

Chapter 1: Data Collection and Exploration (7.5h)

Lab 1 – Session 1 & 2: Introduction to Python (3h + 1.5h)

1. The first notebook (print statements, variable assignment)
2. Variable types in Python (int, float, string, boolean)
3. Mathematical operations
4. Data types: Numbers, Strings, Lists, Dictionaries, Booleans, Tuples, Sets
 - 4.1 Numbers (numeric types, variable assignment)
 - 4.2 Strings (creation, printing, formatting)
 - 4.3 Lists (creation, manipulation, indexing/slicing, mutability, nesting)
 - 4.4 Tuples (construction, basic methods)
 - 4.5 Dictionaries (construction, methods)
 - 4.6 Sets and Booleans
5. Comparison operators (==, !=, >, <, >=, <=, and/or)
6. Conditional statements (if, elif, else)
7. Iteration and loops (for, while, range, break/continue/pass)

Lab 2 – Session 1: Data Collection (3h)

1. Data Collection and Exploration
 - Importing libraries (pandas, numpy, seaborn, matplotlib)
 - Loading the dataset (insurance_claims.csv)
 - Displaying first rows, dimensions, general information
 - Cleaning column names (pyjanitor)
 - Identifying data types and missing values
2. Exploratory Data Analysis (EDA) & Visualization

- Descriptive statistics (.describe())
 - Distribution of the target variable
 - Histograms of numerical variables
 - Correlation matrix (heatmap)
 - Relationships between features and the target
 - Pair plots and bar charts
-

Chapter 2: Data Preprocessing (6h)

Lab 3 – Session 1: Data Cleaning (3h)

1. Detecting missing values
2. Handling techniques (mean/median/mode imputation, "Unknown" labeling, dropping columns)
3. Detecting and removing duplicates

Lab 3 – Session 2: Data Transformation (3h)

1. Importing libraries
2. Loading the data
3. Distribution of the target variable
4. **Handling Missing Values**
 - Detecting missing values
 - Converting special markers ("?", " ") to NaN
 - Dropping rows/columns
 - Imputing numerical values (mean)
 - Imputing categorical values (mode)
5. **Outlier Detection and Treatment**
 - Visualization with boxplots

- IQR-based detection
- Z-score-based detection

6. Feature Selection

- Correlation-based selection (multicollinearity)
- Model-based selection (XGBoost, feature importance)

7. Saving the prepared dataset

Chapter 3: Model Building and Deployment (7.5h)

Lab 4 – Session 1 & 2: Model Building / Model Evaluation (3h + 1.5h)

1. Importing main libraries
2. Reading the data
3. Class distribution of the target variable
4. Identifying features and target variables
5. Train/test split and data scaling
6. **Dummy Classifier** (baseline)
7. **Confusion Matrix**
8. **Building the Model (Decision Tree)**
 - Method 1: using a function
 - Method 2: direct instantiation (DecisionTreeClassifier)
 - Confusion matrix of the model
9. **Model Evaluation**
 - Accuracy score

- Classification report
 - Feature importance (model interpretation)
10. Visualizing the decision tree (Graphviz)

Lab 5 – Session 1: Model Deployment (3h)

1. Installation and setup (Flask, pandas, numpy, scikit-learn)
2. Training and saving the model (Pickle)
3. Creating app.py (loading the model, Flask routes, predictions)
4. Creating HTML templates (index.html, result.html)
5. Deploying and testing the web application