Database Administration
Lab 09: Data Mining.

Andrés Calderón, Ph.D.

April 21, 2025

1 Introduction

Data mining is a cornerstone technique in modern data science, enabling the discovery of meaningful patterns, trends, and relationships within large datasets. As data becomes increasingly abundant across all domains, the ability to analyze and extract insights from this information is critical for informed decision-making. This lab introduces the fundamental ideas of data mining and provides hands-on experience with one of its most commonly used methods: classification using decision trees.

Through a combination of conceptual material, guided tutorials, and individual exploration, you will learn how to apply data mining techniques to real-world datasets. We begin with an overview of key concepts in data mining, followed by a practical tutorial using R and RStudio to build and interpret decision tree models. You will then independently select a dataset from the UC Irvine Machine Learning Repository and perform your own analysis.

This lab emphasizes both technical proficiency and analytical thinking. By the end, you will be able to construct a decision tree model, interpret its structure, and reflect on the outcomes of your analysis in a well-documented report.

2 What is Data Mining

First, we will delve into some definitions and concepts related to data mining to better understand what data mining is, how it is performed, and what some examples look like. We will watch an illustrative video on YouTube via the Cubeware GmbH channel. The video can be seen here. Remember that you can activate automatic English subtitles and then translate them to Spanish if you wish. There is no particular assignment to submit for this section—just take notes and enjoy the video!

3 Classification using Decision Trees

Once we feel more comfortable with the notions of data mining, it is time to explore some techniques through a guided tutorial. An excellent resource is available on Datacamp. We will follow the tutorial titled "Decision Trees in Machine Learning Using R"[1] by Arunn Thevapalan and James Le. In it, we will be introduced to the classification technique of Decision Trees, using the packages and libraries from the R Project to build a model that predicts the median value of houses in the Boston area. Specifically, we will be working in RStudio and using the MASS package. Again, you do not need to submit any evidence of your work, but you must complete the tutorial in order to proceed to the next section.

[1] Thevapalan and Le, 2023. *"Decision Trees in Machine Learning Using R"*. Datacamp.

4 Individual Work

We will apply what we learned in the previous section to mine a new dataset from the UC Irvine Machine Learning Repository. This repository provides a wide variety of datasets from different disciplines and for different purposes. You will filter the output using the following criteria:

- Data Type = Multivariate
- Task = Classification
- Feature Type = Categorical

From the 26 selected datasets, you will choose one and perform a decision tree analysis following the steps described in Section 3.

5 What We Expect

You will submit a well-structured report in **PDF** format, presenting your work and describing the analysis and code you used, along with any relevant information you consider appropriate. Along with the report, we expect you to include a graphical representation of the decision tree and a description of the characteristics of your model. Send your report to my email before **May 05, 2025**, with the subject line: `[DBA] Lab 09`.

Happy Hacking 😎!