Statistical Principle of Data Science - Group 3 Project Proposal

Luka Leko, K51820066
Max-Jonathan Luckow, K12235898
Ronald Moritz, K11718128
Agnes Hinterplattner, K01634183

1. Data Description

We are going to use “Rain in Australia” dataset from kaggle:
https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package

The dataset consists of several attributes concerning the weather on a particular day (such as rainfall, humidity etc.) and it tries to predict whether or not it will rain the next day.
The target variable is RainTomorrow. It is a boolean variable that is equal to 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm on the future day, otherwise it is 0. Below you can find a more precise description of the dataset:

Descriptor	Data Type	Description
Date	string	The date of observation. The data ranges from 2007 until 2017.
Location	categorial	The common name of the location of the weather station
MinTemp	real	The minimum temperature in degrees celsius
MaxTemp	real	The maximum temperature in degrees celsius
Rainfall	real	The amount of rainfall recorded for the day in mm
Evaporation	real	The so-called Class A pan evaporation (mm) in the 24 hours to 9am
Sunshine	real	The number of hours of bright sunshine in the day.
WindGustDir	categorical (16)	The direction of the strongest wind gust in the 24 hours to midnight
WindGustSpeed	integer	The speed (km/h) of the strongest wind gust in the 24 hours to midnight
WindDir9am	categorical (16)	Direction of the wind at 9am
WindDir3pm	categorical (16)	Direction of the wind at 3pm
WindSpeed9am	categorical (16)	Wind speed (km/hr) averaged over 10 minutes prior to 9am
WindSpeed3pm	categorical (16)	Wind speed (km/hr) averaged over 10 minutes prior to 3pm

Humidity9am	integer	Humidity (percent) at 9am
Humidity3pm	integer	Humidity (percent) at 3pm
Pressure9am	real	Atmospheric pressure (hpa) reduced to mean sea level at 9am
Pressure3pm	real	Atmospheric pressure (hpa) reduced to mean sea level at 3pm
Cloud9am	integer	Fraction of sky obscured by cloud at 9am
Cloud3pm	integer	Fraction of sky obscured by cloud at 3pm
Temp9am	real	Temperature (degrees C) at 9am
Temp3pm	real	Temperature (degrees C) at 3pm
RainToday	bool	1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0
RainTomorrow	bool	1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0

2. Motivation

We would like to create a predictor to anticipate if it is likely to rain on the next day. Given the multitude of attributes in our data we are likely able to create a multivariable predictor which will also hold in the future. Given the measurements today we could create a model of influences on the probability of rain the next day. It would result in a very simple predictor, in which only a small number of easily measured attributes would make a forecast for the next day. This is contrary to traditional weather forecasts, where the whole map of surroundings and winds are taken into account, which is far more difficult to measure and collect than what we are trying to achieve.

3. Statistical Methods

First we want to use classification to predict the target variable RainTomorrow. We will use either Random Forests, a Decision Tree or a Bagged Tree as the first method.

The second method will be logistic regression.

We will use Python with the library sklearn (among others) to train and evaluate our models.

4. Rough timeline

Milestone 1: Implementation. This includes the implementation of both statistical methods.
Milestone 2: Draft version of report.
Milestone 3: Final report until deadline

We agreed to meet in between milestones to discuss results and how to proceed (i.e. who does what). In general, the rough goal is to be done with the implementation by the end of the year and to be done with documentation about a week after.