Natural Disaster and Weather Prediction System

CSCI 720: Big Data Analytics
Project Presentation

Milind Kamath
Priyanka Punjabi
Vineet Kamat

Overview

- ❏ Introduction
- ❏ Data Cleaning and Preprocessing
- ❏ Data Exploration and Visualization
- ❏ Data Mining Techniques
  - ❏ K-Means Clustering
  - ❏ DBScan
- ❏ Conclusion
- ❏ Libraries Used

INTRODUCTION

- The goal of this Data Mining task was to analyze and predict weather based on the Tao and Elnino dataset.
- The first step was cleaning the data, pre-processing, followed by extracting meaningful features from it.
- The dataset was analyzed by performing clustering to analyze patterns and correlations in the data which helped us discover useful knowledge.

Data Cleaning and Preprocessing

Creating CSV Files

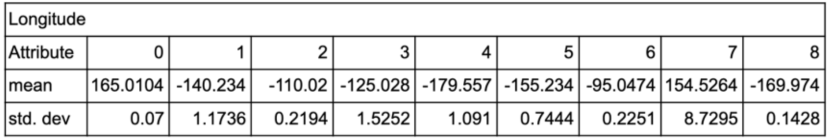
Eliminating Unnecessary Columns

Missing Values

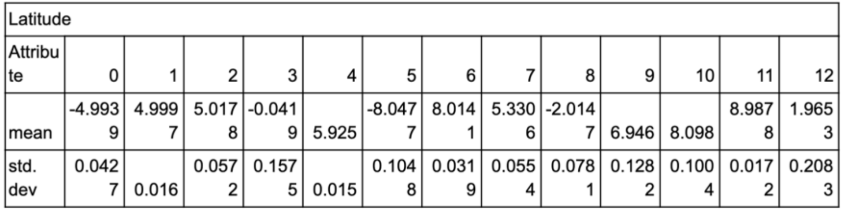
Discretize Data

DATA CLEANING

- The data was thoroughly checked for any illegal characters.
- Missing Values were fixed:
  - Calculated the mean values of similar data (shown in the figure) and replaced the missing values.
  - If all the values for an attribute within the similar data range are missing, then the missing values were replaced with the global mean of that attribute.



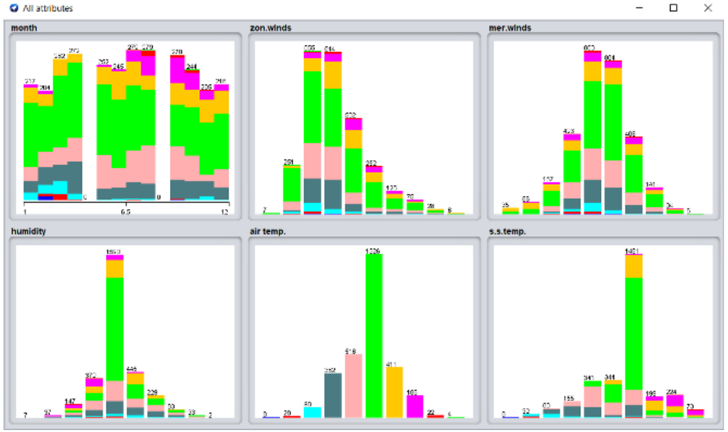
Longitude									
Attribute	0	1	2	3	4	5	6	7	8
mean	165.0104	-140.234	-110.02	-125.028	-179.557	-155.234	-95.0474	154.5264	-169.974
std. dev	0.07	1.1736	0.2194	1.5252	1.091	0.7444	0.2251	8.7295	0.1428



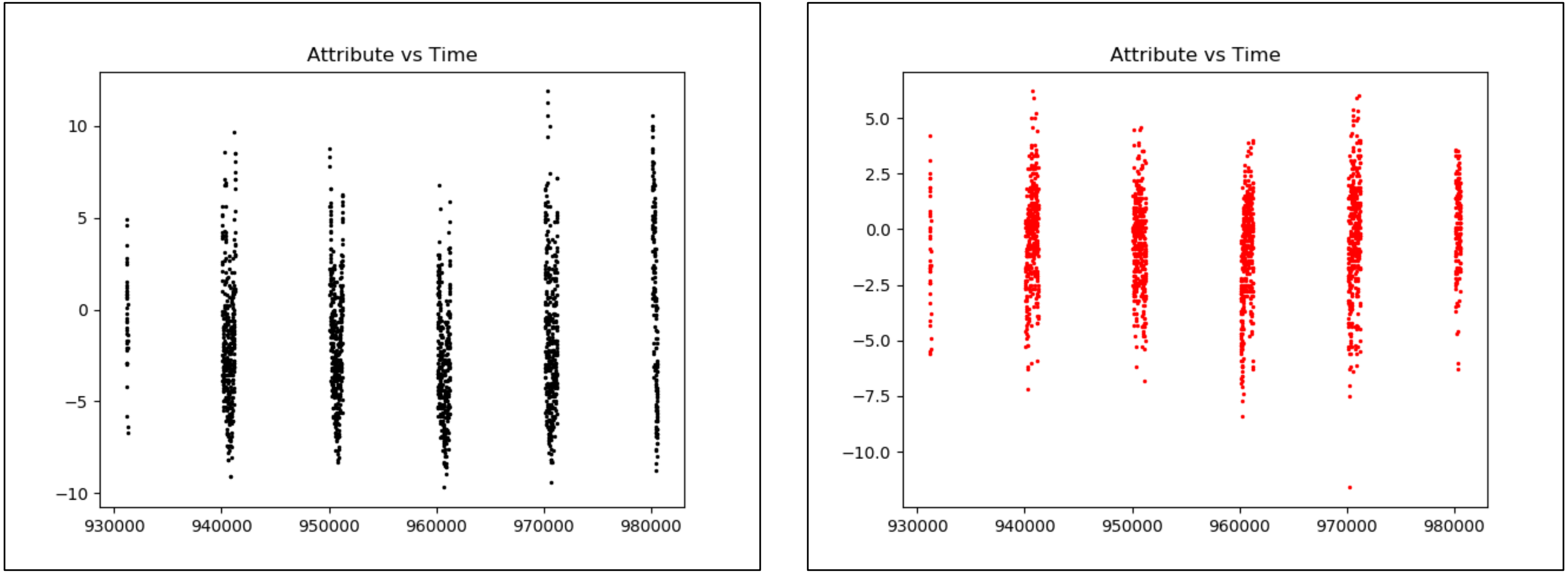
Latitude													
Attribute	0	1	2	3	4	5	6	7	8	9	10	11	12
mean	-4.9939	4.9997	5.0178	-0.0419	5.925	-8.0477	8.0141	5.3306	-2.0147	6.946	8.098	8.9878	1.9653
std. dev	0.0427	0.016	0.0572	0.1575	0.015	0.1048	0.0319	0.0554	0.0781	0.1282	0.1004	0.0172	0.2083

DATA PREPROCESSING

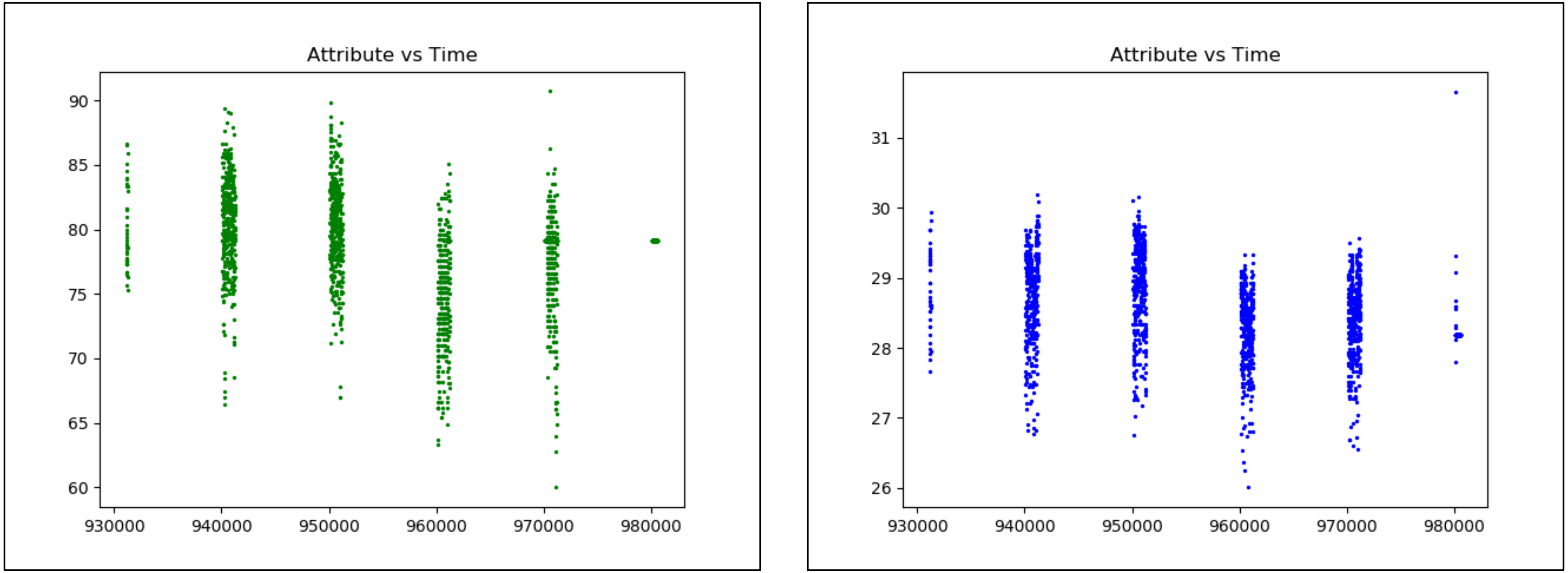
- ❑ Created CSV file using Python script using \s+ as the delimiter.
- ❑ Unnecessary columns such as number, day, year, date were eliminated.
- ❑ In order to efficiently run the algorithms for our dataset, our dataset was discretized into bins. These bins were then converted to nominal values.



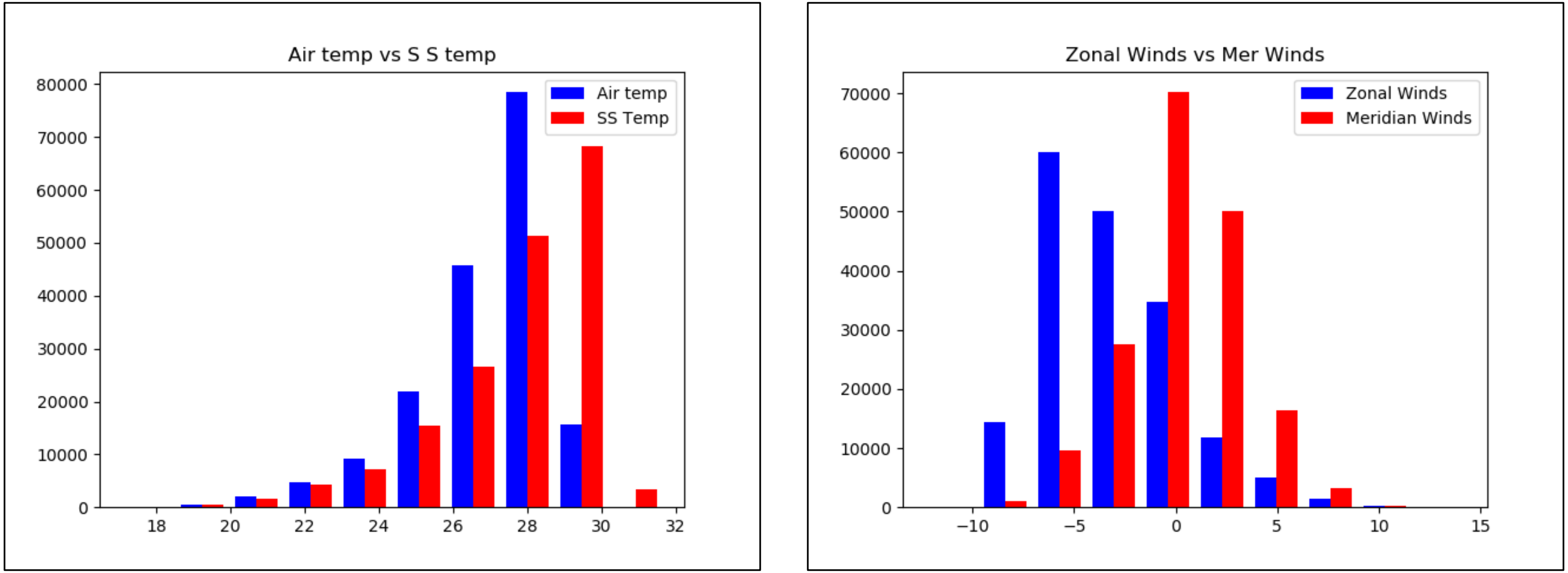
Data Exploration and Visualization



Scatter Plots for buoy 1 with zon. winds, mer. winds, humidity and air temperature



Scatter Plots for buoy 1 with zon. winds, mer. winds, humidity and air temperature



Histograms - relation between Air v/s S.S temperature and zonal winds v/s meridian winds.

Data Mining Techniques

K - Means Clustering

DBScan

Approach

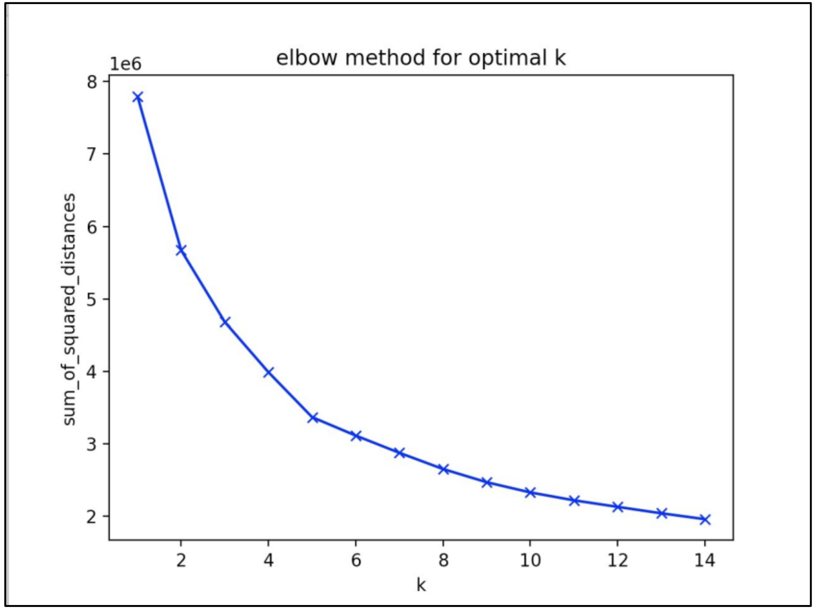
- ❑ We ran various clustering and classification techniques to analyze the datasets:
  - ❑ Rule - Based Classification - PRISM
  - ❑ Tree - Based Classification - Random Forest
  - ❑ Clustering - K-Means and DBScan.
- ❑ Finally, based on the preliminary results obtained, we decided on using Clustering techniques to study patterns in the data.
- ❑ TAO dataset was used form the clusters which were studied using the Elnino dataset.

Approach

- ❑ Initial Setup:
    - ❑ Use the complete dataset.
    - ❑ We discretize the values and then converted them to numeric values.
    - ❑ Result: Due to sparsity the accuracy of the model was not up to the mark.
- ❑ Next Approach:
    - ❑ Grouped database based on seasons.
    - ❑ Result: faced the curse of dimensionality problem. The overall performance was still below mark
- ❑ Finally:
    - ❑ Considered only two highly correlated attributes for clustering and classification purpose.
    - ❑ Result: Drastic improvement in the accuracy and silhouette coefficient observed.

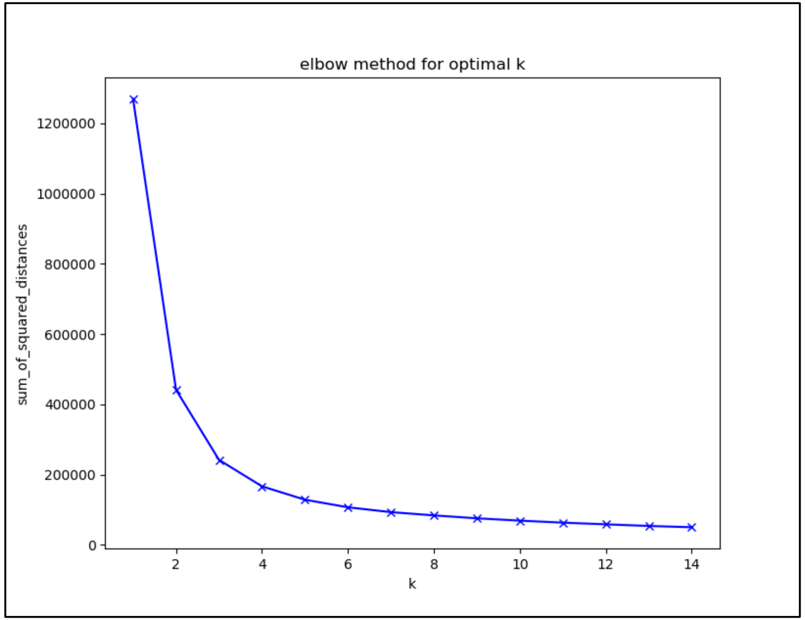
K–MEANS CLUSTERING – OPTIMAL K

- ❏ To achieve optimal value of K and create K clusters, we used the elbow method.
- ❏ The idea of the elbow method is to run k-means clustering on the dataset for a range of values of k (from 1 to 15), and for each value of k calculate the sum of squared errors (SSE).
- ❏ As seen from the graph, our optimal value for k is equal to 4 or 5. Hence, we chose k = 4 for all attributes.



K-MEANS CLUSTERING – OPTIMAL K

- A similar approach was followed for the two highly correlated attributes and the resulted graph gave an optimal value of between 2 and 3.
- So, for the highly correlated attributes, we chose the value of k = 2.



K-MEANS CLUSTERING

- ❑ The elbow method was used to get an optimal k value.
- ❑ We ran k - means to fit the dataset in 4 and 2 clusters (curse of dimensionality problem discussed earlier).
- ❑ To understand the representation of the clusters obtained we used the data points in the Elnino dataset to understand which clusters represent the possibility of a calamity.
- ❑ The higher samples in a cluster represented the higher possibility of a calamity occuring.
- ❑ More number of samples from Elnino were predicted in cluster 1 than in cluster 2.
- ❑ Further, we calculated the multivariate normal distribution of the new data points in order to determine the probability of a datapoint being in cluster 1 and cluster 2.

ALGORITHM COMPARISON

1

K - Means Clustering

No.of clusters: 2

Silhouette Score: 0.6695

2

K - Means Clustering

No.of clusters: 4

Silhouette Score: 0.2590

3

DBScan

Silhouette Score: -0.2982

Conclusion

CONCLUSION

❑ After evaluating both clustering and classification models, for our dataset, clustering seems to an appropriate choice.

❑ K-Means with 2 clusters seemed to be the most promising algorithm resulting in a silhouette score of 0.669 - high enough to be deemed as a good cluster quality.

Thank You

Milind Kamath
mk6715@rit.edu

Priyanka Punjabi
pp4762@rit.edu

Vineet Kamat
vvk1199@rit.edu