DANMARKS TEKNISKE UNIVERSITET



02450 – Introduction to Machine Learning and Data Mining

PROJECT 1

Martin Mikšík s212075





Lukáš Málek s212074

A picture containing icon

Description automatically generated[[1]](#footnote-2)

Table of Contents

[1 Goal of the Project 4](#_Toc97138837)

[1.1 Data set reference 4](#_Toc97138838)

[1.2 Previous work on the data set 4](#_Toc97138839)

[1.3 Problem of interest of this Project 4](#_Toc97138840)

[2 Description of the data set 5](#_Toc97138841)

[2.1 Missing data 5](#_Toc97138842)

[3 Data Visualisation 6](#_Toc97138843)

[4 References 7](#_Toc97138844)

DELETE BEFORE HAND-IN

Check all „DELETE“ or „TODO“ flags before hand in

DELETE: adding citations: references -> insert citation -> date, name, year. Update citation table in the end of the document

TODOS:

1. specify section percentile for each student (one of us always needs to have atleast 60%)
2. solutions for at least 4 exam probles

Upload:

1. report + one .py file (wont be ecaluated, just checked for plagiarism)
2. upload only one member
3. deadline 8 March 13:00

Description:

1. only apply methods you understand and check for possible issues, reason them.
2. thorough understanding of your data and describe how you expect the data can be used in the later reports
3. coverin week 1-4
4. check question in exercises which by answering them can lead to successful report
5. Description of data set

- What are data about, overall problem of interest

- provide reference where we got the data

- summarize results what tehy did to the data

- apply **classification** and **regression task.** So first explain contect and what we want to accomplish. Which attributes: Which class label will we be predicing on which other attributes?

- if transforming data, explan (we will not probably transform data)

estimate sex values on the other attributes except year and island (as they dont carry a meaningful information.

# Goal of the Project[[2]](#footnote-3)

The main goal of this project is to accomplish further analysis of the chosen dataset in Python and with the help of 02250\_Toolbox [1], exercises and lectures from Week1-Week4 of the course [2].

The Project is expected to deepen our understanding of Data mining, data visualisation and Python linear regression and classification implementation preparation.

This project is a work of two directly collaborating students. Percentile contribution of the students are referred in a footnote for each task.

## Data set reference

We have chosen Palmer Penguins Dataset [3] with data provided by Palmer Station Antarctica LTER [4]. The data set contains 8 variables of three penguin species, observed in Antarctira.

## Previous work on the data set

A group of penguins

Description automatically generated with low confidence Previous attempts [5] of data analytics and representation by linear regression and classification were made with accurate results, as the data set contains only few missing variables. The paper was comparing bill depth and length to classify penguin species in programming language R.

[[3]](#footnote-4)

## Problem of interest of this Project

Using linear regression, we will try to predict bill depth for one of the species using the interval attributes in the next report.

For classification problem, we will try to estimate missing sex variables based on all other variables, while omitting Year and Island, as these variables don´t carry meaningful information for this task.

For both problems, no data transformation will be used.

In order for these tasks to be completed, we need to analyze the data, visualize them and make a conclusion of suitable alghorithm options and methods to be further used. Classification tree with Hunts algorithm could be used, adjusted for continous variables. Regression tree is also an option for linear regression.

# Description of the data set[[4]](#footnote-5)

[[5]](#footnote-6)

A picture containing text, clipart, vector graphics

Description automatically generated

Table of attribute description and basic statistics:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Row ID** | **Species[[6]](#footnote-7)** | **Island[[7]](#footnote-8)** | **Bill Length [mm]** | **Bill Depth [mm]** | **Flipper length [mm]** | **Body mass [g]** | **Sex** | **Year** |
| **Data** | Discrete | Discrete | Discrete | Continous | Continous | Continous | Continous | Discrete | Discrete |
| **Attribute** | Nominal | Nominal | Nominal | Interval | Interval | Iterval | Interval | Nominal | Nominal |
| **Missing Values** | No | No | No | 2x N/A | 2x N/A | 2x N/A | 2x N/A | 11x N/A | No |
| **Mean** | - | 0 | Biscoe | 43.922 | 17.1512 | 200.9152 | 4201.7544 | - | 2008 |
| **STD[[8]](#footnote-9)** | - | - | - | 5.4596 | 1.9748 | 14.0617 | 801.9545 | - | - |
| **Median** | - | - | - | 44.4500 | 17.3000 | 187.0000 | 4050.000 | - | - |
| **Range** | 1-345 | 0-2 | 1-3 | 32.1-59.6 | 13.1-21.5 | 172-231 | 2700-6300 | M/F | 2007-2009 |

## Missing data

There are two instances in the data set, where all of the interval attributes, as well as sex were missing. Because these attributes are crucial in further data analytics and other attributes alone don´t contribute by a meaningful way, these to instances were completely removed from the data set analytics, resulting in 342 penguin instances in total, with the only missing/NaN values are present 9 times in Sex attribute, which we might try to fill in in the upcoming report. For now, these instances were also removed.

Were are then working with 333 penguin instances with no missing data:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Row ID** | **Species** | **Island** | **Bill Length [mm]** | **Bill Depth [mm]** | **Flipper length [mm]** | **Body mass [g]** | **Sex** | **Year** |
| **Missing Values** | No | No | No | No | No | No | No | No | No |

# Data Visualisation

## Outliers

Upon careful analysis of the statand visualization of the data set TODO

First, we need to preproccess the data with the accordance of our Project goal. We´re reading the raw *.csv* data with *Pandas* into matrix X, while omitting the first column specifying only the row number. blabla TODO

Text

Description automatically generated

TODO update before hand-in

# References

|  |  |
| --- | --- |
| [1] | DTU Compute, 2022. [Online]. Available: https://www.compute.dtu.dk/. |
| [2] | B. S. Jensen, DTU Compute, 2022. [Online]. Available: http://compute.dtu.dk/courses/02450/. |
| [3] | S. Lopp, 2021. [Online]. Available: https://gist.github.com/slopp/ce3b90b9168f2f921784de84fa445651. |
| [4] | L. a K. Gorman, Environmental Data Initiative, 2007-2009. [Online]. Available: https://environmentaldatainitiative.org/dataset-design/. |
| [5] | A. Horst, 2022. [Online]. Available: https://github.com/allisonhorst/palmerpenguins/. |
| [6] | A. Horst, „Art work,“ 2021. [Online]. Available: https://github.com/allisonhorst/stats-illustrations. |

1. Art work by Allison Horst [6] [↑](#footnote-ref-2)
2. Martin 60%, Lukáš 40% [↑](#footnote-ref-3)
3. Art by Allison Horst [6] [↑](#footnote-ref-4)
4. Martin 60%, Lukáš 40% [↑](#footnote-ref-5)
5. Art by Allison Horst [6] [↑](#footnote-ref-6)
6. Corresponding statistical values reffers to a dictionary: {'Adelie': 0, 'Chinstrap': 1, 'Gentoo': 2} [↑](#footnote-ref-7)
7. Corresponds to three islands: Biscoe, Dream and Torgersen [↑](#footnote-ref-8)
8. Standard Deviation [↑](#footnote-ref-9)