TU Dortmund

Case Studies

Project 1: Stryrian Health Data Analysis

Lecturers:

Dr. Uwe Ligges

Marie Beisemann

Author: Farrukh Ahmed

Group number: 3

Group members: Muhammad Raafey Tariq, Aymane Hachcham, Amirreza Khamehchin Khiabani

June 6, 2023

Contents

1 Introduction

Blood pressure, a vital indicator of cardiovascular health, plays a crucial role in monitoring and managing overall well-being. Accurate and complete collection of blood pressure and health data is vital for understanding our cardiovascular system and identifying potential health risks at an early stage. By tracking blood pressure measurements over time, healthcare professionals can assess the effectiveness of interventions, tailor treatment plans, and empower individuals to make informed lifestyle choices. Collecting blood pressure and health information goes beyond personal health benefits. It also helps researchers and policymakers understand the bigger picture and develop specific plans to fight high blood pressure and lessen the impact of heart disease on society.

The purpose of this project is to fit machine learning models which can explain the response variable with respect to independent variables. At first, a descriptive analysis was done on the data to find the main characteristics and patterns present in the data. Subsequently, data pre-processing was performed, which involved the addition of new variables. These variables were either derived from the existing column or obtained from an external source. After that, Linear regression models were fitted for both *measured_bp_sys* and *measured_bp_dia* as a response variable, and then the coefficients of the model were interpreted to show how the covariates affect the response variables, and the goodness of fit of the regression models were evaluated. Moreover, regression trees and random forests were fitted on the same response variables and afterward, the results of these trees and random forests were discussed for both the response variables. Finally, a comparison was made between the linear models, regression trees, and random forests.

In Section 2, an overview of the dataset is provided and the quality of the data is evaluated. In Section 3, details and explanations of the statistical methods are discussed, which are then used for the analysis of the dataset. It provides information about the formulas and assumptions of the linear regression models and also outlines the details about the Regression trees, Random forests, Best Subset Selection, Cross-validation, and the coefficient of determination. In the subsequent Section 4, the pre-processing of the dataset is discussed and methods explained in Section 3 are applied to the given dataset and results are interpreted. At last, the Summary section briefly summarized the results. It also discusses the potential improvements which can be made to this experiment.

2 Problem Statement

2.1 Data set and Data Quality

The dataset which will be analyzed in this project is collected during *Styrian state exhibition 2006* which was held between April 29, 2006, to October 29, 2006, in the city of *Bruck An Der Mur*. There were three different terminals for data collection with twelve parameters. The data was provided to us by the instructor of Case Studies at TU Dortmund in the summer semester of 2023. The number of observations in the dataset is $n = 16386$ having eighteen columns with six numerical variables and twelve categorical variables. The overall quality of the data is good with some missing values.

All of the dataset is used for the analysis except the variable *id*. It comprises of observations for the variables *time* (Character, it has the time on which the particular observation is taken in AM/PM format), *terminal* (Integer, it has 3 different values), *postal_code* (Integer, 4 digits), *municipality* (Character), *district* (Character), *federal_state* (Character, It has 9 different values), *felt_health_condition* (Integer, it has 5 different values), *gender* (Character, it has two different values), *year_of_birth* (Integer), *is_smoker* (Character, it has two different value), *is_diabetic* (Character, it has two different value), *has_cholestrol* (Character, it has two different value), *in_treatment* (Character, it has two different value).

The dataset consists of five numerical variables. The *year_of_birth* variable represents the birth year of each participant. The variables *self_eval_bp_sys* and *self_eval_bp_dia* correspond to the participant's self-measured systolic and diastolic blood pressures, respectively. On the other hand, the variables *measured_bp_sys* and *measured_bp_dia* represent the systolic and diastolic blood pressures measured by blood pressure monitoring devices.

The dataset contains twelve categorical variables. Among these, the *terminal* variable has three distinct values, each representing the terminal used to record a particular reading. The participant's place of residence is captured using variables such as *postal_code*, *municipality*, *district*, and *federal_state*. The *felt_health_condition* variable is ordinal and denotes the health condition of the participant, ranging from 1 (very good) to 5 (very bad). The remaining four variables are binary, with values of either *true* or *false*. These variables are *is_smoker* (indicating whether the person is a smoker or not), *is_diabetic* (indicating the presence of diabetes), *has_cholestrol* (indicating the

presence of cholesterol), and *in_treatment* (indicating whether the person is receiving hypertension treatment or not).

2.2 Project Objectives

In this project, at first, data pre-processing is done and some new variables are calculated using the existing ones and some were imported from an external data source. Later, two linear regression models were fitted for both the response variables then both models were examined to see which one is more aligned with the assumptions of the linear model. Afterward, coefficients and the confidence interval are interpreted for both models. Additionally, a goodness-of-fit analysis was conducted to evaluate the performance of the models. Finally, regression trees and random forest models were fitted to the response variables, and the findings are extensively discussed in section.4.

3 Statistical Methods

In this section, several statistical methods are discussed which are later used to analyze the data. For all the visualizations and calculations Python software (Van Rossum & Drake, 2009),version 3.8.8 is used with packages numpy (Harris *et al.*, 2020), pandas (McKinney *et al.*, 2010), matplotlib(Hunter, 2007), and seaborn (Waskom *et al.*, 2017).

3.1 Multiple Linear Regression

Multiple Linear regression attempts to model the effect of a vector of k explanatory or independent variables, $x_1, ..., x_k$, on a continuous target or response variable y by fitting a linear equation to observed data. The covariates can be categorical variables or continuous. The response variable is not a deterministic function $f(x_1, ..., x_k)$ of covariates. Instead, it also shows random noise.

$$y_i = f(x_{1i}, ..., x_{ki}) = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \epsilon_i. \tag{1}$$

where x_{ij} represent the value of j^{th} covariate, $j = 1, , , , k$ for the i^{th} observation, $i = 1, ..., n$, and ε is often denoted as the stochastic component or error term. (Fahrmeir *et al.*, 2013, p. 74)

To model the effect of categorical variables x with c categories, among the categories, one category has to be chosen as a reference category. In practice, the category which makes the most sense to interpret the data is chosen as a *reference category*. Here, category c can be considered as a reference category. Therefore, $c-1$ dummy variables can be defined as follows and then included in the model.

$$x_{i,c-1} = \begin{cases} 1 & x_i = c-1, \\ 0 & \text{otherwise} \end{cases}$$

(Fahrmeir *et al.*, 2013, p. 97)

The unknown parameters $\beta_0, \beta_1, ..., \beta_k$ can be represent in $(k+1)$ dimensional vector $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$ where β_0 represents the intercept. Similar to unknown parameters covariates can also be written in $(k+1)$ dimensional vector $\mathbf{x}_i = (1, x_{i1}, ..., x_{ik})'$ for $i = 1, ..n$. The systematic component can be expressed as a vector product. Therefore,

$$y = f(\mathbf{x}) + \varepsilon = \mathbf{x}'\beta + \varepsilon.$$

So the vector $\mathbf{y}$ and ε are defined as:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

and the design matrix $\mathbf{X}$ can be defined as:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} = \begin{pmatrix} \mathbf{x_1'} \\ \vdots \\ \mathbf{x_n'} \end{pmatrix}$$

Then n equations for Eq.1 for n can be summarized as:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon. \tag{2}$$

The Eq.2 is called the linear regression model if the following assumptions hold:

- The errors have zero mean and can be written as $\mathbf{E}(\varepsilon) = 0$.
- The errors have constant variance among them and they are normally distributed.
- The design matrix $\mathbf{X}$ assumed to be full rank.

(Fahrmeir *et al.*, 2013, p. 74-76)

3.1.1 Parameter Estimation

To estimate the $\boldsymbol{\beta}$ parameter vector maximum likelihood method is used. Assuming that errors are normally distributed, the likelihood function can be written as:

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{-1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right).$$

The log-likelihood is given by:

$$ln(L(\boldsymbol{\beta}, \sigma^2)) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

The maximum likelihood estimator of $\hat{\boldsymbol{\beta}}$ for the vector $\boldsymbol{\beta}$ is determined by maximizing the log-likelihood function with respect to $\boldsymbol{\beta}$. We therefore get the estimator $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

The maximum likelihood estimator of $\hat{\boldsymbol{\beta}}$ is same as the least squares estimator under the normality assumptions. (Fahrmeir *et al.*, 2013, p. 107)

To test the significance of a parameter β_j. The null and alternative hypotheses are as follows:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

3.1.2 Confidence Intervals on the Regression Coefficients

To construct a confidence interval for the estimates of regression coefficient $\boldsymbol{\beta_j}, j = 0, .., k$ under the assumption that the errors are independent and normally distributed with zero

mean and variance σ^2. The overall quality of the regression line is also measured by the width of the confidence interval. We use the following test statistic corresponding to the test $H_0 : \beta_j = 0$:

$$\hat{t_j} = \frac{\hat{\beta}_j}{s\hat{e}_j}.$$

where $s\hat{e}_j = \sqrt{\widehat{Var(\hat{\beta}_j)}}$ is the estimated standard deviation error for β_j. The above statistic is t-distributed with $n - p$ degree of freedom. For a defined significance level of $\alpha = 0.05$. The absolute value of the defined t-statistic is compared to the $(1 - \frac{\alpha}{2})^{th}$ quantile of t-distribution with $n - p$ degrees of freedom. The null hypothesis H_0 will be rejected:

$$|\hat{t_j}| > t_{n-p}\frac{1-\alpha}{2}.$$

The confidence interval for the estimator $\hat{\beta}_j$ is given by:

$$\left[\hat{\beta}_j - t_{n-p}(1 - \frac{\alpha}{2}) \cdot s\hat{e}_j, \hat{\beta}_j + t_{n-p}(1 - \frac{\alpha}{2}) \cdot s\hat{e}_j\right].$$

where (1 - α)-confidence interval for β_j. (Fahrmeir *et al.*, 2013, p. 136)

3.2 Mean Squared Error

Mean squared error is a metric that is used to measure the error in statistical models. It evaluates the average squared differences between the predicted and the actual values in a regression problem. MSE is zero when the model has no error. As the error in the model increase, the value of MSE also increases. Mathematically it is given as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i is the actual value and $\hat{y}_i$ is the predicted value for the i^{th} observation, $i = 1, ..., n$. (Fahrmeir *et al.*, 2013, p. 479)

3.3 Decision Trees

A decision tree is a supervised machine-learning model used to deal with regression and classification tasks. The main idea behind the decision tree is to predict a target variable by learning simple decision rules inferred from the data provided. A root node, branches, internal nodes, and leaf nodes make up the tree structure, which has a hierarchical, tree-like structure. Classification and Regression trees are the two main types of decision trees. In this project, regression trees are used as the target variable is continuous.

3.3.1 Regression Trees

A regression tree is a type of decision tree that is used for regression tasks, where the goal is to make numerical predictions. The construction of a regression tree involves partitioning the input space into smaller regions and then fitting a constant model such as the mean value of the data within the region.

Building a Regression Tree First, the whole predictor space of k feature variables with a response variable $y \in R$ for n observations which is (x_i, y_i) for $i \in N$, with $x_i = (x_{i1}, x_{i2}, ..., x_{ik})$ is considered. Then, the predictor space is divided into j partitions $R_1, R_2, ..., R_j$ and a constant model is fitted in each of the regions:

$$f(x) = \sum_{j=1}^{J} c_j I(x \in R_j)$$

Here, the constant model refers to the mean value of each region R_j.

After that, a splitting criterion is adopted which is *mean squared error* in our case is used to calculate the impurity in each region. Starting with the whole data set, consider splitting k^{th} variable and a threshold value t to split the data into two subsets:

$$R_1(k,t) = \{X|X_k \leq t\} \text{ and } R_2(k,t) = \{X|X_k > t\}$$

The goal is to find the variable x_k and the threshold t that minimizes the *mean squared error* within each subset.

$$\min_{k,t} \left[\min_{c_1} \left(\frac{1}{n_1} \sum_{x_i \in R_1(k,t)} (y_i - c_1)^2 \right) + \min_{c_2} \left(\frac{1}{n_2} \sum_{x_i \in R_1(k,t)} (y_i - c_2)^2 \right) \right]$$

where n_1 and n_2 are the number of observations in the region R_1 and R_2 respectively.

The split point t for each variable can be determined rapidly, allowing for a quick scan through all the inputs. This enables the determination of the optimal pair (k, t) to be achievable. Once the best split is identified, the data is divided into two regions, and the splitting process is repeated on each of these regions. This iterative process continues for all the resulting regions until a stopping criterion is met for e.g (minimum node size or maximum depth size etc). (Hastie *et al.*, 2001, p. 307)

3.4 Random Forests

Random forests are a machine-learning technique where multiple regression trees are constructed to make the final prediction. It employs a technique called bagging or bootstrap aggregating which generates multiple subsets of the training data by sampling randomly with replacement, which is then used to train individual regression trees. It effectively reduces the problem of overfitting in the regression trees by training the trees on different subsets of the data.

$$f(x) = \frac{1}{M} \sum_{m=1}^{M} f_m(x).$$

where $f_m(x)$ is the output of the f_m tree given a set of k predictors $x = (x_1, x_2, ..., x_k)$, and the final result is the average of all trees used in the random forest.

Random Forests introduce an additional level of randomness by considering only a subset of features at each node of a decision tree. This helps to decorrelate the trees and improve the overall accuracy. The number of features to be considered at each node is determined by the user. (Murphy, 2013, p. 550 - 551)

3.5 Coefficient of Determination

It is used as a measure of strength for the relationship between two variables. It measures the percentage variability within the y-values that the regression model can explain. It is denoted by by R^2 and has range; $0 \leq R^2 \leq 1$.. For a linear regression model, R^2 (coefficient of determination) is given as (Akinkunmi, 2019, p. 168)

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i{}^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

where $\bar{y}$ is mean value of response variable and $\hat{y}_i$ for $i = 1, ..., n$ are the estimated values

If the value of R^2 is close to 1, the sum of squares of residuals is minimum and the model fit is better. If R^2 is closer to 0, then the sum is larger and the data fit is poor.(Fahrmeir *et al.*, 2013, p. 115)

3.6 Adjusted Coefficient of Determination

The coefficient of Determination is a measure of the goodness of fit to the data but R^2 tends to increase as the new covariate is added. The adjusted coefficient of determination solves this problem by including a penalty term for the number of parameters. It is defined as :

$$\tilde{R^2} \ = 1 - \frac{n-1}{n-p}(1 - R^2)$$

where p is the number of parameters in the model. (Fahrmeir *et al.*, 2013, p. 147)

3.7 Best Subset Selection

The Best Subset selection is used to select a set of k best predictors or features. The goal is to identify the set of predictors that produces the most accurate and interpretable model in terms of chosen statistical metrics, where $k = 1, ..., n$.

The following steps are involved in the best subset selection:

- The algorithm starts with the null model which is denoted by M_0 and contains no predictors which simply predict the mean for every observation.
- for $k = 1, 2, .., n$
  - For k predictors fit $\binom{n}{k}$ models.
  - The model that exhibits the lowest value for a specific criterion (in our case MSE) is chosen as the optimal model for the given subset size and call it M_k.
- The best model among $M_0, .., M_k$ each subset is selected based on the selection criterion. (James *et al.*, 2013, p. 205)

3.8 Cross Validation

Cross-validation is a methodology employed to assess the effectiveness of a predictive model by dividing the available dataset into several subsets. Cross-validation is also called k-fold cross-validation. K-fold CV randomly partitions the data into k different random folds or subsets of data, where $k = 1, .., n$

The following steps are involved in the cross-validation

- The dataset is partitioned into equal-sized folds or subsets. The value of k is determined based on the size of the data. Common choices of k are 5 and 10.
- For every fold, the model is trained on the remaining $k-1$ folds of the dataset for parameter estimation, which implies that usually every data point is used both for training and testing exactly once. The performance of the model is evaluated on the basis of the left-out fold.
- Based on performance metric the model performance is evaluated on each fold.
- Performance metrics obtained from each fold are then aggregated to obtain a single performance estimate of the model. (Fahrmeir *et al.*, 2013, p. 149)

4 Statistical Analysis

Statistical methods explained above are used in this section to analyze the given dataset and to interpret the results.

4.1 Descriptive Analysis

At first, a descriptive analysis was conducted to provide an overview of the dataset. The sample size is $n = 16386$ and the dataset has in total 147 missing values which were not considered in this analysis. The distribution of these missing values per column is given in the Table.1

The variables which give the geographical location of the participant also have 331 null values but those represent the foreigners which tell us that there were 16055 local residents out of 16386 observations.

The dataset comprises twelve categorical variables as described in Section.2. Considering Table.2, the data reveals that female participants outnumbered male participants

Table 1: Missing value distribution per column

Column	**No of missing value**
time	0
terminal	0
postal_code	0
municipality	0
district	0
federal_state	0
felt_health_condition	0
year_of_birth	23
gender	0
is_smoker	0
is_diabetic	0
has_chloestrol	0
in_treatment	0
self_eval_bp_sys	45
self_eval_bp_dia	56
measured_bp_sys	0
measured_bp_dia	0
age	23

by approximately 56.08 percent. Additionally, a significant majority of individuals did not disclose any smoking habits, diabetes, cholesterol issues, or undergoing hypertension treatment and most of them feel good about their health. Furthermore, the data indicates that Terminal 1 and Terminal 2 were the most frequently utilized terminals, accounting for 38.05 percent and 31.71 percent of the total usage, respectively. Finally, the majority of participants, specifically 87.72 percent, were from the state of Steiermark. The second highest representation came from Wien, accounting for 3.06 percent. Additionally, there was a small percentage of foreigners participating, making up 2.02 percent of the overall participation.

The numerical variables were examined as part of the descriptive analysis. Table.3 in the appendix provides relevant information. The data reveals that the average values for self-evaluated systolic blood pressure and measured systolic blood pressure are quite similar, with self-evaluated at 122.41mmHg and measured at 124.10mmHg. The same pattern is observed for self-measured diastolic blood pressure and measured diastolic blood pressure, as the difference between them is not particularly significant. Additionally, the average age of the participants was 42.13.

Table 2: Frequency distribution of the categorical variables.

Categorical Variable	Categorical Values	Frequency	Percentage (%)
gender	f	9190	56.08
	m	7173	43.76
is_smoker	False	13888	84.75
	True	2475	15.10
is_diabetic	False	12276	74.92
	True	4087	24.94
has_cholestrol	False	10636	64.91
	True	5727	34.95
in_treatment	False	14321	87.40
	True	2042	12.46
terminal	1	6235	38.05
	2	5196	31.71
	3a	1568	9.57
	3b	3387	20.67
felt_health_condition	1	5897	35.99
	2	7931	48.40
	3	2240	13.67
	4	198	1.21
	5	97	0.59
federal_state	Steiermark	14373	87.72
	Wien	501	3.06
	Niederösterreich	469	2.86
	Oberösterreich	224	1.36
	Kärnten	188	1.14
	Burgenland	122	0.74
	Salzburg	82	0.50
	Tirol	68	0.41
	Vorarlberg	28	0.17
	Foreigners	331	2.02

Table 3: Frequency distribution of the quantitative variables.

self_eval_bp_sys	self_eval_bp_dia	measured_bp_sys	measured_bp_dia	age
Min. : 34.00	Min. :30.00	Min. : 43.00	Min. : 27.00	Min. :0.00
Mean :122.41	Mean :79.86	Mean : 124.10	Mean : 82.04	Mean :42.13
Max. :299.00	Max. :212.00	Max. :217.00	Max. :197.00	Max. :126.00

4.2 Data Pre-processing

Some new variables were introduced in the dataset before proceeding to data analysis which were *age*, *humidity*, *temp*, *temp_min*, *temp_max*, *month*, *day*, and *hour*. To

calculate the age of the participants of the exhibitions, we used the year 2006 as the reference year. All the observations in the column *year_of_birth* were subtracted from 2006 and we get numeric value for the column *age*. Furthermore, from variable *time* new variables *month*, *day*, and *hour* are created.

Moreover, the variables *humidity*, *temp*,*temp_min*, and *temp_max* were taken from the external source.Visual Crossing (2006), and the reason why these variables were included in the dataset is that they have an influence on the blood pressure as discussed in various studies such as Barnett *et al.* (2007).

Lastly, a new category in the *terminal* column was also introduced as *terminal* 3 was replaced by a different measuring device at the end of May, 2006. So readings taken after May 2006 were marked as *3b* and readings recorded before that were marked as *3a* in the *terminal* column.

Following the creation of new variables, certain data-cleaning procedures were undertaken. Specifically, observations where the age fell below fifteen or exceeded one hundred were eliminated. This was done because individuals with a young age typically do not report the issues that are being addressed in this analysis. Additionally, observations with ages greater than one hundred were considered outliers due to their rarity. Moreover, there were instances where observations were recorded both before and after the exhibition. However, we excluded those observations from our analysis since we focused solely on the data obtained from participants during the exhibition.

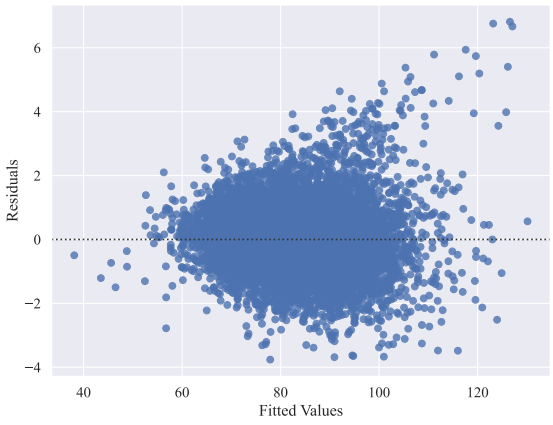
4.3 Fitting Linear Models

Following the data pre-processing stage, we obtained a dataset consisting of $n = 14864$ observations. Subsequently, we divided this dataset into a training set and a testing set using a 70/30 ratio. In this subsection, two linear regression models were fitted; one with *measured_bp_dia* as the response variable and another with *measured_bp_dia*. The target variable is continuous in both cases. Best subset selection is employed to obtain the optimal combination of variables, and the variables obtained after applying this technique are displayed in the Table.4

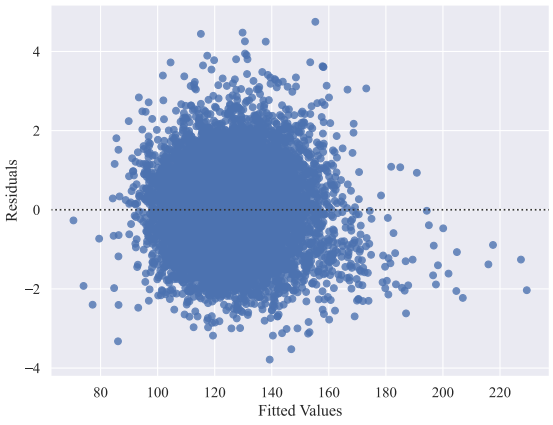
After the two models were fitted, diagnostics plots were used to check which model is in line with the assumptions of the linear model.

Table 4: Variables used in Linear models

Nr	Features
1	terminal
2	federal_state
3	felt_health_condition
4	gender
5	is_smoker
6	is_diabetic
7	has_chloestrol
8	in_treatment
9	measured_bp_sys
10	measured_bp_dia
11	age
12	month
13	hour
14	day
15	temp
16	humidity
17	temp_min
18	temp_max



(a) Residual vs Fitted plot for a model with measured_bp_dia as the response variable.

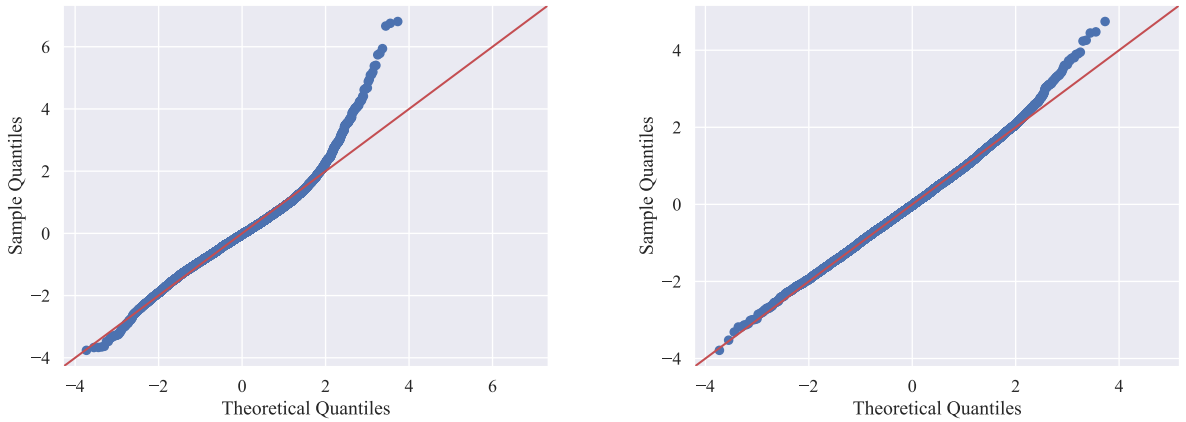


(b) Residual vs Fitted plot for a model with measured_bp_sys as the response variable.

Figure 1: Residual vs Fitted plots

From Fig.1, it can be seen that data points in both the plots are randomly scattered validating the linearity assumption for both the variables. Secondly, for the first plot, a slightly uneven spread of variance can be observed; a small spread of residuals on the left

and then widening up slightly on the right as the fitted values are increased. Whereas, for the second plot spread of the residuals is approximately the same from left to right. Therefore, we can assume constant variance assumption for both variables.



(a) QQ plot for the model with measured_bp_dia as the response variable.

(b) QQ plot for the model with measured_bp_sys as the response variable.

Figure 2: QQ plots for measured_bp_dia and measured_bp_sys

Fig.2 shows the quantiles of the residuals plotted against theoretical normal quantiles. We can see that the data points for *measured_bp_dia* almost follow the reference line but with some deviation in the end. On the other hand, data points for *measured_bp_sys* is also following the straight line but the points are more aligned to the reference line as compared to *measured_bp_sys*, although there is a very little deviation at the end. We can assume that both variables are following the linear model assumptions.

4.4 Results of the Linear Regression Model

In this section, we present the parameter estimates obtained from fitting two models. One model uses *measured_bp_dia* as the response variable, while the other model uses *measured_bp_sys*. The significance level (α) for both models is set to 0.05.

Table 5 displays the parameter estimates specifically for the variable *measured_bp_dia* as the response variable. Based on a significance level of 0.05, the coefficients of certain variables are found to be statistically significant, while others are not.

The statistically significant variables, as indicated by the rejection of the null hypothesis ($H_0 : \beta_j = 0$) and the corresponding non-zero values in their 95% confidence intervals, include dummy variables such as *terminal*, *felt_health_condition_2*,

Table 5: Parameters estimates for measured_bp_dia as the response variable.

	coef	std err	t	P> |t|	[0.025	0.975]
Intercept	18.7946	2.392	7.856	0.000	14.105	23.484
terminal_2	1.9449	0.247	7.876	0.000	1.461	2.429
terminal_3a	-1.2479	0.419	-2.981	0.003	-2.068	-0.427
terminal_3b	-0.6710	0.290	-2.316	0.021	-1.239	-0.103
federal_state_Kärnten	-2.2591	1.471	-1.536	0.125	-5.143	0.625
federal_state_Niederösterreich	-1.6394	1.302	-1.259	0.208	-4.192	0.913
federal_state_Oberösterreich	-1.7391	1.444	-1.204	0.229	-4.571	1.092
federal_state_Salzburg	-0.4074	1.777	-0.229	0.819	-3.890	3.075
federal_state_Steiermark	-1.4835	1.160	-1.279	0.201	-3.757	0.789
federal_state_Tirol	-1.6314	1.891	-0.863	0.388	-5.337	2.074
federal_state_Vorarlberg	0.5143	2.923	0.176	0.860	-5.215	6.243
federal_state_Wien	-0.9295	1.292	-0.720	0.472	-3.462	1.603
federal_state_not_applicable	-1.8763	1.394	-1.346	0.178	-4.609	0.856
felt_health_condition_2	-0.7195	0.229	-3.140	0.002	-1.169	-0.270
felt_health_condition_3	-0.2442	0.327	-0.746	0.456	-0.886	0.398
felt_health_condition_4	-0.3979	0.969	-0.411	0.681	-2.296	1.501
felt_health_condition_5	3.7336	1.613	2.314	0.021	0.571	6.896
gender_m	1.2023	0.208	5.779	0.000	0.795	1.610
is_smoker_True	0.0357	0.282	0.126	0.899	-0.518	0.589
is_diabetic_True	-0.2325	0.291	-0.800	0.424	-0.802	0.337
has_cholestrol_True	0.1427	0.270	0.529	0.597	-0.386	0.671
in_treatment_True	-1.6391	0.329	-4.985	0.000	-2.284	-0.995
month_Aug	0.7421	0.786	0.944	0.345	-0.799	2.283
month_Jul	-0.5602	0.898	-0.624	0.533	-2.321	1.200
month_Jun	0.4529	0.825	0.549	0.583	-1.165	2.071
month_May	0.5684	0.778	0.731	0.465	-0.956	2.092
month_Nov	21.4579	10.456	2.052	0.040	0.962	41.954
month_Oct	-3.2398	0.782	-4.142	0.000	-4.773	-1.707
month_Sep	-2.7086	0.791	-3.425	0.001	-4.259	-1.158
day_Monday	0.1772	0.426	0.416	0.677	-0.658	1.012
day_Saturday	0.4438	0.383	1.157	0.247	-0.308	1.195
day_Sunday	0.2094	0.362	0.578	0.563	-0.500	0.919
day_Thursday	-0.9265	0.401	-2.312	0.021	-1.712	-0.141
day_Tuesday	-1.1554	0.435	-2.659	0.008	-2.007	-0.304
day_Wednesday	-1.0129	0.418	-2.421	0.015	-1.833	-0.193
measured_bp_sys	0.5290	0.006	90.745	0.000	0.518	0.540
age	-0.0892	0.007	-12.168	0.000	-0.104	-0.075
hour	-0.0898	0.046	-1.952	0.051	-0.180	0.000
temp	0.6937	0.155	4.466	0.000	0.389	0.998
humidity	0.0669	0.016	4.136	0.000	0.035	0.099
temp_min	-0.3059	0.101	-3.031	0.002	-0.504	-0.108
temp_max	-0.3587	0.077	-4.634	0.000	-0.510	-0.207

felt_health_condition_4, *gender_m*, *in_treatment_true*, *month_Nov*, *month_Oct*, *month_Sept*, *day_Tuesday*, *day_Wednesday*, and *day_Thursday* with respect to their reference categories, as well as continuous variables like *measured_bp_sys*, *Age*, *Humidity*, *temp_min*, and *temp_max*.

On the other hand, the remaining variables are not statistically significant, as their 95% confidence intervals include the value 0. Therefore, the null hypothesis that these variables do not have a significant impact on determining diastolic blood pressure cannot be rejected.

While interpreting any estimated coefficients, the values for the other ones are assumed to be constant. According to the trained model, if a participant is a smoker it will increase the diastolic blood pressure by 0.04mmHg approximately with respect to the person who doesn't smoke. Similarly, for males, the diastolic blood pressure will increase by 1.2mmHg with respect to the females. People who reported having cholesterol have an increase of 0.143mHg in diastolic blood pressure compared to those who don't have cholesterol. Moreover, participants who are diabetic have a decrease 0.233mmHg approximately in diastolic blood pressure with respect to people who aren't diabetic. Compared to the people who aren't receiving the hypertension treatment, people who received hypertension treatment tend to decrease their diastolic blood pressure by 1.64mmHg approximately.

In addition, a rise of one unit in systolic blood pressure is associated with an increase of approximately 0.53mmHg in diastolic blood pressure. Furthermore, an increase of one unit in weather temperature leads to a rise of approximately 0.7mmHg in diastolic blood pressure. Lastly, an approximate increase of 0.07mmHg in diastolic blood pressure is observed with a one-unit increase in humidity.

The adjusted R-squared value for the model is about 0.471, this tells us that the model is able to explain about 47% of the variation in diastolic blood pressure of the participants. Moreover, the model was also validated against the testing data and an adjusted R^2 squared was calculated which was around 0.452.

Furthermore, Table 6 displays the parameter estimates specifically for the variable *measured_bp_sys* as the response variable. Based on a significance level of 0.05, the coefficients of certain variables are found to be statistically significant, while others are not.

The statistically significant variables, as indicated by the rejection of the null hypothesis ($H_0 : \beta_j = 0$) and the corresponding non-zero values in their 95% confidence intervals, include dummy variables such as *terminal_2*, *terminal_3a*, *felt_health_condition_2*, *gender_m*, *is_diabetic_True*, *in_treatment_True*, *month_Jun*, *month_Jul*, *month_Aug*, *month_Oct*, *day_Saturday*, and *day_Wednesday* with respect to their reference categories, as well as continuous variables like *measured_bp_dia*, *Age*, *Humidity*, and *temp*.

However, the remaining variables lack statistical significance, as their 95% confidence intervals encompass the value 0. Consequently, we cannot reject the null hypothesis that these variables do not significantly influence the determination of systolic blood pressure.

Table 6: Parameter estimates for measured_by_sys as the response variable.

	coef	std err	t	P> |t|	[0.025	0.975]
Intercept	55.8822	2.968	18.828	0.000	50.064	61.700
terminal_2	-3.3801	0.310	-10.912	0.000	-3.987	-2.773
terminal_3a	0.3803	0.527	0.722	0.470	-0.652	1.413
terminal_3b	2.1488	0.364	5.905	0.000	1.436	2.862
federal_state_Kärnten	1.0256	1.851	0.554	0.579	-2.602	4.653
federal_state_Niederösterreich	0.7727	1.638	0.472	0.637	-2.438	3.984
federal_state_Oberösterreich	0.9796	1.817	0.539	0.590	-2.582	4.541
federal_state_Salzburg	-3.6848	2.234	-1.649	0.099	-8.065	0.695
federal_state_Steiermark	0.1215	1.459	0.083	0.934	-2.738	2.981
federal_state_Tirol	-2.1700	2.378	-0.913	0.362	-6.832	2.492
federal_state_Vorarlberg	-3.1340	3.676	-0.852	0.394	-10.340	4.072
federal_state_Wien	-0.4422	1.625	-0.272	0.786	-3.627	2.743
federal_state_not_applicable	-1.0562	1.753	-0.602	0.547	-4.493	2.381
felt_health_condition_2	1.0774	0.288	3.738	0.000	0.512	1.642
felt_health_condition_3	0.0675	0.412	0.164	0.870	-0.740	0.875
felt_health_condition_4	0.3368	1.218	0.276	0.782	-2.051	2.725
felt_health_condition_5	-3.1567	2.030	-1.555	0.120	-7.135	0.822
gender_m	1.6332	0.262	6.242	0.000	1.120	2.146
is_smoker_True	-0.2018	0.355	-0.568	0.570	-0.898	0.495
is_diabetic_True	0.7636	0.365	2.090	0.037	0.047	1.480
has_cholestrol_True	-0.2873	0.339	-0.847	0.397	-0.952	0.378
in_treatment_True	6.0966	0.410	14.880	0.000	5.293	6.900
month_Aug	-2.8868	0.989	-2.920	0.004	-4.825	-0.949
month_Jul	-2.7214	1.129	-2.410	0.016	-4.935	-0.508
month_Jun	-3.2285	1.038	-3.111	0.002	-5.263	-1.194
month_May	-1.1178	0.978	-1.143	0.253	-3.035	0.799
month_Nov	0.6244	13.156	0.047	0.962	-25.163	26.412
month_Oct	1.9373	0.985	1.968	0.049	0.007	3.867
month_Sep	1.0697	0.995	1.075	0.282	-0.881	3.021
day_Monday	-0.7203	0.536	-1.345	0.179	-1.770	0.330
day_Saturday	-1.2049	0.482	-2.498	0.012	-2.150	-0.260
day_Sunday	-0.6766	0.455	-1.486	0.137	-1.569	0.216
day_Thursday	0.3130	0.504	0.621	0.535	-0.675	1.301
day_Tuesday	0.5674	0.547	1.038	0.299	-0.504	1.639
day_Wednesday	1.1153	0.526	2.119	0.034	0.084	2.147
measured_bp_dia	0.8370	0.009	90.745	0.000	0.819	0.855
age	0.2430	0.009	27.089	0.000	0.225	0.261
hour	-0.0651	0.058	-1.125	0.261	-0.178	0.048
temp	-0.3923	0.196	-2.006	0.045	-0.776	-0.009
humidity	-0.0912	0.020	-4.479	0.000	-0.131	-0.051
temp_min	-0.0228	0.127	-0.179	0.858	-0.272	0.226
temp_max	0.1625	0.097	1.668	0.095	-0.029	0.354

For interpreting any coefficient in Table.6, the values for the other ones are assumed to be constant. There is a 1.632mmHg increase in the systolic blood pressure for males with respect to reference category females. Comparing to people who are diabetics have an increase of 0.736mmHg of systolic blood pressure with who doesn't have diabetes. People who do smoke have a decrease of 0.2mmHg in the systolic blood pressure with respect to people who doesn't smoke as reference, plus having cholesterol also an adverse effect on the systolic blood pressure it can be seen that it decreases 0.3mmHg approximately with respect to people do not have cholesterol.

Moreover, one year increase in the age increase the systolic blood pressure by 0.24mmHg. Furthermore, one unit increase in measured diastolic blood pressure will increase the systolic blood pressure by 0.84mmHg approximately. Lastly, weather temperature and humidity have an negative effect on systolic blood pressure.

Finally, the adjusted R-squared value for this model is 0.542, indicating that approximately 54% of the variation in the systolic blood pressure of the participants can be explained by the variables included in the model. Additionally, model was validated against the testing data and adjusted R^2 was calculated for testing data which was around 0.525.

4.5 Fitting Regression Trees and Random Forests

In this particular section, we employed regression trees and random forests to fit models using systolic and diastolic variables as the response variables. Once again, we adopted a 70/30 approach to split the data into training and testing datasets. For each response variable, we fitted two regression trees and two random forests. One set of models used the default parameters, while the other set had tuned parameters. In this section, the identical set of features as those used in the linear model (refer to Table. 4) were employed.

To obtain the optimal set of parameters, cross-validation was employed, as described in Section 3.8. The 10-fold cross-validation technique was utilized to identify the best set of parameters for *measured_bp_dia* and *measured_bp_sys* as a response variable, which are outlined below:

Table 7: Best set of parameters obtained by 10-CV

Parameters	Value for measured_bp_sys	Value for measured_bp_dia
max_depth	10	25
max_features	40	39
min_sample_leaf	70	90
splitter	best	best

To provide an explanation of the parameters, we have the following:

- The parameter *max_depth* determines the maximum depth that a tree can reach. If set to "none," the tree expands until every node becomes a leaf with only one sample.
- The parameter *max_feature* controls the maximum number of features that can be considered for a split at each node.
- The *min_sample_leaf* parameter ensures that each leaf node must contain a minimum number of samples to be considered valid.
- The *splitter* parameter can take two values: "best" and "random." When set to "best," the algorithm selects the best split based on a specific criterion. On the other hand, when set to "random," the algorithm chooses the best random split.
- An extra parameter in the random forest is used and referred to as *n_estimators*. We decided to use the default value of 100 for this parameter, indicating the number of regression trees utilized within the random forest algorithm.
- Lastly, the splitting criteria was set to *squared_error*, which corresponds to the mean squared error. For more details on this criterion, please refer to Section 3.2.

These parameters play a crucial role in controlling the behavior and performance of the regression tree algorithm. Properly adjusting them according to the characteristics of your data and the desired outcome is essential for achieving accurate and reliable results.

4.6 Results of Regression Trees and Random Forests

Once the optimal parameters were identified, we proceeded to fit two regression trees and two random forests using *measured_bp_sys* as the response variable. The resulting outcomes of the regression trees and random forests are presented in the following table:

	Model	Train Mean Sq Error	Test Mean Sq Error	Train R2	Test R2	Train Adjusted R2	Test Adjusted R2
1	Tree (Base)	0.040000	333.930000	1.000000	0.100000	1.000000	0.090000
2	Tree (Fine-tuned)	164.700000	180.690000	0.560000	0.510000	0.550000	0.510000
3	RF (Base)	24.380000	172.040000	0.930000	0.530000	0.930000	0.530000
4	RF (Fine-tuned)	161.030000	172.190000	0.570000	0.530000	0.560000	0.530000

Table 8: Results of Trees/Forests for measured_bp_sys as response variable.

Upon examining Table 8, it is evident that the fine-tuned tree and random forest models, utilizing the best parameters obtained through cross-validation, outperform their respective base versions with default parameters. Notably, the base tree exhibits a perfect adjusted R^2 value for training data, but performs poorly on the testing data,

indicating overfitting. Conversely, the fine-tuned tree displays an adjusted R^2 of 0.55 for training and 0.51 for testing, demonstrating a smaller discrepancy between the two. The fine-tuned random forest also demonstrates improved performance compared to the base version. However, the test mean squared error (MSE) and test adjusted R^2 remain relatively similar for both the base and fine-tuned models. In contrast, the adjusted R^2 for training data experiences a significant drop from 0.93 in the base random forest to 0.56 in the fine-tuned random forest.

For *measured_bp_dia* as a response variable same thing was done i.e fitted two regression tree and two random forests. The result of the fitted models is shown below :

	Model	Train Mean Sq Error	Test Mean Sq Error	Train R2	Test R2	Train Adjusted R2	Test Adjusted R2
1	Tree (Base)	0.180000	218.970000	1.000000	-0.050000	1.000000	-0.060000
2	Tree (Fine-tuned)	101.540000	115.910000	0.490000	0.450000	0.480000	0.440000
3	RF (Base)	14.940000	109.180000	0.920000	0.480000	0.920000	0.470000
4	RF (Fine-tuned)	100.140000	112.030000	0.490000	0.460000	0.490000	0.460000

Table 9: Results of Trees/Forests for measured_bp_dia as response variable.

Based on the information presented in Table.9, it is evident that the base tree model exhibits poor performance with a low training mean squared error (MSE) and a significantly higher test MSE. Likewise, the adjusted training R^2 value is considerably high, while the adjusted testing R^2 value is very low, indicating overfitting.

On the other hand, the optimized tree model performs relatively well compared to the base tree, as the training and testing MSE values are closer in magnitude. The same observation holds true for the training and testing adjusted R^2 values.

Moreover, the random forest model demonstrates a similar pattern, with a notable difference between the training and testing MSE values, as well as the training and testing adjusted R^2 values. However, the base random forest results slightly outperform the base tree model if compare with respect to adjusted R^2 values.

Lastly, the fine-tuned random forest model produces slightly superior results compared to the fine-tuned trees.

4.7 Comparison between the results of Linear models and Trees/Forest

In this section, a comparison analysis was done between the linear model, regression trees, and random forest. At first, we will consider all the models where the response variable is *measured_bp_dia* and following table shows the results of different models:

Table 10: Models(measured_bp_dia) and the results

Model	Train Adjusted R^2	Test Adjusted R^2
LM	0.471	0.452
Tree(Base)	1.000	-0.060
Tree(Fine-tuned)	0.480	0.440
RF(Base)	0.920	0.470
RF(Fine-tuned)	0.490	0.460

The results are presented in Table.10 indicates that the Linear model, Fine-tuned Tree, and Fine-tuned Random Forest exhibit almost identical adjusted R^2 values. In contrast, the Base Tree performs the poorest among all models, with the highest training adjusted R^2 but significantly lower testing adjusted R^2 which is a classic example of overfitting. Base Random forests have slightly better metrics among all the models.

Similarly, all the models which involve *measured_bp_sys* as a response variable are considered, and the results are shown below:

Table 11: Models(measured_bp_sys) and the results

Model	Train Adjusted R^2	Test Adjusted R^2
LM	0.542	0.525
Tree(Base)	1.000	0.090
Tree(Fine-tuned)	0.550	0.510
RF(Base)	0.930	0.530
RF(Fine-tuned)	0.560	0.530

By examining Table.11, we can note a similar pattern as described earlier. The adjusted R^2 values for the Linear Model, Fine-tuned Tree, and Fine-tuned Random Forest are quite similar. However, the Base Tree once again displays signs of overfitting and one again base random forest have slightly better results.

5 Summary

The dataset analyzed contains the data which was collected during an exhibition held in *Bruck An Der Mur* in 2006. The data was provided by the instructors of this course. The data includes twelve parameters and have more than 16000 observations. The aim of this analysis is to fit different machine learning models to explain the response variable with respect to independent variables.

At first, a descriptive analysis was done on the data and the key findings of this analysis were that female participants outnumbered male participants by approximately 56 percent. Most of the participants do not report any smoking habits or cholesterol issues and most of them feel good about their health. Lastly, it was observed that the majority of the participants were from the state of Steiermark. After the descriptive analysis, some data pre-processing was done and some new variables were introduced, some were created from the existing features, and some were taken from external datasets.

After descriptive analysis and data pre-processing, the best subset selection was used to get the best set of features then two linear regression models were trained one with *measured_bp_dia* and one with *measured_bp_sys* as a response variable on those features provided by best subset selection. After that, estimated coefficients were interpreted. Furthermore, variables that are statistically significant or not were discussed. Lastly, confidence intervals and goodness of fit of both models were interpreted.

Afterward, two regression trees and two random forests for each of the response variables were fitted. The first regression tree and random forest were fitted using default parameters for both the response variables then the second regression tree and random forest were fitted with the optimal parameters obtained from the cross-validation technique. Later, the result of these trees and random forests were discussed for both variables.

Lastly, results from the regression tree/forests and linear regression were discussed and a comparative analysis was done. It was found that the Linear model, Tree(Fine-tuned), and random forest (Fine-tuned) for both the response variables have almost the same adjusted R^2 values for both training and testing data.

In future studies, we can use the methods from this project on bigger datasets that include information from many cities or countries. We can also try different machine learning techniques to solve different problems, like figuring out who needs treatment for hypertension based on the given information.

Bibliography

Akinkunmi, Mustapha. 2019. *Introduction to Statistics Using R.* Morgan and Claypool Publishers.

Barnett, Adrian, Sans, Susana, Salomaa, Veikko, Kuulasmaa, Kari, & Dobson, Annette. 2007. The effect of temperature on systolic blood pressure. *Blood pressure monitoring*, **12**(07), 195–203.

Fahrmeir, Ludwig, Kneib, Thomas, Lang, Stefan, & Marx, Brian. 2013. *Regression: Models, Methods and Application.* Springer, Berlin, Heidelberg.

Harris, Charles R., Millman, K. Jarrod, van der Walt, Stéfan J, Gommers, Ralf, Virtanen, Pauli, Cournapeau, David, Wieser, Eric, Taylor, Julian, Berg, Sebastian, Smith, Nathaniel J., Kern, Robert, Picus, Matti, Hoyer, Stephan, van Kerkwijk, Marten H., Brett, Matthew, Haldane, Allan, Fernández del Río, Jaime, Wiebe, Mark, Peterson, Pearu, Gérard-Marchant, Pierre, Sheppard, Kevin, Reddy, Tyler, Weckesser, Warren, Abbasi, Hameer, Gohlke, Christoph, & Oliphant, Travis E. 2020. Array programming with NumPy. *Nature*, **585**, 357–362.

Hastie, Trevor, Tibshirani, Robert, & Friedman, Jerome. 2001. *The Elements of Statistical Learning.* Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Hunter, John D. 2007. Matplotlib: A 2D graphics environment. *Computing in science & engineering*, **9**(3), 90–95.

James, Gareth, Witten, Daniela, Hastie, Trevor, & Tibshirani, Robert. 2013. *An Introduction to Statistical Learning: with Applications in R.* Springer.

McKinney, Wes, *et al.* 2010. Data structures for statistical computing in python. *Pages 51–56 of: Proceedings of the 9th Python in Science Conference*, vol. 445. Austin, TX.

Murphy, Kevin P. 2013. *Machine learning : a probabilistic perspective.* Cambridge, Mass. [u.a.]: MIT Press.

Van Rossum, Guido, & Drake, Fred L. 2009. *Python 3 Reference Manual.* Scotts Valley, CA: CreateSpace.

Visual Crossing. 2006. *Global Forecast and History Data.* `https://www.visualcrossing.com/weather-data`, Last accessed on 2017-11-30.

Waskom, Michael, Botvinnik, Olga, O'Kane, Drew, Hobson, Paul, Lukauskas, Saulius, Gemperline, David C, Augspurger, Tom, Halchenko, Yaroslav, Cole, John B., Warmenhoven, Jordi, de Ruiter, Julian, Pye, Cameron, Hoyer, Stephan, Vanderplas, Jake, Villalba, Santi, Kunter, Gero, Quintero, Eric, Bachant, Pete, Martin, Marcel, Meyer, Kyle, Miles, Alistair, Ram, Yoav, Yarkoni, Tal, Williams, Mike Lee, Evans, Constantine, Fitzgerald, Clark, Brian, Fonnesbeck, Chris, Lee, Antony, & Qalieh, Adel. 2017 (Sept.). *mwaskom/seaborn: v0.8.1 (September 2017).*