CSCI-6401-01 **BLACK SHARKS**

Phase 6

WEATHER PREDICTION

Team name: Black Sharks
Team members:

1. Sudheer Kumar Reddy Reddymalla
   sredd1@unh.newhaven.edu
   00719922
2. Durga Guna SekharReddy Savanam
   dsava2@unh.newhaven.edu
   00755099
3. Vamsi Krishna Regalla
   vrega1@unh.newhaven.edu
   00730353

Research Question:

To predict the weather based on the patterns which were previously collected using data mining techniques. Here, in this case we are using Seattle's weather data.

Data Set:

The data set we choose is taken from Kaggle website. Here is the link for data set: https://www.kaggle.com/datasets/ananthr1/weather-prediction which has 6 column attributes.

Data Mining Techniques:

- Naïve Bayes Classification
- K-Star algorithm
- SMO model
- Decision Tree and
- Random forest

Naïve Bayes Classification

A probabilistic classifier is the Naive Bayes algorithm for classification. It is based on probability models that make substantial assumptions about independence. The independence presumptions frequently do not affect reality. They are therefore viewed as being naive.

K-Star algorithm
The class of a test instance is based on the class of training instance like it, as defined by some clustering algorithm according to the instance-based classifier K star. Its use of an entropy-based distance function sets it apart from other instance-based learners.

SMO model
is a method for resolving the quadratic programming (QP) problem that appears during training support vector machines (SVM). Support vector machines are frequently trained using SMO, which is implemented by the well-liked LIBSVM tool. As a result, that earlier SVM training techniques were far more complicated and needed pricey third-party QP solvers.

Random Forest:
Supervised machine learning algorithms like random forest are frequently employed in classification and regression issues. It creates decision trees from several samples, using the majority vote for classification and the average in the case of regression.

Parameters and Hyperparameters:

In Naive Bayes algorithm, precipitation will look like these:

Attribute	drizzle	rain	Sun	Snow	fog
mean	0	6.5937	0	8.5414	0
Std. dev.	0.0847	8.6186	0.0847	6.8702	0.0847
Weight sum	53	641	640	26	101
precision	0.5082	0.5082	0.5082	0.5082	0.5082

Parameter batch size- 300
useKernekEstimator: False
numDecimalPlaces: 2

SMO model:
Parameter batch size- 300
useKernekEstimator: PolyKernel
numDecimalPlaces: 2
tolerance: 0.001
filter type: Normalize training data

Random Forest:
Batch size: 300
maxDepth: 0
minDepth: 1.0
minVarianceProp: 0.001
numFolds: 0
seed: 1

Decision Table:
Batch size: 300
evaluationMeasure: accuracy
numDecimalPlaces: 2
search: BestFirst
debug: False

Optimized data:

	Correct instances	True positive	False Positive	Precision	Recall
Naïve Bayes Classification	86.8 %	0.869	0.093	0.885	0.869
SMO model	99%	0.995	0.004	0.995	0.995
K-Star algorithm	100%	1.0	0	1.0	1.0
Random Forest	100%	1.0	0.1	1.0	1.0
Decision Table	85.2%	0.852	0.114	0.88	0.852

Conclusion:
Optimization can be a large data component where we collect and manage it efficiently using various techniques. By using these optimizing techniques, K-star has maximum correct instances when compared with others.

GitHub:

https://github.com/sudheerredde/BLACKSHARKS.git