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	Mukhammadmuso Abduzhabbarov, mabduzhabbarov@wiut.uz, Office hours: TBA
Submission deadline	Dec 5, 2025
Results date and type of feedback	Dec 22, 2025, 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 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 design and deliver an end-to-end data analysis and machine learning solution. Beyond a written report and a reproducible Jupyter workflow, you must also deploy an interactive Streamlit application that demonstrates your data pipeline, models, and evaluation. Your work should evidence critical reasoning at each step (choices, trade-offs, assumptions) and connect technical results to a clear business or scientific question.

Datasets

You may select any publicly available dataset with sufficient observations and features for meaningful analysis. Potential sources include: datahub.io, Eurostat, and curated lists of public datasets.

https://www.opensciencedatacloud.org/publicdata/

http://www.kdnuggets.com/datasets/index.html

http://datascience.berkeley.edu/open-data-sets/

http://www.datasciencecentral.com/profiles/blogs/big-data-sets-available-for-free

http://www.datasciencecentral.com/profiles/blogs/great-github-list-of-public-data-sets

http://www.datascienceweekly.org/data-science-resources/data-science-datasets

http://www.statsci.org/datasets.html

http://blog.bigml.com/2013/02/28/data-data-data-thousands-of-public-data-sources/

https://wrds-web.wharton.upenn.edu/wrds/index.cfm

Attention:

To prevent any two persons from 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 the Discussions session: (*will be provided later*). 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 submits 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 your solution is similar to any one can be found on the internet; you will get a 0 mark for the whole coursework. Also, note that there is a list of ***banned*** (not allowed) for the of selected datasets. You ***cannot*** select a dataset to work on from this list. The list of banned datasets is given at the end of the coursework description.

Deliverables

1. **Report**

You are required to provide a comprehensive report detailing each phase of your project. Your report should include the following sections:

A. **Introduction** (business/scientific case, problem framing, dataset origin and licensing)**:**

Here you should consider the business case – you should describe the purpose and origin of the dataset and provide the links where you downloaded it from. You should also 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** (shape, types, distributions, correlations, summary statistics; justified visualizations)**:**

Perform an 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. **For each step of the Exploratory Data Analysis, you should include a clear justification for its importance and what insights it provides about the data in the report.**

C. **Dataset preparation** (Data preprocessing: cleaning, missing values, outliers, normalization/encoding and Feature engineering)**:**

In this section, you need to describe all the data preprocessing that was applied to your dataset. This may include cleaning (e.g., removing duplicates, outliers, impossible data combinations, filling in missing values), instance selection, normalization, encoding, and transformation. **For each step of data preprocessing, you should include a rationale for its importance and a clear justification for your choice of specific approaches or methods in the report.**

D. **Justification of the choice of Machine Learning algorithms** (at least 3 algorithms; hyperparameter choices; validation and metrics; model selection rationale)**:**

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. **You should include a clear justification for your choices regarding model selection, evaluation metrics, and the final model in your report.**

Example:
The table below shows the summary for a classification problem. Several classification algorithms (Decision Tree, Logistic Regression, K-NN, Naïve Bayes) are selected, and five measures (TPR, TNR, AUC, Recall, and 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** (key findings, limitations, future work, and ethical considerations)**:**

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

2. Practical part

You need to work in Jupyter Notebook and use required libraries (e.g., Pandas, Scikit-learn, matplotlib, scipy, …). Your notebook should be organized logically, and 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 training, test and validation sets. You may also consider keeping a validation set to refine your algorithms.

D. Models Training. You need to train at least 3 machine learning algorithms of your choice on the same dataset.

E. Models Evaluation. You need to compare several machine learning algorithms. Provide evaluation metrics to compare your models. Use test set for evaluation. You need to provide proper metrics to compare models.

F. Deployment: A Streamlit app (multi-page) that exposes data exploration, preprocessing choices, model training/inference, and evaluation.

G. Version Control: A Git repository with meaningful commit history matching the weekly milestones below; include an MIT/Apache license and README with setup instructions.

H. Reproducibility: a single `requirements.txt` or `pyproject.toml` and instructions to run both the notebook(s) and the app.

Intranet Submission

You must 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.
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 table of 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: You **CANNOT** use datasets from the **Kaggle** platform for this coursework. Most Kaggle datasets already come with example analyses, which would limit your chance to do original work.

Assessment Criteria

Component	**Mark**
1. Report	**35**
A. Introduction - Full description of the dataset and its purpose should be given	5
B. Exploratory Data Analysis (EDA) - Summary of EDA - Detailed description with clear justification for statistics and correlations. - Detailed description with clear justification for visualization	 1 2 2
C. Dataset Preparation - Summary of data preparation - Detailed description with clear justification for each step of data preparation	 1 2
D. Modeling - Summary of modelling - Detailed description with clear justification for algorithms selection - Detailed description with clear justification for model selection per algorithm - Detailed description with clear justification for evaluation metrics - Detailed description with clear justification for models' comparison and proposing the final model	 1 4 4 4 4
E. Conclusion - Data analysis findings - Interpretation of results	 2 3
2. Practical Part	**65**
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 (4) - Correlation matrix (3) - Other graphs (scatter plots/box plots/histograms/etc.) (3)	10
C. Data Preparation	10

Data should be cleaned and processed for the analysis; the following must be addressed: - Tackle the missing values (2) - scaling (2) - correcting error data (2) - feature engineering (2) - dataset should be split into training and test sets (2)	
D. Model training - Three models, at least, should be used (6) - Hyperparameters tunning should be used when appropriate (4)	10
E. Model evaluation - Proper evaluation metrics	10
F. Deployment - working link (5) - clean solution (5)	10
G. Version Control - weekly commits (3) - readme.md (2)	5
H. Reproducibility - requirements.txt (5)	5