

ADS Assignment 7.

1. Import the data set, 'insurance'. The column 'charges' should be considered as your target label.
2. Explore the data using at least 3 data exploratory tools of your choosing in pandas and interpret your observation in a markdown cell of what form of predictive analysis that can be conducted on the data.
3. Visualize the age distribution for the column 'age' and comment on the results in a markdown cell as well. (Ensure your visualization is of an appropriate size for effective analysis)
4. Isolate all the continuous and discrete columns into their respective lists named 'numerical_continuous' and 'numerical_discrete' respectively.
5. Visually identify if there is presence of any outliers in the numerical_continuous columns and resolve them using a zscore test and a threshold of your choosing.
6. Validate that your analysis above was successful by visualizing the value distribution in the resulting columns using an appropriate visualization method.
7. Isolate all the categorical column names into a list named 'categorical'.
8. Visually identify the outliers in the discrete and categorical features and resolve them using the combined rare levels method.
9. Encode the discrete and categorical features with one of the measures of central tendency of your choosing.
10. Separate your features from the target appropriately. Narrow down the number of features to 5 using the most appropriate and accurate method. Which feature had to be dropped and what inference would you give as the main contributor of dropping the given feature.