Rensselaer

why not change the world?®

Lab 03/ Assignment 2: Exploratory data analysis: examining distributions, linear models, classification and clustering

Ahmed Eleish

ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960 Group 1, Lab 3, October 4th, 2024

Tetherless World Constellation
Rensselaer Polytechnic Institute

Lab 03 / Assignment 2

Files:
https://rpi.box.com/s/auda1sxsniw1f37hfp4jcx9w6js7xwxd

Variable Distributions (2%)

Using the EPI results dataset to perform the following:

1. Derive 2 subsets each for a particular region

    1.1. Plot histograms for a variable of your choice for both regions with density lines overlayed

    1.2. Plot QQ plots for both variables compared to known probability distributions.

Linear Models (3%)

Using the EPI results dataset to perform the following:

2. Fit linear models as follows:

2.1. Choose a subset of 5 variables (excluding EPI) and using the formula EPI~VAR1+VAR2+VAR3+ VAR4+VAR5, fit a linear model and identify which variable most significantly influences EPI. Plot that variable with another and overlay the fitted line.

2.2. Repeat the previous model with a subset of 1 region and in 1-2 sentences explain which model is a better fit and why you think that is the case.

Classification (kNN) (3%)

Using the EPI results dataset to perform the following:

3. Train 2 kNN models using ”region” as the class label as follows:

3.1. Choose a subset of 5 variables and filter the subset by region keeping 3 regions out of 8 (representing 3 classes), then train a kNN model to predict the region based on these variables. Evaluate the model using a contingency matrix and calculate the accuracy of correct classifications.

3.2. Repeat the previous model with the same variables for another set of 3 other regions and evaluate. In 1-2 sentences explain which model is better and why you think that is the case.

Clustering (3%)

Using the EPI results dataset to perform the following:

1. Fit a k-means model for a subset of 5 variables for 2 different groups of regions (3 each)

    1.1. Compare the performance of the models using their within cluster sum of squares.

    1.2. In a loop fit kmeans models for both subsets using multiple values of k. Plot WCSS across k values. In 1-2 sentences explain which model is better and why you think that is the case.

Thanks!
Have a great weekend!*

* Work on your assignments!!