SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Intro to Big Data Science: Project

Due Date: May 28, 2021

✏ **Submission Policy**

This project should be finished in groups. Each group should consist of one or two students. You can find your partners by yourselves. Each group should contribute a final report in the submission.

Your report should be in PDF format. It is suggested to use jupyter notebook to include both your code and your main text, and convert it to PDF file. Your main text should include the following several aspects:

- Background introduction
- Data exploration: data statistics and data visualization;
- Data preprocessing: detecting missing values and outlier samples (if any), data discretization, concatenation, and normalization (if necessary), etc.;
- Model construction: you could use any model you prefer, even the model we did not cover in class;
- Feature selection and model selection if necessary;
- Model evaluation;
- Conclusion.

DO NOT just submit the code file. Necessary statements, analysis, formula, figures, and tables should be included in your report. You should also have a complete set of codes. Your report (typically in pdf format) and codes should be compressed in a zip file. Please use your student ID and name to rename your zip file, e.g., “11600000_ZhangSan”. Then the zip file shall be uploaded to BlackBoard system.

Your project will be graded based on several factors, including the accuracy (e.g., F_1 score, R^2 score, etc.), comparison of different methods, whether you have innovative ideas, the quality of your report, the analysis you made based on you results (e.g., computational efficiency, model interpretability, etc.), and the quality of your codes, but not limited to these.

Last but not the least, we hope you really get familiar with whole data science procedure and discover the new world of your own.

Now enjoy your journey of data science!

PS: No presentation is involved in the project, and both Chinese and English are allowed in your submission files.