# FML Module 2 Part A

1. Given a dataset containing various patient attributes and their corresponding medical conditions, how would you design a classification model to predict the likelihood of a certain disease based on these attributes? What factors would you consider important in making accurate predictions, and how would you evaluate the performance of your model?

- **Approach:** To design a disease prediction model, I would employ classification algorithms such as logistic regression, decision trees, or random forests.
- **Important Factors:** Factors crucial for accurate predictions may include patient attributes like age, gender, medical history, lifestyle choices, and genetic predispositions.
- **Evaluation:** The performance of the model could be evaluated using various metrics such as accuracy, precision, recall, F1 score, and ROC-AUC curve analysis.

---

2. Suppose you are developing a spam filter for an email service provider. How would you build a classification model to distinguish between spam and legitimate emails? What features or characteristics of emails might be indicative of spam, and how would you evaluate the performance of your spam detection system?

- **Model Development:** For building a spam email classifier, I would utilize machine learning algorithms like logistic regression, Naive Bayes, or support vector machines (SVM).

- **Indicators:** Indicators of spam emails could include features like sender's email address, subject line content, email body content, presence of attachments or links, and email metadata.
- **Evaluation:** The performance of the spam detection system can be evaluated using metrics such as accuracy, precision, recall, F1 score, and receiver operating characteristic (ROC) curve analysis.

---

3. You work for an e-commerce company and want to segment customers into different groups based on their purchasing behaviour. How could you use a decision tree approach to identify distinct customer segments? What criteria would you use for splitting nodes in the tree, and how would you interpret the segments generated by the tree?

- **Approach:** Utilizing decision trees for customer segmentation involves recursively splitting nodes based on features such as purchase frequency, total spending, product categories purchased, and demographic information.
- **Criteria:** Splitting criteria could be based on Gini impurity or information gain, where the goal is to maximize homogeneity within segments and heterogeneity between segments.
- **Interpretation:** The segments generated by the decision tree can be interpreted as distinct customer groups with similar purchasing behaviors, allowing for targeted marketing strategies tailored to each segment.

---

4. Suppose you are an HR manager tasked with analyzing employee attrition within your organization. Explain how CART (Classification and Regression Trees) could be applied to identify the factors contributing to employee turnover within an organization, elucidating the process of constructing a decision tree to discern the most influential predictors and their thresholds,

thus facilitating a deeper understanding of the underlying drivers behind employee attrition.

- **Method:** CART can be applied to identify factors contributing to employee turnover by constructing a decision tree that recursively splits nodes based on employee attributes such as job satisfaction, salary, tenure, performance ratings, and work-life balance.
- **Process:** The decision tree is built by selecting the most influential predictors and determining optimal thresholds for splitting nodes, which provides insights into the underlying drivers of employee attrition.
- **Interpretation:** By interpreting the decision tree, HR managers can gain a deeper understanding of the factors influencing employee turnover and formulate targeted retention strategies to address key issues.

---

5. If you're working for a real estate agency and tasked with predicting house prices based on various features such as square footage, number of bedrooms, location, etc., how would you approach this regression problem? What regression techniques would you consider, and how would you evaluate the performance of your predictive model?

- **Approach:** To predict house prices, regression techniques like linear regression, ridge regression, or random forest regression could be employed.
- **Features:** Features such as square footage, number of bedrooms, bathrooms, location (ZIP code or neighborhood), amenities, and proximity to schools or transportation hubs could be considered.
- **Evaluation:** The performance of the predictive model can be evaluated using metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared score.

---

6. You work for a financial institution and need to detect fraudulent transactions. How could logistic regression be used to classify transactions as either fraudulent or legitimate?

- **Logistic Regression:** Logistic regression can be utilized to classify transactions as fraudulent or legitimate based on features such as transaction amount, location, time of day, transaction frequency, and past transaction history.
- **Model Training:** The logistic regression model is trained on historical transaction data, where fraudulent transactions are labeled as positive examples, and legitimate transactions are labeled as negative examples.
- **Prediction:** Once trained, the model can predict the likelihood of a transaction being fraudulent, with a threshold set to balance precision and recall.

7. Imagine you're a healthcare researcher developing a diagnostic model for a particular disease based on patient characteristics and medical tests. What features or diagnostic indicators would you consider in your logistic regression model, and how would you interpret the model's coefficients to understand their significance in disease diagnosis?

- **Features:** Features for the logistic regression model could include patient demographics (age, gender), medical history, symptoms, laboratory test results, genetic markers, and imaging findings.
- **Interpretation:** The coefficients of the logistic regression model indicate the strength and direction of the relationship between each feature and the probability of disease diagnosis. Positive coefficients suggest an increase in the likelihood of disease, while negative coefficients suggest a decrease.

8. Suppose you're working on a binary classification task where the classes are linearly separable. How would you choose and train a linear classification model such as Logistic Regression or Linear SVM?

- **Model Selection:** For linearly separable binary classification tasks, models like Logistic Regression or Linear Support Vector Machines (SVM) are suitable choices.
- **Training:** These models are trained using optimization algorithms such as gradient descent or coordinate descent to find the optimal coefficients (weights) that separate the classes with a hyperplane.
- **Evaluation:** Model performance can be evaluated using metrics such as accuracy, precision, recall, F1-score, and receiver operating characteristic (ROC) curve.

---

9. How can we leverage multiple linear regression to analyze the relationship between multiple independent variables (square footage, number of bedrooms, number of bathrooms, and location) and the dependent variable (selling price) to understand the impact of these house attributes on the selling price of houses within a specific real estate market?

- **Model Setup:** Multiple linear regression models the relationship between multiple independent variables (features) and a dependent variable (selling price).
- **Variables:** Independent variables could include square footage, number of bedrooms, number of bathrooms, location (as dummy variables or ZIP codes), and other relevant features.
- **Analysis:** By analyzing the coefficients of the independent variables, we can understand their impact on the selling price. For example, a positive

coefficient for square footage suggests that larger houses tend to have higher selling prices.

---

10. Choose how logistic regression can be effectively employed to predict the likelihood of a patient developing a specific medical condition based on their demographic information, lifestyle factors, and medical history, aiming to aid healthcare professionals in proactive disease management and personalized treatment strategies.

- **Model Inputs:** Logistic regression can take inputs such as age, gender, lifestyle factors (e.g., smoking, exercise), medical history (e.g., pre-existing conditions), genetic predispositions, and biomarker levels.
- **Training and Prediction:** The logistic regression model is trained on historical data with labeled outcomes (presence or absence of the medical condition). Once trained, it can predict the likelihood of a patient developing the condition based on their input features.
- **Clinical Application:** Healthcare professionals can use these predictions to identify high-risk patients for targeted interventions, early diagnosis, and personalized treatment plans.