University of Cyprus
Department of Computer Science

DSC510: Introduction to Data Science and Analytics
Final Exam

Name/Surname:
ID:
Please try to answer all the subjects. Good Luck!

Exercise 1. Select the correct answers. No justification is required. (30 points, 3 points per question)

What does a histogram visualize?
a. Relationships between two variables
b. Distribution of a single variable
c. Hierarchical data structures
d. Time-series data

What is the purpose of correlation in data analysis?
a. Causation between variables
b. Describing the spread of data
c. Measuring the strength and direction of a linear relationship
d. Identifying outliers

What is the purpose of a box plot in data analysis?
a. Representing the distribution of a single variable
b. Showing the relationship between two variables
c. Identifying outliers and displaying the spread of data
d. Displaying hierarchical relationships

What is the primary goal of hypothesis testing in data analysis?
a. Describing the central tendency of data
b. Identifying outliers
c. Making predictions
d. Drawing inferences about populations based on sample data

How does the term "principal component analysis" (PCA) contribute to dimensionality reduction in data analysis?
a. By clustering similar data points together
b. By transforming features into a lower-dimensional space
c. By creating new features based on existing ones
d. By evaluating feature importance

Explain the concept of "resampling" in data analysis.
a. Aggregating data based on specified criteria
b. Replacing missing values with interpolated values
c. Repeatedly drawing samples from a dataset for statistical analysis
d. Removing outliers from a dataset

How does the term "Ensemble Learning" improve model performance in machine learning?
a. Reducing model complexity
b. Combining predictions from multiple models
c. Handling outliers by giving less weight to extreme values
d. Ensuring that features contribute equally to a model

What is the purpose of the term "Hierarchical Clustering" in clustering analysis?

a. Assessing the correlation between clusters
b. Creating a hierarchy of clusters based on similarities
c. Identifying outliers in clustered data
d. Measuring the similarity within clusters

Which of the following is an example of an unsupervised learning algorithm?

a. Linear Regression
b. K-Means Clustering
c. Decision Trees
d. Support Vector Machines

What is cross-validation used for in machine learning?

a. Cross-training different models
b. Evaluating model performance on multiple datasets
c. Selecting hyperparameters
d. Testing a model's generalization ability

Exercise 2. (30 points, 6 points per question)

You are a data scientist working for a healthcare organization. Your task is to analyze the causal relationship between the implementation of a new drug (Drug X) and its effect on patient recovery time. The organization has conducted a study where two groups of patients were observed:

- **Group 1**: A randomized controlled trial (RCT) where patients were randomly assigned to receive either Drug X or a placebo.
- **Group 2**: An observational study where patients voluntarily chose to take Drug X or not.

The data for each group includes the following variables:

- **Recovery Time** (dependent variable): The number of days it took for patients to fully recover from their condition.
- **Drug X** (independent variable): Whether or not the patient received Drug X (binary: 1 if the patient received Drug X, 0 if the patient received a placebo or no treatment).
- **Age**: Age of the patient.
- **Gender**: Gender of the patient (binary: 1 if male, 0 if female).
- **Pre-existing Conditions**: Whether the patient had any pre-existing conditions that could affect recovery (binary: 1 if yes, 0 if no).
- **Socioeconomic Status (SES)**: Socioeconomic status of the patient (ordinal: low, medium, high).

1. Describe the key differences between the randomized controlled trial (RCT) and the observational study in terms of causal inference. Why is causality easier to establish in the RCT compared to the observational study?

2. Use a directed acyclic graph (DAG) to represent the potential causal relationships between the variables: Recovery Time, Drug X, Age, Gender, Pre-existing Conditions, and Socioeconomic Status. Indicate any confounding variables and describe how these might impact the estimation of the causal effect of Drug X on Recovery Time.

3. Assume you want to estimate the causal effect of Drug X on Recovery Time using the data from the observational study (Group 2). What methods would you use to deal with confounding? Explain their advantages and limitations.

4. Suppose you perform a regression analysis and find that the coefficient for Drug X is negative and statistically significant in both the RCT and the observational study. In the observational study, however, you observe that patients with higher SES are more likely to choose to take Drug X. What can you conclude about the causal effect of Drug X on Recovery Time based on this result? What additional analyses would you perform to strengthen your causal claims?

5. Given that the RCT is considered a gold standard for causal inference, discuss the limitations of using observational data in causal analysis. What are the challenges involved in interpreting results from observational studies in data science, particularly in the presence of confounding or selection bias?

Exercise 3 (20points)
You have a dataset with 500 records to predict Income (annual income in thousands of dollars) based on the following features:

- Age (continuous)
- Education Level (categorical: 0 = High School, 1 = Bachelor's, 2 = Master's, 3 = PhD)
- Experience (continuous)
- Gender (binary: 0 = Female, 1 = Male)

1. Consider using **Age** alone in a **linear regression** model to predict **Income**. How would this affect model performance in terms of underfitting or overfitting? (5 points)

2. If the **linear regression** model shows low training error but high test error, what would you conclude about bias and variance? What changes would you make to the model? (5 points)

3. If you move to a more complex model, like **random forest**, what challenges in terms of variance might arise? How would you address them, and what impact would regularization have? (10 points)

Exercise 4 (20 points, 5 points per question)
You are working with a dataset containing sensitive information about individuals. The dataset includes the following columns:

- Age (continuous): The age of the individual.
- Gender (binary): 0 for Female, 1 for Male.
- Zip Code (categorical): The individual's postal code (5 digits).
- Income (continuous): The annual income of the individual (in thousands).
- Medical History (categorical): A list of previous medical conditions (e.g., diabetes, hypertension, etc.).

Due to privacy concerns, the data must be anonymized before sharing or using it for analysis, especially since the data will be shared with third parties for healthcare research purposes.

1. Explain how k-anonymity could be applied to this dataset. How would you implement k-anonymity for the variables Age, Gender, Zip Code, and Income? What potential challenges could arise when applying k-anonymity to this dataset?

2. Describe how l-diversity improves upon k-anonymity in protecting sensitive attributes. For this dataset, how could you apply l-diversity to Medical History? What challenges might arise in ensuring that l-diversity is maintained in the dataset?

3. How could differential privacy be applied to this dataset to ensure individual privacy? Provide an example of how differential privacy could be used when analyzing the data (e.g., during query execution or statistical analysis). What are the key advantages of using differential privacy over k-anonymity or l-diversity?

4. Discuss the trade-offs between k-anonymity, l-diversity, and differential privacy in terms of data utility and privacy protection. In what situations would you prefer one method over the others?

SOLUTIONS
Exercise 1.

1. **What does a histogram visualize?**
   b. Distribution of a single variable
2. **What is the purpose of correlation in data analysis?**
   c. Measuring the strength and direction of a linear relationship

3. **What is the purpose of a box plot in data analysis?**
   c. Identifying outliers and displaying the spread of data
   (*a* is partially correct, but **c** is the best answer since it captures both main goals.)
4. **What is the primary goal of hypothesis testing in data analysis?**
   d. Drawing inferences about populations based on sample data
5. **How does PCA contribute to dimensionality reduction?**
   b. By transforming features into a lower-dimensional space
6. **Explain the concept of "resampling" in data analysis.**
   c. Repeatedly drawing samples from a dataset for statistical analysis
7. **How does Ensemble Learning improve model performance?**

   b. Combining predictions from multiple models

8. **What is the purpose of Hierarchical Clustering?**
   b. Creating a hierarchy of clusters based on similarities
9. **Which is an example of an unsupervised learning algorithm?**
   b. K-Means Clustering
10. **What is cross-validation used for in machine learning?**
    d. Testing a model's generalization ability

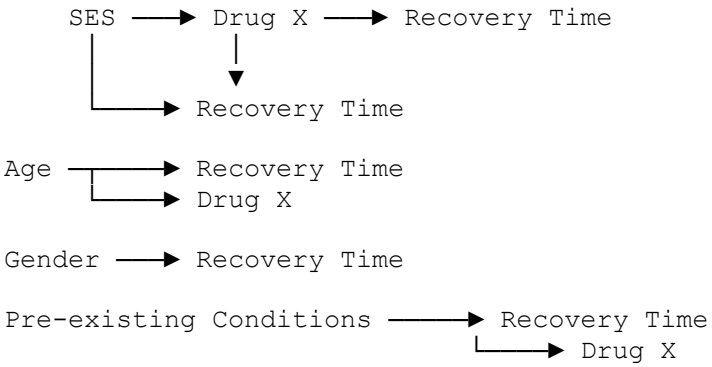
Exercise 2.

1. Differences between the RCT and the Observational Study in terms of causal inference

Aspect	Randomized Controlled Trial (RCT)	Observational Study
Treatment assignment	Random	Based on patient choice
Confounding	Minimized through randomization	Likely present
Causal inference	Strong (high internal validity)	Weak (requires assumptions)
Bias	Reduced	Possible selection bias

Why causality is easier to establish in the RCT:
In an RCT, randomization ensures that both observed and unobserved confounders are equally distributed between treatment and control groups. Thus, any differences in recovery time can be attributed to Drug X. In contrast, observational data is subject to confounding because patients self-select treatment, and their decision may be influenced by factors that also impact recovery (e.g., severity of condition).

2. Causal DAG and Confounding



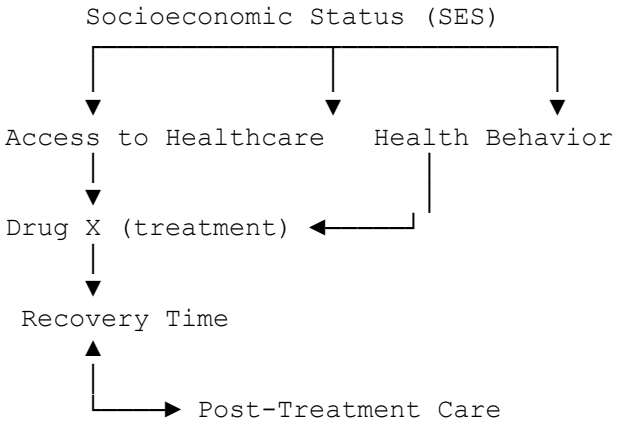
Confounders: Age, Pre-existing Conditions, SES

- They affect both **Drug X** choice and **Recovery Time**, leading to biased estimates if unadjusted.

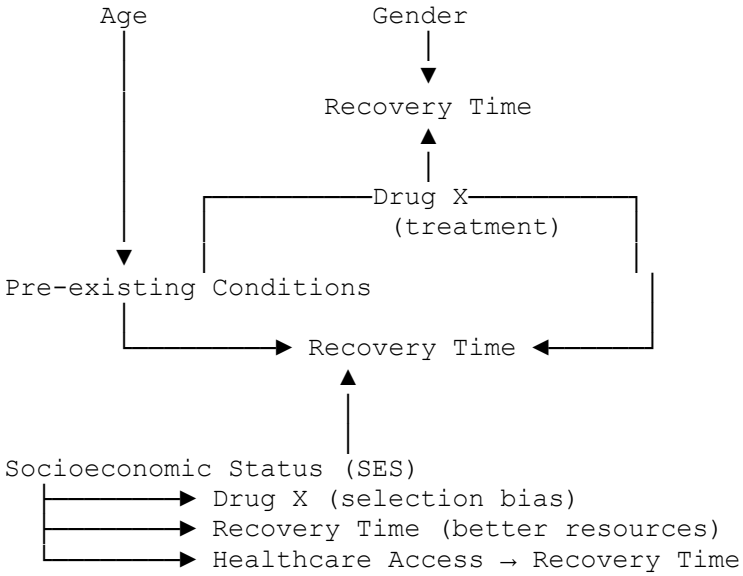
Gender affects recovery but is not expected to directly affect the probability of taking Drug X (unless there's gender-dependent prescribing behavior).

Here is a more complete DAG showing **direct and indirect effects**, confounding paths, and backdoor bias. Arrows indicate direction of influence.

Detailed Causal DAG



Now incorporating **all variables explicitly**:



Explanation of Causal Roles

Variable	Role	Influence Path
Drug X	Treatment (cause)	Direct effect on *Recovery Time*
Age	Confounder	Affects both *Drug X* (doctor preference) and *Recovery Time*
Gender	Covariate	Mainly affects *Recovery Time*
Pre-existing Conditions	Confounder	Influences both *Drug X* usage and recovery
SES	**Major confounder**	Affects treatment choice, access to care, lifestyle, recovery
Healthcare Access	Mediator	*SES → Healthcare Access → Recovery Time*

Variable	Role	Influence Path
Post-Treatment Care	Mediator	*Drug X → Post-Treatment Care → Recovery Time*

Types of Effects

Type of Effect	Path
Direct Effect	`Drug X → Recovery Time`
Indirect (Mediated) Effect	`Drug X → Post-Treatment Care → Recovery Time`
Confounding Path (Bias)	`SES → Drug X` **and** `SES → Recovery Time`
Backdoor Path	`Drug X ← SES → Recovery Time`
Blocked by Randomization (RCT)	All backdoor paths
Open in Observational Study	SES, Age, Pre-existing Conditions

3. Methods to handle confounding in the observational study

Method	Description	Advantages	Limitations
Multivariate regression	Adjust for confounders in a regression model (e.g., linear regression)	Simple, interpretable	Only controls observable confounders
Propensity Score Matching (PSM)	Match treated and untreated patients with similar treatment probabilities	Better balance	Reduces sample size; still sensitive to unobserved confounding
Instrumental Variables (IV)	Use variable correlated with treatment but not outcome (excluding via confounders)	Can handle unobserved confounders	Hard to find valid instruments

4. Interpretation of regression results

You find:

- **Drug X coefficient negative and significant in both RCT and observational study** → Suggests Drug X reduces recovery time.
- **But SES influences treatment choice** → Indicates **selection bias** in observational data.

Conclusion
The RCT supports a causal benefit of Drug X on recovery time. The observational study confirms it but may **overestimate the effect** due to SES confounding (e.g., wealthier patients may have better overall health or access to support).

Additional analyses needed:

- Include SES explicitly as a covariate.
- Use PSM or IPW adjusting for SES.
- Conduct sensitivity analysis for potential unobserved confounding.
- Perform subgroup analysis (e.g., by SES category).
- Check balance diagnostics on confounders after matching/weighting.

5. Limitations of observational data in causal analysis

Limitations	Description
Confounding bias	Treatment decisions influenced by variables that also affect outcome
Selection bias	Patients choosing treatment differ systematically
Unobserved confounding	Variables not measured (e.g., patient motivation) remain unadjusted
Reverse causality	Observational data may not guarantee temporal order
Generalization issues	Observational groups may not match real-world populations

In data science, observational studies are common due to ethical or practical constraints (e.g., you can't force people to take drugs). However, they require careful statistical handling to simulate randomization.

Exercise 3.

1. Using only Age in linear regression

Using **Age alone** ignores important predictors like education, experience, and gender.

Impact on model performance:

- The model is **too simple** → high bias (misses relevant relationships).
- Doesn't capture true variability in income.
- Likely to perform **poorly on both training and test data**.

Conclusion:
The model will underfit (high bias, low variance).

2. Low training error & high test error

This pattern indicates:

Error Type	Training	Testing
Error	Low	High
Bias	Low	❌
Variance	❌	High

Conclusion:
Low bias, high variance → the model is overfitting.

How to fix it:

- Simplify the model.
- Use **regularization** (e.g., Lasso).
- Reduce features or apply feature selection.
- Use **cross-validation**.
- Possibly increase the number of training samples.

3. Moving to a Random Forest model

Random Forest is a more complex, flexible model that can reduce bias but introduces **higher variance**.

Challenges:

- It might **overfit** if the trees are too deep or too many.
- Sensitive to randomness, leading to **high variance** if not controlled.

How to handle variance:

Method	Description
Increase number of trees	Stabilizes predictions
Limit tree depth (`max_depth`)	Reduces overfitting
Minimum samples per split/leaf	Smoothens model
Feature sampling	Reduces correlation between trees
Cross-validation	Helps optimize parameters

Role of Regularization

In tree models, regularization means:

- Restricting depth (**controls complexity**),
- Limiting number of features per split,
- Pruning (in some implementations).

Impact:

- Reduces variance,
- Slightly increases bias (good trade-off),
- Improves generalization.

Final Summary Table

Question	Answer
1. Age-only regression	Underfitting → High bias, low variance
2. Low training & high test error	High variance → Overfitting → Use regularization/simplify
3. Random forest challenges	High variance risk → Controlled via hyperparameter tuning & regularization

Exercise 4

1. Applying k-Anonymity

Concept:
k-anonymity ensures that each record in a dataset is indistinguishable from at least *k–1* other records based on a set of quasi-identifiers (QIs). In this dataset, **Age, Gender, Zip Code, and Income** are quasi-identifiers.

Implementation:

Variable	k-Anonymization Method
Age (continuous)	Generalization into ranges (e.g., 20–25, 26–30).
Gender (binary)	May remain unchanged or potentially suppressed if needed.
Zip Code (categorical)	Truncation (e.g., first 3 digits only: 123**)
Income (continuous)	Generalize into income brackets (e.g., 30–50k, 50–70k).

Example:
Instead of showing *(Age=27, Gender=1, Zip=12345, Income=52)*, show:
→ *(Age=25–30, Gender=1, Zip=123**, Income=50–60)*
With *k=3*, this record must match at least **3 individuals**.

Challenges:

- Loss of granularity → reduces analytical precision.
- Sparse datasets (e.g., rare zip codes) may force excessive generalization.
- Binary gender provides limited diversity.
- Combining QIs may still allow re-identification in small communities.

2. Applying l-Diversity

Concept:
While k-anonymity prevents re-identification from QIs, it *does not* prevent attribute disclosure if all individuals in a k-group share the same sensitive attribute. **l-Diversity requires that within each anonymized group, the sensitive attribute (here Medical History) has at least *l distinct values***.

Implementation:

- After applying k-anonymity to QIs, ensure each anonymized group contains at least *l* different medical conditions.
- For example, with *l=2*, a group must contain multiple medical histories (e.g., hypertension, diabetes).

Challenges:

- Some medical conditions are rare → hard to maintain diversity.
- Highly correlated attributes (age → specific diseases) reduce possible diversity.
- May require merging groups, leading to **further data loss**.

3. Applying Differential Privacy

Concept:
Differential Privacy (DP) adds random statistical noise to query outputs, ensuring that the presence or absence of any individual *does not significantly affect the result*.

Example Use Case:
Suppose you want to compute average income of individuals with diabetes.

Instead of returning:
→ *Average income = 52.3k*,
a DP mechanism returns:
→ *Average income = 52.3k ± Laplace noise based on ε*

Where **ε (epsilon)** measures the privacy level (lower ε → more privacy).

Implementation Scenarios:

- For statistical queries (e.g., average, sum, count).
- For machine learning model training (DP-SGD).
- For publishing synthetic datasets with DP guarantees.

Advantages over k-anonymity and l-diversity:

- Works even against attackers with prior knowledge.
- Does not require grouping or generalization.
- Maintains utility when properly tuned.
- Strong measurable privacy guarantee.

4. Trade-Offs Between the Methods

Aspect	k-Anonymity	l-Diversity	Differential Privacy
Protection Level	Moderate	Better	Strong
Prevents Re-identification?	Yes	Yes	Yes
Prevents Attribute Disclosure?	No	Yes	Yes

Aspect	k-Anonymity	l-Diversity	Differential Privacy
Resistant to Background Knowledge?	Weak	Moderate	Strong
Data Utility	High (if minimal generalization)	Medium	Variable (depends on noise)
Complexity	Low	Medium	High
Suitable For	Sharing raw datasets	Sharing datasets with sensitive attributes	Query-based analytics, ML

When to choose what:

- **k-Anonymity:**
  Use when sharing data *internally* with low re-identification risk and need high data utility.
- **l-Diversity:**
  Use when **sensitive attributes must be protected**, and k-anonymity alone isn't enough.
- **Differential Privacy:**
  Use when **sharing with external third parties**, when queries will be run repeatedly, or when you want **mathematically strong privacy guarantees** (especially for regulatory compliance like GDPR).

Summary

- Use **k-anonymity** to hide individuals among similar others.
- Use **l-diversity** to avoid revealing sensitive medical history.
- Use **differential privacy** for strongest protection in analytical queries or public data releases.