WEATHER REPORT

Final Project Report
ITCS 6190 - Cloud Computing for Data Analysis

SUBMITTED BY:

Shourya Reddy Katkam - 801255429
Manjusha Dondeti - 801261099
Rithesh Reddy Manchi Reddy - 801208216

PROJECT OVERVIEW:

This project aims to predict the climatic conditions based on the data which is available. We have a dataset named "weatherHistory.csv" which includes various attributes of climatic conditions on an hourly basis for a few years. So our goal is to predict a particular column in our dataset named "Summary" using other known attributes. We came up with this project as it would be interesting to predict the weather based on different criteria. And also to check the accuracy levels of each of those algorithms which we implemented on this particular dataset.

DATASET:

For this project, the dataset which we used is weatherHistory. This dataset consists of 96,454 rows and 12 columns for each row. We obtained this dataset from Kaggle. This dataset consists of weather information on an hourly analysis during the time period 2006-2016. This dataset consists of various climatic condition information such as Time, Temperature, Apparent Temperature, Humidity, Wind Speed, Visibility and Pressure.

SOURCE: https://www.kaggle.com/muthuj7/weather-dataset?select=weatherHistory.csv

Formatted D	Summary	Precip Type	Temperature	Apparent Te	Humidity	Wind Speed	Wind Bearin	Visibility (km	Loud Cover	Pressure (m	Daily Summary
2006-04-01 (	Partly Cloudy	rain	9.47222222	7.38888889	0.89	14.1197	251	15.8263	0	1015.13	Partly cloudy throughout the day.
2006-04-01 (	Partly Cloudy	rain	9.35555556	7.22777778	0.86	14.2646	259	15.8263	0	1015.63	Partly cloudy throughout the day.
2006-04-01 (	Mostly Cloud	rain	9.37777778	9.37777778	0.89	3.9284	204	14.9569	0	1015.94	Partly cloudy throughout the day.
2006-04-01 (	Partly Cloudy	rain	8.28888889	5.94444444	0.83	14.1036	269	15.8263	0	1016.41	Partly cloudy throughout the day.
2006-04-01 (	Mostly Cloud	rain	8.75555556	6.97777778	0.83	11.0446	259	15.8263	0	1016.51	Partly cloudy throughout the day.
2006-04-01 (	Partly Cloudy	rain	9.22222222	7.11111111	0.85	13.9587	258	14.9569	0	1016.66	Partly cloudy throughout the day.
2006-04-01 (	Partly Cloudy	rain	7.73333333	5.52222222	0.95	12.3648	259	9.982	0	1016.72	Partly cloudy throughout the day.
2006-04-01 (	Partly Cloudy	rain	8.77222222	6.52777778	0.89	14.1519	260	9.982	0	1016.84	Partly cloudy throughout the day.
2006-04-01 (	Partly Cloudy	rain	10.8222222	10.8222222	0.82	11.3183	259	9.982	0	1017.37	Partly cloudy throughout the day.
2006-04-01 (	Partly Cloudy	rain	13.7722222	13.7722222	0.72	12.5258	279	9.982	0	1017.22	Partly cloudy throughout the day.
2006-04-01 1	Partly Cloudy	rain	16.0166667	16.0166667	0.67	17.5651	290	11.2056	0	1017.42	Partly cloudy throughout the day.
2006-04-01 1	Partly Cloudy	rain	17.1444444	17.1444444	0.54	19.7869	316	11.4471	0	1017.74	Partly cloudy throughout the day.
2006-04-01 1	Partly Cloudy	rain	17.8	17.8	0.55	21.9443	281	11.27	0	1017.59	Partly cloudy throughout the day.
2006-04-01 1	Partly Cloudy	rain	17.3333333	17.3333333	0.51	20.6885	289	11.27	0	1017.48	Partly cloudy throughout the day.
2006-04-01 1	Partly Cloudy	rain	18.8777778	18.8777778	0.47	15.3755	262	11.4471	0	1017.17	Partly cloudy throughout the day.
2006-04-01 1	Partly Cloudy	rain	18.9111111	18.9111111	0.46	10.4006	288	11.27	0	1016.47	Partly cloudy throughout the day.
2006-04-01 1	Partly Cloudy	rain	15.3888889	15.3888889	0.6	14.4095	251	11.27	0	1016.15	Partly cloudy throughout the day.
2006-04-01 1	Mostly Cloud	rain	15.55	15.55	0.63	11.1573	230	11.4471	0	1016.17	Partly cloudy throughout the day.
2006-04-01 1	Mostly Cloud	rain	14.2555556	14.2555556	0.69	8.5169	163	11.2056	0	1015.82	Partly cloudy throughout the day.
2006-04-01 1	Mostly Cloud	rain	13.1444444	13.1444444	0.7	7.6314	139	11.2056	0	1015.83	Partly cloudy throughout the day.
2006-04-01 2	Mostly Cloud	rain	11.55	11.55	0.77	7.3899	147	11.0285	0	1015.85	Partly cloudy throughout the day.
2006-04-01 2	Mostly Cloud	rain	11.1833333	11.1833333	0.76	4.9266	160	9.982	0	1015.77	Partly cloudy throughout the day.
2006-04-01 2	Partly Cloudy	rain	10.1166667	10.1166667	0.79	6.6493	163	15.8263	0	1015.4	Partly cloudy throughout the day.
2006-04-01 2	Mostly Cloud	rain	10.2	10.2	0.77	3.9284	152	14.9569	0	1015.51	Partly cloudy throughout the day.
2006-04-10 (	Partly Cloudy	rain	10.4222222	10.4222222	0.62	16.9855	150	15.8263	0	1014.4	Mostly cloudy throughout the day.
2006-04-10 (	Partly Cloudy	rain	9.91111111	7.56666667	0.66	17.2109	149	15.8263	0	1014.2	Mostly cloudy throughout the day.
2006-04-10 (	Mostly Cloud	rain	11.1833333	11.1833333	0.8	10.8192	163	14.9569	0	1008.71	Mostly cloudy throughout the day.
2006-04-10 (	Partly Cloudy	rain	7.15555556	5.04444444	0.79	11.0768	180	15.8263	0	1014.47	Mostly cloudy throughout the day.
2006-04-10 (	Partly Cloudy	rain	6.11111111	4.81666667	0.82	6.6493	161	15.8263	0	1014.45	Mostly cloudy throughout the day.
2006-04-10 (	Partly Cloudy	rain	6.78888889	4.27222222	0.83	13.0088	135	14.9569	0	1014.49	Mostly cloudy throughout the day.
2006-04-10 (	Mostly Cloud	rain	7.26111111	5.15555556	0.85	11.1734	141	6.1985	0	1014.52	Mostly cloudy throughout the day.
2006-04-10 (	Mostly Cloud	rain	7.8	5.52777778	0.83	12.8156	150	8.05	0	1014.16	Mostly cloudy throughout the day.
2006-04-10 (	Mostly Cloud	rain	9.87222222	7.93333333	0.78	13.7494	160	9.982	0	1014.24	Mostly cloudy throughout the day.

ALGORITHMS IMPLEMENTED:

We initially preprocessed the data a little by adding an extra column named labelIndex to store the summary values in the numerical format. Once it's done, the data is split into two parts: train data and test data. The train data is run against the test data to predict the climatic condition. The predictions made are stored in a separate column named Prediction. As both labelIndex and Prediction columns refer to the same test data climatic conditions, by comparing these two columns we do understand the level at which we could rely on this. We could also get this using the accuracy values which we found for each algorithm.

DEFINITE GOAL:

We have used Spark framework to implement our project. The in-built algorithms which we used in this project are:

1. **RANDOM FOREST ALGORITHM:** Random Forest is a classifier algorithm which uses a large number of decision trees on various subsets from the dataset. This algorithm calculates the mean to present better results. The basic logic of Random Forest Algorithm lies with the point that it takes the majority votes of prediction from various decision trees instead of relying its decision on a single decision tree.

   OUTPUT:

```
Random Forest Classifier :
+----------+----------+
|labelIndex|prediction|
+----------+----------+
|       0.0|       2.0|
|       1.0|       1.0|
|       2.0|       1.0|
|       1.0|       1.0|
|       1.0|       1.0|
|       2.0|       2.0|
|       2.0|       1.0|
|       1.0|       1.0|
|       2.0|       1.0|
|       2.0|       1.0|
|       2.0|       2.0|
|       2.0|       2.0|
|       2.0|       2.0|
|       2.0|       2.0|
|       2.0|       2.0|
|       2.0|       2.0|
|       2.0|       1.0|
|       1.0|       1.0|
|       1.0|       1.0|
|       1.0|       1.0|
|       2.0|       1.0|
|       2.0|       1.0|
|       2.0|       0.0|
|       2.0|       1.0|
|       2.0|       2.0|
+----------+----------+
only showing top 25 rows

Accuracy = 0.5026025798150677
Test Error = 0.49739742018493227
```

2. **LOGISTIC REGRESSION:** Logistic Regression uses a set of independent variables to predict the categorical dependent variables. As this algorithm predicts the output of a categorical dependent variable, the outcome of this algorithm must be a categorical or discrete value.

OUTPUT:

```
Logistic Regression :
+----------+----------+
|labelIndex|prediction|
+----------+----------+
|       0.0|       2.0|
|       1.0|       2.0|
|       2.0|       1.0|
|       1.0|       2.0|
|       1.0|       2.0|
|       2.0|       2.0|
|       2.0|       1.0|
|       1.0|       1.0|
|       2.0|       1.0|
|       2.0|       1.0|
|       2.0|       0.0|
|       2.0|       2.0|
|       2.0|       2.0|
|       2.0|       2.0|
|       2.0|       2.0|
|       2.0|       2.0|
|       1.0|       1.0|
|       1.0|       1.0|
|       1.0|       1.0|
|       1.0|       1.0|
|       2.0|       1.0|
|       2.0|       0.0|
|       2.0|       1.0|
|       2.0|       1.0|
|       2.0|       2.0|
+----------+----------+
only showing top 25 rows

Accuracy = 0.423444577244844
Test Error = 0.576555422751559
```

3. **DECISION TREE CLASSIFIER:** Decision Tree Classifier is a Supervised learning technique. Although Decision Tree is majorly used for solving Classification problems, this classifier can be used for both Classification and Regression problems. In this classifier, internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome which makes this classifier a tree structured classifier.

OUTPUT:

```
Decision Tree Classifier :
+----------+----------+
|labelIndex|prediction|
+----------+----------+
|       0.0|       1.0|
|       1.0|       1.0|
|       2.0|       1.0|
|       1.0|       1.0|
|       1.0|       1.0|
|       2.0|       1.0|
|       2.0|       1.0|
|       1.0|       1.0|
|       2.0|       1.0|
|       2.0|       1.0|
|       2.0|       1.0|
|       2.0|       2.0|
|       2.0|       2.0|
|       2.0|       2.0|
|       2.0|       2.0|
|       2.0|       2.0|
|       1.0|       1.0|
|       1.0|       1.0|
|       1.0|       1.0|
|       1.0|       1.0|
|       2.0|       1.0|
|       2.0|       1.0|
|       2.0|       1.0|
|       2.0|       1.0|
|       2.0|       1.0|
+----------+----------+
only showing top 25 rows

Accuracy = 0.44940658772218506
Test Error = 0.5505934122778149
```

- As planned earlier we have implemented our project on three different algorithms and compared the accuracy levels of these algorithms to decide the best one among three. This was our definite goal to be met.

LIKELY GOAL:

- Apart from this, we have planned to work on one other algorithm but we practiced on two other algorithms to compare the accuracy levels for better results. Those two algorithms are:

1. **KNN ALGORITHM:** KNN Algorithm is one of the supervised Machine Learning algorithms which can be used for both Classification and Regression problems. Though it can be used for both of these algorithms, it is majorly used for Classification problems just like Decision Tree Classifier. It used the phenomenon of “feature similarity” for predicting the values of various data points.

KNN Algorithm

In [71]:
```
model = KNeighborsClassifier(n_neighbors=5, metric='minkowski', p=2 )
model.fit(x_train, y_train)
y_pred = model.predict(x_test)
Accuracy = accuracy_score(y_test,y_pred)*100
print(" accuracy is :  ",Accuracy)
```

```
 accuracy is :   48.01285595797623
```

2. **NAIVE BAYES CLASSIFIER:** Naive Bayes Classifier can be considered as a collection of various classification algorithms named Bayes Theorem. The basic assumption in this particular classifier is that each and every feature makes an independent and equal contribution to the final outcome.

OUTPUT:

Naive Bayes Classifier

:
```
model = GaussianNB()
model.fit(x_train,y_train)
y_pred = model.predict(x_test)
Accuracy = accuracy_score(y_test,y_pred)*100
print(" accuracy is :  ",Accuracy)
```

```
 accuracy is :   47.26983688139342
```

TEAM MEMBERS CONTRIBUTION:

This project is a collective effort of all 3 of us. We divided the tasks equally for getting the project done.

- Manjusha worked on getting outputs from Random Forest Classifier and also Naive Bayes Classifier.
- Shourya worked on getting outputs using Decision Tree Classifier and also KNN Algorithm.
- Rithesh Reddy worked on Logistic Regression and took the responsibility of preparing the Project Report.
- We worked together for the rest of the project work.

SUMMARY

We have gone through the three main types of machine learning algorithms mainly Decision Tree Classifier, Random Forest Classifier and Logistic Regression. The code snippets to implement the algorithms in Pyspark were also discussed and we saw that the Random Forest Classifier outperformed both Decision Tree Classifier and Logistic Regression in terms of accuracy. Logically speaking, since Random Forest Classifier involves growing more trees than the single tree of Decision Tree Classifier and Logistic Regression which is used for predicting the categorical dependent variable using a given set of independent variables. Therefore Random Forest Classifiers will be able to make more accurate predictions/classifications. However, this is not always the case as it also depends on the dataset as well. As there is a common saying in Machine Learning, no one algorithm works best for every prediction task. It's all about exploring and investigating to know which algorithm performs better.

PROJECT GITHUB LINK: [shaurya Reddy/GroupProject_cloud_Weatherreport (github.com)]