CISC520 Data Engineering and Mining

HW1: Cluster Analysis, Decision Tree Induction and naive Bayes Classification

In this homework assignment, you are going to use Cluster Analysis and Decision Tree induction algorithm on a weather forecast problem. It is a binary classification problem to predict whether or not a location will get rain the next day.

Information about the dataset (*Weather Forecast Training.csv*):

- Location: The location name of the weather station
- MinTemp: The minimum temperature in degrees celsius
- MaxTemp: The maximum temperature in degrees celsius
- Rainfall: The amount of rainfall recorded for the day in mm
- Evaporation: The so-called Class A pan evaporation (mm) in the 24 hours to 9am
- Sunshine: The number of hours of bright sunshine in the day.
- WindGustDir: The direction of the strongest wind gust in the 24 hours to midnight
- WindGustSpeed: The speed (km/h) of the strongest wind gust in the 24 hours to midnight
- WindDir: Direction of the wind
- WindSpeed: Wind speed (km/hr) averaged over 10 minutes
- Humidity: Humidity (percent)
- Pressure: Atmospheric pressure (hpa) reduced to mean sea level
- Cloud: Fraction of sky obscured by cloud This is measured in “oktas”, which are a unit of eigths. It records how many eigths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.
- Temp: Temperature (degrees C)
- RainTodayBoolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0
- **RainTomorrow: The target variable. Did it rain tomorrow?**

Organize your report using the following template (with the section breakdown and grading rubrics):

Section 1: Data preparation (30%)

- Discuss the potential data quality issues you identify about the dataset and how you apply various data preprecessing techniques to cope with those issues and perform Exploratory Data Analysis (EDA). Specifically discuss the type of techniques you carry out in order to prepare the dataset for the machine learning algorithms you use in the next section. Whenever appropriate, enhance your EDA with the effective data visualization.

Section 2: Build, tune and evaluate cluster analysis and decision tree models (50%)

- Apply both clustering algorithm (*kmeans* and *HAC*), *decision tree induction algorithm* and *naive Bayes classifier* to the weather forest training data and construct models. Perform extensive model experiments with hyper-parameters' tuning. Discuss your choice of hyper-parameters for each algorithm and produce tables summarizing the best performing models and their corresponding model specifications (i.e. the combination of hyper-parameters). Also explain your choice of model performance evaulation methods and metrics in order to produce unbiased and low variance estimates.

- In your analysis writeup, include the discussion regarding how to repurpose the unsupervised learning algorithms like clustering for classification and how to judge the performance of the algorithms.

- For decision tree induction algorithm model performance evaluation, generate Receiver Operating Characteristics (ROC) curve and calculate Area Under Curve (AUC) metric for the identified best performing model. Include a decision tree output visualization and interpret the decision tree model.

- Include detailed explanation of your modeling process and interpretation of the results in your analysis writeup (with markdown language) and structure such writeup in an easy-to-follow layout. Please limit your program output only to the most relevant part which is used to support your analysis. Excessive amount of less relevant outputs (e.g. display the whole dataset) in your report will have a negative effect on the grade.

Section 3: Prediction and interpretation (20%)

- After building the classification models, apply them to the test dataset (*Weather Forcast Testing.csv*) provided to predict if each location will rain tomorrow.
- Please submit your prediction results as a CSV file with four columns (ID, kmeans, HAC, DT) for the classfication results out of the pre-specified machine learning algorithms respectively.

Please submit your jupyter notebook together with knitted report (HTML format) and prediction CSV file.