IKT215 Exercise 1: Preprocessing Data for a Pattern Recognition System

January 15th, 2025

Exercise sessions, while not mandatory, are an integral part of your learning experience and exam preparation. Some exam questions will be based on tasks covered during these sessions. Although you have the flexibility to complete exercises at home and ask questions later (preferably within one week), I strongly recommend attending the sessions in person for immediate support and guidance. This allows you to benefit from direct interaction and real-time problem-solving. I respond promptly to emails (manuel.s.mathew@uia.no), and for urgent matters, you're welcome to contact me by phone (available on Lecture 1 slides).

Case description:

Avocado King, is an avocado supplier that sells conventional and organic avocados all over the U.S. The client has given you historical data on the avocado prices/sales in the US market and they want to know how they could use this data and predictive models to gain a competitive edge. They are particularly interested in predicting the prices of avocados and the number of avocados sold.

Data:

Two Excel files are provided:

1. Price and sales data: Data on the average prices and volumes sold
2. Feature explanations: Data dictionary with more detailed info on each variable

Exercise:

You are the assistant data analyst and expected to help the senior analyst by preprocessing data. This is your first day at the job. Good luck!

Tasks:

1. Install Anaconda (available here)
2. Install Pandas (available here, instructions here)
3. Explore the given data
    a. Import required libraries and data
    b. Print ten random samples
    c. Print first and last ten samples
    d. Find out the size of dataset

    e. Find out what are the type of variables in dataset
    f. Find out the number of missing values in each column
    g. Get a statistical overview of your dataset
    h. Find out the correlation between different variables in dataset.
4. Visualize data
    a. Univariate analysis of continuous variables:
       Pick any two continuous variables and plot their distribution.
    b. Bivariate analysis of categorical variables:
       Pick any continuous variable and plot their variation across years.
    c. Bivariate analysis of every variable against every variable
    d. Generate a heatmap of correlation matrix
5. Find and deal with missing values
    a. Fill with interpolation
    b. Remove missing values
6. Find and replace outliers
    a. IQR Method
    b. Z-score Method
7. Make categorical features useful
    a. One Hot Encoding
    b. Ordinal Encoding
8. Normalize the features using
    a. Min-Max Normalization
    b. Standard Normalization