What is the average solar energy per square meter at this location?

To choose a good location for deploying solar panels, a company would like to classify locations according to their received average hourly solar energy per unit area. However, the company only has history data from some locations, and needs to classify new locations. The company uses the following classification according to the average hourly extraterrestrial radiation (ETR), i.e., energy at top-of-atmosphere:

- Class 1: ETR below 320 Wh/sqm;
- Class 2: ETR above 320 Wh/sqm and below 340 Wh/sqm;
- Class 3: ETR above 340 Wh/sqm.

out comes

You are hired to design a system which can determine the class of new locations for which the average hourly ETR is unknown, based on the above classification. You are given the following data to develop the system:

1) Location (longitude and latitude)
2) Average hourly solar energy per square meter throughout the year 2005 (Extraterrestrial radiation (ETR), i.e., energy at top-of-atmosphere)

inputs

The data can be found in the file 'Solar_energy' and is for 925 locations in the USA. The company will reveal their measurements for Canadian data only once you demonstrate the proof-of-concept. To this end, you are required to:

1) Split the data set into two parts. The first part will be used for training, and should consist of 600 data samples. The second part contains the remaining samples and should be used for testing purposes. You should make sure that the training set spans all (or most) of the range of longitudes or latitudes of the data set.
2) Use the training data to train a classifier, which classifies the locations into the 3 classes above based on their location.
3) Use the trained classifier to predict the classes of the locations of the samples in the test set.
4) Compare the results with the ground-truth (actual class obtained from the file); and calculate the number of errors.
5) The company then asks you to run a similar classification using average hourly METSTAT data (METeorological/STATistical solar irradiance per square meter). The METSTAT models the irradiance after taking the effects of clouds and precipitation into account. The company now defines the three classes using the intervals [0,60), [60,70),[70,infinity). Train a new classifier for the METSTAT data and test it similar to above. Then calculate the number of errors
6) The performance of the METSTAT classifier is expected to be worse than the ETR classifier. Explain to the company why this is the case, and what can be done to improve METSTAT classification

Data sources:

1) NSRDB: National Solar Radiation Database, https://nsrdb.nrel.gov/data-sets/archives.html

Acknowledgement:
This project is developed in collaboration with Dr. Morad Abdelaziz, whose research focus is on smart grids, power systems operation and decarbonisation, electrification of the transportation system, electric vehicles charging infrastructure, and high performance computing.
https://engineering.ok.ubc.ca/about/contact/morad-abdelaziz/