**Integrated Capstone Project**

**This Case Study has four check points defined in it.**

|  |  |  |
| --- | --- | --- |
| **Check Point Topics** | **Remarks** | **Max Marks** |
| 1.1 Data manipulation using Python ( 50 marks)  1.2 Analysis using SQL Queries (50 Marks) | **Checkpoint 1** | **100** |
| 2.1Visualization using Power-BI (50 marks)  2.2 Data Analysis using Big Data Tools (50 marks) | **Check Point 2** | **100** |
| 3.1 Data Analysis + ML Model Training and Deployment on Cloud (100 Marks) | **Checkpoint 3** | **100** |
| 4.1Final Presentation and Viva( 50 marks) | **Check point 4** | **50** |

**Domain:**

Healthcare

**About:**

The liver is one of the most critical organs of the human body. It plays an essential role in the body’s function. Primary purposes include removing toxins from the body, fighting against infections, and balancing the hormones and secretion of bile juice. Liver disease is one of the most chronic and threatening diseases globally that can cause various side effects if not treated early. With advancements in biomedical sciences, the health care system has significantly improved by predicting disease using machine learning techniques.

**Challenges:**

The given challenge here is to predict whether a patient has liver disease or not based on certain features. You have to figure out whether any of the given factors are associated with the liver disease.

Hence, given the dataset, the first step of this analysis is to assess what data is available and perform some exploratory and descriptive analytics to identify interesting and useful patterns, trends, and insights.

And the next step is to build a predictive model on the given data.

**What is Expected?**

Being a data analyst, you must come up with a first step document that lists output of your exploratory analysis, any issues or problems you may see with data that need follow up, and some basic descriptive analysis that you think highlights important outcomes/findings from the data. Based on your findings, the next level of analysis will be charted out.

Also, you need to build an appropriate predictive model to predict if the given person has a liver disease or not.You can perform comparative study of several predictive models with various approaches and give your inferences accordingly.

**Data Dictionary:**

1. **ID:** Unique id of the patient
2. **Age** : Age of the patient
3. **Gender** : Gender of the patient
4. **TB** : Total Bilirubin
5. **DB** : Direct Bilirubin
6. **Alkphos**: Alkaline Phosphotase
7. **Sgpt:** Alamine Aminotransferase
8. **Sgot**: Aspartate Aminotransferase
9. **TP**: Total Protiens
10. **ALB**: Albumin
11. **A/G Ratio** : Albumin and Globulin Ratio
12. **State\_Code:** Patient state
13. **City\_Code:** Patient city

**Target variable:**

**Status :** 1 indicates Liver Patient, 2 indicates not Liver Patient

**Check Point 1**

**Task 1.1(Data Manipulation using Python)**

Here are some indicative types of analysis you can perform. Please note that this is not an exhaustive list, you may add more

* Come up with appropriate results for the following:
* Identify if there are any missing values in the data.
* Any affect of age group on the occurrence of liver disease.
* Any relationship between different factors and with the presence of liver disease given in the dataset.
* Is the dataset balanced? Any missing values? Are there duplicate records?

**Task 1.2 (SQL-Oracle)**

**Stage 1:**

1. Construct and ER-Diagram for the above-mentioned Requirement
2. Construct Tables as per the ER-Diagram.
3. Identify the relationships between tables and use appropriate standards for the same where applicable
4. Insert the appropriate data into the identified tables from the sample dataset provided.

**Stage 2:**

1. Retrieve the liver patients with a direct bilirubin level higher than the average direct bilirubin level for all patients.
2. Retrieve average alkaline phosphatase level for liver patients under the age of 50
3. Retrieve the average age and average total proteins for each gender
4. Retrieve the gender with the highest total protein level
5. Retrieve the gender with the highest average total protein level among liver patients below the age of 50
6. Retrieve the gender with the highest average total bilirubin and direct bilirubin levels

**Deliverables/Submission guidelines of Checkpoint 1**

1. You have to prepare a power point presentation with screenshots of outputs (10 -15 slides) for each check point
2. Mention Problem Statement and Your approach to the problems
3. You need to submit all the code files - Task 1.1
4. The code file(html file for Task 1.1) should contain the Batch Name and the group name, group members (One of the group member) at the top (in Jupyter Notebook).
5. All comments/inferences/insights/reasons for doing a particular tasks etc should be written as a ‘markdown text’, but **NOT** using a comment lines with # or ‘’’.
6. Submit the code file as HTML file format (you have an option in Jupyter Notebook to save the file as HTML).

Name of the file must be in the form of:

*BatchName\_FirstName\_SecondName.html*

1. Task 1.2 SQL code to be copied in the word doc
2. The presentation file should have the Batch name, group name, Project name, Group members, their responsibilities
3. Upload all the deliverables in the UNext LMS

**Check point 2 (Visualization using Power-BI, Data Analysis using Big Data Tools)**

**TASK 2.1(Visualization using Power-BI)**

**Connect the data with Power BI desktop and perform Data Manipulation using Power Query Editor. Perform the below tasks in Power BI Desktop.**

**Recommendations:**

* As a data analyst, what are the approaches do you suggest the marketing team to identify ideal target group to make the campaign successful? Recommend based on your analysis.

**NOTE:** Results and graphs must be backed with appropriate inferences and insights.

**TASK 2.2** **Data Analysis using Big Data Tools**

**What is Expected?**

Big Data technologies like HDFS (Hadoop Distributed File System), Hive, and PySpark need to be used as the historical data increases in size. As part of this task, the following activities need to be done.

* Develop PySpark routines to load data from Spark DataFrames and save it into Hive tables in an optimized format on a Hadoop cluster, de-normalizing the data if required.
* Write PySpark applications to cleanse the data, prepare the data to handle missing values, and the data transformations identified in task 1.1, making sure that the data is written into Hive tables in an efficient format as well.
* Ensure that the best practices are followed, and the design & code use the features of Spark and take advantage of them.

**Deliverables/Submission guidelines of Checkpoint 2**

1. You have to prepare a power point presentation with screenshots of outputs (10 -15 slides) for each check point.
2. Mention Problem Statement and your approach to the problems
3. Task 2.1
   * 1. PowerBI .pbix file to be submitted.
     2. Have all comments written properly in the .pbix file.
     3. The .pbix file should contain the Batch Name and the Group Number, Group member names at the top.

Task 2.2

* + Submit Jupyter code file in html format. The code file(html file for Task 2.2) should contain the Batch Name and the group name, group members (One of the group member) at the top (in Jupyter Notebook).
    1. All comments/inferences/insights/reasons for doing a particular tasks etc should be written as a ‘markdown text’, but **NOT** using a comment lines with # or ‘’’.
    2. Submit the code file as HTML file format (you have an option in Jupyter Notebook to save the file as HTML).
    3. Name of the file must be in the form of:
    4. *BatchName\_\_GroupNumber\_FirstName\_SecondName.html*
* Put all Tasks 2.1 & 2.2 as zip file (Mentioning batch name, Group number and your name) and upload it on the LMS.

**CheckPoint 3**

**Task 3.1 - Data Analysis + ML Model Training and Deployment on Cloud**

**AWS**

1. Redshift to PowerBI Connectivity
2. Move the Datasets to AWS s3
3. Create Redshift Instance
4. Ensure you create required tables in Redshift
5. Create a data pipeline/copy command to move the data from storage to datawarehouse(Redshift). You are allowed to use other copy commands as well to move the data from storage to data warehouse.
6. Connect the Redshift data to PowerBI
7. Dynamodb to s3 bucket confgiure SNS notifications for any new records added in the Dynamodb
8. Transfer the AWS s3 data to AWS Quicksight perform the same analysis doe using powerBI(Any 5 core reports)
9. 10 mb of storage of S3 exceeds then cloud watch has to trigger the alaram
10. Write a Lambda function which logs in cloud trail about S3 file type and size.
11. Build appropriate ML model/s on the data using AWS Sagemaker , Identify the right metric to evaluate the performance of the model **and Deploy on AWS Sagemaker.**

**AZURE**

1. Azure Synapse to PowerBI Connectivity
   1. Move the DataSet to Azure Synapse Storage Gen2
   2. Create a serverless SQL pool to query the data from Storage gen2
   3. Create a Linked service to PowerBI
   4. Ensure you have sufficient privileges on Synapse to access the serverless sql pool.
   5. Perform various analytics on PowerBI
   6. Perform the tasks mentioned in Task 2.1(Only 4-5 core reports)
2. Enable Azure blob storage monitoring by adding sample data and upon processing if storage receives more than 20 bytes of data
3. Move Raw data from Azure Storage to Azure Databricks, perform 2-3 visualization using PySpark command, transfer the databack to the same storage account in contaziner/blob.from new container to move the data to Azure SQL using copy command in data factory. Connect Azure SQL to PowerBI to perform the visualization.
4. Write Azure functions to trigger to trigger when blob storage exceeds 20 bytes of data.
5. Build appropriate ML model/s on the data using Azure Machine Learning , Identify the right metric to evaluate the performance of the model **and Deploy on Azure Machine Learning.**

**GCP**

1. BigQuery to PowerBI Connectivity
   1. Move the Datasets to Google Storage (Bucket)
   2. Create Bigquery Instance
   3. Ensure you create required tables in Bigquery
   4. Create a data pipeline/copy command to move the data from storage to data warehouse. You are allowed to use other copy commands as well to move the data from storage to data warehouse.
   5. Connect the BigQuery to PowerBI.

1. Write Cloud Function by adding sample data in the cloud storage and upon processing if storage storage receives more than 20 bytes of data as inbound or outbound
2. Transfer the data from bucket to Looker and perform any 5 reports.
3. Configure GCP monitoring services when storage exceeds 20 bytes of data, notify using pub/sub.
4. Configure Google Big Query and enable monitoring services (Cloud Logging) for every record insertion or deletion.
5. Build appropriate ML model/s on the data using Google Big Query Models/Vertex AI , Identify the right metric to evaluate the performance of the model **and Deploy the model on GCP Machine Learning.**

**Deliverables/Submission guidelines of Checkpoint 3**

Task 3.1

Complete all the above tasks on your respective Cloud Platform allotted and for submission take screenshots of each task specified with step-by-step flow in a Word document with proper caption mentioned along with your Batch/Group/Team member names convert as a PDF file and submit the PDF document on the LMS

**CheckPoint 4**

**Task 4.1**

Prepare crisp Final presentation including all three Checkpoint achievements and appear for Q&A session

**Deliverables/Submission guidelines of Checkpoint 4**

* You have to prepare a power point presentation with screenshots of outputs (10 -15 slides)
* Submit the ppt.

**The above four Checkpoints completes UNext Capstone Project**