WESTMINSTER
INTERNATIONAL UNIVERSITY IN TASHKENT
An Accredited Institution of the University of Westminster (UK)

Module name and code	**6COSC017C-n, Machine Learning and Data Analytics**
CW weighting	50%
Lecturer setting the task with contact details and office hours	Dilshod Ibragimov, dibragimov@wiut.uz Office hours: Wednesday 16:00 – 17:00
Submission deadline	03/12/2021
Results date and type of feedback	28/12/2021, Written
The CW checks the following learning outcomes:	
1. Critically justify the use of data mining and machine learning techniques for Data Science applications. 2. Critically reflect on how different data mining and machine learning algorithms operate and their underlying design assumptions and biases in order to select and apply an appropriate algorithm to solve a given problem. 3. Implement, encode and test data mining/machine learning projects, focused on problem analysis, data pre-processing, data post-processing by choosing/implementing appropriate algorithms. 4. Critically analyze the output of data mining and machine learning algorithms by drawing technically appropriate and justifiable conclusions resulting from the application of data mining and machine learning algorithms to real-world data sets. 5. Perform critical evaluation of performance metrics for data mining and machine learning algorithms for a given domain/application.	

Introduction

The aim of this coursework is to develop a complete data analysis solution for the given dataset using several machine learning algorithms. You should prepare the dataset for the analysis and apply appropriate machine learning algorithms. You should be able to critically evaluate the performance and output of your models using appropriate metrics.
The datasets for the analysis can be found at http://datahub.io, http://ec.europa.eu/eurostat/, https://github.com/caesar0301/awesome-public-datasets, and many other websites. You may find other ideas through the links below:

1. https://www.opensciencedatacloud.org/publicdata/
2. http://www.kdnuggets.com/datasets/index.html
3. http://datascience.berkeley.edu/open-data-sets/
4. http://www.datasciencecentral.com/profiles/blogs/big-data-sets-available-for-free
5. http://www.datasciencecentral.com/profiles/blogs/great-github-list-of-public-data-sets
6. http://www.datascienceweekly.org/data-science-resources/data-science-datasets
7. http://www.statsci.org/datasets.html
8. http://blog.bigml.com/2013/02/28/data-data-data-thousands-of-public-data-sources/
9. https://wrds-web.wharton.upenn.edu/wrds/index.cfm

You can use other sources to find the dataset. The dataset of your choice should be publicly available. The dataset should have sufficient number of records and variables for the analysis.

Attention: To prevent any two persons working on the same dataset you should **post (!)** the *name* of the topic of your dataset and the *URL address* where you downloaded the dataset from to the webpage of the module in Discussions session: (https://intranet.wiut.uz/LearningMaterial/Discussion/Details/1966?moduleId=610). If another student selects the dataset of your choice (and the corresponding post about it already exists) you are ***not allowed*** to work on the same dataset. If more than one student submit a report over the same dataset, the first person who selected the dataset will receive full marks, others will be ***reported to the Academic Misconduct Panel***. You need to note, you will be asked to present your work during the viva voce session. In case the solution similar to yours will be found on Internet, you will get 0 for the whole coursework. Also note that there is a list of ***banned*** (not allowed) to the selection datasets. You ***cannot*** select a dataset to work on from this list. The list of banned datasets is given in the end of the coursework description.

Deliverables

1. Report

A. **Introduction**. Here you should consider the business case – you have to describe the purpose and origin of the dataset and provide the links where you downloaded it from. You have to describe what kind of analysis is suitable for the given dataset and what you are trying to investigate.

B. **Description of the Exploratory Data Analysis.** Perform initial investigation of your data. Give a detailed description of the dataset (dataset shape, observations, characteristics, data types, variables correlation). You also need to describe your dataset with measures of central tendency (mean, variance, standard deviation). You may use graphs to illustrate your analysis.

C. **Dataset preparation: Data preprocessing and Feature engineering.**
In this section you need to describe all the data preprocessing that was applied to your dataset. This may include cleaning (removing impossible data combinations, editing, standardizing), instance selection, normalization, transformation.

D. **In this section you need to justify the choice of Machine Learning algorithms.** You need to specify what type of ML algorithms you selected (Supervised/Unsupervised) and compare several algorithms. You need to specify and explain all the hyper-parameters of your model, if applied. You need to explain the use of metrics that you used to evaluate your models.

For example, below is the summary table for classification problem. Several classification algorithms (Decision Tree, Logistic Regression, K-NN, Naïve Bayes) are selected, five measures (TPR, TNR, AUC, Recall, Precision) are used to compare these algorithms.

Algorithm	TPR	TNR	AUC	Recall	Precision
Decision Tree					
Logistic Regression classifier					
K-Nearest Neighbors					
Naïve Bayes					

E. **Conclusion**. In this section you need to present the results of your findings.

2. Practical part

You need to work in Jupyter Notebook and use Pandas, Scikit-learn libraries. Your notebook should be organized logically, use appropriate headings. Comments should be provided to explain your code.

A. Load dataset. The quantity and quality of your data dictate how accurate your model will be. Select your dataset thoroughly and load it to the dataframe. You can use one or several sources.

B. Exploratory Data Analysis. Provide the summary statistics for your dataset.

C. Data Preparation. You need to prepare your data for training. You may need to clean the data, remove duplicates, correct errors, deal with missing values, normalization and data types conversions. You may also engineer features during this step, depending on the needs of your analysis. The dataset should be split into the training and evaluation sets.

D. Models training. You need to train several machine learning algorithms.

E. **Models Evaluation**. You need to compare several machine learning algorithms. Provide evaluation metrics to compare your model. Use test set for evaluation. You need to provide classification metrics and confusion matrix to compare models.

Intranet Submission

You have to submit an electronic version of your work on intranet.wiut.uz. Under the 'Lectures and Seminars' section find the 'Courseworks and Assignments' section. Select the Machine Learning and Data Analytics module from the list (https://intranet.wiut.uz/Coursework/ViewCourseWorks?moduleID=517).
Upload the zipped file containing your report, Jupyter notebook with your solution and dataset. Name your report file according to the following pattern:

MLDA.CW1.IDnumber.zip

Example: **MLDA.CW1.5678.zip**

Do not include leading zeroes in your Id.

Format

1. Word-processed Times New Roman/ Arial 12, single-spaced and printed single-sided on A4 paper.
2. The cover sheet should state your ID number, module title and marker's name.
3. Include a contents page giving the headings and page numbers of each section.
4. Pages should be numbered.
5. Please do not submit any loose pages.
6. Use Harvard method of referencing.

General notes:

Please ensure that you work individually on this coursework. Plagiarism and close collaboration will not be tolerated and it will be considered an assessment offence. According to Essential Information Handbook of Academic Regulations, any students may be invited for oral viva. (Please, see the regulations for full details)
Check that the CW has a standard cover page, table of contents, page numbers and bibliography. Your name should not appear on the cover page or anywhere else. Put your ID number on the cover page and on every other page.
This is your responsibility to put CW through the anti-plagiarism software before submission.

Banned Datasets

1. http://archive.ics.uci.edu/ml/datasets/Bank+Marketing
2. https://www.kaggle.com/aparnashastry/building-permit-applications-data
3. https://www.kaggle.com/ishandutta/early-stage-diabetes-risk-prediction-dataset
4. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

5. https://catalog.data.gov/dataset/energy-and-water-data-disclosure-for-local-law-84-2017-data-for-calendar-year-2016
6. https://www.kaggle.com/tarunpaparaju/apple-aapl-historical-stock-data
7. https://www.kaggle.com/arpina/passenger-satisfaction
8. https://www.kaggle.com/anthonypino/melbourne-housing-market
9. https://www.kaggle.com/doaaalsenani/usa-cers-dataset
10. https://www.kaggle.com/dinhanhx/studentgradepassorfailprediction
11. https://www.kaggle.com/ishansingh88/europe-hotel-satisfaction-score
12. https://www.kaggle.com/osmi/mental-health-in-tech-survey
13. https://www.kaggle.com/imdevskp/corona-virus-report?select=country_wise_latest.csv
14. https://www.kaggle.com/joshmcadams/oranges-vs-grapefruit
15. https://www.kaggle.com/rohanchreddy/advertsuccess
16. https://www.kaggle.com/aayushmishra1512/netflix-stock-data
17. https://www.kaggle.com/rtatman/chocolate-bar-ratings
18. https://www.kaggle.com/haithemhermessi/breast-cancer-screening-data-set
19. https://www.kaggle.com/loveall/clicks-conversion-tracking
20. https://www.kaggle.com/jsphyg/weather-dataset-rattle-package
21. https://www.kaggle.com/janiobachmann/bank-marketing-dataset
22. https://www.kaggle.com/shaijudatascience/loan-prediction-practice-av-competition
23. https://www.kaggle.com/barun2104/telecom-churn
24. https://www.kaggle.com/ajay1735/hmeq-data
25. https://www.kaggle.com/zhiruo19/covid19-symptoms-classification
26. https://www.kaggle.com/ronitf/heart-disease-uci
27. https://www.kaggle.com/ruiqurm/lianjia
28. https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe
29. https://www.kaggle.com/sonujha090/insurance-prediction
30. https://www.kaggle.com/uciml/pm25-data-for-five-chinese-cities
31. https://www.kaggle.com/uciml/mushroom-classification
32. https://www.kaggle.com/uciml/glass
33. https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants
34. https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction
35. https://www.kaggle.com/shivam2503/diamonds
36. https://www.kaggle.com/sagnikpatra/edadata
37. https://archive.ics.uci.edu/ml/datasets/Electrical+Grid+Stability+Simulated+Data+
38. https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset
39. https://www.kaggle.com/sakshigoyal7/credit-card-customers
40. https://www.kaggle.com/jessemostipak/hotel-booking-demand
41. https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction
42. https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016
43. https://www.kaggle.com/uciml/student-alcohol-consumption?select=student-mat.csv
44. https://www.kaggle.com/sriharipramod/bank-loan-classification
45. https://www.kaggle.com/shubh0799/churn-modelling
46. https://www.kaggle.com/aiaiaidavid/cardio-data-dv13032020
47. https://www.kaggle.com/shantanuss/banknote-authentication-uci
48. https://www.kaggle.com/nikdavis/steam-store-games
49. https://www.kaggle.com/spscientist/students-performance-in-exams
50. https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks

Assessment Criteria

Component	Mark
1. Report	**40**
A. Introduction - Full description of the dataset and its purpose should be given	5
B. Exploratory Data Analysis - Statistics with appropriate visualizations to describe the dataset are provided	5
C. Dataset Preparation - Description of data preprocessing steps	5
D. Model selection - Algorithms discussion (5) - Metrics discussion (5) - Comparison of all models using appropriate metrics (5)	15
E. Conclusion - Data analysis findings (5) - Interpretation of results (5)	10
2. Practical Part	**60**
A. Data load - Data should be loaded from one or several sources; In case the data is loaded from several sources joins should be applied correctly.	5
B. Exploratory Data Analysis - Statistical summary data (5) - Correlation matrix (5) - Other graphs (scatter plots/box plots/histograms/etc.) (5)	15
C. Data Preparation Data should be cleaned and processed for the analysis, the following has to be addressed: - missing values (3) - scaling (3) - correcting error data (4) - feature engineering (10) - dataset should be split into training and evaluation sets; Test set should be isolated (5)	25
D. Model training - Three models at least should be used. (6) - Hyperparameters should be used when appropriate. (4)	10
E. Model evaluation - Classification metrics (5) - Confusion matrix (5)	10