Assignment

This assignment consists of the two parts. Team work, each team has 4-5 students. All students must be involved in programming. The maximum score for each task is 100. You must submit report (pdf) file containing your project and txt file with url to your GitHub repository. The defense is obligatory. During defense any team member maybe asked questions related to any part of the project and the topics. The defense a group call involving all team members. Student who refused to defend gets 0.

Dataset Description:

- types.csv - reference of transaction types
- codes.csv - reference of transaction codes
- transactions.csv - transactional data on banking operations
- train_set.csv - training set with client gender marking (0/1 - client gender)
- test_set.csv - no need to use.

transactions.csv columns description:

- client_id - client is id
- datetime -transaction date (format - ordered day number hh:mm:ss - 421 06:33:15)
- code - transaction code
- type - transaction type
- sum - sum of transaction

TASKS: WARNING! Here is the written Minimum evaluation criteria.

Assignment (part 1, Unsupervised, 25%):

I. Explore the dataset. Do the descriptive statistics.
II. Explanatory data analysis. Exploring the features, visualizations etc. (https://www.kaggle.com/learn/data-visualization, https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15, https://www.mastersindatascience.org/learning/what-is-exploratory-data-analysis/)
III. Feature engineering. Encodings, generating the features from date-time, sum and from other columns. (https://www.kaggle.com/learn/feature-engineering, https://www.kaggle.com/learn/data-cleaning)
IV. Unsupervised learning. Do the Cluster analysis. Segment the customers. K-means, Hierarchical Clustering. With different metrics, linkages. Visualize the clusters etc. Look for the optimal number of the clusters
V. Analyzing the results.
VI. Conclusion.

Assignment 2 (part 2, Supervised, 25%):

I. Explore the dataset. Do the descriptive statistics.

II. Explanatory data analysis. Exploring the features, visualizations etc. (https://www.kaggle.com/learn/data-visualization, https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15, https://www.mastersindatascience.org/learning/what-is-exploratory-data-analysis/)
III. Feature engineering. Encodings, generating the features from date-time, sum and from other columns. (https://www.kaggle.com/learn/feature-engineering, https://www.kaggle.com/learn/data-cleaning)
IV. Supervised learning. Build model for prediction the gender of the clients. Decision Trees, KNN, Random Forest. Tune the hyper parameters, grid search, cross validation etc. Visualization of the models etc..
V. Analyze models, Result comparison, ROC/AUC, precision and recall curves, deep analyzing.
VI. Conclusion.

GRADING:

90-100
- Work would be worthy of further dissemination under appropriate conditions
- Mastery of advanced methods and techniques at a level beyond that explicitly taught
- Ability to synthesize and deploy in an original way idea from across the subject

80-89
- Excellent range and depth of attainment of intended outcomes
- Mastery of a wide range of methods and techniques
- Evidence of study and originality of what has been taught

70-79
- Attained all the intended learning outcomes for a unit
- Able to use well a range of methods and techniques to come to conclusions

60-69
- Some limitations in attainment of learning objectives, but has managed to grasp most of them
- Able to use most of the methods and techniques taught
- Evidence of study and comprehension of what has been has been taught but grasp insecure
- Some grasp of the issues and concepts underlying the techniques and material taught, but weak and incomplete

50-59
- Attainment of only a minority of the learning outcomes
- Able to demonstrate a clear but limited use of some of the basic methods and techniques taught
- Weak and incomplete grasp o1'what has been taught
- Deficient understanding of the issues and concepts underlying the techniques and material taught

25-49
- Attainment of nearly all the intended learning outcomes deficient
- Lack of ability to use at all or the light methods and techniques taught
- Inadequately and incoherently presented

- Wholly deficient grasp of what has been taught
- Lack of understanding of the issues and concepts underlying the techniques and material taught