DS 5110: Introduction to Data Management and Processing (Fall 2024) Roi Yehoshua
Student name: (Due) December 9, 2024

Final Project

1 Project Description

In this project, you will build an entire data science pipeline, which includes data collection, database design, data modeling, data processing, visualization, analysis, and predictive modeling. The project involves working with a database, but also expands to include machine learning modeling and data analysis. Project completion will involve the submission of a project proposal, a class presentation, and a comprehensive final report.

2 Project Requirements

Your data science pipeline should include the following components:

1. **Database**:
   (a) Design and create a database with at least 10 tables (for a relational database) or 3 collections (for a NoSQL database). Each table or collection should contain at least 10 rows (objects).
   (b) Most of the data in the tables or collections should be retrieved from external sources. Use techniques such as web scraping, web APIs, or publicly available datasets to gather relevant data.
   (c) Ensure that the data is clean and properly formatted before inserting it into the database. You may need to perform data preprocessing steps such as handling missing values, normalizing text, or converting data types.
   (d) For relational databases, ensure that the database is normalized to eliminate data redundancy and define appropriate integrity constraints, such as primary keys (PK) and foreign keys (FK) to maintain data consistency.
2. **Interactive Application**:
   (a) Develop an application in a high-level programming language (e.g., Python) that demonstrates the basic functions of your system.
   (b) The application should allow the user to run queries on the tables (collections) and insert/update/delete data into the tables (collections).
   (c) The application can be command-line based (i.e., a graphical user interface (GUI) is not required).
3. **Data Analysis**:
   (a) Extract data from your database to analyze using SQL queries.
   (b) Perform exploratory data analysis (EDA) using Pandas, including tasks such as:

i. Calculating summary statistics (e.g., mean, median, standard deviation) for relevant columns.
ii. Using Pandas functions to find correlations, missing data, and apply group-by operations.
iii. Applying custom functions or lambda functions to columns for data transformations.

(c) Create at least two visualizations (e.g., bar charts, scatter plots) based on the analysis, using Pandas plotting or other visualization libraries (e.g., Matplotlib, Seaborn).

(d) Summarize your findings and include them in your final project report.

4. **Machine Learning Modeling:**

   (a) Use a simple machine learning algorithm (e.g., linear regression, decision tree) on the data extracted from your database.

   (b) Train the model using part of the data (training set) and test it on the remaining part (test set).

   (c) Report relevant performance metrics (e.g., accuracy, precision, mean squared error).

   (d) Store the predictions and performance metrics back into the database.

   (e) Discuss the model's performance in the final report and explain what insights or patterns about the data were revealed by the model.

3 Sample Topics

This section provides some examples for systems that you may develop, but you are welcome to choose anything else that interests you.

- **Hospital Management System.**

  Design and implement an automated system that handles doctors' and patients' data in a hospital. The system enables an admin to register a patient for the hospital and store their disease details into the database. Doctors can add, view, or update their patients' data, including their subscriptions, diagnosis and lab results. Patients can search for availability of doctors and make appointments. Admins can access statistics about the hospital such as room capacity, number of doctors/patients in each department, etc.

  Data Analysis/Machine Learning: Perform analysis to predict the likelihood of patients being diagnosed with certain conditions based on historical patient data, or use machine learning to predict hospital bed occupancy rates.

- **Movie Theater Reservation System.**

  Design and implement a system for reserving tickets to a movie theater. Users of the system can perform tasks such as searching for a movie, viewing movie show details and schedules, booking a movie show, card registration and receiving tickets. Admins can use the system to insert, update and delete data such as movie descriptions and movie schedules. Admins can also access statistics about the movie theater, such as what are the most popular movies and the monthly revenue.

Data Analysis/Machine Learning: Analyze ticket sales trends and predict which upcoming movies will be most popular based on past user preferences and ratings.

- **Restaurant Management System.**

  Design and implement a system that automates the day-to-day activity of a restaurant. After a successful login to the system, customers can browse through the menu of the restaurant, and look at the various food options available along with the price of each item. Then they can select items from the menu and add them to their order. Customers should also be able to reserve tables at the restaurant or join a reservation waiting list. Chefs of the restaurant can view the current queue of orders, and update the order status to ready once it is prepared. Restaurant managers should be able to view statistics about the restaurant such as weekly sales and current inventory.

  Data Analysis/Machine Learning: Analyze order data to forecast future demand or use machine learning to optimize inventory management by predicting which items will be most ordered during certain periods.

- **Library Management System.**

  Design and implement a system that keeps track of books and their checkouts, as well as member accounts and subscriptions. Users of the system will be able to search for books, issue and return books, and check fines (if any). Librarians can read information about any member, track the books issued by a particular member, and update the availability status of books. Admins should be able to view, update or delete all members records, and update or delete book records.

  Data Analysis/Machine Learning: Perform analysis on book checkout patterns to predict future book popularity or use machine learning to recommend books to members based on their checkout history.

4 Teams

Teams should have 2–3 members. Teams of size 1 or 4 are not restricted, but will only be allowed per request (i.e., a reason must be provided and found reasonable). Each student in the team should work on a different part of the project and what each student will work on should be explicitly listed in the project proposal.

5 Project Proposal

The project proposal is a short document (2-3 pages) that describes the system you intend to build and includes an initial design of the data science pipeline. The proposal must specify all the following items:

- Project title.

- Project participants: Who is on the team? Is there a clear division of labor between the team members?
- System description: What are the objectives of the system you are going to develop? Describe the system requirements and its main features.
- Database design: Initial schema of the database that depicts the main tables and their relationships.
- Data sources: Where will the data come from? Which preprocessing steps will be needed to clean and prepare the data?
- Machine learning and data analysis: Briefly describe any planned machine learning models or data analysis methods you intend to use. What insights or predictions are you hoping to generate?
- Libraries and tools: What libraries/platforms are you going to use in order to accomplish the project? Provide references where applicable.

6 Project Presentation

Towards the end of the semester, each team will present its project to the class. The presentation should include the following items:

1. A Power Point slide deck, containing highlights, to showcase the project.
2. The final schema of the database.
3. A demo of the application running on some sample inputs.
4. At least two reports or graphs that show relevant information from the database, including results from your data analysis or machine learning model.

The presentation materials should be uploaded to Canvas on the day before the presentation. The presentation should be 5 minutes long, with an additional 1-2 minutes for questions. All members of the team should take part in the presentation.

7 Project Report and Deliverables

Your final submission should include a report that describes the system you have built. The report should have the following sections:

1. Introduction: A general description of the system and its objectives.
2. Database design: The final schema of the database. Explain your key design decisions. Describe any normalization procedures you have taken.
3. Application description: Describe the main features provided by the application, and how the application stores and retrieves data from the database.

4. Data collection: Describe how the data was acquired and from where (with proper references). Describe any processing procedures used to prepare the data (e.g., cleaning procedures, aggregating data from different sources, etc.).

5. Data analysis/Exploratory Data Analysis (EDA): Describe the steps taken to analyze the data using SQL queries and Pandas. Summarize key findings from the analysis, including any trends, patterns, or anomalies. Include at least two visualizations (e.g., bar charts, scatter plots) and explain what insights they provide.

6. Machine learning model: Explain which machine learning algorithm you used and why it was chosen. Include key performance metrics (e.g., accuracy, precision, mean squared error). Discuss the model's performance and any insights gained from the data based on the model's predictions.

7. Reports: At least five reports or graphs that show relevant information from the database, including insights from the data analysis and predictions from the machine learning model.

8. Conclusions / future directions: What have you learned from this project? If you had more time to spend on the project, what would you have liked to do next? What advice about the project would you give to future DS 5110 students?

The report should be written using the AAAI Conference on Artificial Intelligence format: AAAI Author Kit.

In addition to the report, your final submission should include the following deliverables:

1. The SQL statements for creating the database and populating it with data.

2. The database schema diagram. This should be generated from the final database.

3. Source code of the application (or a link to your GitHub repository). The source code should be well commented and organized, and run without errors.

8 Academic Integrity

The system should be built entirely by yourself and your teammates from the ground up, and must not contain any part of someone else's work (e.g., copying code from online sources is totally forbidden).

While tools like ChatGPT or other AI systems can be used for brainstorming ideas or clarifying concepts, they must not be used to generate substantial parts of the code or written sections of the project. If you do use ChatGPT for any part of the project (e.g., asking for guidance on a specific issue), this usage must be explicitly disclosed in your report. All code, analyses, and reports must be your own original work.

Failure to adhere to this guideline will result in a grade of 0 and a report to the university administration (OSCCR) for academic dishonesty.

9 Timeline

All deliverables must be submitted to Canvas on the due date before 11:59 PM. The submission should be made by all team members.

- The project proposal is due 10/24.
- Project presentations will take place during the final class of the semester.
- **The final project report / deliverables is due on 12/9.**
  There can not be any extensions on the final project, because final grades are due soon after that, and we need time to ensure every project receives a proper assessment.