WEATHER PREDICTION USING MACHINE LEARNING

CLEANING TECHNIQUES:

- Locating missing data
- Missing data resolution
- Duplicate check
- Outlier detection
- Outlier resolution
- Normalize casing

CLEANING TECHNIQUES CONTINUATION:

- Remove noisy data
- Sampling data group by monthly, daily

DATA VISUALIZATION

- Bar plot
- Line plot
- Pair plot
- Heatmap

PREPARE DATA FOR TRAINING

- Datatype conversions
- Summary classes to numeric
- Feature scaling
- Split training and testing data into 33.3% and 66.6% respectively.

DECISION TREE CLASSIFIER

- Creates predictive model to draw conclusions based on scenarios and situations
- Function : DecisionTreeClassifier
- Hyperparameters: max_leaf_nodes, random_state
- Accuracy at 10 nodes - 48%
- Accuracy at 250 nodes – 52%
- Accuracy at 27 nodes - ~50%

K-NN CLASSIFIER

- Object is classified by vote of its neighbours, with most common class amoing its k nearest neighbor's.
- Function : KNeighborsClassifier
- Hyperparameters: n_neighbours
- Accuracy at N-neighbours 250 : 50%

RANDOM FOREST ALGORITHM

- Algorithms constructs multitude of decision trees at training time, output if random forest is the output that is selected by most of the decision trees.
- Function : RandomForestClassifier
- Hyperparameter: n_estimators
- Accuracy at n_estimators as 250: 58.4%

GAUSSIAN DISTRIBUTION

- This algorithm is based on continuous value probability distribution also known as normal distribution.
- Function: GaussianMixture
- Prediction: 42.9%

GRADIENT BOOSTING CLASSIFIER

- Prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. – weaker decision tree
- Function: GradientBoostingClassifier
- Hyperparameters: n_estimators, learning_rate, max_depth, random_state
- Accuracy at (n_estimators:100),(learning_rate:0.1),(max_depth:1),(random_state:0) = 57.9%

K-MEANS

- Select random k points centroid, assign each point to closest centroid which will form predefined k clusters,
- Update new centroid based on variance, repeat above bullets until centroid position are not modifying.
- Function: k-means
- Accuracy: 12.9%

OPTICS

- Stands for **Ordering points to identify the clustering structure,** algorithm is based on detecting meaningful cluster based on varying density. Internally implemented by – points of db ordered such that spatially closest point becomes neighbour in the ordering. - represented as dendrogram.
- Function: cluster_optics_dbscan
- Hyperparameters: reachability, core_distances, ordering, eps
- Accuracy at (reachability: reachability_), (core_distances: core_distances_),(ordering: ordering_), (eps:0.5) : 25.1%

NEAREST CENTROID CLASSIFICATION

- How it works: Centroid for each class is calculated while training, while predicting for point x, nearest centroid class is assigned to input under test.
- Function: NearestCentroid
- Hyperparameters: non
- Accuracy : 18.9%

SUMMARY

	Decision tree classifier	K-NN Classifier	Random forest algorithm	Gaussian Distribution	Gradient boosting classifier	K-Means	OPTICS	NEAREST CENTROID CLASSIFICATION
Algorithm	Supervised learning	Supervised learning	Supervised learning	clustering	classification	clustering	clustering	clustering
Accuracy	52%	50%	58.4%	42.9%	7.5%	12.9%	25.1%	18.9%
Precision	0.53	0.531	0.584	0.173	0.25	0.211	0.261	0.372
Recall	0.526	0.499	0.584	0.026	0.075	0.097	0.251	0.189
F-score	0.513	0.469	0.576	0.041	0.08	0.128	0.175	0.231