Mock exam:

Question 1:

Which of the following is NOT a type of machine learning task according to “samenvatting.pdf”?

a) Predictive tasks

b) Descriptive tasks

c) Analytical tasks

d) Supervised learning

Question 2:

What is the primary goal of probabilistic models in machine learning, as described in “samenvatting.pdf”?

a) To represent complex relationships using geometric functions.

b) To reduce uncertainty by utilizing probability distributions.

c) To create flowchart-like structures for decision-making.

d) To group data points based on distance metrics.

Question 3:

According to “samenvatting.pdf,” which of the following is an example of a grading model?

a) Tree model

b) Naive Bayes model

c) Linear model

d) Decision tree model

Question 4:

Which feature transformation technique involves converting numerical features into categorical features, as explained in “samenvatting.pdf”?

a) Feature construction

b) Discretisation

c) Feature selection

d) Feature transformation

Question 5:

In the context of binary classification, what does the True Negative Rate (TNR) or Specificity measure?

a) The proportion of correctly identified positive instances out of all positive instances.

b) The proportion of correctly identified negative instances out of all negative instances.

c) The proportion of incorrectly identified positive instances out of all negative instances.

d) The proportion of incorrectly identified negative instances out of all positive instances.

Question 6:
What is the purpose of using a validation set in machine learning model development, as described in “samenvatting.pdf”?

a) To train the initial model parameters.

b) To fine-tune the model’s hyperparameters.

c) To evaluate the final model’s performance.

d) To identify and handle missing data points.

Question 7:
What does the area under the ROC curve (AUC) represent in machine learning model evaluation?

a) The accuracy of the model on the training dataset.

b) A measure of the model’s performance, with values closer to 1 indicating better performance.

c) The degree of overfitting present in the trained model.

d) The computational cost of training the machine learning model.

Question 8:
According to “samenvatting.pdf”, what is the primary difference between inherently binary and inherently non-binary algorithms for multi-class classification?

a) Inherently binary algorithms are more computationally expensive than inherently non-binary algorithms.

b) Inherently non-binary algorithms are specifically designed to handle more than two classes, while inherently binary algorithms need adaptation.

c) Inherently binary algorithms are more prone to overfitting than inherently non-binary algorithms.

d) Inherently non-binary algorithms require a larger amount of training data compared to inherently binary algorithms.

Question 9:
What is the purpose of the “One-vs-One” approach in adapting a binary classifier for multi-class classification?

a) It trains a single classifier to distinguish all classes simultaneously.

b) It trains separate binary classifiers for each possible pair of classes.

c) It converts the multi-class problem into a regression task.

d) It reduces the number of features used to simplify the classification process.

Question 10:
How does “samenvatting.pdf” define residuals in the context of regression analysis?

a) The difference between the predicted values and the actual (true) values.

b) The variance of the data points around the regression line.

c) The slope of the regression line representing the relationship between variables.

d) The coefficient of determination (R-squared) of the regression model.

Question 11:
What is the key characteristic of distance-based algorithms in machine learning, as explained in "samenvatting.pdf"?

a) They utilize probability distributions to model uncertainty.

b) They classify instances by calculating distances to stored exemplars.

c) They build tree-like structures to partition data based on feature values.

d) They rely on finding linear relationships between variables for prediction.

Question 12:
What is the "curse of dimensionality" as described in "samenvatting.pdf" in the context of distance-based models?

a) The phenomenon where distances become less meaningful and informative as the number of dimensions increases.

b) The exponential increase in computational cost with the growth of data features.

c) The difficulty in visualizing high-dimensional data for model interpretation.

d) The problem of overfitting when the number of features exceeds the number of data points.

Question 13:
What is the function of a linkage function in hierarchical clustering, as described in "samenvatting.pdf"?

a) It assigns data points to the nearest cluster centroid.

b) It determines the distance between two clusters.

c) It calculates the silhouette score for cluster evaluation.

d) It visualizes the clusters in a scatter plot.

Question 14:
How does "samenvatting.pdf" define regularization in the context of linear models?

a) A technique to normalize feature values into a standard range.

b) A method to handle missing data points in the dataset.

c) A technique to prevent overfitting by adding a penalty term to the loss function.

d) An approach to transform non-linear data into a higher-dimensional space.

Question 15:
According to "samenvatting.pdf", what is the primary goal of a Support Vector Machine (SVM)?

a) To minimize the sum of squared errors between predicted and actual values.

b) To find the hyperplane that maximizes the margin between different classes.

c) To cluster data points based on their distances from each other.

d) To build a tree-like structure for hierarchical classification.

Question 16:
Which statistical measure is less sensitive to outliers in a dataset: the mean or the median?

a) Mean

b) Median

Question 17:

What does a positive skewness value indicate about the distribution of a dataset, as explained in "samenvatting.pdf"?

a) The data is skewed to the right, with a longer tail on the right side of the distribution.

b) The data is skewed to the left, with a longer tail on the left side of the distribution.

c) The data is symmetrically distributed around the mean.

d) The data has high kurtosis, indicating heavy tails.

Question 18:

According to "samenvatting.pdf," what is the primary purpose of one-hot encoding in feature engineering?

a) To normalize numerical feature values to a standard scale.

b) To convert categorical data into a numerical representation using binary vectors.

c) To reduce the number of features by selecting the most relevant ones. d) To handle missing data points using imputation techniques.

Question 19:

In the context of evaluating multiple machine learning algorithms on a single dataset, what statistical test is recommended in "samenvatting.pdf"?

a) Paired t-test

b) Wilcoxon's signed-rank test

c) Friedman test

d) Nemenyi test

Question 20:

What is the primary goal of AutoML (Automated Machine Learning), as described in "samenvatting.pdf"?

a) To develop new machine learning algorithms from scratch.

b) To interpret and explain the predictions of black-box models.

c) To automate the process of applying machine learning to real-world problems.

d) To collect, clean, and prepare data for machine learning tasks.

Answer Sheet:

1. **c)** Analytical tasks
2. **b)** To reduce uncertainty by utilizing probability distributions.
3. **c)** Linear model
4. **b)** Discretisation
5. **b)** The proportion of correctly identified negative instances out of all negative instances.
6. **b)** To fine-tune the model's hyperparameters.
7. **b)** A measure of the model's performance, with values closer to 1 indicating better performance.
8. **b)** Inherently non-binary algorithms are specifically designed to handle more than two classes, while inherently binary algorithms need adaptation.
9. **b)** It trains separate binary classifiers for each possible pair of classes.
10. **a)** The difference between the predicted values and the actual values.
11. **b)** They classify instances by calculating distances to stored exemplars.
12. **a)** The phenomenon where distances become less meaningful and informative as the number of dimensions increases.
13. **b)** It determines the distance between two clusters.
14. **c)** A technique to prevent overfitting by adding a penalty term to the loss function.
15. **b)** To find the hyperplane that maximizes the margin between different classes.
16. **b)** Median
17. **a)** The data is skewed to the right, with a longer tail on the right side of the distribution.
18. **b)** To convert categorical data into a numerical representation using binary vectors.
19. **a)** Paired t-test
20. **c)** To automate the process of applying machine learning to real-world problems