Sarve Khorramshahi
Ayan Ashkenov
CS699: Data Mining
02/16/2023

Project Proposal

Dataset:

https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00014739/detail
1461 rows, 45 columns
Class attribute in the dataset: “WEATHER_CONDITION”

* The weather.csv dataset file will be attached to the submission
**To read it in R, please save to your desktop and run the following

```
file_path <- "~/Desktop/weather_data.csv"
weather_data <- read.csv(file_path)
head(weather_data)
```

Dataset Description:

This dataset contains meteorological data from the National Centers for Environmental Information (National Oceanic and Atmospheric Administration) collected from a weather radar placed at Boston Logan International Airport. We were able to select a custom date range of data which we limited to January 2019 to December 2022 and obtain daily meteorological data including but not limited to precipitation, snow, maximum and minimum temperatures, snow depth, windspeed, and weather types. This date range can certainly be expanded if we choose to pull in more data.

The challenging and simultaneously interesting aspect of this dataset is its recency. This data can be accessed almost to the current week which would allow us to develop our classification algorithms and make inferences on when and how much snow we should expect given current weather conditions. We believe limiting our dataset to the most recent years will provide the least margin for error as it will take into consideration the current state of the climate and effects of global warming. Our intention is to be able to classify meteorological conditions to determine when there will and will not be snow. We may also expand our scope to classify the various levels of snow with more detail to distinguish between days of heavy snow and periods of light snow. Research, discussion, and testing will be done to determine the best method for this dataset on our part.

In regard to the algorithms, it is likely that the selection of 5 will start off with classical, simple algorithms, such as the vanilla decision tree, and finish with very complex, industry level types.

Data Attributes:

- STATION: Alphanumeric station code, unique identifier
- NAME: Name of the station where the meters are located

- LATITUDE: Numeric latitude where the station meter is located
- LONGITUDE: Numeric longitude where the station meter is located
- ELEVATION: Numeric elevation above mean sea level in tenths of a meter
- DATE: Date of recording in the format yyyy-mm-dd
- AWND: Average daily wind speed (miles per hour)
- AWND_ATTRIBUTES: Wind attribute measurement, quality, and source flags (below)
- PGTM: Peak gust time (hours and minutes)
- PGTM_ATTRIBUTES: gust attribute measurement, quality, and source flags (below)
- PRCP: Precipitation (inches)
- PRCP_ATTRIBUTES: precipitation attribute measurement, quality, and source flags (below)
- SNOW: Snowfall (inches)
- SNOW_ATTRIBUTES: snow attribute measurement, quality, and source flags (below)
- TAVG: Average temperature (Farenheit)
- TAVG_ATTRIBUTES: average temp attribute measurement, quality, and source flags (below)
- TMAX: Maximum temperature (Farenheit)
- TMAX_ATTRIBUTES: max temp attribute measurement, quality, and source flags (below)
- TMIN: Minimum temperature (Farenheit)
- TMIN_ATTRIBUTES: min temp attribute measurement, quality, and source flags (below)
- WDF2: Direction of average 2 minute wind (degrees)
- WDF2_ATTRIBUTES: direction of avg 2 minute wind attribute measurement, quality, and source flags (below)
- WDF5: Direction of average 5 second wind (degrees)
- WDF5_ATTRIBUTES: direction of avg 5 second wind attribute measurement, quality, and source flags (below)
- WSF2: Fastest 2 minute wind speed (miles per hour)
- WSF2_ATTRIBUTES: fastest 2 minute wind speed attribute measurement, quality, and source flags (below)
- WSF5: Fastest 5 second wind speed (miles per hour)
- WSF5_ATTRIBUTES: Fastest 5 second wind speed attribute measurement, quality, and source flags (below)
- WT01: Weather type where there is fog, ice fog, or freezing fog
- WT01_ATTRIBUTES: Fog weather type attribute measurement, quality, and source flags (below)
- WT02: Weather type where there is heavy fog or heavy freezing fog
- WT02_ATTRIBUTES: Heavy fog weather type attribute measurement, quality, and source flags (below)
- WT03: Weather type where there is thunder
- WT03_ATTRIBUTES: Thunder weather type attribute measurement, quality, and source flags (below)
- WT04: Weather type where there are ice pellets, sleet, snow pellets, or small hail

- WT04_ATTRIBUTES: Ice pellet weather type attribute measurement, quality, and source flags (below)
- WT05: Weather type where there is hail
- WT05_ATTRIBUTES: Hail weather type attribute measurement, quality, and source flags (below)
- WT06: Weather type where there is glaze or rime
- WT06_ATTRIBUTES: Glaze or rime weather type attribute measurement, quality, and source flags (below)
- WT08: Weather type where there is smoke or haze
- WT08_ATTRIBUTES: Smoke or haze weather type attribute measurement, quality, and source flags (below)
- WT09: Weather type where there is blowing snow
- WT09_ATTRIBUTES: Blowing snow weather type attribute measurement, quality, and source flags (below)

For attribute flags:

Measurement Flag

Blank = no measurement information applicable

A = value in precipitation or snow is a multi-day total, accumulated since last measurement

(used on Daily Form pdf file)

B = precipitation total formed from two twelve-hour totals

D = precipitation total formed from four six-hour totals

H = represents highest or lowest hourly temperature (TMAX or TMIN)

or average of hourly values (TAVG)

K = converted from knots

L = temperature appears to be lagged with respect to reported

hour of observation

O = converted from oktas

P = identified as "missing presumed zero" in DSI 3200 and 3206 T = trace of precipitation, snowfall, or snow depth

W = converted from 16-point WBAN code (for wind direction)

Quality Flag

Blank = did not fail any quality assurance check D = failed duplicate check

G = failed gap check

I = failed internal consistency check

K = failed streak/frequent-value check

L = failed check on length of multiday period M = failed mega-consistency check

N = failed naught check

O = failed climatological outlier check

R = failed lagged range check

S = failed spatial consistency check

T = failed temporal consistency check

W = temperature too warm for snow

X = failed bounds check

Z = flagged as a result of an official Datzilla investigation

Source Flag

Blank = No source (i.e., data value missing)

W = WBAN/ASOS Summary of the Day from NCDC's Integrated Surface Data (ISD).

Data Mining Goal:

To determine which days will have snow (to classify weather types by meteorological factors).