DS 301 – Machine Learning Techniques – Lab 2 (Case Study)

Supervised Machine Learning Techniques and Data Preprocessing including Feature Extraction

Max Marks: 100

Student Names: Lucas Yuki Nishimoto **Student IDs: 2024000017**

Description:

You are provided with a list of Machine Learning problems below. Work in groups of at least 4 members to brainstorm on the following problems and answer the following questions for each of the problems.

Problems:

Problem 1: Employee Attrition Prediction

A multinational company is experiencing high employee turnover. The HR department wants to develop a predictive model to identify employees who are likely to leave the organization in the next six months.
They have access to employee data including demographic information, performance scores, job satisfaction, salary, and years at the company.

Problem 2: Disease Diagnosis (Diabetes Prediction)

A healthcare clinic wants to help doctors predict whether a patient has diabetes based on health indicators collected during routine check-ups.

Problem 3: Sales Forecasting for a Retail Store

A large retail chain wants to predict next month's total sales for each branch based on past sales patterns and other external factors.

Problem 4: Vehicle Fuel Efficiency Prediction

An automobile manufacturer wants to predict a car's fuel efficiency (in miles per gallon) based on its engine and design specifications.

Questions:

For each of the problems above answer the following questions:

1. Which class of supervised machine learning problems (Classification or Regression) does the problem belong to and why? Provide rationale behind your answer.
2. What sort of data should be collected to solve the problem
3. Which feature extraction technique(s) (LDA or PCA) is/are beneficial to apply for that problem and why?
4. What data preprocessing steps are essential in order to solve the problem.

Suggested Discussion Hints for Students:

- **Classification vs Regression:** Think about whether the target variable is categorical or continuous.
- **Data Collection:** What sensors, records, or systems could provide the data?
- **Feature Extraction:** PCA helps when reducing many continuous variables; LDA helps when separating known classes.
- **Preprocessing:** Consider handling missing values, scaling, encoding categorical variables, and removing outliers.

Answers:

Problem 1:

1 - Classification

2 - To solve this problem, the following employee-related data should be collected:

- **Demographic information (age, gender, marital status, education level)**
- **Job-related characteristics (job role, department, years at company, promotions)**
- **Performance metrics (performance ratings, awards, evaluations)**
- **Compensation data (salary, bonuses, benefits)**
- **Work environment indicators (job satisfaction, work–life balance, workload)**

- **Behavioral data (absences, lateness, overtime hours)**
- **Manager-related information (feedback scores, team climate)**

3 - LDA

Because this is a supervised classification problem.

4 - Check for missing values and duplicates and resolve them.

Impute missing data using the median for numerical variables and the mode for categorical variables.

Encode categorical variables such as job role, department, education level, and marital status using one-hot encoding.

Scale numerical features because algorithms like Logistic Regression, SVM, and LDA require standardized data.

Detect and handle outliers, especially in salary and performance metrics.

Balance the dataset if attrition cases are highly imbalanced, using methods like SMOTE or undersampling.

Apply LDA if dimensionality reduction is needed, since it improves class separation.

Perform a train–test split to evaluate the model properly.

Problem 2:

1 - Classification

2 - To build a diabetes prediction model, the following patient health data is required:

- **Medical measurements (glucose levels, blood pressure, insulin levels)**
- **Anthropometric data (age, BMI, weight, height)**
- **Family history of diabetes**
- **Lifestyle indicators (diet, physical activity)**
- **Previous diagnoses or health conditions**
- **Lab test results**

3 - LDA

This is also a supervised classification problem.

4 - Check for missing values and duplicate entries.

Impute missing values using the median for continuous medical measurements such as glucose or BMI, and the mode for categorical features.

Scale numerical features to ensure consistent magnitude across variables, improving performance for algorithms and LDA.

Encode categorical variables such as lifestyle indicators or family history of diabetes.

Handle outliers, especially extreme glucose, insulin, or BMI values.

Check for class imbalance and apply suitable techniques if necessary, such as SMOTE or class weighting.

Apply LDA if dimensionality reduction is required to enhance class separation.

Split the data into training and testing sets.

Problem 3:

1 - Regression

2 - To predict next month's sales for each store, the following data should be gathered:

- **Historical sales data (daily/weekly/monthly sales records)**
- **Store information (location, size, number of employees)**
- **Inventory levels**
- **Promotions and discounts**
- **Seasonal factors (holidays, events)**
- **External factors (weather, economic indicators)**
- **Online vs in-store sales**
- **Competitor activity (if available)**

3 - PCA

LDA cannot be used for regression because it requires categorical class labels.

4 - Check for missing values and duplicate records, especially in time-series data.

Impute missing values using the median or forward-fill techniques when appropriate.

Encode categorical variables such as store location or promotion types.

Scale numerical features including sales amounts, weather indicators, and economic variables.

Handle outliers, since extreme promotional spikes may distort the model.

Create time-based features such as month, week, season, or lagged sales values.

Use PCA if the dataset contains many correlated numerical predictors, such as economic indicators or historical features.

Perform a train–test split respecting chronological order to preserve the time-series structure.

Problem 4:

1 - Regression

2 - To predict a vehicle's fuel efficiency (MPG), the following car-related data is needed:

- **Engine specifications (horsepower, displacement, cylinders)**
- **Transmission type (manual/automatic)**
- **Vehicle weight**
- **Acceleration performance**
- **Aerodynamic properties (drag coefficient)**
- **Tire specifications**
- **Fuel type**
- **Vehicle dimensions (length, width, height)**

3 - PCA
Since this is a regression task with a continuous target (MPG), LDA is not applicable.

4 - Check for missing values and duplicates.
Impute missing numerical values using the median and impute categorical missing values using the mode.
Encode categorical features such as transmission type or fuel type.
Scale numerical features such as horsepower, weight, acceleration, and displacement.
Detect and handle outliers such as unrealistic MPG values or irregular engine measurements.
Apply PCA if the dataset contains many correlated engine and design-related features, improving model stability.
Split the data into training and testing sets to evaluate model performance.

Submission Instructions:
Answer all questions in the document and convert it into PDF. Submit the PDF file in the Lab 2 submission on Google Classroom by the deadline assigned on Google Classroom.