Assignment 1: Machine Learning

1. Used-Car Listing: Regression or Classification?

This problem is a **classification** task.

Justification:

The goal is to predict "whether a car will sell above its listing price." The target feature is a binary, categorical value (e.g., 'Yes/Above Price' or 'No/Not Above Price'). Classification is used when the target feature is categorical. If the goal were to predict the exact final selling price, that would be a continuous value, making it a regression problem.

Comparison of Formulations:

- **As a Classification Problem (Chosen):**
  - **Model Output:** The model would output a discrete class label.
  - **Evaluation Metrics:** Performance would be measured using metrics like **accuracy** (total correct predictions) , **precision** (reliability of positive predictions) , **recall/sensitivity** (ability to find all positive cases) , and **specificity** (ability to find all negative cases).
- **As a Regression Problem (Alternative):**
  - **Model Output:** The model would output a continuous numerical value, such as the predicted final price.
  - **Evaluation Metrics:** Performance would be measured by the error, or **residual**, which is the distance between the predicted price and the actual selling price.

2. Handling Missing BMI and Glucose Data

To handle missing values for quantitative (numeric) attributes like BMI and fasting glucose, several data remediation strategies can be employed.

The choice of method depends on the nature of the data and the number of missing records:

- **Dropping Records:** This strategy involves eliminating any record that has a missing value. This is most appropriate **if the proportion of data elements with missing values is "within a tolerable limit"** and the "quantum of data left... is sizeable". If too many records are dropped, it can reduce the dataset's utility.
- **Mean Imputation:** This strategy involves replacing the missing value with the **mean** (average) of all the remaining, non-missing values for that attribute. The mean is "likely to get shifted drastically" by outliers. Therefore, mean imputation is most appropriate **when the data distribution is relatively symmetric (not skewed) and does not contain significant outliers**.

- **Median Imputation:** This strategy involves replacing the missing value with the **median** (the middle value of an ordered list) of the remaining values. The median is not easily affected by outliers. This method is most appropriate **when the data distribution is skewed or when outliers are present**, which is common in clinical measurements like BMI and glucose.

3. Spam Classifier Modeling Error

The model achieves 48% accuracy on training data and 50% on test data. On a binary problem like spam detection, this performance is equivalent to random guessing.

This is a clear case of **underfitting**. Underfitting occurs when a data model "is unable to capture the relationship between the input and output variables accurately".. This results in "poor performance with training data and test data", as seen here.

Several steps could be taken to improve performance:

1. **Use More Training Data:** The model may not have enough data to learn the patterns.
2. **Feature Engineering:** The current features (e.g., word counts) may be insufficient. New, more relevant features could be created. The source also suggests "Reducing features by effective feature selection" as a prevention step.
3. **Algorithm Selection:** The chosen algorithm (e.g., a simple linear model) may be too simple. A more complex model, such as a Naïve Bayes classifier (common for text), Decision Tree, or Support Vector Machine, could be selected.
4. **Hyperparameter Tuning:** The model's parameters may be poorly tuned, for example, by "Stop[ping] growing the tree before it reaches perfection" (pre-pruning). Adjusting these parameters, or "Tuning the model", can improve performance.

4. One-Hot Encoding for 'Department' Column

One-hot encoding is a preprocessing step used to convert nominal categorical data into a numeric format for machine learning models. It creates new binary (0 or 1) columns for each unique category.

Original DataFrame:

Name	Department
Alice	HR
Bob	Engineering
Charlie	HR
Dana	Sales

Transformed Table after One-Hot Encoding:

Name	Department_HR	Department_Engineering	Department_Sales
Alice	1	0	0
Bob	0	1	0
Charlie	1	0	0
Dana	0	0	1

5. K-NN Prediction for Student E (k=3)

The k-Nearest Neighbour (KNN) algorithm predicts the class of a new data point based on the "predominantly present" class label of its 'k' nearest neighbors. Distances are commonly measured using Euclidean distance.

Task: Predict the label for Student E (6, 7) using k=3.

1. **Calculate Euclidean Distances**:
    - Distance to A (8, 6) [Pass]:
      $\sqrt{(6-8)^2 + (7-6)^2} = \sqrt{(-2)^2 + 1^2} = \sqrt{4+1} = \sqrt{5} \approx 2.236$
    - Distance to B (5, 4) [Fail]:
      $\sqrt{(6-5)^2 + (7-4)^2} = \sqrt{1^2 + 3^2} = \sqrt{1+9} = \sqrt{10} \approx 3.162$
    - Distance to C (7, 5) [Pass]:
      $\sqrt{(6-7)^2 + (7-5)^2} = \sqrt{(-1)^2 + 2^2} = \sqrt{1+4} = \sqrt{5} \approx 2.236$
    - Distance to D (3, 2) [Fail]:
      $\sqrt{(6-3)^2 + (7-2)^2} = \sqrt{3^2 + 5^2} = \sqrt{9+25} = \sqrt{34} \approx 5.831$
2. Identify k=3 Nearest Neighbors:
   The three closest neighbors are:
    - Student A (Distance $\approx$ 2.236) - **Pass**
    - Student C (Distance $\approx$ 2.236) - **Pass**
    - Student B (Distance $\approx$ 3.162) - **Fail**
3. Predict Label:
   The labels of the 3-nearest neighbors are {Pass, Pass, Fail}. The "predominantly present" label is Pass.

6. PCA and Dimensionality Reduction

Principal Component Analysis (PCA) is a statistical technique used for **dimensionality reduction**.

Its goal is to reduce the number of features (dimensions) in a dataset while retaining as much information as possible. It does this by transforming the original, often correlated, variables into a new set of uncorrelated variables called **principal components**.

These new components are linear combinations of the original variables and are "orthogonal to each other". PCA selects these components in a specific order: the first principal component is the one that "capture[s] the maximum amount of variability in the data". The second component captures the next most variance, and so on.

By keeping only the first few principal components (e.g., those corresponding to the largest eigenvalues of the covariance matrix), PCA can "significantly reduce... computational cost" while retaining most of the data's original variance.

7. Logistic Regression for Customer Churn

Transformation:

Logistic regression predicts the probability of a categorical outcome (e.g., churn=1, no-churn=0). It uses the logistic function (also known as the sigmoid function) to transform the linear combination of input features (tenure, monthly charges, support calls). This function is used because it "always takes values between zero and one" , effectively converting the linear model's output into a probability.

Coefficient Interpretation:

The model predicts the probability that Y=1 (customer churns). A positive coefficient for 'support calls' indicates a "strong positive linear relationship" between that feature and the outcome. This means that as the number of 'support calls' (X) increases, the probability that the customer will churn (Y=1) also increases.

8. Disease-Prediction Model Evaluation

(a) Compute Metrics:

Based on the confusion matrix:

- **True Positive (TP):** 50
- **False Negative (FN):** 20
- **False Positive (FP):** 15
- **True Negative (TN):** 85
- **Total:** 50 + 20 + 15 + 85 = 170

The performance metrics are:

- **Accuracy:** $\frac{TP+TN}{Total} = \frac{50+85}{170} = \frac{135}{170} \approx 79.4\%$
- **Precision:** $\frac{TP}{TP+FP} = \frac{50}{50+15} = \frac{50}{65} \approx 76.9\%$
- **Recall (Sensitivity):** $\frac{TP}{TP+FN} = \frac{50}{50+20} = \frac{50}{70} \approx 71.4\%$
- **Specificity:** $\frac{TN}{TN+FP} = \frac{85}{85+15} = \frac{85}{100} = 85.0\%$
- **F1-Score (F-measure):** $\frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times 0.769 \times 0.714}{0.769 + 0.714} \approx 0.740$

(b) Justify Metric Choice:

For a critical health monitor, **Recall (Sensitivity) is more important**.

Justification: Recall measures the "proportion of TP examples or positive cases which were correctly classified". In this medical context, it represents the model's ability to correctly identify patients who *actually have the disease*.

A **False Negative (FN)** (where the model predicts 'Negative' but the patient is 'Positive') means the disease goes undetected. This is a highly critical and dangerous error. Therefore, minimizing false negatives by maximizing recall is the primary goal, even if it means increasing false positives (precision).

9. Smartphone Resale Price: Regression or Classification?

This problem is a **regression** task.

Justification:

The goal is to predict the "final resale price". Price is a continuous value, not a discrete category. Regression models are used to predict a continuous value. If the task were to predict a category (e.g., 'High Price', 'Medium Price', 'Low Price'), it would be a classification problem.

Comparison of Formulations:

- **As a Regression Problem (Chosen):**
    - **Model Output:** The model would output a continuous numerical value representing the predicted price.
    - **Evaluation Metrics:** Performance would be evaluated based on the "marginal or residual error", which is the difference between the predicted price and the actual price.
- **As a Classification Problem (Alternative):**
    - **Model Output:** The model would output a discrete class label (e.g., 'Good Deal').
    - **Evaluation Metrics:** Performance would be measured using metrics like **accuracy**, **precision**, and **recall**.

10. Handling Missing Clinical Trial Data

To handle missing values for quantitative (numeric) attributes like serum creatinine and hemoglobin, several data remediation strategies can be used.

The choice of method depends on the data's distribution and clinical context:

- **Record Deletion (Eliminate records):** This strategy involves removing the records of patients with missing data. This is only appropriate **if the proportion of missing data is "within a tolerable limit"** and the remaining dataset is "sizeable". In clinical trials, this is often undesirable as it reduces statistical power.
- **Mean Imputation:** This involves replacing the missing value with the **mean** (average) of all other patients for that attribute. The mean is "likely to get shifted drastically" by outliers. This method is most appropriate **when the data is normally distributed and does not have extreme outliers**.
- **Median Imputation:** This involves replacing the missing value with the **median** (the 50th percentile) of the remaining values. The median is robust to outliers. This method is most appropriate **when the clinical data is skewed or contains outliers** (e.g., a few patients with extremely high creatinine levels), as it will provide a more representative "central" value.

11. Sentiment Classifier Modeling Error

The model achieves 54% accuracy on the training set and 56% on the validation set. This performance is extremely low, barely better than a 50/50 random guess.

This indicates **underfitting**. The model "is unable to capture the relationship between the input and output variables accurately" , leading to "poor performance with training data and test data".

To improve performance, you could use the following techniques:

1. **Feature Extraction:** The current text features may be poor. Refining them (e.g., using different n-grams, TF-IDF) or "Reducing features by effective feature selection" might help.
2. **Algorithm Selection:** The chosen algorithm may be too simple. Selecting a different model (e.g., Naïve Bayes, which is strong for text classification) is a key step.
3. **Hyperparameter Tuning:** The model's parameters may be limiting its ability to learn. "Tuning the model" is a necessary step for performance improvement.
4. **Use More Training Data:** The model may not have seen enough examples to learn the patterns in customer reviews.

12. One-Hot Encoding for 'Category' Column

One-hot encoding is a preprocessing step used to convert nominal categorical data into a numeric format for machine learning models. It creates new binary (0 or 1) columns for each unique category.

Original DataFrame:

Item	Category
Pen	Stationery
Notebook	Office Supplies
Eraser	Stationery
Marker	Art

Transformed Table after One-Hot Encoding:

Item	Category_Stationery	Category_Office Supplies	Category_Art
Pen	1	0	0
Notebook	0	1	0
Eraser	1	0	0
Marker	0	0	1

13. K-NN Prediction for Student P5 (k=3)

The k-Nearest Neighbour (KNN) algorithm predicts the class of a new data point based on the "predominantly present" class label of its 'k' nearest neighbors. Distances are commonly measured using Euclidean distance.

Task: Predict the label for Student P5 (7, 4) using k=3.

1. **Calculate Euclidean Distances**:
    - Distance to P1 (9, 5) [Pass]:
      $\sqrt{(7-9)^2 + (4-5)^2} = \sqrt{(-2)^2 + (-1)^2} = \sqrt{4+1} = \sqrt{5} \approx 2.236$
    - Distance to P2 (4, 3) [Fail]:
      $\sqrt{(7-4)^2 + (4-3)^2} = \sqrt{3^2 + 1^2} = \sqrt{9+1} = \sqrt{10} \approx 3.162$
    - Distance to P3 (8, 6) [Pass]:
      $\sqrt{(7-8)^2 + (4-6)^2} = \sqrt{(-1)^2 + (-2)^2} = \sqrt{1+4} = \sqrt{5} \approx 2.236$
    - Distance to P4 (2, 4) [Fail]:
      $\sqrt{(7-2)^2 + (4-4)^2} = \sqrt{5^2 + 0^2} = \sqrt{25} = 5$
2. Identify k=3 Nearest Neighbors:
   The three closest neighbors are:
    - Student P1 (Distance $\approx$ 2.236) - **Pass**

    - Student P3 (Distance $\approx$ 2.236) - **Pass**
    - Student P2 (Distance $\approx$ 3.162) - **Fail**
3. Predict Label:
   The labels of the 3-nearest neighbors are {Pass, Pass, Fail}. The "predominantly present" label is Pass.

14. PCA and Principal Component Selection

Principal Component Analysis (PCA) is a **dimensionality reduction** technique. It is a statistical method used to convert a set of correlated variables into a new set of "transformed, uncorrelated variables called principal components".

PCA selects these components (axes) by identifying the "linear combination of the original variables" that "capture[s] the maximum amount of variability in the data". The first principal component is the axis that accounts for the most variance. The second principal component, which is "orthogonal" (perpendicular) to the first, accounts for the next largest amount of variance, and so on.

By selecting only the top components (those associated with the largest eigenvalues of the covariance matrix), PCA reduces the number of dimensions while retaining the maximum possible information (variance) from the original data.

15. Logistic Regression for Patient Complications

Transformation:

Logistic regression uses the logistic function (also known as the sigmoid function). This function's role is to take the linear predictor (the weighted sum of age, BMI, and blood sugar) and convert it into a probability. This output "always takes values between zero and one" , which is necessary for representing the probability that the patient develops complications (Y=1).

Coefficient Interpretation:

A negative coefficient for 'BMI' indicates a negative linear relationship between BMI and the log-odds of the outcome. This means that as the value of 'BMI' (X) increases, the probability that the patient will develop complications (Y=1) decreases.

16. Fraud-Detection Model Evaluation

(a) Compute Metrics:

Based on the confusion matrix:

- **True Positive (TP):** 40
- **False Negative (FN):** 10

- **False Positive (FP):** 20
- **True Negative (TN):** 130
- **Total:** 40 + 10 + 20 + 130 = 200

The performance metrics are:

- **Accuracy:** $\frac{TP+TN}{Total} = \frac{40+130}{200} = \frac{170}{200} = 85.0\%$
- **Precision:** $\frac{TP}{TP+FP} = \frac{40}{40+20} = \frac{40}{60} \approx 66.7\%$
- **Recall (Sensitivity):** $\frac{TP}{TP+FN} = \frac{40}{40+10} = \frac{40}{50} = 80.0\%$
- **Specificity:** $\frac{TN}{TN+FP} = \frac{130}{130+20} = \frac{130}{150} \approx 86.7\%$
- **F1-Score (F-measure):** $\frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times 0.667 \times 0.80}{0.667 + 0.80} \approx 0.727$

(b) Justify Metric Choice:

For real-time fraud alerts, **Recall (Sensitivity) is more critical**.

Justification: Recall measures the "proportion of TP examples... which were correctly classified". In this context, it is the proportion of *actual fraud transactions* that the model *correctly identified*.

A **False Negative (FN)** (where the model predicts 'Legitimate' but the transaction is 'Fraud') means a fraudulent charge is successful. This is the most costly and critical error for the bank and the customer. Therefore, maximizing recall to minimize false negatives is the priority, even if it means lowering precision (i.e., flagging more legitimate transactions for review, which are False Positives).

17. Rental Apartment Price: Regression or Classification?

This problem is a **classification** task.

Justification:

The goal is to predict "whether an apartment will rent above the median market rate." The target feature is a binary, categorical value (e.g., 'Above Median' or 'Below/At Median'). Classification is used when the target feature is categorical. If the goal were to predict the exact rental price (a continuous value), it would be a regression problem.

Comparison of Formulations:

- **As a Classification Problem (Chosen):**
    - **Model Output:** The model would output a discrete class label (e.g., 'Above Median').
    - **Evaluation Metrics:** Performance would be measured using metrics like **accuracy**, **precision**, **recall/sensitivity**, and **specificity**.

- **As a Regression Problem (Alternative):**
    - **Model Output:** The model would output a continuous numerical value, such as the predicted rental price.
    - **Evaluation Metrics:** Performance would be measured by the "marginal or residual error", the difference between the predicted and actual prices.

18. Handling Missing Cardiology Data

To handle missing values for quantitative (numeric) attributes like cholesterol and triglycerides, several data remediation techniques can be applied.

The most suitable method depends on the data's characteristics:

- **Removing Records (Eliminate records):** This involves deleting the records of participants with missing data. This is generally the least preferred method in clinical studies as it reduces the dataset size, but it is appropriate **if the proportion of missing data is "within a tolerable limit"** and the remaining "quantum of data... is sizeable".
- **Mean Imputation:** This replaces the missing value with the **mean** (average) of all other participants for that attribute. This method is most suitable **under the condition that the data is normally distributed and does not contain significant outliers**, as the mean is "likely to get shifted drastically" by extreme values.
- **Median Imputation:** This replaces the missing value with the **median** (the middle value) of the remaining data. Because clinical data like cholesterol and triglycerides can often be skewed or contain outliers (e.g., a few participants with very high levels), the median is often the most suitable choice as it is robust to these extreme values.

19. Sentiment Analysis Model Error

The model records 55% accuracy on its training corpus and 57% on a validation set. This performance is extremely poor, indicating the model has failed to learn any meaningful patterns.

This is a state of **underfitting**. An underfit model "is unable to capture the relationship between the input and output variables accurately" , which is demonstrated by the "poor performance with training data and test data".

Corrective actions to boost accuracy include:

1. **Refining Text Features:** The current features may be inadequate. A different approach to feature engineering or "effective feature selection" could be necessary.
2. **Selecting Different Algorithms:** The chosen algorithm may be too simple. "Model selection" (e.g., trying a Naïve Bayes classifier or a more complex model) is a critical step.
3. **Tuning Hyperparameters:** The model's existing parameters might be incorrectly set. "Tuning the model" can optimize its ability to learn.
4. **Use More Training Data:** The model may not have "sufficient training data" to identify

the nuances of sentiment.

20. One-Hot Encoding for 'Genre' Column

One-hot encoding is a preprocessing step used to convert nominal categorical data into a numeric format for machine learning models. It creates new binary (0 or 1) columns for each unique category.

Original DataFrame:

Title	Genre
Inception	SciFi
Titanic	Romance
Joker	Drama
Up	Animation

Transformed Table after One-Hot Encoding:

Title	Genre_SciFi	Genre_Romance	Genre_Drama	Genre_Animation
Inception	1	0	0	0
Titanic	0	1	0	0
Joker	0	0	1	0
Up	0	0	0	1

21. K-NN Prediction for Student X5 (k=3)

The k-Nearest Neighbour (KNN) algorithm predicts the class of a new data point based on the "predominantly present" class label of its 'k' nearest neighbors. Distances are commonly measured using Euclidean distance.

Task: Predict the label for Student X5 (6, 6) using k=3.

1. **Calculate Euclidean Distances**:
    - Distance to X1 (7, 8) [Pass]:
      $\sqrt{(6-7)^2 + (6-8)^2} = \sqrt{(-1)^2 + (-2)^2} = \sqrt{1+4} = \sqrt{5} \approx 2.236$
    - Distance to X2 (4, 3) [Fail]:
      $\sqrt{(6-4)^2 + (6-3)^2} = \sqrt{2^2 + 3^2} = \sqrt{4+9} = \sqrt{13} \approx 3.606$
    - Distance to X3 (9, 6) [Pass]:

$\sqrt{(6-9)^2 + (6-6)^2} = \sqrt{(-3)^2 + 0^2} = \sqrt{9} = 3$
    - Distance to X4 (2, 5) [Fail]:
      $\sqrt{(6-2)^2 + (6-5)^2} = \sqrt{4^2 + 1^2} = \sqrt{16+1} = \sqrt{17} \approx 4.123$
2. Identify k=3 Nearest Neighbors:
   The three closest neighbors are:
    - Student X1 (Distance $\approx$ 2.236) - **Pass**
    - Student X3 (Distance = 3) - **Pass**
    - Student X2 (Distance $\approx$ 3.606) - **Fail**
3. Predict Label:
   The labels of the 3-nearest neighbors are {Pass, Pass, Fail}. The "predominantly present" label is Pass.

22. PCA and Axis Selection

Principal Component Analysis (PCA) is a **dimensionality reduction** technique. It is a statistical method used to convert a set of correlated variables into a new set of "transformed, uncorrelated variables called principal components".

PCA selects these new axes (components) by identifying the "linear combination of the original variables" that "capture[s] the maximum amount of variability in the data". The first principal component is the axis that accounts for the most variance. The second principal component, which is "orthogonal" (perpendicular) to the first, accounts for the next largest amount of variance, and so on.

By selecting only the top components (those associated with the largest eigenvalues of the covariance matrix), PCA reduces the number of dimensions while retaining the maximum possible information (variance) from the original data.

23. Logistic Regression for Subscriber Cancellation

Transformation:

Logistic regression uses the logistic function (also known as the sigmoid function). This function is essential because it takes the linear predictor (the weighted sum of usage frequency, subscription length, and customer service calls) and transforms it into a probability. This output "always takes values between zero and one" , which is required to represent the probability that a subscriber cancels (Y=1).

Coefficient Interpretation:

A "large positive coefficient" for 'customer service calls' indicates a "strong positive linear relationship" between that feature and the outcome. This means that as the number of

'customer service calls' (X) increases, the probability that the subscriber will cancel (Y=1) also increases significantly.

24. Credit Card Fraud Detector Evaluation

(a) Compute Metrics:

Based on the confusion matrix:

- **True Positive (TP):** 45
- **False Negative (FN):** 5
- **False Positive (FP):** 25
- **True Negative (TN):** 125
- **Total:** 45 + 5 + 25 + 125 = 200

The performance metrics are:

- **Accuracy:** $\frac{TP+TN}{Total} = \frac{45+125}{200} = \frac{170}{200} = 85.0\%$
- **Precision:** $\frac{TP}{TP+FP} = \frac{45}{45+25} = \frac{45}{70} \approx 64.3\%$
- **Recall (Sensitivity):** $\frac{TP}{TP+FN} = \frac{45}{45+5} = \frac{45}{50} = 90.0\%$
- **Specificity:** $\frac{TN}{TN+FP} = \frac{125}{125+25} = \frac{125}{150} \approx 83.3\%$
- **F1-Score (F-measure):** $\frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times 0.643 \times 0.90}{0.643 + 0.90} \approx 0.750$

(b) Justify Metric Choice:

For an automated fraud alert system, **Recall (Sensitivity) is more crucial**.

Justification: Recall measures the "proportion of TP examples... which were correctly classified". In this context, it represents the proportion of *actual fraud transactions* that the system *correctly identified*.

A **False Negative (FN)** (where the model predicts 'Legit' but the transaction is 'Fraud') means a fraudulent charge is approved and goes undetected. This is the most critical and costly error. Therefore, maximizing recall to minimize the number of missed frauds is the highest priority, even if it comes at the cost of lower precision (i.e., flagging more legitimate transactions as alerts, which are False Positives).