Assignment #3

Monday, September 25, 2023 4:17 PM

11260658

Data Mining

Assignment Three — **10 points**

Due: Thursday, October 5 @ 11:50 PM

Modify Tutorial_4_Data_Preprocessing.ipynb to do a few steps with the attached university salary and faculty size data set (aaup.csv).

Here is a description of the variables in the dataset.

Univ_id: id number
Univ_name: Name of institution
State: 2 letter state code
Type: (I, IIA, or IIB)
fp_sal: Average salary - full professors
ac_sal: Average salary - associate professors
at_sal: Average salary - assistant professors
to_sal: Average salary - all ranks
fp_com: Average compensation - full professors
ac_com: Average compensation - associate professors
at_com: Average compensation - assistant professors
to_com: Average compensation - all ranks
fp_#: Number of full professors
ac_#: Number of associate professors
at_#: Number of assistant professors
in_#: Number of instructors
to_#: Number of faculty - all ranks

Tasks

Replicate the preprocessing steps applied to the breast cancer example as guided below:

1. Input the data into a Pandas dataframe; create the data columns of your choice; print the number of observations and attributes.

2. Recode the missing values to NaN. This dataset uses *. Print the counts of missing values across the attributes.

3. How do you handle missing values in this dataset? Explain your selection. (put your answer in the box below)

The missing values in the dataset are not considered and therefore dropped. By doing so, the attributes we are analyzing prov ide some numeric attribute that would be helpful in finding meaningful data within the dataset. I think using median did not feel appropriate for NaN in this circumstance. Because the q uestion is implying over the entire dataset and not particular attributes … but if that were the case, then, particular attributes could benefit from using the median, such as the salary o r compensation. But, the compensation or salary varies based on what state the institution is in, considering their own tax laws among other things. Therefore, grouping the data to their respect ive state before applying any techniques to the NaN values would be smart to do, as well as avoid using the mean because there were outliers present in every attribute, which would misrepresent the data. The techniques is dependent upon what attributes we want to analyze.

4. Explore for outliers. Apply the boxplot display. Are there any outliers? (put responses in the box below)

Yes, there are outliers present in the dataset across all attributes available. The outliers are beyond the upper bound within a boxplot.

5. Are there any duplicate records? No, there are not any duplicate records present in the dataset.

6. Can you aggregate the institutions within each state using the grouping operation from Pandas? So you should end up with ~50 observations. Which statistics are you aggregating on? (describe in the box below)

Yes, you can aggregate the number of institutions within each state. After doing so, I computed the mean to find the average salary or compensation, in each of the respective attributes that were selected.

7. Explore some sampling from the original data set, not the aggregate. What did you find to be the best? Why? (put answer in the box below)

Considering the dataset's characteristics, the most suitable sampling method would be fractional sampling without replacement . This approach involves randomly selecting a certain percentage of rows from the dataset for analysis. Since the dataset comprises 1100 rows, which is not con sidered large in the machine learning realm, and practically most devices can read that much data from a file effortlessly, this method is the best for our given dataset. We also know that our dataset does not contain any duplicate rows, but that does not play much of an influence in deciding what sampling method to use. However, it's worth noting that the idea of using sampling with replacement could be considered in specific scenarios where you want to simulate a larger dataset by allowing rows to be obse rved more than once. Sampling with replacement might be helpful if we find that our dataset is not large, as it intends to 'increase the sample size' by observi ng data points more than once.

8. Pick a salary column to discretize and pick a count to discretize. Why did you choose your type of discretization? (put in the box below)

I decided to apply two discretization methods to the 'fp_sal' and 'fp_#' columns: equal width and equal frequency. I used equ al width discretization to create bins with consistent salary ranges, showing uniformity in the width of each category. I also experimented with equal frequency dis cretization to distribute data points more evenly across bins, making it a valuable for identifying groups with balanced representation. Equal width prioritizes un iformity and equal frequency makes sure that each category contains a similar number of data points. My decision to use one method over the other depended on wh ether I wanted consistent range widths or balanced representation in my analysis. If I had to choose, I would use equal width discretization as it provides a simpler and more intuitive way to understand salary distributions by creating consistent bins with fixed ranges.

Submission: Submit both the. ipynb file and the word document.

Good luck!