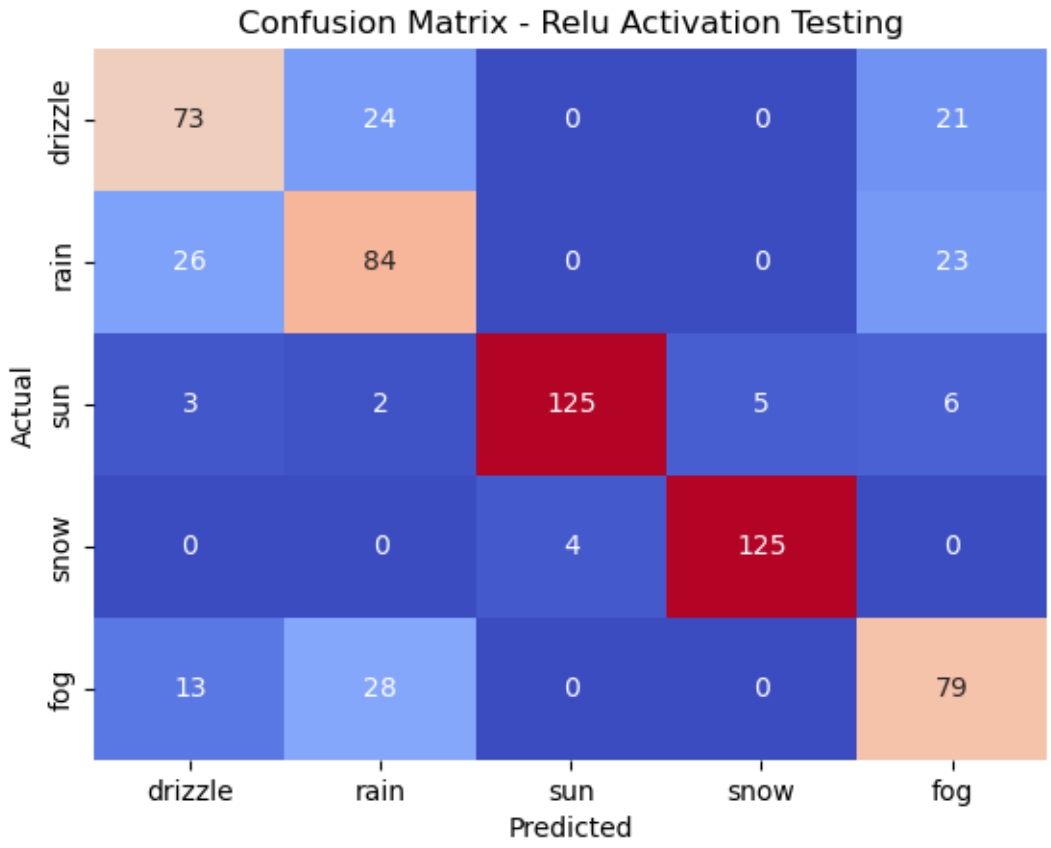
Precision: 0.764454930067779

Recall: 0.7581903276131046

F1-score: 0.7606982949315416

Log Loss: 0.5872464254959496




Confusion Matrix - Relu Activation Testing
drizzle
rain
sun
snow
fog
Actual
73
24
0
0
21
26
84
0
0
23
3
2
125
5
6
0
0
4
125
0
13
28
0
0
79
drizzle
rain
sun
snow
fog
Predicted