1.

a. A machine learning approach called nearest neighbor classification seeks to identify previously unknown query items while differentiating two or more destination classes. It, like any other classifier, requires some training data with predetermined labels and is therefore an example of supervised learning. The label of the query object is inherited from the training set's nearest sample object. It is like saying “Birds of a feather flock together”.

We can use nearest neighbor classification since we have labeled data.

b.

Day (A,F) = 29.15

Day (B,F) = 25.495

Day (C,F) = 26.077

Day (D,F) = 29.496

Day (E,F) = 19.950

c. Nearest neighbour for day F is : E

Class is : Yes

1. 3 nearest neighbors are: E, B, C

Predicted class labels are for these 3 NN are: Yes, No, Yes

Therefore, predicted class label for 3-NN is: Yes

2.

a. Overfitting is a term used in data science to describe when a statistical model fits its training data perfectly. When this happens, the algorithm is unable to perform accurately against unknown data, negating the goal of the method.

b. Measuring the accuracy does not provide information about the model's performance when applied to new data. Overfitting can occasionally result in a better output, but if the model is overfitted for fresh data, the outcome will be quite inaccurate. As a result, simply testing the accuracy of the models on the training dataset would not be sufficient to evaluate the performance of nearest neighbour classifiers.

c. During the algorithm's training, a portion of the dataset is "held out set." It's used to see how well a classification model performs with new data. If a set of data is held back during the performance test, it will aid in determining the model's true performance.

d. In this scenario, we have weather data, and we can choose the best K value. However, in a real-world circumstance, we must predict because we do not have the data to look at initially. As a result, it's possible that the model's genuine accuracy will be overstated on future days.

1. A validation set is a type of "Held Out" data that is utilized to fine-tune the hyper-performance, parameter's such as the value of K. To do so, we must train the model on a variety of K values and then evaluate it on the validation set.
2. We can use clustering to understand the data from days when universities were open in a certain location. We can then predict the weather based on the data in a particular location. This allows us to alert students to any type of inclement weather so that they can take appropriate precautions.