**Statistical Learning Lab**

**End Semester Examination Spring 2024**

**Time: 3 hrs**

**Full marks: 60**

***Answer all questions***

***Submit a report with your code and output with proper explanation***

For the dataset given, the response variable (*Y*) is the willingness to pay (WTP) for the gamble (Definition: The game where agents get to pick an envelope is termed as the gamble/lottery). The experimenter collects 92 observations with the following information in each observation. During the experiment, the subject makes a choice of an envelope having various levels of risky proposals which offer monetary rewards subject to probability (1 being the least risky envelope and 6 being the most risky envelope).

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| Age | Age of the respondent |
| Gender | Gender of the respondent |
| Family Type | Type of family the respondent lives in (e.g., Nuclear, Extended, Single-parent, etc.) |
| Total Siblings | Total number of siblings in the respondent's family, including the respondent themselves |
| Position Among Siblings | Position of the respondent among their siblings (e.g., 1st, 2nd, 3rd, etc.) |
| Number of Sisters | Number of sisters the respondent has |
| Number of Brothers | Number of brothers the respondent has |
| Area Type | Description of the area the respondent lives in (e.g., Urban, Suburban, Rural) |
| Home State | State in which the respondent's home is located |
| Educational Institute | Name of the current educational institute attended by the respondent |
| Highest Education Attained | Highest level of education attained by the respondent (e.g., High School, Bachelor's, Master's) |
| Major/Field of Study | Area of study or major pursued by the respondent |
| Father's Education Level | Highest level of education attained by the respondent's father/guardian |
| Mother's Education Level | Highest level of education attained by the respondent's mother/guardian |
| Father's Employment Status | Employment status of the respondent's father/guardian (e.g., Employed, Unemployed, Retired) |
| Mother's Employment Status | Employment status of the respondent's mother/guardian (e.g., Employed, Unemployed, Homemaker) |
| Annual Family Income (in INR) | Annual family income in lakhs of rupees |
| Housing Situation | Current housing situation of the respondent's family (e.g., Own house, Rented house, etc.) |
| Chosen Envelope | Envelope chosen by the respondent based on the given data |
| **Response variable** |  |
| Maximum Amount (in INR) | Maximum amount in Indian Rupees that the respondent would choose |

1. On the data do the following [20 marks]
   1. Some of the responses corresponding to Annual income is not specified in Lakhs. Convert the same into lakhs of rupees. Also, assume that education variable can be converted to numerical data by taking (12th sd=12 year, diploma=14 year, graduate=15, postgraduate=17, doctoral degree=25 years, high school or lesser=10). Change the dataset for educational variables to quantitative type.
   2. Do a preliminary data processing on the dataset and remove outliers?
   3. Perform exploratory data analysis to depict how willingness to pay is associated with the other variables?
2. Find the significant factors affecting someone’s Willingness to pay (*Y*). Write an inference about how these factors influence the response and interpret the same. Perform diagnostics and model adequacy checking and comment on the results. [20 marks]
3. The WTP greater than or equal to 1000 is considered high and that less than 1000 is considered low. Divide the response variable into two categories High (>=1000) and Low (<1000). Divide the dataset in 80% train and 20% test using random sampling. Use SVM, Decision tree, and Logistic regression and compare the performance of these models on the test data (using confusion matrix, AIC, BIC, F1 score) and comment on the results.

[20 marks]