|  |  |
| --- | --- |
| **Instructions** | |
|  | This is an open book exam, but any evidence of copying from third party resources or plagiarised contents will lead to suspension and other stringent actions. All the answers must be paraphrased and customised. |
|  | All questions are mandatory to attend. |
|  | You may use any tool (please mention the tools) to answer the questions but do submit well-commented answers. |
|  | Answers should be brief and to the point. You may submit a zipped folder with;   1. Word document 2. Well commented scripts, and, 3. PPT/Excel. |

|  |  |  |
| --- | --- | --- |
| **Section A** | | |
| Instructions | **Case Study - Mandatory** | Marks |
|  | Data scientist is the sexiest job in the world. How many times have you heard that? Analytics India Annual Salary Study which aims to understand a wide range of trends in data science says that the median analytics salary in India for the year 2017 is INR 12.7 Lakhs across all experience levels and skill sets. So, given the job description and other key information can you predict the range of salary of the job posting? What kind of factors influence the salary of a data scientist?  **Introduction**: You are provided with a dataset containing information about job listings, including job descriptions, job titles, location, and key skills. Your task is to predict the salary for these job listings based on the given features. This case study explores how natural language processing (NLP) techniques can be applied to solve real-world challenges in the job market.  **Dataset:**  experience (yrs): The number of years of experience required for the job.  job\_description: A text description of the job.  job\_desig: The job title or designation.  job\_type: The type or category of the job.  key\_skills: Keywords or skills associated with the job.  location: The job location.  salary: The target variable, representing the salary for the job.  company\_name\_encoded: Encoded company name. |  |
| Q1. | **Text Preprocessing and Feature Extraction:**   * Implement text preprocessing techniques (e.g., tokenization, stop word removal, and stemming/lemmatization) on the job\_description and key\_skills columns. * Extract relevant features from the text data for model training, such as TF-IDF or word embeddings. How can these features improve the prediction of salary? | 10 |
| Q2. | **Predictive Modeling:**   * Build a predictive model to estimate the salary based on the provided features (excluding salary itself). Choose an appropriate algorithm and evaluate the model's performance using regression metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). How well does the model perform? | 15 |
| Q3. | **Job Location Impact:**   * Analyze the impact of job location (from the 'location' column) on salary predictions. Are there specific locations associated with higher or lower salaries? Use NLP techniques to uncover location-related insights. | 5 |
| Q4. | **Skill Requirements:**   * Investigate the significance of key skills (from the 'key\_skills' column) in determining salary levels. Identify the most influential skills and their effect on salary predictions. | 5 |
| Attachments |  |  |

|  |  |  |
| --- | --- | --- |
| **Section B** | | |
| Instructions | **MCQs**  Write a code and use a suitable library such as hugging face pipeline or NLTK or Spacy to find the answer and tick the right option for each of these questions.  You need to submit both code and the right answer for each of the below question.  Marks Distribution for each question:  Marks for correct Code: 2.5  Marks for correct Option: 2.5 | Marks |
| Q1. | "While the film's cinematography was captivating, the storyline left much to be desired".  What is the overall sentiment of the text, considering both positive and negative aspects?  A) Positive  B) Negative  C) Neutral | 5 |
| Q2. | "Economics is the study of how societies allocate limited resources to meet their diverse needs and desires. It examines production, consumption, and distribution of goods and services. Economists analyze markets, government policies, and individual behavior to understand and improve economic systems, addressing crucial issues like inflation, unemployment, and income distribution".  What can be inferred about the document's primary topic?  A) Economics  B) Business  C) Accounting  D) It's equally distributed between multiple topics | 5 |
| Q3. | "On April 26, 1986, the Chernobyl Nuclear Power Plant in Pripyat, Ukraine, suffered a catastrophic explosion".  What is the recognized entity type for "Chernobyl Nuclear Power Plant"?  A) Person  B) Organization  C) Date  D) Location  E) None of the above | 5 |

\*\*\*\*\*\*End\*\*\*\*\*\*