Technical Report: Programming for Data Science Coursework

COMScDS252P-013

---

1. Executive Summary

This technical report presents the implementation and findings of three major components: an Object-Oriented University Management System, an E-commerce Data Analysis, and an AI Ethics evaluation in healthcare. The coursework demonstrates practical application of Python programming, data handling, statistical analysis, and ethical reasoning.

The University Management System was developed using core OOP principles such as inheritance, encapsulation, and polymorphism to simulate real-world academic structures. The E-commerce analysis involved web scraping, preprocessing, statistical exploration, and predictive modeling on book data. Several practical challenges were encountered, particularly in data cleaning and feature engineering, which were resolved through iterative improvements.

The ethics section assesses algorithmic bias, privacy laws, and stakeholder impact in healthcare AI systems critically. All things considered, the coursework demonstrates a strong integration of technical proficiency with critical thinking and moral consciousness. Improved data preprocessing techniques, comprehension of practical data challenges, and the significance of responsible AI development are some of the main results.

---

2. Question 1: OOP Implementation (University Management System)

The University Management System was designed using a structured object-oriented approach. To represent common attributes like name, ID, email address, and phone number, a base class Person was created. In order to create a distinct two-level inheritance hierarchy, this class was expanded into three derived classes: Staff, Faculty, and Students.

A major design decision was to ensure code reusability through inheritance. The use of `super()` allowed child classes to inherit and extend base attributes efficiently. Each subclass also implemented specialized attributes, such as `student_id` for students and `department` for faculty and staff.

Encapsulation was enforced using property decorators. For example, GPA was implemented as a read-only computed property rather than a directly modifiable attribute. Input validation was applied to ensure grades remained within the valid 0.0–4.0 range, and enrollment was restricted to a maximum of six courses per semester. Invalid inputs triggered exceptions, improving system robustness.

Polymorphism was demonstrated by overriding the `get_responsibilities()` method across all subclasses. This allowed different object types to respond uniquely while maintaining a consistent interface. A heterogeneous list of objects was used to validate this behavior dynamically.

To mimic actual academic structures, additional classes like Course and Department were added. Relationships like faculty assignments and course enrollment were covered in these classes. The system effectively simulated real-world interactions, such as managing department-level data, assigning instructors, and enrolling students.

Overall, the implementation reflects strong adherence to OOP principles and demonstrates scalability, maintainability, and clarity in design.

---

3. Question 2: Data Analysis (E-commerce Data Analysis)

Methodology and Tools

The data analysis task focused on extracting and analyzing book data from an online source. Web scraping was implemented using `requests` and `BeautifulSoup`. A delay mechanism and error handling ensured ethical and stable scraping. Approximately 100 books were collected across multiple pages.

Initially, the dataset was scraped without explicitly including the book category. While this worked for basic analysis, it became problematic when performing grouped analysis. To address this, the scraping process was redesigned from the beginning to explicitly extract category information. Although time-consuming, this significantly improved data quality and analytical depth.

Another key challenge involved data cleaning. Even after removing the "£" symbol from prices, errors persisted due to hidden or additional special characters. This required a second round of cleaning where all unwanted symbols were stripped more rigorously. This step was crucial for accurate numerical conversion and further analysis.

Additionally, the scraping function (getSoup) was manually adjusted to retrieve approximately 100 books, as different page ranges produced inconsistent results. This ensured compliance with coursework requirements.

Data Cleaning and Preprocessing

Several preprocessing steps were applied:

- Prices were converted to numeric values after removing special characters
- Ratings were transformed from text format into numerical scale (1–5)
- Missing values and duplicates were handled
- New features such as price_category and in_stock were created

These transformations improved data consistency and enabled deeper analysis.

Statistical Analysis

Descriptive statistics revealed key insights into pricing patterns. The mean and median prices indicated a moderate pricing distribution, while standard deviation highlighted variability among categories.

Outlier detection using the IQR method identified unusually high-priced books. Correlation analysis between price and rating showed a weak relationship, suggesting that higher prices do not necessarily imply better ratings.

A hypothesis test comparing Fiction and Non-Fiction categories was conducted using a t-test. The results indicated no statistically significant difference in average prices at a 0.05 significance level.

Data Visualization

Several visualizations were created:

- Histogram showing price distribution
- Boxplot comparing prices across categories

- Scatter plot of price vs rating
- Bar chart of average rating by category

These visualizations provided intuitive insights and supported statistical findings.

Predictive Analysis

To predict book prices based on category and rating, a linear regression model was created. Implementing the model presented difficulties during this phase. In order to get around this, more online references were consulted, and the process was guided by AI.
Moderate predictive performance was demonstrated by the model evaluation. According to feature importance analysis, price was more influenced by category than by rating.

Insights

- Book category plays a significant role in pricing strategy
- Ratings are not a strong determinant of price
- Data preprocessing is critical for reliable analysis

---

4. Question 3: Ethics Analysis (AI in Healthcare)

The ethics component focused on privacy, bias, and stakeholder impact in healthcare AI systems.

Healthcare data requires strict protection due to its sensitive nature. Regulations such as HIPAA and GDPR provide frameworks for safeguarding patient data. While HIPAA focuses on healthcare-specific data in the US, GDPR offers broader protections including the “right to be forgotten.” However, compliance across regions remains challenging.

A key issue identified is the limitation of anonymization. Simple de-identification is insufficient because individuals can be re-identified through indirect attributes. This highlights the need for stronger data protection mechanisms.

Algorithmic bias is another major concern. Bias can arise from underrepresentation in datasets, geographic imbalances, and historical inequalities. A real-world example demonstrated how biased algorithms underestimated the needs of certain racial groups, leading to unequal healthcare access.

To mitigate bias, both technical and process-based strategies are necessary. These include using diverse datasets, fairness-aware algorithms, and involving multidisciplinary teams in development.

An ethical decision framework was proposed, including questions related to data fairness, privacy, interpretability, accountability, and patient well-being. The concept of the “right to explanation” emphasizes the need for transparency in AI-driven decisions, especially in critical domains like healthcare.

Stakeholder analysis showed that:

- Patients benefit from improved diagnosis but face privacy risks
- Healthcare providers gain efficiency but bear responsibility for decisions
- Developers and AI Engineers must ensure fairness and transparency

A key recommendation is the establishment of interdisciplinary oversight boards to ensure ethical compliance and balanced decision-making.

---

5. Technical Implementation

Code quality was maintained by following PEP 8 guidelines and including meaningful naming conventions and docstrings. Modular design was used to separate functionalities into different scripts, improving readability and maintainability.

Basic error handling was implemented throughout the system, particularly in web scraping and data validation. Testing involved running different scenarios, including invalid inputs, to ensure system stability.

One major challenge was handling inconsistent data during preprocessing. This was resolved through iterative cleaning and validation steps. Another challenge was model implementation, which required external references and AI support to successfully complete.

Version control practices were followed(Git, Github, Branching), ensuring organized development and easy tracking of changes.

---

6. Reflection

Both programming and analytical skills were greatly enhanced by this course. While the data analysis component offered practical experience with real-world data challenges, the OOP task improved comprehension of software design principles.

The significance of appropriate data collection from the start was one of the main learning outcomes. Ignoring the category feature at first caused issues later, which made careful planning even more important. Managing messy data also brought to light the intricacy of real-world datasets.

Implementing more sophisticated machine learning models and developing interactive dashboards for improved visualization are possible future enhancements. Reliability would be improved by incorporating automated testing.

Overall, this coursework highlighted the significance of fusing technical proficiency with moral responsibility in real-world situations and illustrated the practical application of data science concepts.