Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

Midterm Exam			
Topic:	**Data Analysis and Visualization** **Probability and Visualization** **Linear Regression** **Logistic Regression**	**Week No.**	10
Course Code:	**CSEL302**	**Term:**	2nd Semester
Course Title:	**Introduction to Intelligent Systems**	**Academic Year:**	2023-2024
Student Name		**Section**	
Due date		**Points**	

Datasets:

a. Boston Housing Dataset

- **Description:** Contains information about the housing values in suburbs of Boston.
- **Use Case:** Ideal for Linear Regression to predict continuous outcomes like house prices based on various features (e.g., crime rate, number of rooms, age) and Logistic Regression for classification tasks (e.g., predicting whether a house's value is above or below a certain threshold).
- **Link:** https://www.kaggle.com/datasets/nancyalaswad90/review

b. Titanic Dataset

- **Description:** Includes passenger information from the Titanic.
- **Use Case:** Can be used for Logistic Regression to predict binary outcomes such as survival, and for exploratory data analysis to understand correlations between variables (e.g., age, class, fare) through probability and statistics concepts.
- **Link:** https://www.kaggle.com/datasets/brendan45774/test-file

c. Diabetes Dataset

- **Description:** Comprises diagnostic measurements for a set of patients diagnosed with diabetes.
- **Use Case:** Suitable for both Linear Regression (predicting a quantitative measure of disease progression) and Logistic Regression (classifying patients into categories based on diagnostic measurements).
- **Link:** https://www.kaggle.com/datasets/shantanudhakadd/diabetes-dataset-for-beginners

d. Marketing Campaign Dataset

- **Description:** Features data from marketing campaigns, including customer interactions, previous purchases, and social media engagement.
- **Use Case:** Linear Regression could model the relationship between campaign features and amount spent, while Logistic Regression could predict the likelihood of a purchase.

- **Link:** https://www.kaggle.com/datasets/pkdarabi/bank-marketing-dataset

e. Credit Scoring Dataset

- **Description:** Contains individual's financial history, credit usage, loan details, and whether they defaulted on loans.
- **Use Case:** Use Linear Regression to predict continuous variables such as credit score, and Logistic Regression to classify individuals into "default" or "no default" categories.
- **Link:** https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

1. Objective

- **Overview**: Integrate Probability and Statistics, Linear Regression, and Logistic Regression theories into a cohesive case study that demonstrates their application in predicting an outcome based on a given dataset.

- **Goal**: Predict [Outcome] using the dataset provided, utilizing Linear and Logistic Regression models to analyze and interpret the results.

2. Dataset Preparation

- **Description**: Briefly describe the dataset used for the case study, including the source, variables (both dependent and independent), and any preprocessing steps (e.g., cleaning, normalization) performed to prepare the data for analysis.

3. Exploratory Data Analysis (EDA)

- **Probability and Statistics Concepts**: Utilize descriptive statistics to summarize the dataset, showcasing measures such as mean, median, variance, and standard deviation. Employ probability distributions to understand the data's distribution.

- **Visualization**: Create visualizations (histograms, scatter plots) to explore relationships between variables and identify patterns or outliers.

4. Linear Regression Model

- **Theory Recap**: Highlight key concepts from the Linear Regression lecture, emphasizing the model's assumptions and the importance of variable relationships.

- **Implementation**: Demonstrate the process of fitting a Linear Regression model using Python in Google Colab, including selecting independent variables and interpreting the model's coefficients.

- **Evaluation**: Use metrics like R-squared, MSE (Mean Squared Error), and RMSE (Root Mean Squared Error) to evaluate model performance.

5. Logistic Regression Model

- **Theory Recap**: Reiterate the foundational concepts of Logistic Regression, focusing on its application in binary classification problems.
- **Implementation**: Show the steps to build a Logistic Regression model, detailing feature selection, model fitting, and coefficient interpretation.
- **Evaluation**: Discuss model evaluation techniques specific to classification problems, such as Accuracy, Precision, Recall, F1 Score, ROC Curve, and AUC (Area Under the Curve).

6. Model Comparison and Selection

- Compare the Linear and Logistic Regression models based on their performance metrics, discussing each model's suitability for different types of prediction problems.
- Explain the decision-making process for choosing one model over the other, considering factors like model accuracy, interpretability, and assumptions.

7. Conclusion and Insights

- Summarize the key findings from the case study, highlighting how the applied statistical and machine learning methods facilitated data-driven decision-making.
- Discuss potential applications of these models in various fields, as outlined in the lectures, and reflect on the importance of understanding underlying assumptions and model limitations.

8. References

- Cite all sources, including datasets, libraries used in Google Colab (e.g., pandas, NumPy, scikit-learn, matplotlib), and any additional resources that informed your analysis.

This structure serves as a comprehensive guide for combining theoretical knowledge with practical application, providing a holistic view of how data science techniques can be applied to solve real-world problems. You can adapt each section based on the specifics of your dataset and the particular outcomes you wish to predict.

Submission Instruction:

- Share the Google Collab Activity to markbernardino@lspu.edu.ph
- Filename Format: **2A-BERNARDINO-MIDTERM**

Inability to follow this instruction will be deducted 5 points each for filename format and late submission per day. Also, cheating and plagiarism will be penalized.