

# **Data Science and Machine Learning**

Vienna University of Economics and Business  
Januray 23rd, 2026

Name: \_\_\_\_\_

Student ID: \_\_\_\_\_

| Question     | Points |
|--------------|--------|
| <b>1</b>     | 35     |
| <b>2</b>     | 20     |
| <b>3</b>     | 20     |
| <b>4</b>     | 25     |
| <b>Total</b> | 100    |

## Question 1: Multiple Choice

Which statements about confusion matrices are true? [5 pts]

- ☐ Overall accuracy is a reliable measure for judging the performance of a classification model.
- ☐ Decreasing the value of the cut-off point will increase sensitivity but decrease specificity.
- ☐ Decreasing the value of the cut-off point will increase sensitivity but decrease precision.
- ☐ Cohens Kappa indicates how high the trade-off between sensitivity and specificity is.
- ☐ Increasing the value of the cut-off point will increase both precision and sensitivity.

Which statements about the Grammar of Graphics are true? [5 pts]

- ☐ The Grammar of Graphics is a software library built upon D3.
- ☐ Aesthetics map variables in the data to visual properties such as position, color or size.
- ☐ Aesthetics refer to stylistic and decorative design choices.
- ☐ Scales control how data values are transformed into aesthetic values (e.g. logarithmic axes, color gradients).
- ☐ The Grammar of Graphics divides up charts into a fixed set of types (e.g. scatter plot, bar chart, isotype chart).

Which statements about webscraping are true? [5 pts]

- ☐ A websites terms of service may prohibit scraping.
- ☐ Inspecting network requests in the browser can reveal underlying data sources without scraping the HTML.
- ☐ HTML structure changes can break webscraping scripts.
- ☐ Websites may block IP addresses that send too many requests (HTTP error 429).
- ☐ Executing JavaScript in the browser can help access data that is not rendered directly in the DOM.

Which statements about Classification Trees are true? [5 pts]

- ☐ Trees are robust to small changes in the training data.
- ☐ Trees cannot handle missing values in predictors.
- ☐ Random forests are ensembles of classification trees.
- ☐ A split in a classification tree is chosen to maximize the homogeneity of the resulting nodes.
- ☐ Classification trees can only be used for binary target variables.

Which statements about ROC and PR curves are true? [5 pts]

- ☐ The area under the ROC and PR curves provide reliable measures for judging the performance of a classification model.
- ☐ The ROC curve is especially informative for imbalanced datasets.
- ☐ The PR curve is especially informative for imbalanced datasets.
- ☐ ROC and PR curves cannot be used to compare the performance of different types of models (e.g. neural networks and classification trees)
- ☐ Each different point on the ROC and PR curves represent a different confusion matrix for a different cut-off point.

Which statements about Neural Networks are true? [5 pts]

- ☐ Neural networks achieve lower MSEs than other estimators even in data sets with a low number of observations.
- ☐ Adding more layers generally improves model performance.
- ☐ The activation function introduces non-linearity into a neural network.
- ☐ The higher the learning rate, the faster the convergence toward the optimum.
- ☐ The bias term shifts the weighted sum before it is passed to the activation function.

Which statements about Statistical Matching are true? [5 pts]

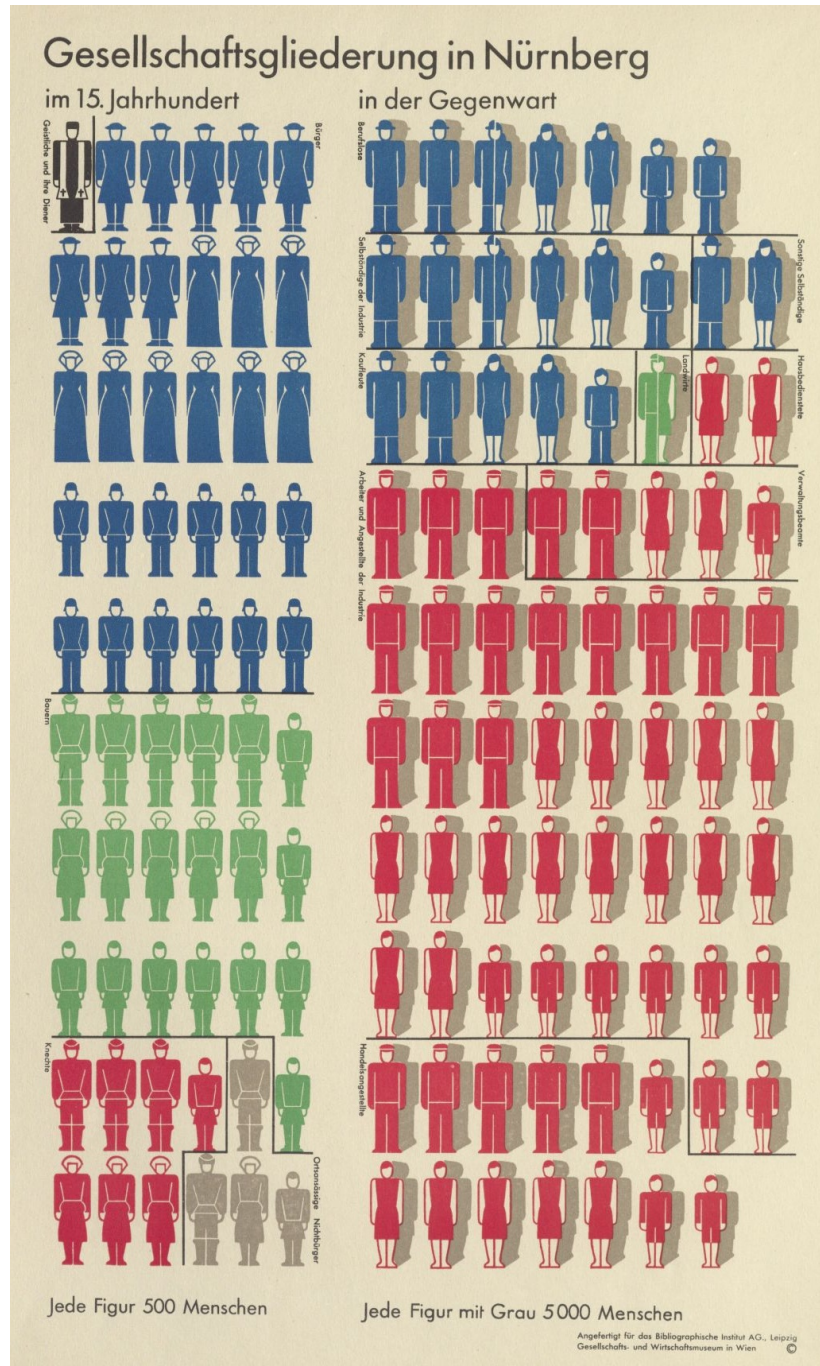
- ☐ Continuous variables can not be used to define donor classes in hot deck or statistical matching procedures.
- ☐ Hot deck matching methods impute missing values in a recipient dataset by using observed values from similar donor records in the donor dataset.
- ☐ Predictive mean matching is a semi-parametric method that combines model-based prediction with donor-based hot deck-style imputation.
- ☐ For statistical matching, it is generally preferable that the donor dataset contains less observations than the recipient dataset.
- ☐ Under the conditional independence assumption, the joint distribution of  $Y$  and  $Z$  can be modeled using  $f(Y | X)$ ,  $f(Z | X)$ , and  $f(X)$ .

## Question 2: The Grammar of Graphics

(a) Explain in one sentence what the following concepts mean in the context of the Grammar of Graphics. [12 pts]

- aesthetics
- geometries
- scales

(b) What are the aesthetics in the plot below? Specify at least four as a mapping data dimension → visual dimension [8 pts]



### **Question 3: Neural Networks**

What is the structure and functioning of an artificial neural network? Draw a sketch and try to cover the following questions: [15 pts]

- What are the differences between the different layers and how do they relate to each other?
- What is the role of weights and biases? Where are they located in the graph?

What is backpropagation in Artificial Neural Networks? How is it connected to forward propagation and gradient descent? Explain. [5 pts]

**Question 4: Cross Validation and the Bias-Variance Trade-Off**

- (a) Explain what the bias-variance trade-off is and how it relates to cross validation. (Hint: State the decomposition of the MSE and draw a graph of performance over complexity of a model) [15 pts]
- (b) How does this relate to shrinkage methods such as LASSO and Ridge Regression? [10 pts]