2.2

Data Discretization Strategy:

From the observation of the data, we classified the features into 3 categories, which are “Exact Ordered Label”, “Estimate Ordered Label” and “Unordered Label”. For the “Exact Ordered Label”, the exact value of the feature is given, and the greatness of these values can be compared; there is only one feature in this dataset belonging to this type, that is “YOB”. For the “Estimate Ordered Label”, its value can be compared but the values these features give are the estimate one instead of exact one; for examples, “Income” and “EducationLevel” are such labels. For the “Unordered Label”, it does not have greatness so we treat it just as labels; The remaining features are all “Unordered Labels”.

2.3:

Strategy: As mentioned in 2.2, we have three kinds of features. When filling missing data, we treat different types of features with different strategies. For the “Exact Ordered Label”, we adopt the “mean” strategy to fill the missing data; for the “Estimate Ordered Label”, we choose “median” strategy; for “Unordered Label”, we select “most frequent one” strategy.

For the case that “isClassified == False”, we calculate the “mean”, “median” or “most frequent label” according the strategy we choose from the samples that have values, and fill it to all the corresponding feature space of the sample with missing data.

Different from the above one, for “isClassified == True”, we use the decision tree given by the specification to cluster data, and use statistic values of the corresponding groups to replace the missing data. For samples that have missing “Gender” or “EducationLevel” data, we simply drop them.

We have considered to build the decision tree using the “greedy” strategy. However, there are so many missing data for each of the feature, which makes calculating the entropy more complicated. Thus, we do not adopt this strategy. Another possibility is to find the nearest neighbor of the sample with miss data, and fill this miss data with the corresponding value of its nearest neighbor. However, there are miss data for any sample of any features, which makes calculating the distance between samples a tough task. As a result, we adopt the decision tree given in the specification finally.