greatlearning
Power Ahead

AIML | MODULE PROJECT

©Great Learning. Proprietary content. All Rights Reserved. Unauthorised use or distribution prohibited

greatlearning
Power Ahead

5 Takeaways

1. AIML module projects are designed to have a detailed hands on to integrate theoretical knowledge with actual practical implementations.

2. AIML module projects are designed to enable you as a learner to work on realtime industry scenarios, problems and datasets.

3. AIML module projects are designed to enable you simulating the designed solution using AIML techniques onto python technology platform.

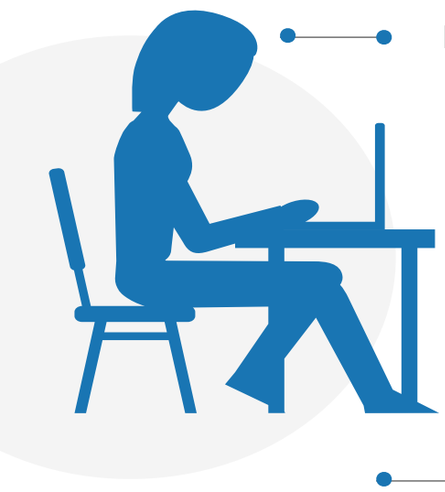
4. AIML module projects are designed to be scored using a predefined rubric based system.

5. AIML module projects are designed to enhance your learning above and beyond. Hence, it might require you to experiment, research, self learn and implement.

AIML | MODULE PROJECT

©Great Learning. Proprietary content. All Rights Reserved. Unauthorised use or distribution prohibited

UNSUPERVISED LEARNING



PART I

AIML module project part I consists of industry based problems statement which can be solved using clustering techniques.

PART II

AIML module project part II consists of designing a synthetic data generation model for a company which has a predesigned dataset.

PART III

AIML module project part III consists of industry based problems statement which can be solved using dimensional reduction techniques

PART IV

AIML module project part IV consists of designing a data driven ranking model for a sports management company.

PART V

AIML module project part V consists of implementing dimensionality reduction on multimedia dataset.

TOTAL SCORE 60

©Great Learning. Proprietary content. All Rights Reserved. Unauthorised use or distribution prohibited

PART ONE | PROJECT BASED | TOTAL SCORE | 25

- **DOMAIN:** Automobile
- **CONTEXT:** The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes
- **DATA DESCRIPTION:** The data concerns city-cycle fuel consumption in miles per gallon
- Attribute Information:
    1. mpg: continuous
    2. cylinders: multi-valued discrete
    3. displacement: continuous
    4. horsepower: continuous
    5. weight: continuous
    6. acceleration: continuous
    7. model year: multi-valued discrete
    8. origin: multi-valued discrete
    9. car name: string (unique for each instance)
- **PROJECT OBJECTIVE:** Goal is to cluster the data and treat them as individual datasets to train Regression models to predict 'mpg'

  Steps and tasks: **[Total Score: 25 points]**

    1. Import and warehouse data: [Score: 3 points]
        - Import all the given datasets and explore shape and size.
        - Merge all datasets onto one and explore final shape and size.
        - Export the final dataset and store it on local machine in .csv, .xlsx and .json format for future use.
        - Import the data from above steps into python.
    2. Data cleansing: [Score: 3 points]
        - Missing/incorrect value treatment
        - Drop attribute/s if required using relevant functional knowledge
        - Perform another kind of corrections/treatment on the data.
    3. Data analysis & visualisation: [Score: 4 points]
        - Perform detailed statistical analysis on the data.
        - Perform a detailed univariate, bivariate and multivariate analysis with appropriate detailed comments after each analysis.

        Hint: Use your best analytical approach. Even you can mix match columns to create new columns which can be used for better analysis. Create your own features if required. Be highly experimental and analytical here to find hidden patterns.
    4. Machine learning: [Score: 8 points]
        - Use K Means and Hierarchical clustering to find out the optimal number of clusters in the data.
        - Share your insights about the difference in using these two methods.
    5. Answer below questions based on outcomes of using ML based methods. [Score: 5 points]
        - Mention how many optimal clusters are present in the data and what could be the possible reason behind it.
        - Use linear regression model on different clusters separately and print the coefficients of the models individually
        - How using different models for different clusters will be helpful in this case and how it will be different than using one single model without clustering? Mention how it impacts performance and prediction.
    6. Improvisation: [Score: 2 points]
        - Detailed suggestions or improvements or on quality, quantity, variety, velocity, veracity etc. on the data points collected by the company to perform a better data analysis in future.

PART TWO | PROJECT BASED | TOTAL SCORE | 5

- **DOMAIN:** Manufacturing
- **CONTEXT:** Company X curates and packages wine across various vineyards spread throughout the country.
- **DATA DESCRIPTION:** The data concerns the chemical composition of the wine and its respective quality.

  Attribute Information:
    1. A, B, C, D: specific chemical composition measure of the wine
    2. Quality: quality of wine [Low and High]
- **PROJECT OBJECTIVE:** Goal is to build a synthetic data generation model using the existing data provided by the company.

  Steps and tasks: **[Total Score: 5 points]**

    1. Design a synthetic data generation model which can impute values [Attribute: Quality] wherever empty the company has missed recording the data.

©Great Learning. Proprietary content. All Rights Reserved. Unauthorised use or distribution prohibited

PART THREE | PROJECT BASED | TOTAL SCORE | 20

- **DOMAIN:** Automobile
- **CONTEXT:** The purpose is to classify a given silhouette as one of three types of vehicle, using a set of features extracted from the silhouette. The vehicle may be viewed from one of many different angles.
- **DATA DESCRIPTION:** The data contains features extracted from the silhouette of vehicles in different angles. Four "Corgie" model vehicles were used for the experiment: a double decker bus, Cheverolet van, Saab 9000 and an Opel Manta 400 cars. This particular combination of vehicles was chosen with the expectation that the bus, van and either one of the cars would be readily distinguishable, but it would be more difficult to distinguish between the cars.
  - All the features are numeric i.e. geometric features extracted from the silhouette.
- **PROJECT OBJECTIVE:** Apply dimensionality reduction technique – PCA and train a model using principal components instead of training the model using just the raw data.

Steps and tasks: **[Total Score: 20 points]**

1. Data: Import, clean and pre-process the data
2. EDA and visualisation: Create a detailed performance report using univariate, bi-variate and multivariate EDA techniques. Find out all possible hidden patterns by using all possible methods.

   For example: Use your best analytical approach to build this report. Even you can mix match columns to create new columns which can be used for better analysis. Create your own features if required. Be highly experimental and analytical here to find hidden patterns.
3. Classifier: Design and train a best fit SVM classier using all the data attributes.
4. Dimensional reduction: perform dimensional reduction on the data.
5. Classifier: Design and train a best fit SVM classier using dimensionally reduced attributes.
6. Conclusion: Showcase key pointer on how dimensional reduction helped in this case.

PART FOUR | PROJECT BASED | TOTAL SCORE | 5

- **DOMAIN:** Sports management
- **CONTEXT:** Company X is a sports management company for international cricket.
- **DATA DESCRIPTION:** The data is collected belongs to batsman from IPL series conducted so far. Attribute Information:
  1. Runs: Runs score by the batsman
  2. Ave: Average runs scored by the batsman per match
  3. SR: strike rate of the batsman
  4. Fours: number of boundary/four scored
  5. Six: number of boundary/six scored
  6. HF: number of half centuries scored so far
- **PROJECT OBJECTIVE:** Goal is to build a data driven batsman ranking model for the sports management company to make business decisions.

Steps and tasks: **[Total Score: 5 points]**

1. EDA and visualisation: Create a detailed performance report using univariate, bi-variate and multivariate EDA techniques. Find out all possible hidden patterns by using all possible methods.
2. Build a data driven model to rank all the players in the dataset using all or the most important performance features.

PART FIVE | QUESTION BASED | TOTAL SCORE | 5

- **Questions**: **[Total Score: 5 points]**
  1. List down all possible dimensionality reduction techniques that can be implemented using python.
  2. So far you have used dimensional reduction on numeric data. Is it possible to do the same on a multimedia data [images and video] and text data ? Please illustrate your findings using a simple implementation on python.

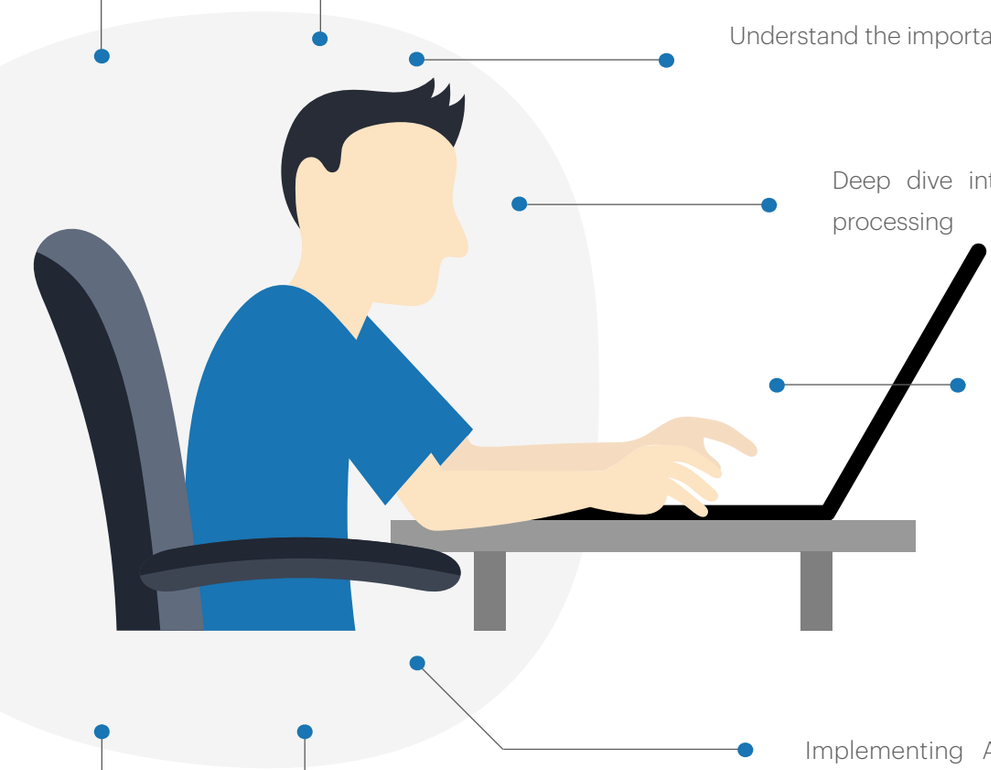
©Great Learning. Proprietary content. All Rights Reserved. Unauthorised use or distribution prohibited

LEARNING OUTCOME

Hands-on understanding on implementing dimensional reduction technique on an industry scaled dataset to enable train an AIML supervised learning model.

Hands-on experience on importing data into python from different extensions [excel, csv, json, txt, html etc.] and later merging them to form a singular data frame to train AIML models.

Understand the importance of dimensional reduction



Deep dive into extensive data cleansing and pre-processing

Extracting the best features out of data by Employing dimensional reduction algorithm to build a ranking system.

Implementing AIML clustering techniques on an industry dataset.

Using AIML unsupervised learning clustering techniques to generate synthetic data. It can be used to generate the target data and later train an AIML supervised learning model.

Designing an AIML sequential production pipeline consisting of unsupervised learning followed by supervised learning model. Understanding other types of dimensionality reduction techniques available for different variety of datasets.

©Great Learning. Proprietary content. All Rights Reserved. Unauthorised use or distribution prohibited

" Put yourself in the shoes of an actual "

DATA SCIENTIST

THAT's YOU

Assume that you are working at the company which has received the above problem statement from internal/external client. Finding the best solution for the problem statement will enhance the business/ operations for your organisation/project. You are responsible for the complete delivery. Put your best analytical thinking hat to squeeze the raw data into relevant insights and later into an AIML working model.

PLEASE NOTE

Designing a data driven decision product typically traces the following process:

1. Data and insights:
   Warehouse the relevant data. Clean and validate the data as per the the functional requirements of the problem statement. Capture and validate all possible insights from the data as per the the functional requirements of the problem statement. Please remember there will be numerous ways to achieve this. Sticking to relevance is of utmost importance. Pre-process the data which can be used for relevant AIML model.
2. AIML training:
   Use the data to train and test a relevant AIML model. Tune the model to achieve the best possible learnings out of the data. This is an iterative process where your knowledge on the above data can help to debug and improvise. Different AIML models react differently and perform depending on quality of the data. Baseline your best performing model and store the learnings for future usage.
3. AIML end product:
   Design a trigger or user interface for the business to use the designed AIML model for future usage. Maintain, support and keep the model/ product updated by continuous improvement/training. These are generally triggered by time, business or change in data.

©Great Learning. Proprietary content. All Rights Reserved. Unauthorised use or distribution prohibited

IMPORTANT POINTERS

Project should be submitted as a single ".html" and ".ipynb" file. Follow the below best practices where your submission should be:

- ".html" and ".ipynb" files should be an exact match.
- Pre-run codes with all outputs intact.
- Error free & machine independent i.e. run on any machine without adding any extra code.
- Well commented for clarity on code designed, assumptions made, approach taken, insights found and results obtained.

Project should be submitted on or before the deadline given by the program office.

Project submission should be an original work from you as a learner. If any percentage of plagiarism found in the submission, the project will not be evaluated and no score will be given.

greatlearning
Power Ahead

HAPPY LEARNING

©Great Learning. Proprietary content. All Rights Reserved. Unauthorised use or distribution prohibited