Answer the following questions briefly (in one line).

Q1. Define Feature Engineering in the context of improving model accuracy.
Q2. True or False: Filter methods for feature selection rely on the specific machine learning algorithm (e.g., Random Forest) to select the best features.
Q3. A "Rolling Mean" is a feature extraction technique primarily used for what type of data?
Q4. In image processing, what specific structural information does Edge Detection capture?
Q5. Why is One-Hot Encoding generally *not* recommended for a "Zip Code" column with 30,000 unique values?
Q6. What is the purpose of a Logarithmic Transformation on a highly skewed "Income" feature?
Q7. Name one automated tool mentioned in the syllabus for Time-Series feature extraction.
Q8. True or False: PCA (Principal Component Analysis) preserves the original meaning of the features (e.g., "Age", "Salary").
Q9. In text analysis, what does TF-IDF penalize to highlight important words?
Q10. Give one example of a domain-specific feature you might calculate for a financial trading dataset (e.g., stock prices).

Answer the following questions based on the provided scenarios. Justify your choice of technique.

Q11. You are predicting house prices. The "Lot Area" feature mostly ranges from 500 to 5,000 sq ft, but contains three valid but extreme outliers (mansions) with 500,000 sq ft.
Identify the most appropriate feature scaling technique for this specific distribution.
Justify why is it a better choice compared to the other techniques.
Q12. You have a dataset with only 200 patients (rows) but 50,000 gene features (columns). You need to select the top 50 relevant genes. Which class of feature selection method (Filter or Wrapper) is practically feasible here? Justify your answer.
Q13. You are training a model to distinguish between photos of wheels and photos of bricks. The model using raw pixel colors is failing because the object colors vary. Recommend a specific feature descriptor/extraction technique. Justify your choice.
Q14. You have a dataset containing only the "Closing Price" of a stock for the last 365 days. You cannot use the "Date" column directly in a regression model.
Propose a specific temporal feature engineering technique that utilizes past data (e.g., previous days' prices) to create a predictive feature for tomorrow's price. Explain the logic of your chosen feature.

Answer the following questions in detail, addressing all parts.

Q15. You have a dataset of 28x28 pixel images of handwritten digits (0-9). You want to build a classifier (e.g., SVM).

1. The raw data has 784 features (pixels) per image. Would applying PCA (Principal Component Analysis) be useful before training the SVM?
2. Suggest a technique to capture the *structure* or *edges* of the digits. Why is this better than raw pixels?
3. Why would "Min-Max Scaling" the pixel values (0-255) be useful, whereas "Standardization" might be less intuitive for image pixel intensity?

Q16. You are building a spam detector using a dataset of 50,000 emails.

- (a) Which would be more useful - TF-IDF or Boolean Bag-of-Words ?
- (b) Why would using One-Hot Encoding for every unique word in the dataset be a bad idea?