Assignment 3: **Analyzing Sensor Data for Weather Prediction**

Overview:

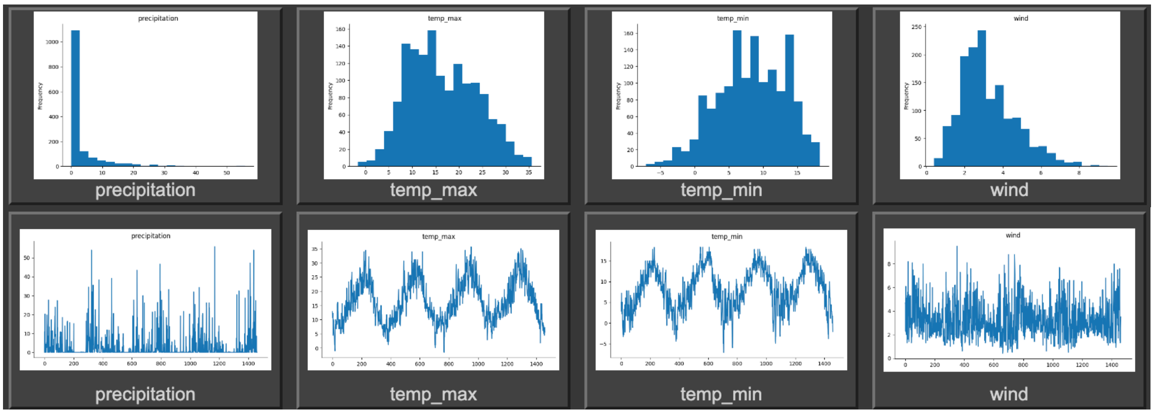
Students are expected to work with a dataset say, representing sensor data related to weather conditions (temperature, humidity, wind, etc.) over a period. The goal is to use various probabilistic models and inference methods to predict weather conditions and understand the underlying processes.

Resources: https://pgmpy.org/

Task 1: Bayesian Network for Weather Prediction.

Data Description:

a. Name: seattle-weather.csv [Kaggle - CC BY-NC-SA 4.0]
b. Number of features: 5
   'precipitation' - All forms in which water falls on the land surface and open water bodies as rain, sleet, snow, hail, or drizzle.
   'temp_max' - Maximum Temperature,
   'temp_min' - Minimum Temperature,
   'wind' - Wind speed,
   'weather' - (Rain: 44%; Sunny: 44%; Other (180): 12%).



1.1 Construct a Bayesian Network that models the relationships between different weather variables. See the precode.ipynb for this. We have provided the precode for the following relation (Figure 1):

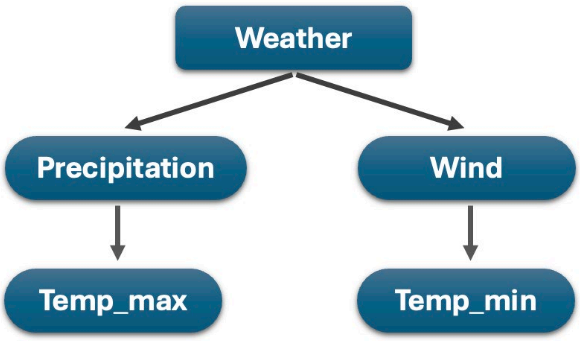


Fig1: Sample Hierarchy in consideration for the given data

1.2 Implement exact inference to answer the following questions. Use Variable Elimination (VE) for this.

Question 1.2.1: (a) What is the probability of high wind when the weather is sunny? (b) What is the probability of sunny weather when the wind is high?

Question 1.2.2: (a) Calculate all the possible joint probability and determine the best probable condition. Explain your results? (b) What is the most probable condition for precipitation, wind and weather, combined?

Question 1.2.3. Find the probability associated with each weather, given that the precipitation is medium? Explain your result.

Question 1.2.4. What is the probability of each weather condition given that precipitation is medium, and wind is low or medium? Explain your method and results. How does the result change with the addition of the wind factor compared to Question 1.2.3?

1.3 Use approximate inference methods to answer all four questions in Task 1.2. You can use techniques such as sampling, direct sampling, rejection sampling, likelihood weighting, Markov chain simulation, Gibbs sampling, and so on. Make sure to use a different technique for each question. So, if you used rejection sampling for Question 1.2.1, don't use it for any other questions? Compare the results of approximate and exact inference and explain any differences or similarities.

1.4 Exploring other hierarchies:

1.4.1 Create a Bayesian Network on the same data with two different hierarchies given below (ref. Figure 2 & 3).

1.4.2 Compare the joint probabilities from all the three-hierarchy using either exact or approximate inference? Discuss your results?

Hierarchy 1:

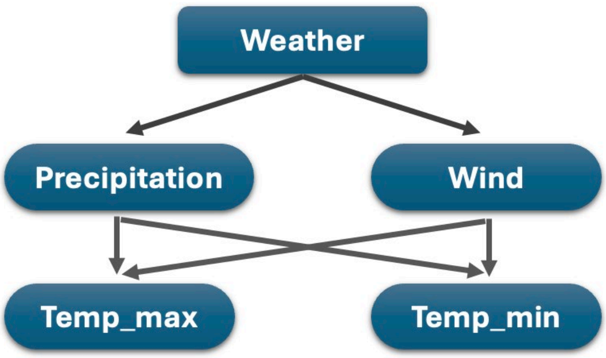


Fig2: Hierarchy 1 for the given data.

Hierarchy 2:

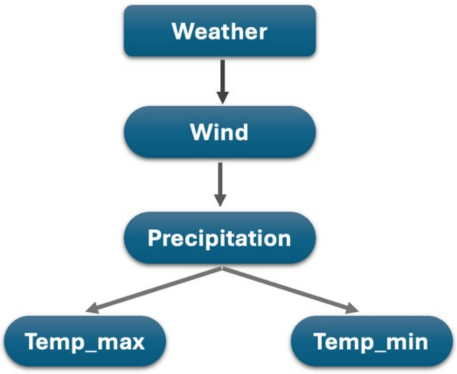


Fig 3: Hierarchy 2 for the data.

Task 2: Hierarchical Feature Analysis in Bayesian Networks

Data Description:

Spray icing modelling and decision support in the Norwegian maritime sector (SPRICE)

Funding Source - Research Council of Norway, funded under the MAROFF-2 Programme [Grant No: 320843]

a. Name: SPRICE_Norwegian_Maritime_Data.csv [Responsible: Sushmit Dhar]
b. Number of features: 30
c. Important Features: You are free to use as many variables as you want, but we've highlighted a few key features from the data for you to consider.
   'TIMESTAMP' - Timestamp of the data record (1 minute) [dd-mm-yy hh:mm]
   'Air_temp_Act' - Actual air temperature (°C),
   'Rel_Humidity_act' - Actual relative humidity (%),
   'Rel_Air_Pressure' - Relative air pressure (hPa),
   'Wind_Speed_avg' - Average wind speed (m/s),
   'Wind_Direction_vct' - Wind direction vector (degrees 0-360°),
   'Precipitation_Type' - Type of precipitation (Type code),
   'Precipitation_Intensity' - Intensity of precipitation (mm).

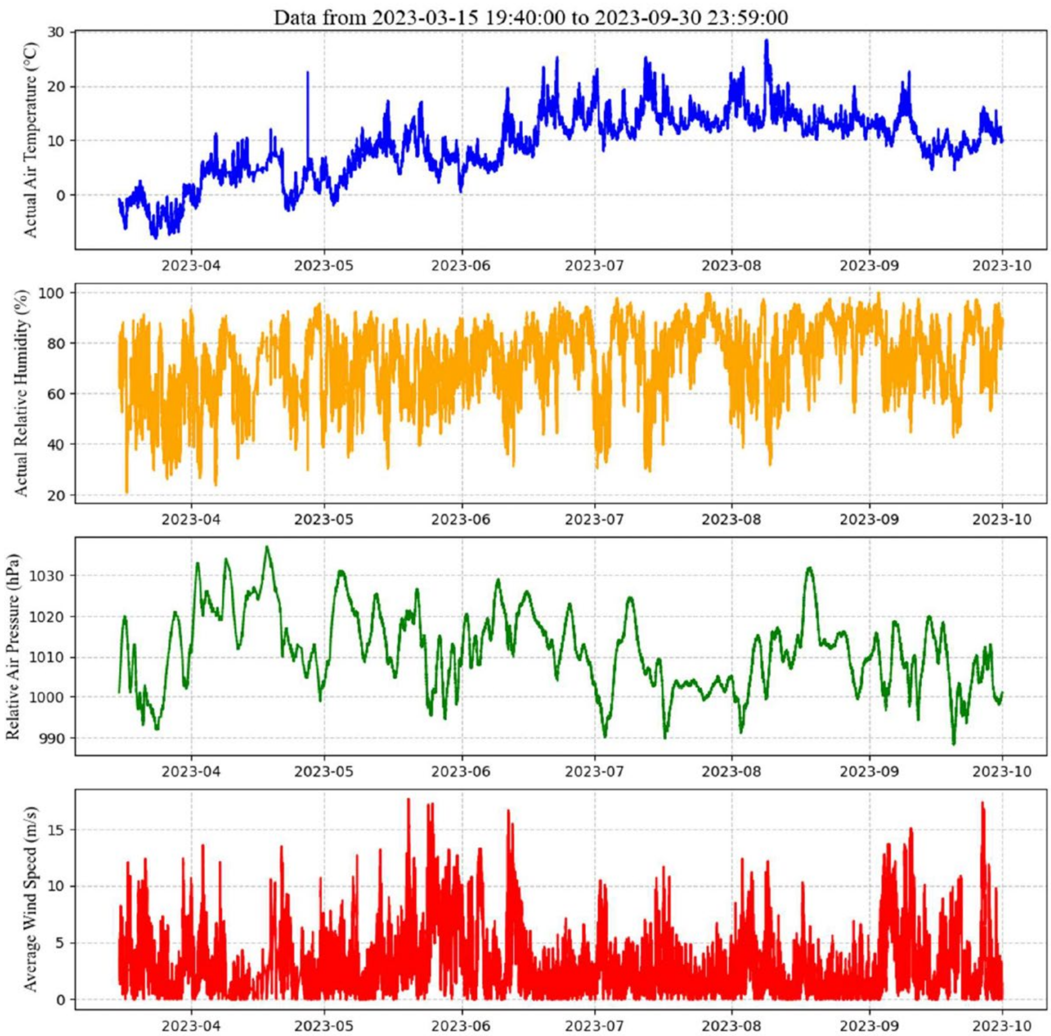


Fig 4: Dataplot for four variables from dataset- SPRICE_Norwegian_Maritime_Data.csv

You are expected to use correlation analysis, like the Pearson correlation coefficient, and additional statistical methods to identify relationships and potential hierarchies of your choice between features in this new dataset?

2.1 Bayesian Network Structure Learning

Based on your choice of variables for the new data provided, provide insight on what criteria you are using to determine the feature relationships and hierarchy between variables. Given a derived feature hierarchy, how can you employ Bayesian Network structure learning algorithms to construct a Bayes Net model? Explain the specific algorithm(s) you would select and justify your choice.

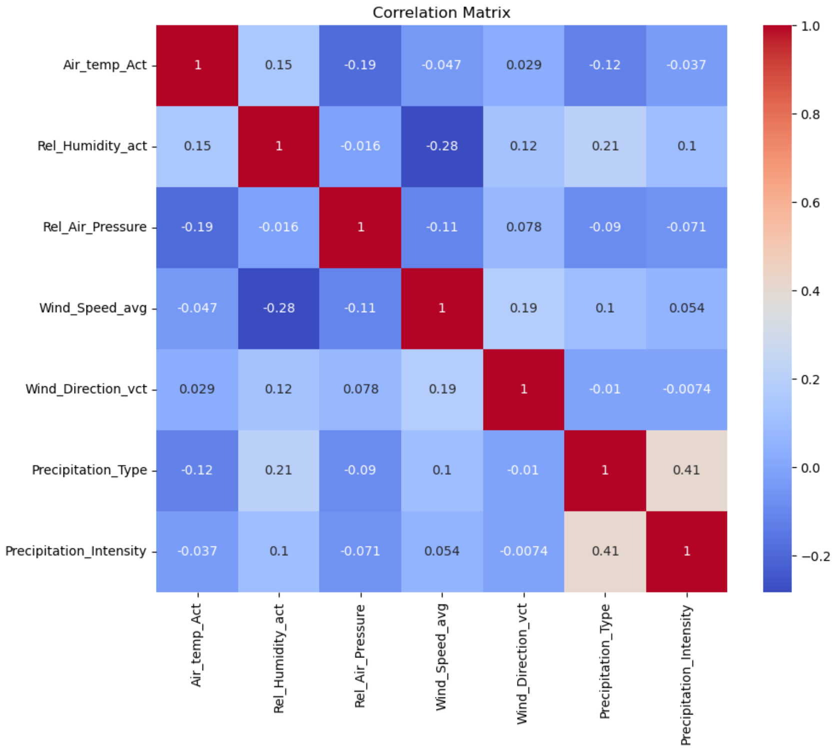


Fig 5: Sample correlation matrix for the important features in consideration.

2.2 Describe the techniques you would use to learn the conditional probability distribution (CPDs) for each possible combination of the features relationship for your Bayesian network.

2.3 Bayesian Inference and Analysis

With a constructed Bayesian network and learned parameters, perform similar inference Task 1.1, 1.2, and 1.3 for the new data with your choice of variables based on your constructed network you are analyzing.

Bonus Task:

3.1 How will you measure the accuracy or performance of your Bayesian network model? Outline appropriate metrics for evaluation.

3.2 Timestamps and Dynamic Bayesian Networks

Could you identify specific temporal patterns in the data using timestamp information? How could you incorporate these patterns into your dynamic Bayesian network structure?

For example, create nodes in your network to represent the weather variables at each time interval (e.g., Air_temp_Act_t, Rel_Humidity_act_t', 'Rel_Air_Pressure_t', 'Wind_Speed_avg_t). Establish conditional probability distributions (CPDs) to quantify the relationships between variables at different time points. For instance, the CPD for Wind_Speed_avg_t+1 might consider the current wind speed ('Wind_Speed_avg_t), humidity amount at the previous time step (Rel_Humidity_act_t-1), and historical temperature data (Air_temp_Act_t). This allows your network to adapt to changing weather patterns.

Possible Insights:

a. Understanding how temperature, humidity, pressure, and wind speed influence each other over time.
b. Identifying common temporal patterns and their potential drivers.
c. Can your dynamic Bayesian network be designed to detect significant changes in weather patterns over time?

3.3 Kalman Filtering with Noisy Measurements

Assuming the sensors measuring these variables have inherent measurement noise, how can you integrate Kalman filtering techniques to improve the estimation of the true values of air temperature, humidity, pressure, and wind speed over time?

Deliverables:

Students are expected to submit the following:

- Code implementations for each task and sub-task. The project can be divided into several Python files.

- A report that includes:
  - An overview of each method used and its relevance to the assignment.
  - A detailed analysis of the results obtained from each task, including insights and interpretations of the model predictions.