**Course Work Project Description and Rubric**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Semester** | **202420** | | **Division** | | CIS |
| **Assessment title in Syllabus** | **Project** | | **Program** | | **IT and IS** |
| **1** |  | |  | |  |
| **Course Code** | **CIS 2423** | | | | |
| **Course Title** | **Programming for Data Analytics** | | | | |
| **CLOs** | **All CLOs** | | **Accreditation Body** | | **CAA & CIPS** |
| **Course Instructor** |  | | **CRN** | |  |
| **Assessment Weight** | **40%** | | **Submission Date** | | **Week 14** |
| **For Group Work submissions an additional individual assessment will be conducted.**  **Grades for the students in one group will vary based on the individual performance in the additional assessment.** | | | | | |
|  | | | | | |
| **Student Declaration**:  **Academic Integrity Statement**  In accordance with the HCT Academic Integrity Policy  • Students are required to refrain from all forms of academic integrity breaches as defined and explained by HCT.  • A student found guilty of having committed acts of academic integrity breach(es) will be subject to the relevant sanctions as outlined by HCT.  إفادة النزاهة الأكاديمية  **وفقًا لسياسة كليات التقنية العليا للنزاهة الأكاديمية**  **• على الطلبة الإلتزام بلوائح وقواعد النزاهة الأكاديمية، كما هو مبيّن وموضح في السياسات والإجراءات الخاصة بكليات التقنية العليا.**  **• في حالة ارتكاب الطالب أي شكل من أشكال الإخلال بالنزاهة الأكاديمية، سيتعرض الى العقوبات الموضحة في السياسات ذات الصلة.**  This assignment is entirely my own work except where I have duly acknowledged other sources in the text and listed those sources at the end of the assignment.  I have not previously submitted this work to the HCT, or any other entity. I understand that I may be orally examined on my submission.  **Student (s) Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | | | | | |
|  | | | | | |
| **Student Name(s):** | Mulook sultan Alshehhi | Nada Ahmed Alshehhi | |
| **Student HCT ID(s):** | H00494605 | H00502799 | |

**For Examiner’s Use Only**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Group (50%)** | | | | | **Individual (50%)** |  |  |
| **C**LO | **1** | **2** | **3** | **4** | **Report Formatting** | **Oral Defense** | **Total** | **%** |
| **Marks Allocated** | 10 | 10 | 42 | 26 | 12 | **50** | **100** | **4**0 |
| **Marks Obtained** |  |  |  |  |  |  |  |  |

# Project Objectives

This is an intensive project-based course. It enables students to perform data analysis using Python programming. Students should select their dataset from a free source and conduct a methodical data analysis using Machine Learning algorithms. The project's primary objectives are:

* Generate a data summary using descriptive analysis
* Create a sample and visualize sample data using graphs/charts and remove the unwanted outliers;
* Investigate the correlation between the variables;
* Perform hypothesis testing if you have any assumptions about your dataset;
* Perform data preprocessing prior to building a data model;
* Create and optimize the regression model for the selected dataset in order to predict the values; and
* Develop and Optimize the classification model for the selected dataset in order to predict the values.
* Analyze the data for patterns or groups based on clustering and optimize the model in order to obtain the desired output.

# Project Description

You are assigned to work on the data analysis for chosen dataset. The list of datasets is available in a Kaggle data source(<https://www.kaggle.com/datasets>). The project carries 25% of your coursework marks. You are required to work in a team of maximum FOUR (4) members. It is important that you need to collaborate in working on the project within your team. The collaboration between the team members will be recorded, tracked, and monitored.

**For CLO1, CLO2, and CLO3 – Regression same dataset should be used**

**If required, then CLO3 – classification and Cluster different dataset can be used.**

# Project Tasks/Questions

|  |  |  |
| --- | --- | --- |
| CLO | Deliverable Learning Outcomes | Marks |
| 1 | 1. Define the purpose of data analysis for the chosen dataset   The purpose of data analysis for the Heart Disease Statlog dataset is to develop a reliable and data-driven approach for diagnosing heart disease by analyzing patient attributes. The key objectives include: Predicting Heart Disease – Build a classification model to determine whether a patient has heart disease based on medical features. Identifying Key Risk Factors – Analyze the dataset to discover significant attributes (such as age, cholesterol levels, and blood pressure) that contribute to heart disease. Improving Early Diagnosis – help medical professionals detect heart disease early, leading to better treatment and preventive measures. | 2 |
| 1. Identify and Justify the type of programming used for data analysis   Python is the most suitable choice for data analysis in this scenario. 1- Rich Ecosystem of Libraries: Python boasts a vast collection of libraries specifically designed for data analysis, including: Pandas: Powerful for data manipulation, cleaning, and analysis. NumPy: Provides efficient array operations and mathematical functions. 2- Versatility: Python is not only limited to data analysis. It can be used for various tasks, including data preprocessing, model building, and even web development. 3- Readability: Python's syntax is known for its simplicity and readability, making it easier to learn and maintain code. | 2 |
| 1. Identify the type and purpose of the machine learning algorithm to be implemented for the chosen dataset   Since we're aiming to predict whether a person has heart disease (a yes/no outcome), we use a classification algorithm. This type of algorithm helps in sorting data into categories—in this case, determining if heart disease is present or not based on the person's health information. | 3 |
| 1. Identify and Justify the independent and dependent variables for the chosen dataset.   **Dependent Variable (What we're trying to predict):**  Presence of heart disease (Yes = 1, No = 0).  **Independent Variables (Factors used to make the prediction):**  **Age**: The person's age.  **Sex**: Male or female.  **Chest Pain Type**: Different kinds of chest pain experienced.  **Resting Blood Pressure**: Blood pressure while at rest.  **Serum Cholesterol**: Cholesterol level in the blood.  **Fasting Blood Sugar**: Blood sugar level after fasting.  **Resting ECG Results**: Results from a heart activity test at rest.  **Maximum Heart Rate Achieved**: Highest heart rate during exercise.  **Exercise-Induced Angina**: Chest pain caused by physical activity.  **Oldpeak**: Measure of changes in heart activity during exercise.  **Slope of the Peak Exercise ST Segment**: Describes the slope of the peak exercise ST segment.  **Number of Major Vessels Colored by Fluoroscopy**: Number of major blood vessels visible after a special dye test.  **Thalassemia**: A blood disorder affecting red blood cells. | 3 |
| **Total** | **10** |
| 2 | 1. Justify why you want to perform the descriptive analysis for the chosen dataset.   Descriptive analysis helps us understand the basic features of the data. By summarizing and visualizing the dataset, we can identify patterns, detect anomalies, and gain insights into the relationships between different variables. This foundational understanding is crucial before diving into more complex analyses or building predictive models. | 1 |
| 1. Create a script to develop a Python function for descriptive statistics. The input for the function should be the sample and the field to perform the descriptive statistics. | 1 |
| 1. Create a program to random sampling of size 150 and find the descriptive statistics for the dependent variable from the sample [Apply the descriptive function which you created]. | 1 |
| 1. Create a script for systematic sampling by giving certain conditions and finding the desc stat for the dependent variable from the sample [Apply the descriptive function which you created]. | 1 |
| 1. Create a detailed descriptive statistics report about the dependent variable of the chosen dataset. | 1 |
| 1. Visualize the dependent variable by the Graph/Chart of the following using Python Program:    1. Scatter plot    2. Box Plot    3. Histogram    4. Heat Map   Hint: Use Matplot or Ski-learn library | 3 |
| 1. Perform the hypothesis test to find the correlation (Pearson and Spearman for numerical variable and chi-square test for categorical variable) between the independent variable and the dependent variable.   Note: If you have more than one independent variable, then chose any one of the independent variables. | 1 |
| 1. Assess the performance of the dependent variable to know whether the sample is representative of the normal population by a one-sample t-test. | 1 |
| **Total** | **10** |
| 3 | 1. Build, Train, Develop and Evaluate using Simple Regression for chosen dataset. | 5 |
| 1. Develop a script to forecast the value of the dependent variable from all the relevant independent variables using Multiple Linear Regression | 5 |
| 1. Predict the value of the dependent variable from the different classifier such as Logistic Regression, KNN, Naïve-Bayes and Decision Tree. | 17 |
| 1. Evaluate the performance of each model using confusion matrix and accuracy and identify the best fit classifier for the chosen dataset.     Thus, Logistic Regression model is best fit model for the given dataset (Heart disease statlog). Highest accuracy - Lowest number of misclassifications - Most balanced performance across both classes | 9 |
| 1. Predict the dependent variable by using best-fit classifier. | 1 |
| 1. Perform the cluster analysis such as K-means and Horizontal for any field from the chosen dataset.   K-means Clustering    Hierarchical Clustering (Dendrogram) | 8 |
| 1. Explain the strategy for improving the system after viewing the cluster diagram.   After looking at the cluster diagram, we noticed that the patients in our dataset were grouped into different clusters based on their health features, like age, cholesterol level, chest pain type, and more. These groupings gave us useful ideas about how to make our system better.  Here are some simple strategies to improve the system:  Focus on High-Risk Groups  Some clusters clearly had more patients with heart disease. We can use this to build special models that focus only on those high-risk groups. This can help the system predict more accurately.  Use Important Features  The clusters showed us that some features, like thalassemia and chest pain type, played a big role in how the patients were grouped. We can now focus more on those features when building our prediction model.  Create Smaller Models for Each Group  Instead of one big model for everyone, we can create smaller models for each cluster. This way, each model is more accurate because it’s made for people with similar symptoms.  Handle Unusual Cases  Clustering also helped us spot outliers—patients who don’t fit into any group. These special cases can be handled differently or looked at more carefully so the system doesn't get confused.  Keep Improving the System  As we get more data, we can run clustering again to see if any new patterns appear. The system can keep learning and improving over time.  Make It Easier for Doctors to Understand  Showing the clusters visually can help doctors understand which group a patient belongs to and why the system made a certain prediction. This can make the tool more useful in real life. | 2 |
| **Total** | **42** |
| 4 | 1. Create a new repo for project in Git Hub   We created a new repo for project in Git Hub and named it "Heart\_disease\_project". | 3 |
|  | 1. Upload all the project files created for CLO1,CLO2 and CLO3 to the Git Hub repo | 4 |
|  | 1. Configure Git with GitHub | 5 |
|  | 1. Clone Git hub repo to Git | 4 |
|  | 1. Pull any file from Git Hub repo to Git | 5 |
|  | 1. Modify the pulled file and push the modified file to Git Hub | 5 |
|  | **Total** | **26** |

*Please link each question/task to its corresponding CLO’s and assign marks according to the CAP.*

*Please note that a task might address many CLOs.*

# Project Deliverables

Project Report (50%)

1. **Deliverable 1**: A complete report about the purpose of data analysis, programming language chosen for data analysis, types of machine language algorithm to be analysed and the list of variables chosen for analysis [CLO1]
2. **Deliverable 2**: A detailed report about the summary of the data, sampling, graphs/charts to analyze the data, relationship between variables, evaluating assumptions using hypothesis testing, predicting the variables using the regression model [CLO 2,3]
3. **Deliverable 3**: A comprehensive description about the data model created using classification and clustering algorithm of machine learning. It should involve the narrative about the data model is optimize to predict the variables and bow the best fit model has been chosen. [CLO 2,3]
4. **Deliverable 4**: Complete narration about data versioning using Git. [CLO 4]
5. **Written Communication:** Complete report with specified format and structure [12 points].

Project Oral (50%)

1. **Oral Communication**: Each student will be assessed in the form of individual oral defense with PowerPoint presentation. [All CLOs] [10 Marks]
2. **Follow-up Questions and Discussion** [All CLOs] [30 Marks]
3. **[Collaboration]** [10 Marks].

Note: For oral presentations, a slide should be dedicated for each student to present their collaboration and lesson learnt. This will allow each student to showcase their individual contributions and reflect on the overall group experience. It also provides a structured format for sharing key takeaways and insights gained from working together.

*Points 1 to 4 are a part of group grading [50%] and points 5,6, 7 and 8 contribute to individual grading.*

# Rubric

*Please note that the Project rubric should reflect the project description and be CAP-compliant. Please feel free to customize the descriptors as per the project requirements and course level.*

## Group Project Rubric [Task-specific RUBRIC]

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Absent (F)** | **Insufficient (1-59%)**  **(F)** | **Emerging (60-69%)**  **(D/D+/C-)** | **Satisfactory (70-76%)**  **(C/C+)** | **Competent (77-86%)**  **(B-/B/B+)** | **Mastering (87-100%)**  **(A-/A)** |
| * **Group Grading-Critical Thinking: Analysis and Evaluation [45%]** | **Deliverable 1: understanding of data analytics [CLO1]**  **(5%)** |  | None of the following:   * Purpose of data analysis not mentioned. * Selection of programming language is not referenced. * Type of machine learning algorithm to be analyzed is not defined * Variables for data analysis are not mentioned   . | Some but not all the following:   * Purpose of data analysis mentioned but not clear. * Selection of programming language is referenced. * Required machine learning algorithm to be analyzed is ill-defined [Too many models mentioned] * Variables for data analysis are mentioned without justification | Most but not all the following:   * Purpose of data analysis is clearly mentioned. * Selection of programming language is referenced. * Required machine learning algorithm to be analyzed is defined, but without justification. * Variables for data analysis are mentioned with justification but not clear | All of the following:   * Purpose of data analysis is clearly mentioned. * Sound rationale for choosing the programming language * Required machine learning algorithm to be analyzed is well-defined with appropriate justification. * Variables for data analysis are mentioned with clear justification | All of the following:   * Purpose of data analysis is clearly mentioned with appropriate explanation * Thoroughness of programming language chosen is evident. * Student synthesizes information about machine learning algorithm from multiple disciplines and sources and presents them with clarity. * Variables for data analysis are precisely defined with appropriate justification * Employ outstanding knowledge about the data analysis from different sources |
|  | **Deliverable 2: Apply data analysis and visualization techniques [CLO2]**  **(25%)** |  | * No evidence of justification to perform descriptive statistics * Function is not created for descriptive statistics * Sampling techniques are not designed * Exploratory analysis are not performed * Hypothesis tastings are not achieved * Data preprocessing is not devised * Regression models are not functioned * Optimization for regression model is not operated * Best fit model is not enacted | Some but not all the following:   * Evidence of justification to perform descriptive statistics exist but not clear * Function is created for descriptive statistics having major syntactical errors. * Sampling techniques are designed but not appropriate * Exploratory analysis are performed but ill-designed * Hypothesis tastings are achieved with flaw * Evidence for data preprocessing exist but with inappropriate techniques * Regression models are functioned with huge faults * Optimization for regression model is operated with inaccuracies   Best fit model is enacted without justification. | Most but not all the following:   * Evidence of justification to perform descriptive statistics exist with clear elucidation * Function is created for descriptive statistics without flaws. * Sampling techniques are designed with appropriate rationalization * Exploratory analysis are performed with appropriate legends * Hypothesis tastings are achieved appropriate assumption * Evidence for data preprocessing exist but with applicable techniques * Regression models are functioned without mistake * Optimization for regression model is operated with appropriate metrics * Best fit model is enacted appropriate interpretation | All of the following:   * Appropriate description and exploratory analysis provided. * Sound rationale for chosen hypothesis testing * High accuracy obtained for regression model * Rigor is evident. | All of the following:   * Appropriate description and exploratory analysis provided along with exemplary analysis * Sound rationale for chosen hypothesis testing and infer statically in commendable way * Other regression model are handled obtained from different sources * Rigor is evident. |
|  | **Deliverable 3: The data model created using classification and clustering algorithm of machine learning [CLO2, CLO3]**  **(15%)** |  | * Incomplete classification models are presented. * No optimization and recommendation are performed for classification model. * No evidence of any cluster analysis * No strategies are produced from cluster model. | Some but not all the following:   * Comprehensive classification models are presented but with significant errors * Thorough optimization and recommendation are performed for classification model but with poor validation * Evidence of cluster analysis exists with poor performance * Strategies are derived from cluster model which are insignificant | Most but not all the following:   * Comprehensive flawless classification models are presented * Thorough optimization and recommendation are performed with appropriate validation for classification model * Evidence of cluster analysis exists with better performance * Strategies are derived from cluster model which are significant but not consistent | All of the following:   * Implemented all classification models * Major validation works demonstrated on optimization and recommendation of classification model * Appropriate cluster analysis was performed and good strategies are postulated and well-documented. | All of the following:   * All classification models are accurate * Perform more classification models from different sources * All cluster models are accurate * Perform more cluster models from different sources * Appropriate interpretation and recommendations are provided |
|  | **Deliverable 4: Data Versioning**  **[CLO 4]**  **(5%)** |  | * No evidence of creating a new repository and uploading project files to Git Hub * Git is not configured with Git Hub * No evidence for cloning Git hub repo to Git * Push and Pull are not performed | Some but not all the following:   * Repository model is created but files are not uploaded to repo * Git is not configured with Git Hub * Evidence of cloning exists with appropriate options * Push and Pull are performed which are insignificant | Most but not all the following:   * Repo are created and project files are uploaded to repo * Git configuration are performed with accordance of Git Hub * Evidence of cloning exists with proper options * Push and Pull are significant but not consistent | All of the following:   * Repo are created and project files are uploaded * Advanced configuration is performed on Git * Many functions related to cloning is performed apart from basic operations * Push and pull with different scenario are mentioned | All of the following:   * Scenario is created and explained about significance of repo creation * Advanced configuration is performed on Git * Different cloning techniques should be elucidated. * Push and pull with different scenario are mentioned * Branching should be exhibited |
| **Group Grading:**  **Report Quality  [5%]** |  |  | * Incomplete report with missing most of the deliverable components. * Too many typographical errors. | Some but not all of the following:   * Complete report with required deliverables. * Clear table of contents showing all required sections. * Free from formatting and typographical errors. | Most but not all of the following:   * Complete report with all required format and deliverables. * Some minor formatting and/or typographical errors. * Table of contents presented with some missing information. | All of the following:   * Complete report with required format and deliverables. * Clear table of contents showing all required sections. * Some minor formatting and/or typographical errors. | All of the following:   * Complete report with required format and deliverables. * Clear table of contents showing all required sections. * Free from formatting and typographical errors. |

## Individual Oral Presentation

## Student (1) Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Student HCT ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Absent** | **1-59.49%** | **59.5% - 69.49%** | **69.5% - 76.49%** | **76.5% à86.49%** | **86.5% à100%** |
| **Collaboration (10%)** |  | * Does not partake in the task or is so frequently distracted that he/she produces. * Having made no contribution to the group's objective. | * Participates in the task but does not collaborate with others or contribute to the group process. * Participates in discussions. * Expresses own opinion and viewpoints. * Remains dedicated to the subject. * Performs independently specific tasks | * Cooperates with the group process, but does not coordinate contributions with those of others. * Does not interrupt while listening. * Actively seeks the input of others. * Accepts duties that have been assigned. * Complies with the group consensus. * Utilizes the ideas of others | * Coordinates processes and products with those of teammates, but does not resolve significant conflicts. * Actively pays attention to what is being said. * Offers and accepts constructive feedback. * Adapts ideas and/or processes to the needs of colleagues. * Seeks consensus. * Effectively resolves trivial disputes | * Student coordinates collaborative processes and outputs. * Effectively resolves both significant and minor conflicts. * Expresses disagreements in an open and diplomatic manner. * Supports group decisions even if they are not unanimously supported. * Compromised and negotiated to attain an agreement |
| **Oral Communication (Presentation)**  **(12 Marks)** **[All CLOs]** |  | * Communicates with a limited sense of audience and purpose * No eye contact, no body language, and no decorum * Communicates with limited clarity * Uses language with limited accuracy and effectiveness. * Tension and anxiety are palpable; * Has difficulty recovering from errors. | Not all of the foregoing are true:   * Communicates with a clear sense of audience and purpose * Eye contact, body language, and decorum) * Communicates information and ideas with considerable clarity * Uses language with considerable accuracy and efficacy * Makes insignificant errors, but rapidly recovers * Displays minimal or no tension. | Most of the following, but not all:   * Communicates with a clear sense of audience and purpose * Eye contact, body language, and decorum) * Communicates information and ideas with substantial clarity * Uses language with substantial accuracy and effectiveness. * Makes simple mistakes but recovers swiftly; * Displays minimal or no tension. | Each of the subsequent:   * Communicates with a clear sense of audience and purpose * Eye contact, body language, and decorum * Communicates information and ideas with considerable clarity * Uses language with considerable accuracy and efficacy. * Makes insignificant errors, but recovers rapidly * Demonstrates minimal or no tension. | Each of the subsequent:   * Communicates with a strong awareness of audience and purpose * Maintains audience's attention by using direct eye contact and rarely glancing at notes. * Communicates information and ideas with a high degree of precision and precision * Student appears at ease and confident, with no errors. |
| **Follow-up questions and discussion (28 Marks)** **[All CLOs]** |  | Unable to answer questions from the examining board | Able to answer some but not all questions from the examining board | Able to answer most but not all questions from the examining board | Capable of answering all of the examining board's queries. | Capable of responding to all inquiries and demonstrating a thorough understanding of the material. |