Wrocław University
of Science and Technology

Faculty of Pure and Applied Mathematics
Field of study: Mathematics and Statistics
Specialty: Statistics and Data Analysis

Bachelor's Thesis

MACHINE LEARNING USING METEOROLOGICAL DATA

Jan Solarz

keywords:
machine learning, statistics, prediction, meteorological data, agriculture.

short summary:
The purpose of the work is an attempt to predict rainfalls with the use of machine learning. The problem of the importance of water is shown based on one of the branches of the economy which is agriculture. Knowing that the knowledge about rainfalls can contribute to proactive strategies in agriculture, the goal of this work was to evaluate the performance of this task based on an acquired set of meteorological data. In this thesis, several paths of approach to the problem are chosen, using the previously illustrated tools of statistical domains in order to obtain the best results.

Supervisor	dr inż. Radosław Michalski	....	....
	Title/degree/name and surname	grade	signature

*For the purposes of archival thesis qualified to:**

a) category A (perpetual files)

b) category BE 50 (subject to expertise after 50 years)

** delete as appropriate*

stamp of the faculty

Wrocław, 2021

Contents

Introduction 3

1 Introduction to the subject of meteorology 5
1.1 Meteorology and prediction in agriculture . . . 5
1.2 Water as a power parameter for plants . . . 6

2 Goals, conceptions, assumptions 11

3 Theoretical part 13
3.1 Supervised methods in Machine learning . . . 13
3.2 Overview of methods used in analyzes . . . 14
3.2.1 Regression models . . . 14
3.3 Classification methods . . . 22
3.3.1 Decision Trees . . . 22
3.3.2 Naive Bayes Classifier . . . 33

4 Set of meteorological data of Wroclaw 37
4.1 Data acquisition . . . 37
4.2 Preview and initial analysis . . . 38

5 Experimental part 43
5.1 Regression models . . . 44
5.1.1 Simple Model 1. . . . 44
5.1.2 Model 2- deleting two variables . . . 46
5.1.3 Deleting given observations- Cook's Distance . . . 46
5.1.4 Future data prediction . . . 48
5.1.5 Summary of regression models . . . 51
5.2 Classification Methods . . . 52
5.2.1 Decision Trees . . . 53
5.2.2 Naive Bayes . . . 55
5.3 Methods Comparison and Summary . . . 56

6 Summary and conclusion 59

Bibliography 60

Introduction

The weather is a phenomenon that surrounds us and changes every day. Moreover, it is very difficult to predict. The weather influences significantly many our actions and living conditions as well as all other animals and plants. Today's world has an innumerable number of tools thanks to which it is able to conduct weather forecasting. An ordinary man is having contact with weather forecasts only when watching television or using a smartphone. This information is needed to plan our actions, for instance as a choice of outfit when leaving home or when planning a holiday trip.

A better understanding of the weather through prediction of its behavior could contribute to the prevention of some unexpected or harmful events, among others droughts, floods, extension of the frost period, or hail. In the long term, very negative symptoms of weather changes are visible, which should be noted in advance so that we could plan long-term actions accordingly.

Compared to data from 1961, actual performance when it comes to animal culture and plant cultivation decreased by 21 percent. Everything because on earth gets warmer.This is the finding of conducted research in Cornell University [12]. These years, the decrease in the amount of food is very well covered by technological progress, the use of new, more efficient fertilizers and globalization of trading. Thanks to this, agricultural performance still keeps up with our growing demand for food. The UN predicts that by 2050 the number of people living in the land will increase to over 9 billion, and therefore the production of food will have to increase by about 70 percent to feed us. This means that plant production will have to increase every year by almost one billion tonnes. Achieving a critical point seems to be only a matter of time [10].

> "In studies assessing the influence of global warming on agricultural performance, the following methodology was applied: all factors consisting of performance, i.e. a working force, use of fertilizers and equipment (including technology) were measured. The share of these factors was analyzed in the time interval from 1961 to 2019. - Impact of climate change turned out to be bigger than I expected"*Ariel Ortiz-Bobea, economist from Cornell University.* [12]

This statement is not obvious. In recent decades, agricultural performance increased. Farmers are working better, but it takes place in worse conditions caused by the waves of heat, long-lasting drought, extreme weather phenomena, which are destroying crops. Another problem is that the intensification of agriculture to increase production in itself caused serious environmental damage. Progressive soil erosion as a result of agricultural processes, environmental pollution used in agriculture with pesticides, increasing consumption of drinking water resources and emissions of a huge amount of greenhouse gases.

These are elements additionally driving global warming and contributing to the natural decline in the efficiency of our planet when it comes to agricultural production. The use of nitrogen fertilizers increased by an average of 800% all over the world from 1961. Water consumption for irrigation of plants increased by 100%.

The changing climate is one of the more serious problems in which we live. The use of the most optimal method for predicting weather conditions could contribute not only to improving the efficiency of actions in various sectors of the economy, but would also have a positive impact on the environment.

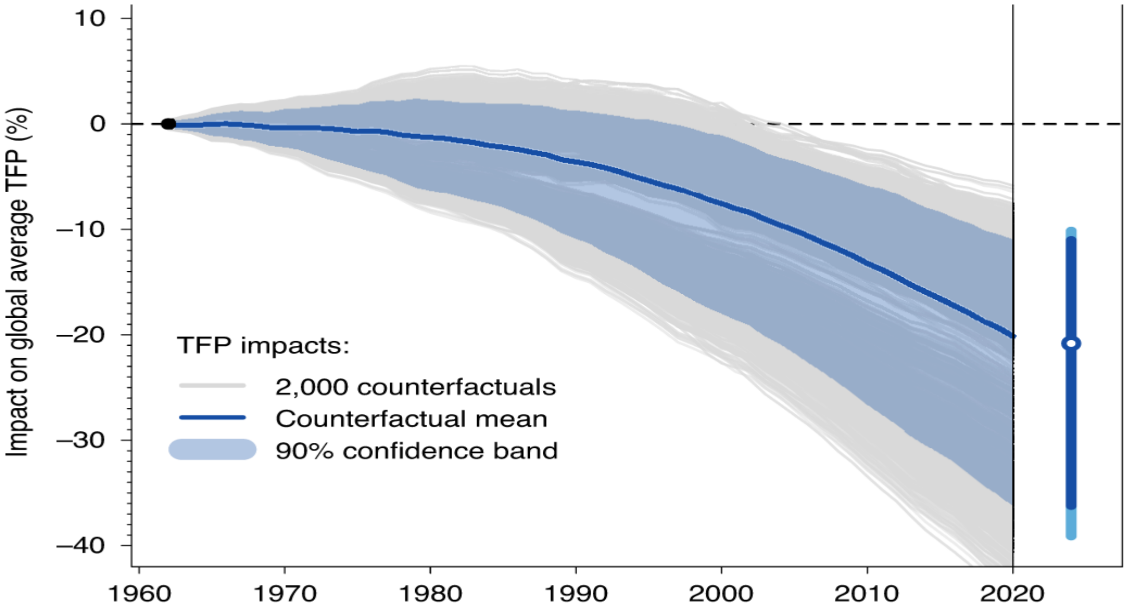


Figure 1: Chart of the impact o global average TFP in years 1961-2020 [13].

In the Figure 1 above is shown the impact of Anthropogenic climate change (ACC) on Total Factor Productivity (TFP) in years 1961-2020 in the world. The blue line represents mean of alternative pathways of different factors affecting on the TFP. The error bars on the right indicate 90% and 95% confidence intervals for the impact on 2020. A downward trend is definitely visible.

Chapter 1

Introduction to the subject of meteorology

Practicing agriculture consists in investing capital in a field on which the cultivation of a particular species of plants will lead in a given season. The farmer through appropriate treatments can influence the development of plants. This can be achieved by using various types of chemicals, for instance as spraying or fertilizers, or work aimed at fertilizing the soil itself. One of the frequent problems during high temperatures and lack of rain for a long time is the poor access of plants root systems to oxygen and water, however there are methods of preventing such situations with the help of specialized mechanical treatments.

In agriculture there are many independent variables that affect the final results. If there was a chance to be prepared earlier for the change, it would make it easier for agriculture to act proactively, would reduce over investing farmlands which translates into the saving of capital but also protecting the environment (reduction of chemicals). People got used to treating weather changes as something completely random, typically they are not aware that it can be possible to interpret some symptoms-seasonal trends.

1.1 Meteorology and prediction in agriculture

Weather information, which have on a daily basis are based on the occurring phenomena such as the way of moving clouds or the quality of atmospheric fronts. However, the weather parameters from a given day are the key on which forecasts for next period of time are based. With each further period, the prediction is less reliable and burdened with greater uncertainty. In this work problem of prediction will come in an analytical manner based only on historical data.

Agricultural economy is largely based upon weather conditions. For analyzing the crop productivity, rainfall prediction could be very useful for all farmers, but also for other economy fields like energetic. Rainfall Prediction is the application of science and technology to predict the state of the atmosphere. It is important to exactly determine the rainfall for effective use of water resources, crop productivity and preplanning of water structures. Observed climate change is already affecting on amount of food through increasing temperatures, changing precipitation patterns, and greater frequency of some extreme events. Studies that separate out climate change from other factors affecting crop yields have shown that yields of some crops (e.g., maize and wheat) in many lowerlatitude regions have been affected negatively by observed climate changes, while in many higher-latitude regions, yields of some crops (e.g., maize, wheat, and sugar beets) have been affected positively over recent decades [10].

Machine learning techniques is used to discover new patterns from large data sets and has had a profound impact on the society by solving real-life problems. The tools aims to extract useful knowledge and present it in understandable way. This knowledge can be utilized for future.

Climate change analysis explore the behavior of weather during a specific period of time. The key characteristic behind climate change lies in the nature of its data that is captured in time point manner. Most techniques tend to be supervised learning methods, where the key point behind them is selecting an appropriate model with appropriate features.

Weather forecasting is a challenging task to predict factors associated with specific variables such as pressure, temperatures, duration of storm etc. Since there are many different supervised learning techniques, different performances could be gained from them. In addition, weather data could be formed in different forms including long-terms (e.g. months) and short-terms (e.g. daily). Therefore, selecting an appropriate technique for a specific duration of rainfall is a crucial task.

In the forecasting weather at the regional and national level, a wide range of precipitation methods is applied. Basically, two approaches are used to predict precipitation: empirical and dynamic [4]. The empirical approach is based on the analysis of historical data regarding rainfall and their relationship with various atmospheric and oceanic variables in different parts of the world.
The work will be presented an empirical approach based on supervised methods of machine learning.

1.2 Water as a power parameter for plants

Care of plants during the vegetation period require a different amount of water available for plants. The weather has a huge impact on plant growing. Reaction of various plant species, and even varieties for water deficiency are different in individual development phases. The amount of water in the soil and its availability for plants affects vegetation and yielding.

In the soil water there are food ingredients necessary for the proper growth of plants and only with it can be collected by the plant. In recent years, a decisive factor about the height of agricultural plants becomes water. The amount and quality of water in the soil are varied depending on various factors: way of use, applied agrotechnical treatments, area sculpture, soil construction and its properties. A basic body in plants responsible for water and nutrients retrieving are the roots. The water available for the roots of plants comes from atmospheric precipitation and a slopes, from the level of groundwater mirror by an unsaturated soil profile zone.

Water supplied in a timely manner and suitable quantities brings tangible benefits in the form of better health of plants, which translates into a larger and better quality yield. This is confirmed by research carried out by employees and in Skierniewice [8], which evidently show the legitimacy of building irrigation systems. For comparison, the size of the yield from a certain species of plants cultivated in Poland, which were irrigated by about 20-30% more compared to plants without irrigation increased by over 50 %. Delivery by 200% more water than it results from real demand raised a yield only by about 10% relative to the average yield. This means a lot for agriculture development. Understanding the needs of plants and using methods that would use most of rainwater in proper way could have multi-dimensional benefits.

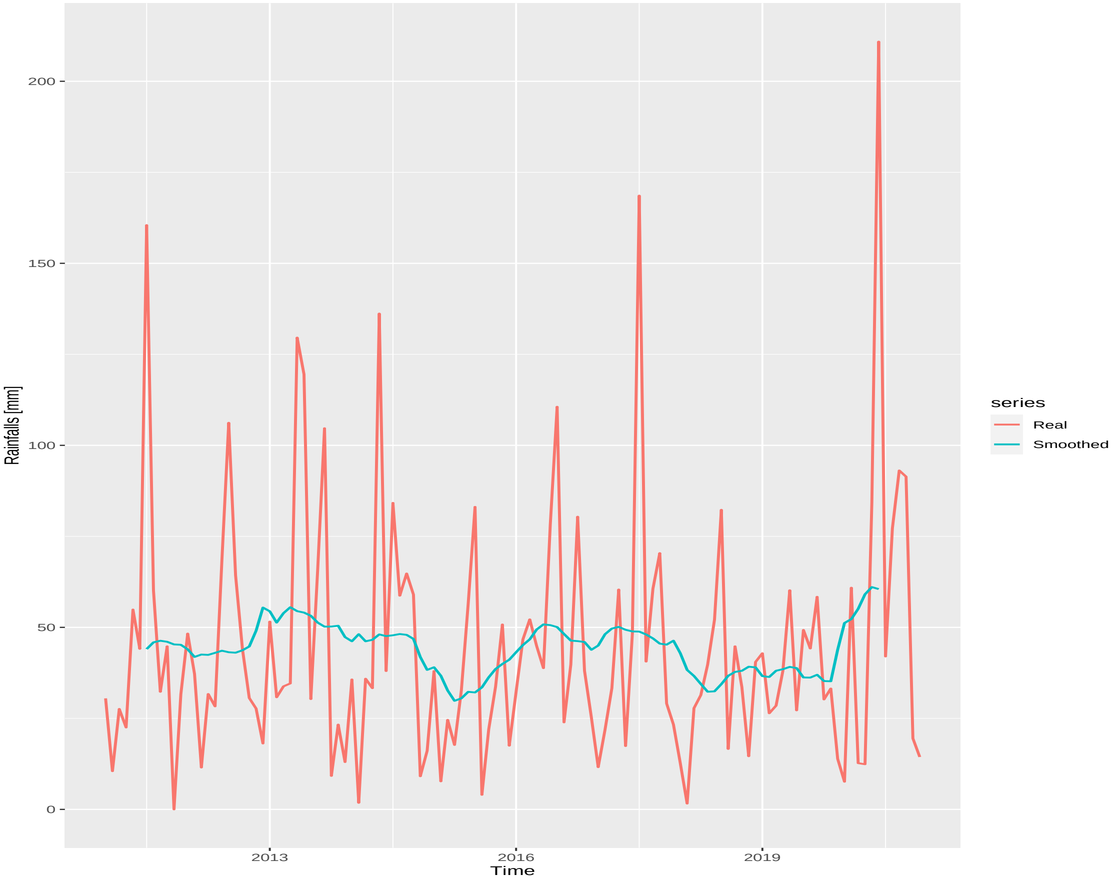


Figure 1.1: Chart of monthly rainfalls in Wrocław in years 2011-2020

Based on reports drawn up by the Central Statistical Office [15] on the estimated results of cereal yields and weather data in Poland can be considered that there is a relationship between rainfall and plant yield.

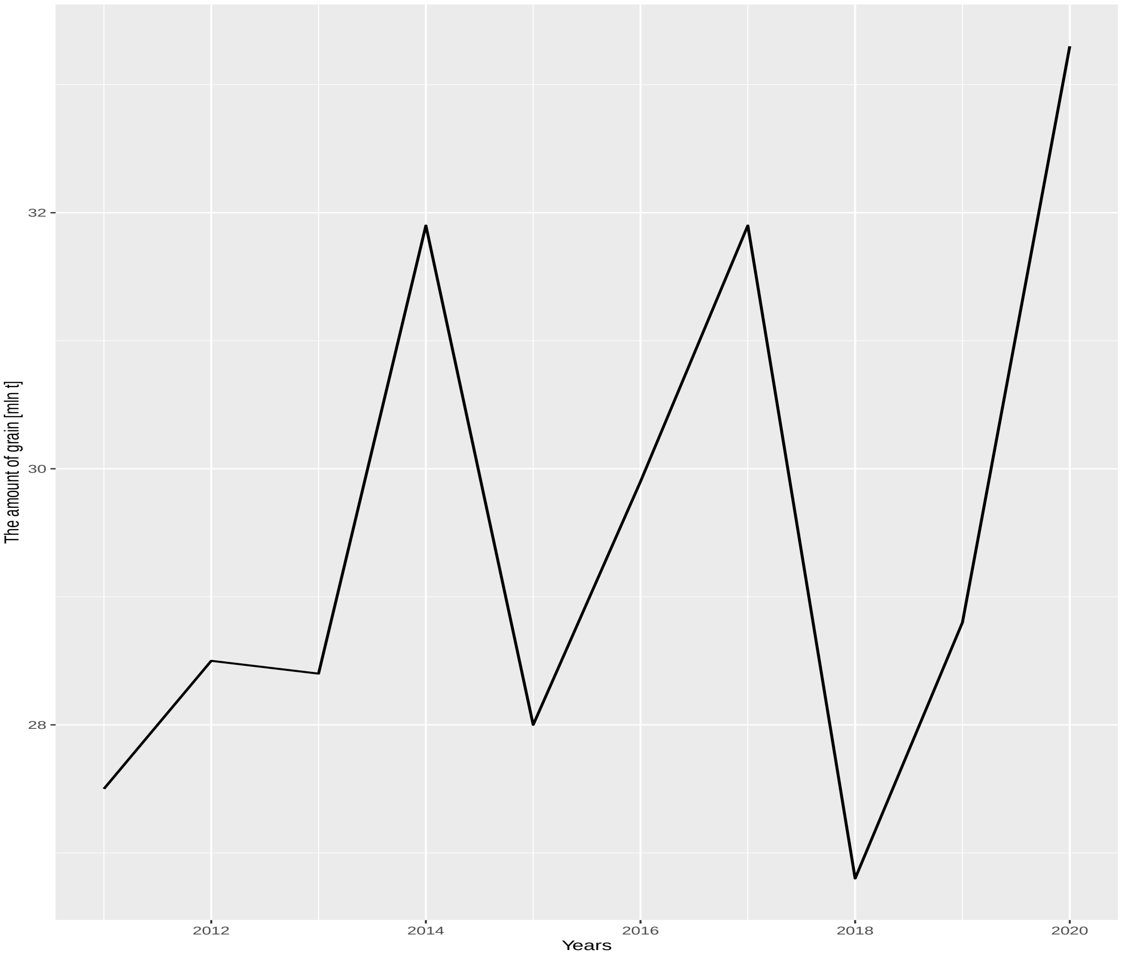


Figure 1.2: Chart of harvested cereals in Poland in the years 2011-2020

Year 2020 was a breakthrough for Polish agriculture after three years of drought. It is estimated that the collections of total cereals were about 16 % greater than harvest in 2019 and amounted to approximately 33.5 million tonnes. The total cereal set defines as harvest of all cereal species grown in Poland such as wheat, barley, rye, etc. and cereal blends. Total rainfall in 2020 in May, June, July amounted to 76.8 mm, 119.6 mm, 54.5 mm, respectively. Concerning standards, i.e. average rainfall from 1981-2010, this is another 117.2, 164.4, 63.4 percent.

In the Figures 1.1 and 1.2 the comparison between rainfalls and the results of cereal collections in years 2011-2020 in Wrocław is shown. The Figure 1.1 shows the data of the monthly precipitation as a time series smoothed by the Moving Average Method. In the beginning in original data we can observe the occurring seasonality in rainfall, which depends on the seasons in Poland. Periods of the years are noticeable where rain was definitely too low like in years 2015 and 2017-2019. Such a phenomenons can be very catastrophic not only for agriculture but also for the whole country economy. In the Figure 1.2 one can notice the specific years where collected cereals are in a satisfied level. In that stage we can assume the similarity, that rainfalls have on of the biggest impact for cereal grow.

In the EEA (European Environment Agency) report [16], attention was paid to the fact that at a time when the countries of Southern Europe still struggle with huge problems of water deficiency, its deficit also grow in the north of the continent. In addition, along with the climate change, the severity and frequency of the occurrence of drought in the future increases, which further increases the water deficit, especially in the summer months.

> "When it comes to water, we live above the state. A short-term solution to the problem of water shortage consisted in taking more water from the surface of the earth and resources accumulated under its surface. Water resources are excessively exploited, which exerts a strong impact on the quality and amount of remaining water and ecosystems that are dependent on it. We must reduce demand, limit the amount of water taken to a minimum and increase the efficiency of its consumption."[16]

In Europe, 44% of the booted water consumes for energy production, 24% for agriculture purposes, 21% for public deliveries, and 11 % for the needs of industry. However, there are significant differences in the sectoral water consumption throughout the continent behind these numeric data. In Southern Europe, for example, agriculture consumes 60% of the total amount of water consumed, and in some areas it reaches the level of 80%.
In all of Europe, surface water, such as lakes and rivers, provide 81% of the total amount of fresh water consumed and constitute the main source of water for the needs of industry, energy and agriculture.

Conclusion about water

During vegetation, the demand for plants on water changes from the development phase. It is usually the largest during the fastest increment of biomass, which generally falls at the end of the vegetative development phase and the beginning of generative bodies.

> "In Poland, potential plants constraints caused by water deficiencies reach up to 60 % [7]."

The most important stage of growth and development of plants occurs during the filling grain period. This provides how much the proper mass will have a grain and in what large extent will be used potential of the plant which have obtained during the entire vegetation period.

Chapter 2

Goals, conceptions, assumptions

In this chapter the goals of this work are presented.

An overview of selected supervised machine learning techniques will be presented. They are based on mathematical models, their details will be presented in the next chapter. The way they work will be demonstrated on an exemplary selected data set. In this thesis we concentrated on one data set that covers data collected in Wrocław.

Assumptions

- The analyses are based on historic weather parameters from Wrocław's data set. It is based on the climate that occurs in Poland, in the temperate zone.
- The main assumption of the work is that water is the most important parameter in the context of agriculture development, that is why the analyses focus on that parameter.
- The data refer to monthly values, which are averaged. There is no need to look at data in daily interval.
- A more general approach to the problem will increase the effectiveness of the result. The key is to obtain information whether the rains in May-June will be at a sufficiently high level.

Goals

- We decide to pick *Sum of Rainfall* as explained variable. That variable, which we want to predict.
- In the analyses will be used only techniques of machine learning, which will be described earlier.
- Not only prediction exact amount of rainfalls is important, but also learning about the behavior of the weather. We want to learn more about the dependencies between weather data.
- Comparison of the effectiveness of known methods of machine learning.

Suppositions

- It is impossible to make a long-term forecast several months in advance in day or even month intervals.
- There is a monthly seasonality in rainfall each year. There are also noticeable trends in precipitation in annual periods.
- The use of statistical treatments to improve the performance of the model, such as averaging values, getting rid of outliers, etc., significantly improves the predictive properties.
- There are high correlations between rainfall values and parameters related directly to water such as cloudiness, precipitation time, evaporation, etc. Models constructed on models without water related parameters will have much lower efficiencies.
- There is a large randomness of historical variables, therefore it is difficult to build an effective model on a specific training set. More parameters from previous years may reduce efficiency

Chapter 3

Theoretical part

3.1 Supervised methods in Machine learning

Supervised learning depends on training the model on based on examples marked with certain labels (classification problem) or numbers (in our case regression). The training set (X, Y) is delivered at the input of the system, where X is the set of vectors describing the training examples, and Y is the set of labels (or numbers) assigned to successive vectors. The goal of the learning process is to find the best approximation of the function:

$$h(x_i) = y_i$$

Finding such a function enables us to predict, what label should be assigned to any new x_k example that is not included at the learning stage. That is why for reliable checking how the algorithm works it is needed to test on the independent *Test sample*.

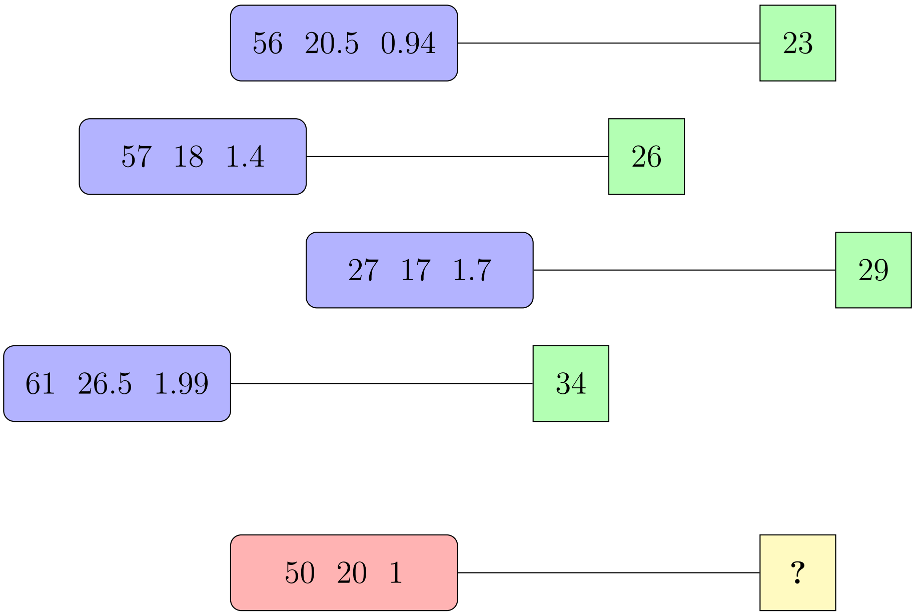


Figure 3.1: A diagram depicting the ideas of supervised learning

3.2 Overview of methods used in analyzes

3.2.1 Regression models

Regression models are one of the most popular methods of analyzing statistical data. The main idea of regression is to predict, forecast data for a certain variable based on other variables. In other words, what value will a given variable take when we know the value of another variable. Of course, in order to be able to "search" for the value of one variable on the basis of another variable, we must use regression analysis to construct a regression model, a model that will predict the value and level of a given feature with an assumed statistical error.

In regression models, it should be emphasized that we do not manipulate independent variables, the idea of regression is to measure **independent variables** (called predictors, explanatory or explained variables) and dependent variables (explained). In regression, one cannot speak strictly about the influence of one variable on another, because we do not manipulate the values of the variables. In regression with a variable or a set of variables, we want to explain some other variable, not affect it, because the significant relationship between the predictors and the dependent variable may only be in the coexistence of variables and not in the actual influence of one variable on the other.

Regression allows to estimate the **conditional expected value** of a random variable, called the **dependent variable**, for the given values of the vector of random variables - **explanatory variables**.

The use of regression in practice comes down to two phases:

- Constructing a model - building the regression model, i.e. a function describing how the expected value of the dependent variable depends on the explanatory variables. The model is constructed so as to best fit the data from the sample, containing both explanatory and explained variables (training set).

- Applying the model (scoring) - using the calculated model for data in which we know only the explanatory variables in order to determine the expected value of the dependent variable.

Multiple linear regression models

Linear regression model is an example of Multiple linear regression (there is just one independent variable)

Definition 3.1 (Multiple regression function)**.** In the model, we assume that the relationship between the dependent variable Y and the explanatory variables $X_1, \ldots, X_{p-1}$ has the form:

$$r(x) := \mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_{p-1} + \varepsilon.$$

Relationships between independent and dependent variables will be presented in analytical and matrix forms below.

$$\begin{cases} Y_1 &= \beta_0 + \beta_1 X_{1,1} + \cdots + \beta_p X_{1,p-1} + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_{1,2} + \cdots + \beta_p X_{2,p-1} + \varepsilon_2 \\ \vdots \\ Y_n &= \beta_0 + \beta_1 X_{1,n} + \cdots + \beta_p X_{n,p-1} + \varepsilon_n \end{cases}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{1,1} & \dots & X_{1,p-1} \\ 1 & X_{1,2} & \dots & X_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1,n} & \dots & X_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- The vector $\beta = (\beta_0, \dots, \beta_{p-1})^T$ is unknown and we want to estimate it using the least squares method.
- The random errors $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ are not observed
- It is assumed that: $\mathbb{E}(\varepsilon) = 0$
- Residuals determine the accuracy of matching the estimated values: $(e_1 = y_1 - \hat{y_1}, e_2 = y_2 - \hat{y_2}, \dots, e_n = y_n - \hat{y_n})$

In Gauss-Markov model [5] we have a vector Y , which:

$$Y = X\beta + \varepsilon,$$
$$\mathbb{E}(Y) = X\beta,$$
$$Cov(Y) = \sigma^2 \mathbb{I}$$

Prediction
The least squares estimator $\hat{\beta}$ is used to forecast the future value of the dependent variable Y corresponding to the value of $x = (x_0, \dots, x_{p-1})^T$ of the vector of explanatory variables: $\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_{p-1} x_{p-1}$

Model fitting

One of the criteria for selecting a model is **forward selection**. It starts with a model without explanatory variables, only with an intercept. Then it finds the explanatory variable (with the highest correlation with explanatory variable), which addition to the model will cause the greatest decrease in the SSE coefficient. Then, using the appropriate version of the F test, it checks whether this variable should be added to the model. We add it when p-value does not exceed the selected significance level. We repeat these steps until it is not possible to add another variable to the model.

Definition 3.2. The Pearson correlation coefficient is perhaps the most commonly used measure of the linear relationship between two normally distributed variables and is therefore often referred to simply as the "correlation coefficient". Typically, the Pearson coefficient is obtained using the least squares fit.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}}$$

$|r|$ is close to 1, we find that there is a strong linear relationship between the features X and Y.

- $|r| \leqslant 1$, where $|r| = 1$ means that all points $(x_1, y_1), \ldots, (x_n, y_n)$ lie on a straight line.
- If $|r|$ is close to 1, we find that there is a strong linear relationship between the features X and Y .
- The r coefficient is sensitive to outliers.

Let's denote:

$$\mathbf{Y} = [y_1, \ldots, y_n]^T,\ \overline{Y} = [\overline{y}, \ldots, \overline{y}]^T,\ \hat{Y} = [\hat{y_1}, \ldots, \hat{y_n}]^T$$

Main indicators of model fit:

- **The total sum of squares**: $SST = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
  It is the sum of the squared errors of the model that contains only the constant.
- **Regression sum of squares**: $SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
  Describes by how much the SST decreases when all explanatory variables are included in a regression model containing only a constant.
- **Sum of squared errors**: $SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2$
  This is the sum of the squared errors in the model that includes the constant and all explanatory variables.

Lemma 3.3. *The following property exists:*

$$SST = SSR + SSE$$

- **Multiple determination coefficient:**

$$R^2 = \frac{SST - SSE}{SST} = \frac{SSR}{SST}$$

  This coefficient describes the variability explained by a given model by the total variability of the data.
  $SSR = SST - SSE$ measures by how much SST decreases when all explanatory variables are included in a regression model containing only the constant, the multiple determination coefficient R^2 is an indicator of the adequacy of the model.
  $0 \leq R^2 \leq 1$. The closer R^2 to 1, the better fit the model to the data. Unfortunately, R^2 increases with p (number of independent variables). The Reason is that when another explanatory variable is added to the model, the SSE will definitely not increase, so R^2 increases with p.

- **Adjusted multiple determination coefficient:**

$$\text{adj}R^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{(n-1)(1-R^2)}{n-p}$$

  $adjR^2 \leq 1$ (usually not negative). The closer the $adjR^2$ is to 1, the better the model fits the data.
  The $adjR^2$ is slightly smaller than R^2 and does not need to increase with p, so it is more suitable for model selection than R^2. $(\frac{SST}{(n-1)})$ is constant, so the $adjR^2$ will decrease as a new variable is added, only as the error mean square $(\frac{SSE}{(n-p)})$ decreases.

Hypothesis testing

In this scenario we want to create a test where is checked the lack of linear influence of the explanatory variables on the explained variable. The main goal here is to check whether any of the explanatory variables has a linear impact on the dependent variable- if the linear regression model makes sense.

We want to compare nested models. It means, that the more complex model M_1 contains all and at least one more explanatory variable that are in the simple model M_0.

- Model M_0 contains only intercept, so:

$$\mathbb{E}(y_i) = \beta_0,\ SEE_0 = SST$$

- Model M_1 contains intercept and $p-1$ independent variables, so:

$$\begin{gathered}\mathbb{E}(y_i) = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1}, \\ SSE_1 = SSE\end{gathered}$$

We take into consideration two hypothesis: *null* (H_0) and *alternative* (H_1).

$$\begin{gathered}H_0 : \beta_1 = \cdots = \beta_{p-1} = 0 \\ H_1 : \beta_1 \neq 0 \lor \ldots \beta_{p-1} \neq 0\end{gathered}$$

Significance level (α) is a measure of the strength of the evidence that must be present in your sample before you will reject the null hypothesis and conclude that the effect is statistically significant. It is the probability of rejecting the null hypothesis when it is true. Lower significance levels indicate that you require stronger evidence before you will reject the null hypothesis [6].

In that thesis we are using $alpha = 0.05$

Definition 3.4. If $N_1, \ldots, N_k$ are independent, standard normal random variables with variance equals 1, then the sum of their squares has **chi-square distribution**. That means:

$$\begin{gathered}Q = \sum_{i=1}^{k} N_i^2, \\ Q \sim\ \chi_k^2\end{gathered}$$

, where positive integer k that specifies the number of degrees of freedom.

Definition 3.5. If X and Y are IID where $X \sim \chi^2_{d_1}$, $Y \sim \chi^2_{d_2}$, then:

$$\frac{X/d_1}{Y/d_2} \sim F(d_1, d_2).$$

, where F is a random variable with Snedecor distribution with d_1, d_2 degrees of freedom.

We consider normally distributed errors, so: $\varepsilon \overset{D}{=} N(0, \sigma^2 I_N) \implies Y \overset{D}{=} N(X\beta, \sigma^2 I_N)$
The matrix $(X^T X)$ is invertible $\implies \hat{\beta} \overset{D}{=} N(\beta, \sigma^2 (X^T X)^{\text{-}1})$

If hypothesis H_0 is true using Cochran theory[1] following facts can be in use:

- $\dfrac{SSE_0 - SSE_1}{\sigma^2} = \dfrac{SST - SSE}{\sigma^2} = \dfrac{SSR}{\sigma^2} \overset{D}{=} \chi^2_{(n-1)-(n-p)} = \chi^2_{p-1}$
  The final number of degrees of freedom: We subtract the number of variables independent of the number of observations in each model.
- $\dfrac{SSE_1}{\sigma^2} = \dfrac{SSE}{\sigma^2} \overset{D}{=} \chi^2_{n-p}$
- $SSE_0 - SEE_1$ and SEE_1 are independent

To get the chi-square distribution it is needed to have sum of squares normal distribution with variance equals 1, that is why in the formulas we divide by σ^2

Value of F statistic:

$$F = \frac{(SSE_0 - SSE_1)/\sigma^2(p-1)}{SSE_1/\sigma^2(n-p)} = \frac{SSR/(p-1)}{SSE/(n-p)}$$

In other hand the observed value of testing statistic is defined:

$F_{obs} \sim F_{p-1,n-p}$ which represents Snedecor distribution with $p-1, n-p$ degrees of freedom. During the test we read the value from statistic table.

Interpretation of inference using the F test statistic:

- $SSE_0 - SSE_1 = SST - SSE = SSR$ show the influence of adding the independent variables to the model of hypothesis H_0. The higher the value of F, the more confidently we can reject the null hypothesis.
- The **p-value** tells how often we would expect to see a test statistic as extreme or more extreme than the one calculated by the statistical test if the null hypothesis of that test was true. If the p-value is less than the significance level, we can reject the null hypothesis and conclude that the effect is statistically significant.

$$\text{p-value}=Pr_{H_0}(F \geq F_{obs}))$$

[1] In statistics, Cochran's theorem, devised by William G. Cochran, is a theorem used to justify results relating to the probability distributions of statistics that are used in the analysis of variance.

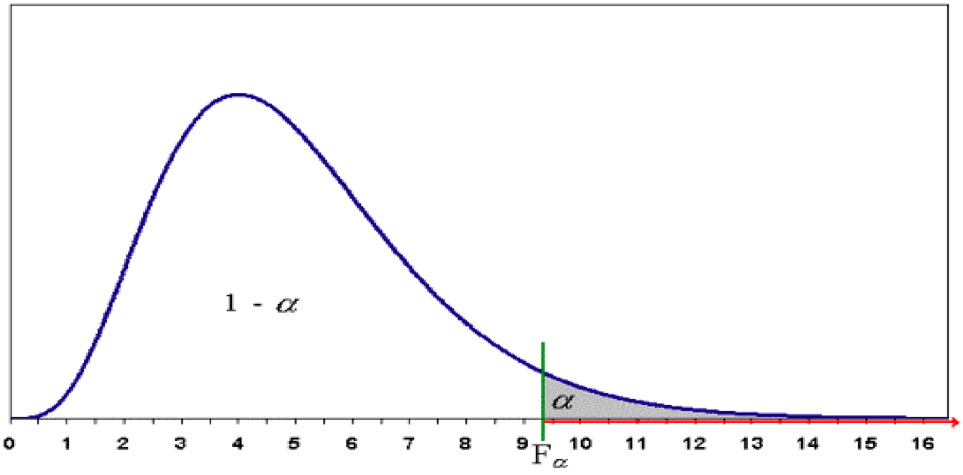


Figure 3.2: Chart [9] of Snedecor distribution with significance level.

At the significance level α, we reject the hypothesis H_0 when $F \geq F_{(p-1,n-p)}(\alpha)$. So if the computed statistic F value is greater or equals to the critic value $F(\alpha)$ then the hypothesis H_0 is rejected. It is important to remember, that it depends on the previously adopted α significance level .
If the $F(\alpha)$ is lower than the critic value $F(\alpha)$, it falls within the confidence interval at the $1 - \alpha$ level of the given distribution. This means that it is in the main mass of the distribution and then we have no grounds to reject the null hypothesis.

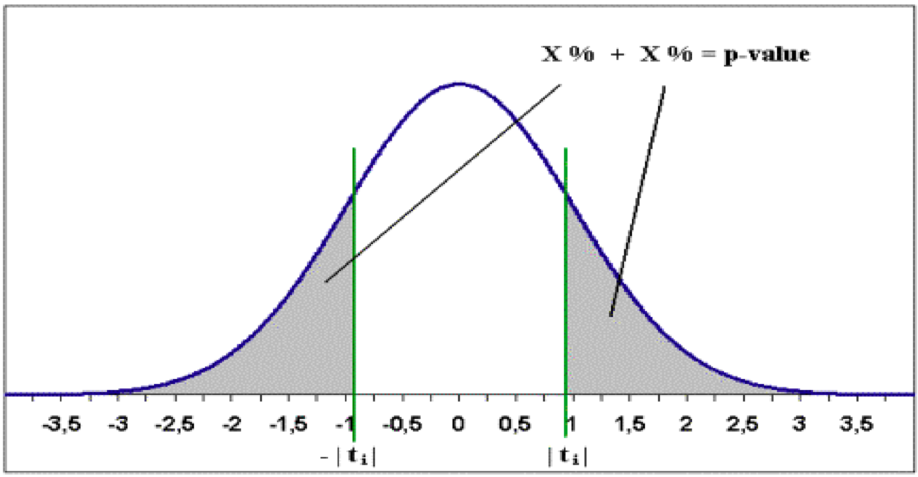


Figure 3.3: Interpretation of p-value in the chart

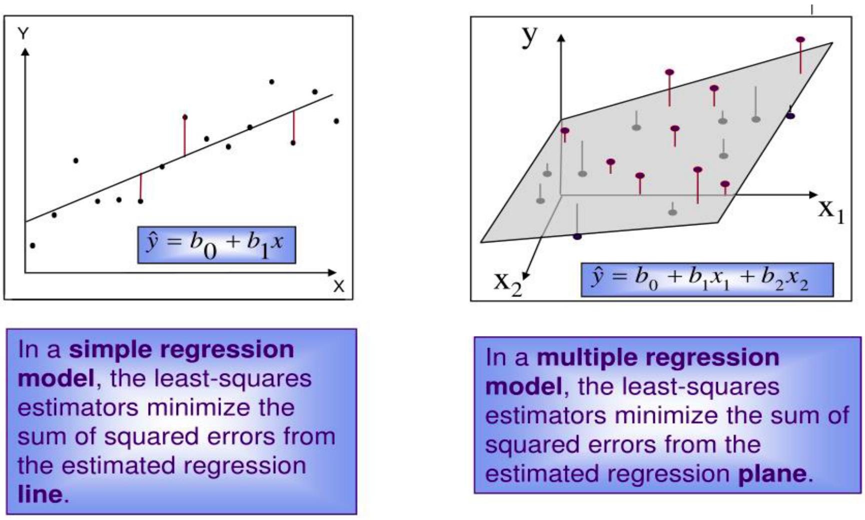
The chart above shows that the *p-value* corespondents to the total critical surface.

Creating appropriate coefficients - Least Square Method

The matrix form of a regression model allows us to analyze and present many properties of the regression model more conveniently and efficiently.
The main problem in regression models is to obtain the appropriate coefficients that will properly affect individual independent variables and describe the dependent variable with the greatest possible accuracy.

The idea of the least squares estimator consists in choosing $(\hat{\beta}_1, \ldots, \hat{\beta_N})$ in such a way that, the sum of squared residual $\sum\limits_{i=1}^{N} e_i$ in the sample is as small as possible. Mathematically this means that in order to estimate the β we have to minimize the whole sum of residuals which in matrix notation is nothing else than:

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\ [(y - X\beta)^T(y - X\beta)]$$



In a **simple regression model**, the least-squares estimators minimize the sum of squared errors from the estimated regression **line**.

In a **multiple regression model**, the least-squares estimators minimize the sum of squared errors from the estimated regression **plane**.

Figure 3.4: Comparison of the least-squares estimators in simple and multiple regression models

Theorem 3.6. *The coefficients of the multivariate linear regression function $y = \beta X + \varepsilon$ are computed using the least squares method as follows (assuming, that the columns of the X matrix are linearly independent vectors):*

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Proof

We begin of the optimisation problem from the beginning of the chapter. We assume that X has full rank. In order to find the least squares estimators, we need to minimize the sum of squared residuals, which is equivalent to solving the following equation:

$$e^T e = \begin{bmatrix} e_1 & e_2 & \ldots & e_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \sum_{i=1} e_i^2$$

$$\min_{\beta}(e, e^T)[(y - X\hat{\beta})^T(y - X\hat{\beta})] = \min_{\beta}(e, e^T)[(y^T - \hat{\beta}^T X^T)(y - X\hat{\beta})] =$$

$$\min_{\beta}(e, e^T)[y^T y - \hat{\beta}^T X^T y - y^T X\hat{\beta} + \hat{\beta}^T X^T X\hat{\beta}] = \min_{\beta}(e, e^T)[y^T y - 2\hat{\beta}^T X^T y + \hat{\beta}^T X^T X\hat{\beta}]$$

Terms $\hat{\beta}^T X^T y$ and $y^T X\hat{\beta}$ have the same dimension: 1 × 1, so they are scalars. That is why $y^T X\hat{\beta} = (y^T X\hat{\beta})^T = \hat{\beta}^T X^T y$

To minimize (e, e^T) it is needed to differentiate the expression above with respect to $\hat{\beta}$. Following mathematical statements are needed to be used.

$$\frac{\partial}{\partial\hat{\beta}}[\hat{\beta}^T X^T y] = X^T y,$$
$$\frac{\partial}{\partial\hat{\beta}}[\hat{\beta}^T X^T X\hat{\beta}] = 2X^T X\hat{\beta}$$

So minimizing (e, e^T) gives us:

$$\frac{\partial}{\partial\hat{\beta}}(e, e^T)[y^T y - 2\hat{\beta}^T X^T y + \hat{\beta}^T X^T X\hat{\beta}] = 0$$
$$-2X^T y + 2X^T X\hat{\beta} = 0$$
$$X^T X\hat{\beta} = X^T y$$

Finally we are getting the $\hat{\beta}$ least squares estimator:

$$\hat{\beta} = (X^T X)^{\text{-1}} X^T y$$

Cook's distance

An **influencing observation** is an observation whose removal from the data set significantly changes the vector of regression coefficients. The regression model constructed without this observation is significantly different from the regression model constructed with this observation.

An **outlier** is an observation of x or y that has an atypical value compared to the other n - 1 observations.

Lemma 3.7. *Let's define two concepts:*

- *Hat matrix of dimension $n \times n$:*

$$H := X(X^T X)^{\text{-1}} X^T$$

this matrix satisfies the conditions:

$$H = H^T, H^2 = H$$

- *The Unbiased estimator of variance:*

$$\hat{\sigma} = MSE = \frac{||Y - X\hat{\beta}||}{n-p} \text{ ,where X is a full-order matrix. If not } p = rank(x)$$

Leverage of i observation is the h_{ii} element of matrix H:

$$h_{ii} = x_i^T (X^T X)^{\text{-1}} x_i$$

h_{ii} measures how much y_i influences the prognosis of $\hat{y}_i$. If h_{ii} is high it means that the observation x_i is an outlier.

If $p = tr(H) = h_{11} + h_{22} + \cdots + h_{nn}$ then then the following property holds.

The thumb rule:

$$h_{ii} > \frac{3p}{n}$$

If the above inequality holds then i observation is influential due to x_i

If $\varepsilon \sim N(0, \sigma^2 I_n)$ then random variable'r $r_i = \dfrac{y_i - \hat{y}_i}{\sigma\sqrt{1-h_{ii}}}$ variance equals 1, because $e \sim N(0, \sigma^2(I-H))$. σ is not known, that's why it is needed to be estimated.
Studentized residual:

$$r_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

Studentized residual have a distribution similar to $N(0,1)$, but are not independent.

Cook's distance [1] is a measure of the impact of an observation, using information contained in studentized residuals and in the leverages. It allows detecting outliers due to $x = (x_1, x_2, \ldots, x_{p1})$ or due to y.

Let $\hat{\beta}_i$ denote the least squares estimator of the vector β, constructed on the basis of the sample $(y_1, x_1), \ldots, (y_{i-1}, x_{i-1}), (y_{i+1}, x_{i+1}), \ldots, (y_n, x_n)$. Very important is the fact, that it does not contain the observation (y_i, x_i). Cook's distance measures like this the reduction of the data set affects the estimator.

Cooks distance for i observation will be defined:

$$D_i = \frac{(\hat{\beta}_i - \hat{\beta})^T (\hat{Var}(\hat{\beta}))^{\text{-}1} (\hat{\beta}_i - \hat{\beta})}{p} = \frac{(\hat{\beta}_i - \hat{\beta})^T (X^T X)^{\text{-}1} (\hat{\beta}_i - \hat{\beta})}{p\hat{\sigma}^2}$$

$$\iff$$

$$D_i := {r_i}^2 \left(\frac{h_{ii}}{1-h_{ii}}\right)\frac{1}{p} = \frac{(y_i - \hat{y}_i)^2 h_{ii}}{p\hat{\sigma}^2(1-h_{ii})^2}$$

- The factor r_i^2 is high when the i observation is unusual due to y.
- The factor $(\frac{h_{ii}}{1-h_{ii}})$ is high when the i observation is unusual due to x.
- **Rule of thumb**: The observation can be influential when: $D_i \geqslant \dfrac{4}{n-p}$

3.3 Classification methods

After demonstrating linear regression approach to try predict the values which interest us we can use classification methods. In our case we want to forecast rainfalls, so in this conception at first is needed to build specific rain classes. This section presents techniques for classifying rainfall into different classes based on relevant variables.

3.3.1 Decision Trees

Decision trees are the most intuitive and clear classifier. It is very often used as a graphic method of supporting the decision-making process used in decision theory.

Structure of the decision tree and the main idea

The considerations are based on the **Classification and Regression Trees (CART) [3]**

Definition 3.8. Decision tree Decision Tree is a acyclic[1], directed[2] and consistent[3] graph.

- Each classification tree consists of **nodes**[4]. The initial node is called the **root** of the tree.
- The edges, that connect nodes are called **branches**.
- There is only one way from the root leads to each **leaf** (the nodes at the end).

In this thesis we will be limited to binary trees, i.e. those for which two edges come out of each node, except the leaves.

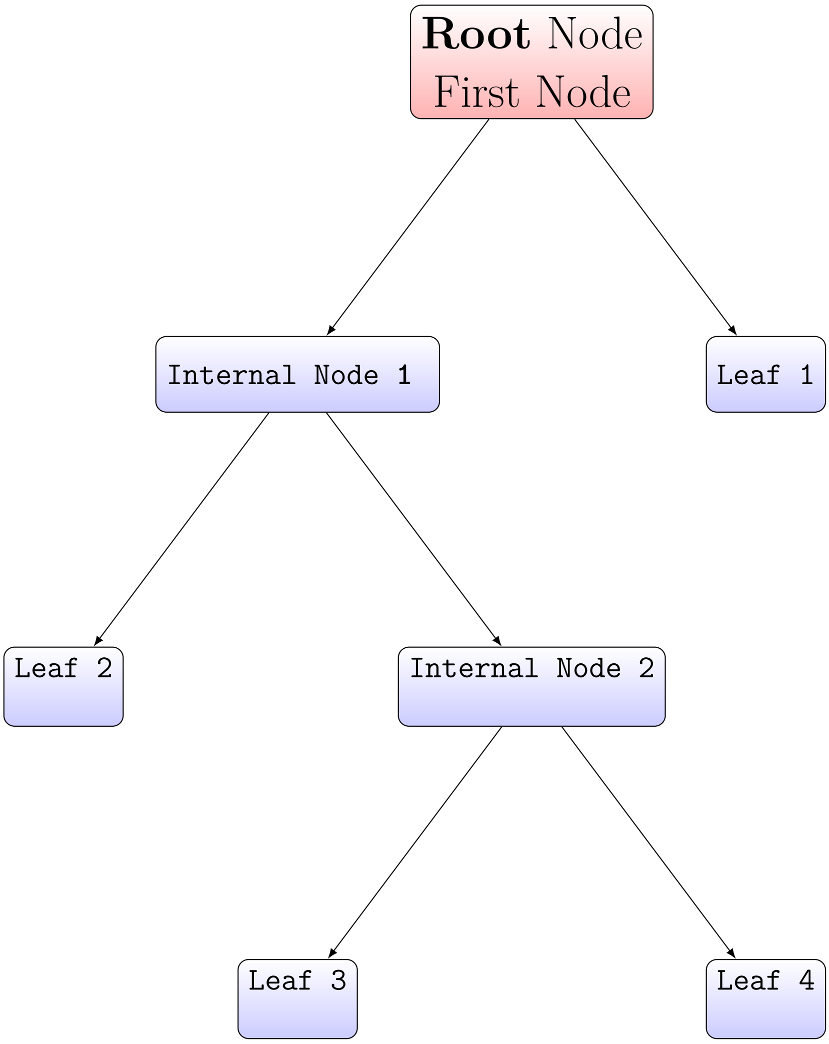


Figure 3.5: The structure of the Classification Tree

The classification tree is built on the basis of a learning sample, which is concentrated in the root. Subsequent elements of the test are moving along the branch (from top to bottom): **if the condition in the node is met we go left, if not – to the right.**

[1] there is no edge string connecting the vertex to itself

[2] the edges between the vertices have a designated direction

[3] each pair of vertices is connected by an edge string

[4] these are the vertices in CRAT, where decision tests are carried out

The elements in the leaves are labeled by the class from which comes the highest amount of learning samples elements, those which have reached the proper leaf.
The aim is to obtain as homogeneous leaves as possible (i.e. containing most observations of the same class)

Definition 3.9 (Classification rule)**.** We define that:

- We are using training sample $\Big\{(x_i, y_i)\Big\}$, where x_i means proper observation, and y_i is its right class.
- **C** represents the number of classes. **c** in this case means the proper class in node.
- n_m is the number of x_i in m node. $n_{m,c}$ means number of x_i in m node with c class.
- M is the subset of $\Big\{(x_i, y_i)\Big\}$ in the m node.

$$\hat{p}_{m,c} = \frac{1}{n_m} \sum_{x_i \in M} \mathbb{1}_{\{y_i = c\}} = \frac{n_{m,c}}{n_m}$$

$$\hat{k}(m) = \underset{c \in C}{\operatorname{argmax}}\ \hat{p}_{m,c}$$

The definition just means that, we are counting all classes of observations in each node and choosing that particular one which is the most popular in the proper node. We are trying to get the largest amount of the same classes in each node- impurity of a node.

Measures

One of the most used techniques of *Impurity measure* is **Gini Impurity Index** [**3**].

$$Gini(Leaf_k) = 1 - \sum_{class=1}^{C} (probability(class|Leaf_k))^2 = 1 - \sum_{c=1}^{C} (p(c|L_k))^2$$
$$Gini(Node_j) = \sum_{K} \left(\frac{count(L_k)}{count(L_1, \dots, L_k)} \right) Gini(L_k)$$

To compute the impurity of specific node it is needed to check both of the leaves. The application of the equations above will be shown in the example.

Example- how the tree really works?

For better understanding the process and how exactly the decision trees are "growing" let's present an simple data set, where decision test can take place.

The goal is to predict what class of rain it will be. In our example **A** class means a weak rain, **C** class represents a heavy rainfall.

We want to show here step by step how really looks like the process using those six observations in the data set below.

Table 3.1: Table of a meteorological data example

Temperature	Cloudiness	Days with foge	Class
23	1	5	**A**
24	3	2	**A**
20	5	4	**A**
21	7	4.5	**B**
18	5.5	5.5	**B**
19	8	10	**C**

Source: Own Study

There are several important questions regarding the order and significance of features, which are needed to be clarify. The following problems:

- Which features should be used in the classification tree?
- In what order should be build the nodes?
- What are the specific values of variables to make some thresholds for a split?

In this case the variables have discrete values, so it is not known how to split the examples. At the beginning let's sort the feature by ascending or descending. Then the technique is to iterate over each row of observation, then find the average of the value of the current row plus the value of the next row. Such a process is done for each variable.

$$Temp\ Treshehold = \Big\{18.5, 19.5, 20.5, 22, 23.5\Big\},\ Cloud\ Treshehold = \Big\{2, 4, 5.25, 6.25, 7.5\Big\},$$
$$Fog\ Treshehold = \Big\{3, 4.25, 4.75, 5.25, 7.75\Big\}$$

The next step is to choose best value from each set. Criterion on the basis of which we can determine which threshold is best is Gini Impurity Index.

$$Gini(Leaf_{2k-1}) = 1 - \sum_{class=1}^{C} (p\Big(w_i < C|Leaf_{2k-1}\Big))^2$$
$$Gini(Leaf_{2k}) = 1 - \sum_{class=1}^{C} (p\Big(w_i > C|Leaf_{2k}\Big))^2$$

, where w_i means the specific value of i observation. We are computing the probabilities in cases where the value of variable it is correspondingly smaller or larger than the threshold in the proper Leaf.

1. ***Cloudiness* Variable**
   At first we take the *Cloudiness* variable to compute the final gini impurities for each of these thresholds below. Every time it is needed to split the learning set into 2 specific sets- the Leafs.

   - Treshold=2, $Leaf1 : 2 \times A, 2 \times B, 1 \times C|Leaf2 : 1 \times A$
     $Gini(L_1) = 1 - \left(\frac{1}{1}\right)^2 = 0$
     $Gini(L_2) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{2}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = 0.64$
     $Gini(I_1) = \left(\frac{1}{6}\right) 0 + \left(\frac{5}{6}\right) 0.64 = 0.534$

- Treshold=4, $Leaf1 : 1 \times A, 2 \times B, 1 \times C | Leaf2 : 2 \times A$
  $Gini(L_1) = 1 - \left(\frac{2}{2}\right)^2 = 0$
  $Gini(L_2) = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{2}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0.375$
  $Gini(I_1) = \left(\frac{2}{6}\right) 0 + \left(\frac{4}{6}\right) 0.375 = 0.4167$
- Treshold=5.25, $Leaf1 : 2 \times B, 1 \times C | Leaf2 : 3 \times A$
  $Gini(L_1) = 1 - \left(\frac{3}{3}\right)^2 = 0$
  $Gini(L_2) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.44$
  $Gini(I_1) = \left(\frac{3}{6}\right) 0 + \left(\frac{3}{6}\right) 0.44 = 0.22$
- Treshold=6.25, $Leaf1 : 1 \times B, 1 \times C | Leaf2 : 3 \times A, 1 \times B$
  $Gini(L_1) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0.375$
  $Gini(L_2) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$
  $Gini(I_1) = \left(\frac{4}{6}\right) 0.375 + \left(\frac{2}{6}\right) 0.5 = 0.4167$
- Treshold=7.5, $Leaf1 : 1 \times C | Leaf2 : 3 \times A, 2 \times C$
  $Gini(L_1) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$
  $Gini(L_2) = 1 - \left(\frac{1}{1}\right)^2 = 0$
  $Gini(I_1) = \left(\frac{5}{6}\right) 0.48 + \left(\frac{1}{6}\right) 0 = 0.4$

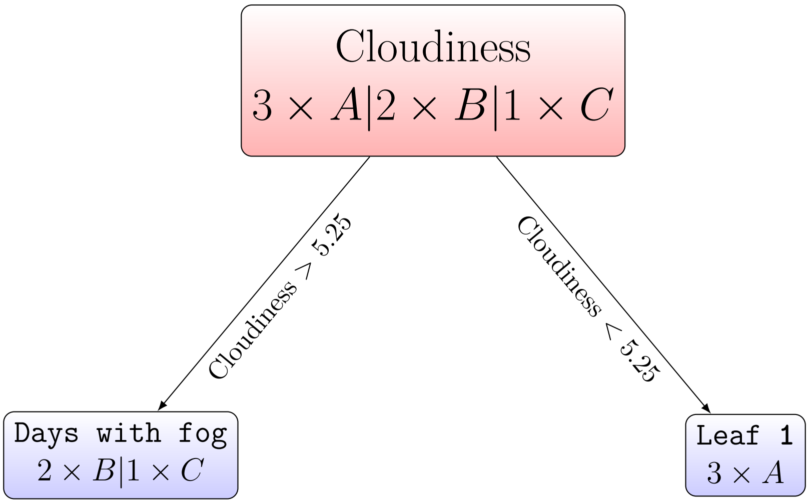


Figure 3.6: Tree with *Cloudiness* as the root

For all variables in each level of nodes the way of computing Gini Index Impurity looks always the same, the next calculations are being skipped.

2. ***Temperature* Variable**

- Threshold=18.5, $Gini(I_1) = 0.467$
- Threshold=19.5, $Gini(I_1) = 0.4167$
- Treshold=20.5, $Gini(I_1) = 0.56$

- Treshold=22, $Gini(I_1) = 0.4167$
- Treshold=23.5, $Gini(I_1) = 0.534$

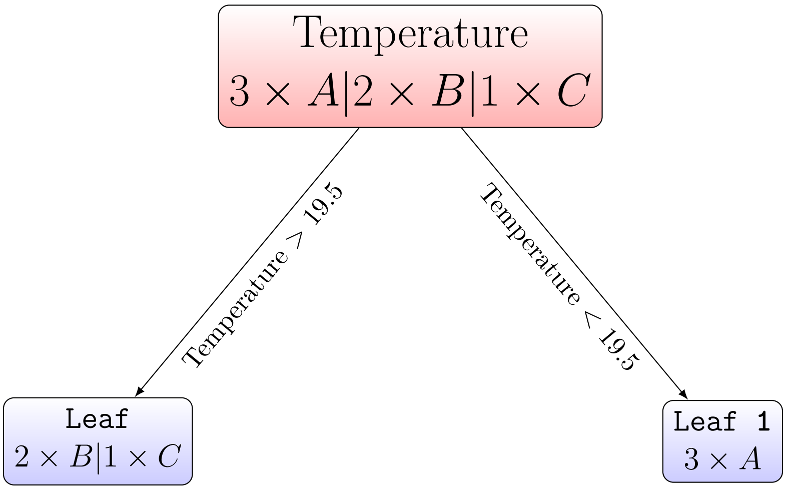


Figure 3.7: Tree with *Temperature* as the root

3. ***Days with fog*** **Variable**

- Threshold=3, $Gini(I_1) = 0.534$
- Threshold=4.25, $Gini(I_1) = 0.467$
- Threshold=4.75, $Gini(I_1) = 0.44$
- Threshold=5.25, $Gini(I_1) = 0.467$
- Threshold=7.75, $Gini(I_1) = 0.4$

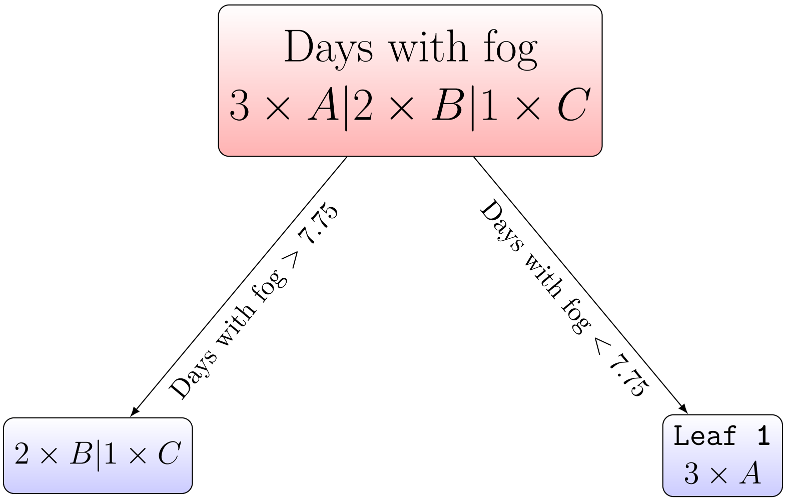


Figure 3.8: Tree with *Days with fog* as the root

After checking all Thresholds of each variable it is needed to check the Gini Impurity.

Table 3.2: Table of a Gini Impurity Index

Temperature	Cloudcover	Days with glaze
0.467	0.534	0.534
0.4167	0.467	0.467
0.56	**0.22**	0.44
0.4167	0.467	0.467
0.534	0.4	**0.4**

Source: Own Study

The lowest Gini Impurity represent the best threshold for each variable. Also it means what is the most important feature for algorithm in that specific moment of splitting data. After the comparison the conclusion is that the best properties for value separation has *Cloudiness* with the **threshold = 5.25**. It means *Cloudiness* is the root of the tree.

On the right side we have separated perfect the *A class*- Leaf 1. On the left node, we have now a new data set. It is needed to decide which variable should be chosen in the next internal node. Necessary is to try to split the new data set and compute the Gini Index again. So the same process as previously takes place.

In that case, we already know what Thresholds are best for each split.

First we will try to use variable *Days with fog* as the second Node.
Our new set: $\Big\{B, B, C\Big\}$

1. ***Days with fog*** **Variable for Node 2**

   - Treshold=7.75, $Leaf2.1 : 1 \times C | Leaf2.2 : 2 \times B$
     $Gini(L_{2.1}) = 1 - \left(\frac{1}{1}\right)^2 = 0$
     $Gini(L_{2.2}) = 1 - \left(\frac{2}{2}\right)^2 = 0$
     $Gini(I_2) = \left(\frac{1}{3}\right) 0 + \left(\frac{2}{3}\right) 0 = 0$

   The Gini Index is zero, so it means that the classes are perfectly separated, which is shown in the tree below.

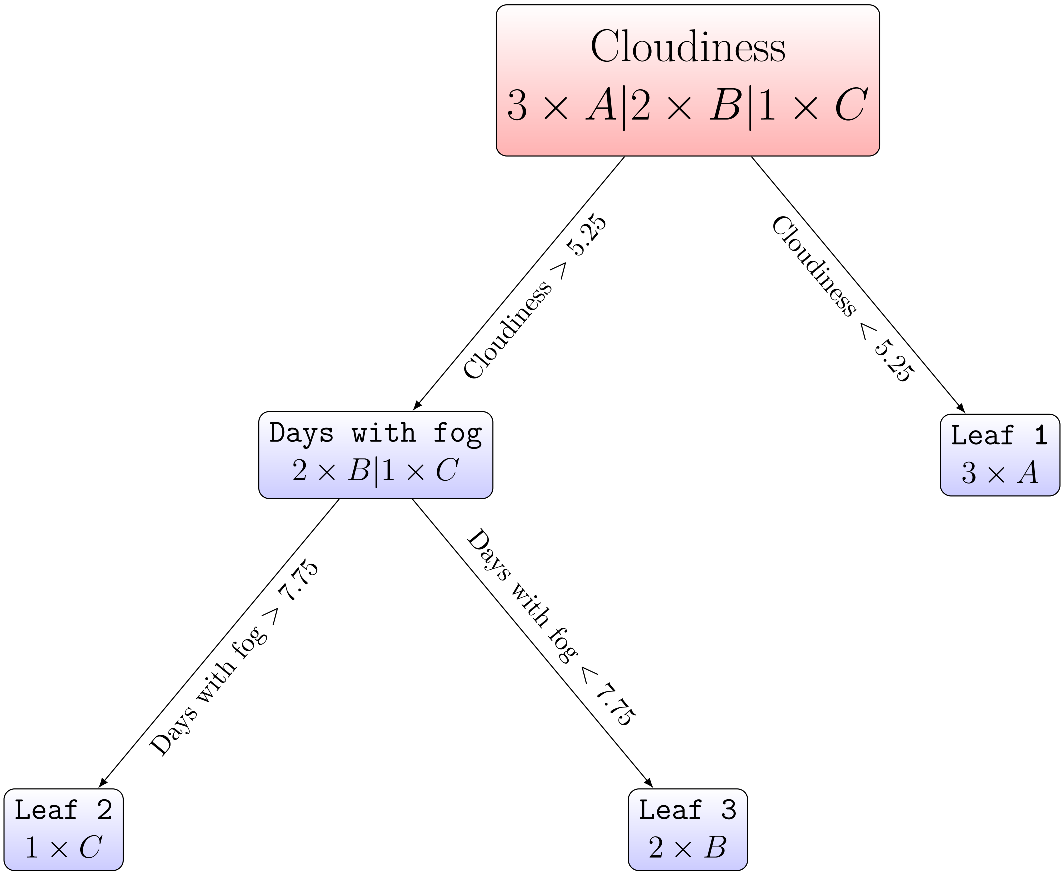


Figure 3.9: The right final tree with *Days with fog* as a internal node

2. ***Temperature* Variable for Node 2**
   The "growing" process is already done. If the Gini index would be not used, the tree structure could look like this:

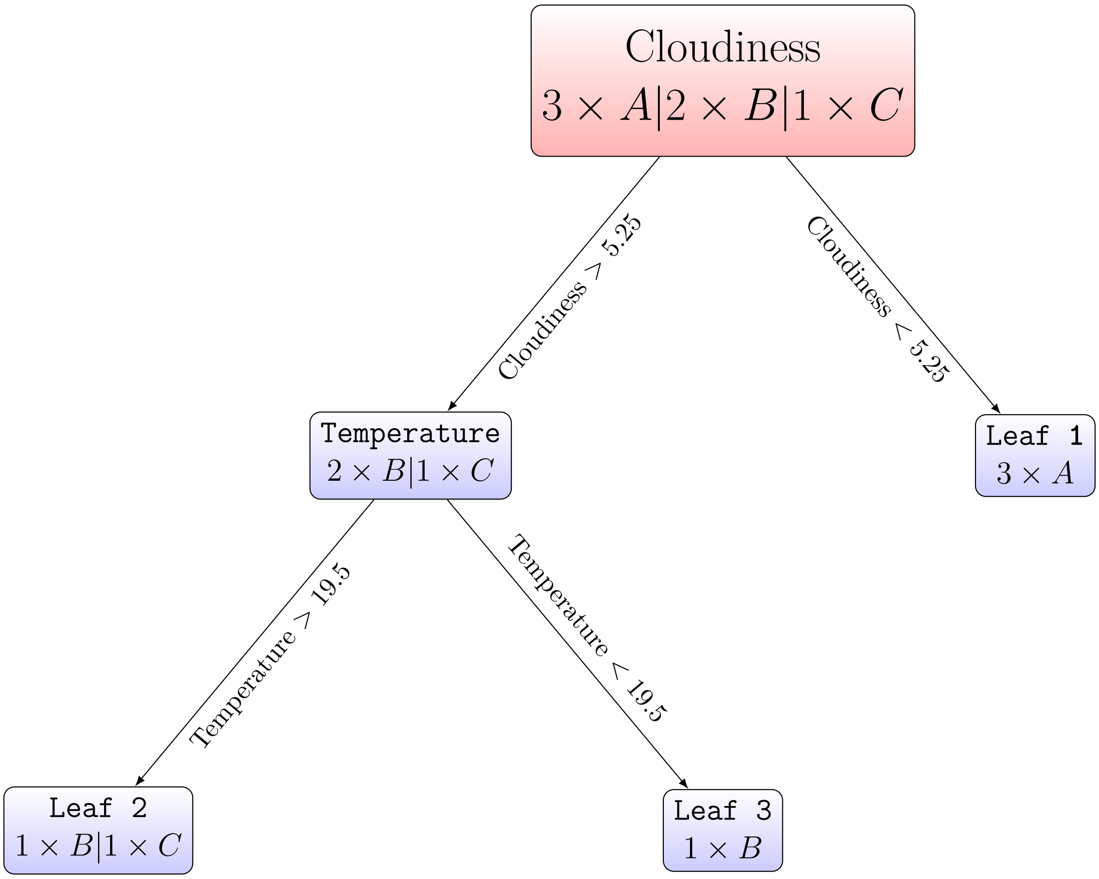


Figure 3.10: The wrong tree with *Temperature* as a internal node

In that example only two from three variables were used. The Impurity of class is perfect- all classes have been correctly separated- in practice such a situation in very rarely.

Complexity

Let's formalise and define the measure of tree's complexity.

An assumption takes place, that all the data comes from the same source, it means that every x example is generated by the same distrubtion D, which is unknown. The X family is IID since every example x originate from the same distribution D. $c(x)$ function is defined, which generates the correct class labels of $x_1, \ldots, x_n$ examples. $h(x)$ function represents the prediction's output of the algorithm classification. The prediction rule is called *hypothesis*.

Now it is easy to present the measure of goodness of the classifier by determining the generalization error. It is defined by the propability that the random x_i example will be misclassified-

$$\text{generalization error} = \underset{x \in D}{P}(c(x) \neq h(x)) = error(h)$$

H describes the height of a tree. H_n is set of all trees that contains n nodes. H_i is bigger (more complex class), than H_{i-1}, because it contains all trees from that class. There is a dependency: $|H_i| > |H_{i-1}|$

Definition 3.10. Let's be $ln|H|$ be a measure of the complexity which is proportional to the size of the tree.

$$complexity = ln|H| = O(n)$$

, where n represents number of nodes in the tree.

Theorem 3.11. *The hypothesis is consistent* $\iff$ *there are no mistakes on the* m *example training set. In that case error on the training set equals to zero.*
When the algorithm A *finds a hypothesis* $h_A \in H$ *then with probability greater than* $1 - \delta$ *occurs:*

$$error(h_A) \leq \frac{ln|H| + ln\frac{1}{\delta}}{m} \tag{3.1}$$

There is always a small chance that the alghorytm will perform not as good as expected from him. That is why δ *as a constant is included, the algorytm cannot be perfect.*

The inequality above says two important things. When the amount of m examples in training set increases the error decreases. Second observation is that if the complexity increases, the error increases as well.

Proof

Let $h_A \in H$ and $\varepsilon = \frac{ln|H| + ln\frac{1}{\delta}}{m}$. Is it known also that hypothesis h_A is consitent with all m training examples.

The goal of the proof is to show that:

$$error(h_A) \leq \frac{ln|H| + ln\frac{1}{\delta}}{m}$$
$$\text{h is } \varepsilon - bad \iff error(h_A) > \varepsilon$$
$$P(h_A \text{ is not } \varepsilon - bad) \geqslant 1 - \delta \iff P(h_A \text{ is } \varepsilon - bad) \leqslant \delta$$

$P(h_A \text{ is } \varepsilon - bad) = P(h_A \text{ is } \varepsilon - bad \text{ and consistent})$, because this is due to the assumption, that every hypothesis, in this case h_A is consitent.

If the condition of this probability holds, then there is some hypothesis that is consistent and ε-bad. Thus,

P(h_A is consistent and ε-bad) $\leqslant$ P($\exists h \in H$: h_A is consistent and ε-bad)

It is known that, that probability of an existing, random hypothesis is higher than a specific one like here h_A

Let set F include all hypothesis from H space, which are ε-bad:

$$\text{F} = \Big\{\ h \in H : h \text{ is } \varepsilon - bad\Big\} = \Big\{h_1, \ldots, h_k\Big\}$$

It follows, that

P($\exists h \in F$: h is consistent)= P(h_1 is consistent $\vee$ h$_2$ is consistent $\vee \cdots \vee$ h$_n$ is consistent)

From the propability's theory is also known the fact below:

$$P(A \vee B) \leqslant P(A) + P(B)$$

Based on the inequalities above, it can be concluded that

$$\text{P}(h_1 \text{ is consistent } \vee h_2 \text{ is consistent } \vee \dots \vee h_n \text{ is consistent}) \leqslant$$
$$\text{P}(h_1 \text{ is consistent}) + \text{P}(h_2 \text{ is consistent}) + \dots + \text{P}(h_n \text{ is consistent})$$

In the same time let's compute the probability that any $h \in F$ is consistent in the training set.

$$\text{P}(h \text{ is consistent}) = P\Big((\text{h}(\text{x}_1) = c(x_1) \wedge h(x_2) = c(x_2) \wedge \dots \wedge h(x_n) = c(x_n)\Big)$$

As it was said at the beginning of the chapter, the x examples are *IID*, so we can transform the equation as a product.

$$= P\Big(\text{h}(\text{x}_1) = c(x_1)\Big) \times P\Big(h(x_2) = c(x_2))\Big) \times \dots \times P\Big(h(x_n) = c(x_n)\Big)$$

For each x:

$$P\Big(h(x) = c(x)\Big) = 1 - \underset{x \in D}{P}(c(x) \neq h(x)) = 1 - error(h) \leqslant 1 - \varepsilon$$

For the $x_1, x_2, \dots, x_n$ family we have:

$$P(h \text{ is consistent}) \leqslant (1-\varepsilon)^n$$

Additionally it is needed to remember, that:

$$\forall x : 1 + x \leqslant e^x \implies 1 - x \leqslant e^{-x},$$

$$F \subseteq H \implies |F| \leqslant |H|$$

Using the above reasoning it is possible to move to last stage.

$$\text{P}(h_A \text{ is consistent and } \varepsilon\text{-bad}) \leqslant \text{P}(h_1 \text{ is consistent}) + \dots + \text{P}(h_m \text{ is consistent})$$
$$\leqslant |F|(1-\varepsilon)^m \leqslant |H|(1-\varepsilon)^m \leqslant |H|e^{-\varepsilon m} = |H|e^{-(ln|H| + ln\frac{1}{\delta})} = \delta$$

That is why, that at the beginning an assumption has been made:

$$\varepsilon = \frac{ln|H| + ln\frac{1}{\delta}}{m}$$

We have shown that $P(h_A$ **is** ε**-bad**$) \leqslant \delta$

Pruning and Overfitting

Classification trees, due to their high flexibility in modeling decision limits, are susceptible to overfitting.

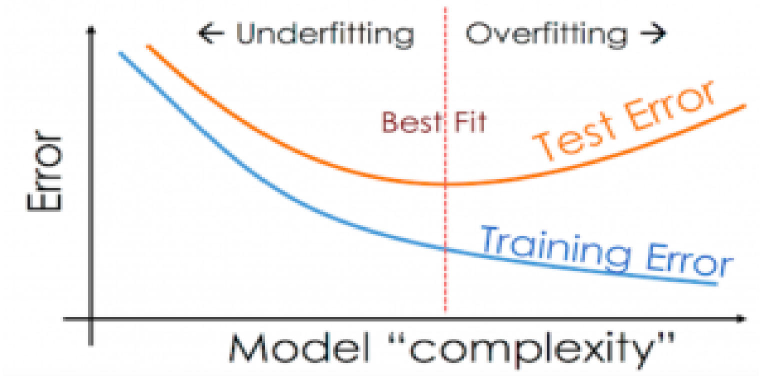


Figure 3.11: Chart presenting classification errors during the process of tree "growing"

As the tree size increases, training error decreases. However, as the tree size increases, testing error decreases at first since we expect the test data to be similar to the training data, but at a certain point, the training algorithm starts training to the noise in the data, becoming less accurate on the testing data. At this point we are no longer fitting the data and instead fitting the noise in the data. Therefore, the shape of the testing error curve will start to increase at a certain point at which the tree is too big and too complex to perform well on testing data.

There is a tradeoff between training error and tree size [2].

One of the Rules for pruning trees is an algorithm based on the Cost-Complexity Pruning. In practise the **One Standard Error** rule is often used. As the optimal tree, we choose the tree with the smallest number of divisions (the smallest size) for which the fraction of misclassifications is distant by no more than one standard deviation from the minimum fraction of misclassifications.

3.3.2 Naive Bayes Classifier

Naïve Bayes is a classification technique being used in Machine Learning algorithms based on the Bayes Theorem. The classifiers are based on the assumption that the predictors are independent of each other. It means that the outcome of a model depends on a set of independent variables. The main goal is to calculate the probability of each class and then pick the one with the highest probability. Usually in data the predictors can be correlated or dependent by each other. The classifier ignores the assumption. Since Naive Bayes considers each predictor variable to be independent of any other variable in the model, and in practice cases it works well that is why it is called 'Naive'.

Theorem 3.12. *The a priori probability is the probability computed before the random experiment is carried out, that is, the classical probability, as opposed to the a posterior probability, computed on the basis of the experimental results, that is frequency.*

Theorem 3.13. *Let $X_i = (x_{i,1}, x_{i,2}, \dots, x_{i,n})$ be a vector of n explanatory variables in i case and y_k be the k class of explained variable.*

$$P(y_k \mid X_i) = \frac{P(X_i \mid y_k)P(y_k)}{P(X_i)}$$

where from the conditional probability results the following equality $P(y \mid X) = \dfrac{P(y \cap X)}{P(X)}$

The values of the features $X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n})$ are given and the denominator $P(y \mid X)$ does not depend on y, so it is effectively constant.

At the beginning we use the vector of independent variables $X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n})$ representing n parameters to build a conditional probability model for classifying to which of y_k classes the X_i vector should belong. Using conditional probability we have:

$$\begin{gathered}
P(y_k|X_i) = P(X_i|y_k)P(y_k) = P((x_{i,1}, x_{i,2}, \ldots, x_{i,n})|y_k)P(y_k) = \\
\mathrm{P}(\mathrm{x}_{i,1}, x_{i,2}, \ldots, x_{i,n}, y_k) = P(x_{i,1}|x_{i,2}, \ldots, x_{i,n}, y_k)P(x_{i,2}, \ldots, x_{i,n}, y_k) = \\
\mathrm{P}(\mathrm{x}_{i,1}|x_{i,2}, \ldots, x_{i,n}, y_k)P(x_{i,2}|x_{i,3}, \ldots, x_{i,n}, y_k)P(x_{i,3}, \ldots, x_{i,n}, y_k) = \ldots = \\
P(x_{i,1}|x_{i,2}, \ldots, x_{i,n}, y_k)P(x_{i,2}|x_{i,3}, \ldots, x_{i,n}, y_k)(x_{i,n-1}|x_{i,n}, y_k)P(x_{i,n}|y_k)P(y_k)
\end{gathered}$$

In the above reasoning we use the chain rule for repeated applications of the definition of conditional probability: $P(X_1, X_2, \ldots, X_n) = P(X_1|X_2, \ldots, X_n)P(X_2, \ldots, X_n)$ where X_i is a random variable.

The next step is to use the naive conditional independence. The model assume that all features in X_i are independent, so the "common part" of variables $x_i, x_{i+1}, \ldots, x_n, y_k$ is y_k. It means that:

$$P(x_i|y_k, x_{i+1}, x_n) = P(X_i|y_k)$$

Summing up we get following equation:

$$P(y_k|X_i) = \frac{P(y_k, x_1, \ldots, x_n)}{P(x_1, \ldots, x_n)} = \frac{1}{C}P(y_k)P(x_1|y_k)P(x_2|y_k)\cdots P(x_n|y_k) = \frac{1}{C}P(y_k)\prod_{i=1}^{n} p(x_i|y_k)$$

,where $C = P(X_i) = \sum_k p(y_k)\ P(X_i|y_k)$

C is a scaling factor that depends only on $x_1, \ldots, x_n$

In the above reasoning the derivation of the independent feature model is discussed, which is a naive probabilistic Bayesian model.

Algorithm constructed on the basis of naive Bayesian classifier

The classifier combines this model with the decision rule. One general rule is to reveal the most probable hypothesis, it calls *maximum a posteriori.*

Definition 3.14 (Bayesian classifier)**.** Let's suppose the class $y_k \in 1, \ldots, K$.We define the Bayesian classifier as :

$$d_B(x) = \underset{k \in \{1, \ldots, K\}}{\operatorname{argmax}}\ P(y_k|X_i)$$

- $f_{k,i}$ denotes the density of distribution the x_i variable for k class.
- π_k means the propablity *a priori* of k classes

Equivalent form:

$$d_B(x) = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \; \pi_k f_k(x)$$

Theorem 3.15 (Bayesian classifier optimality)**.** *The Bayesian d_B classifier is optimal when $e(d_B) \leqslant e(d_t)$, where $e(d_t)$ is the actual error rate of the any other classifier d_t Additional fact: $e(d_B) = 1 - \int_2 \max_{\{k \leqslant k \leqslant K\}} (\pi_k f_k(x)) dx$*

By estimating π_k and $_k$ we can estimate the Bayesian classifier. However, direct estimation of the density f_k for a large dimension may not be easy, especially for a non-parametric approach. In the Naive Bayes approach we assume the independence of x_i variables thanks to which the problem of estimation is simplified.

For every k class we find the estimator f_k of X_i random variable density, using only those elements of the training sample that belongs to the k class.It is possible to define the multivariate density estimator for the k class and a priori propability estimator.

$$\hat{f}_k(X_i) = \hat{f}_k(x_1, \ldots, x_n) = \prod_{i=1}^{n} \hat{f}_{k,i}(x_i)$$

$$\hat{\pi}_k = \frac{n_k}{n},$$

In the equation above n means the number of all elements and n_k denotes the number of set belonging to the k class.

Definition 3.16 (Naive Bayesian classifier)**.** Using the above facts, we can determine the main classifier:

$$\hat{d}_{NB}(x) = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \; \hat{\pi}_k \hat{f}_k(x)$$

Chapter 4

Set of meteorological data of Wroclaw

4.1 Data acquisition

The data used during the analysis come from the *National Research Institute of meteorology and water management* [11]. In this part of the analysis, we use a data set with a monthly resolution. We investigate 31 quantitative variables where **rainfall is the dependent variable**. The data refer to city of Wrocław in Poland and covers the period of years 1966-2020.

The data set consists from 660 observations, it was created by merging appropriate parameters in each separate year set. Categories of features were splitted by rainfall, climate and synoptic meteorological data. In month and day interval. The data include parameters related to temperatures in Celcius degrees, snow cover, pressures, humidity, cloudiness and wind condition and many time duration of specific weather phenomena like fog, hail, kind of wind speed, snow blizzard etc. The predicted rainfalls are expressed in particular units - liters per square meter, which can be used interchangeably with millimeters of rainfall.

The whole data set was created by merging data from each year for Wrocław region. Only weather parameters have been selected.

There are no missing observations in the data set.

For some classification methods, an additional variable *Rainfall class* was created. It depends on the rain level in a given month.

Table 4.1: Table of constructed rainfall classes

Rainfall level	Rain precipitation range [mm]	Rainfall Class
Low	0 - 35.0	**1**
Moderate	35.0 - 60.0	**2**
High	60.0 - 90.0	**3**
Very high	90.0 or higher	**4**

Source: Own Study

4.2 Preview and initial analysis

At the beginning is important to check the correlations between each variable.

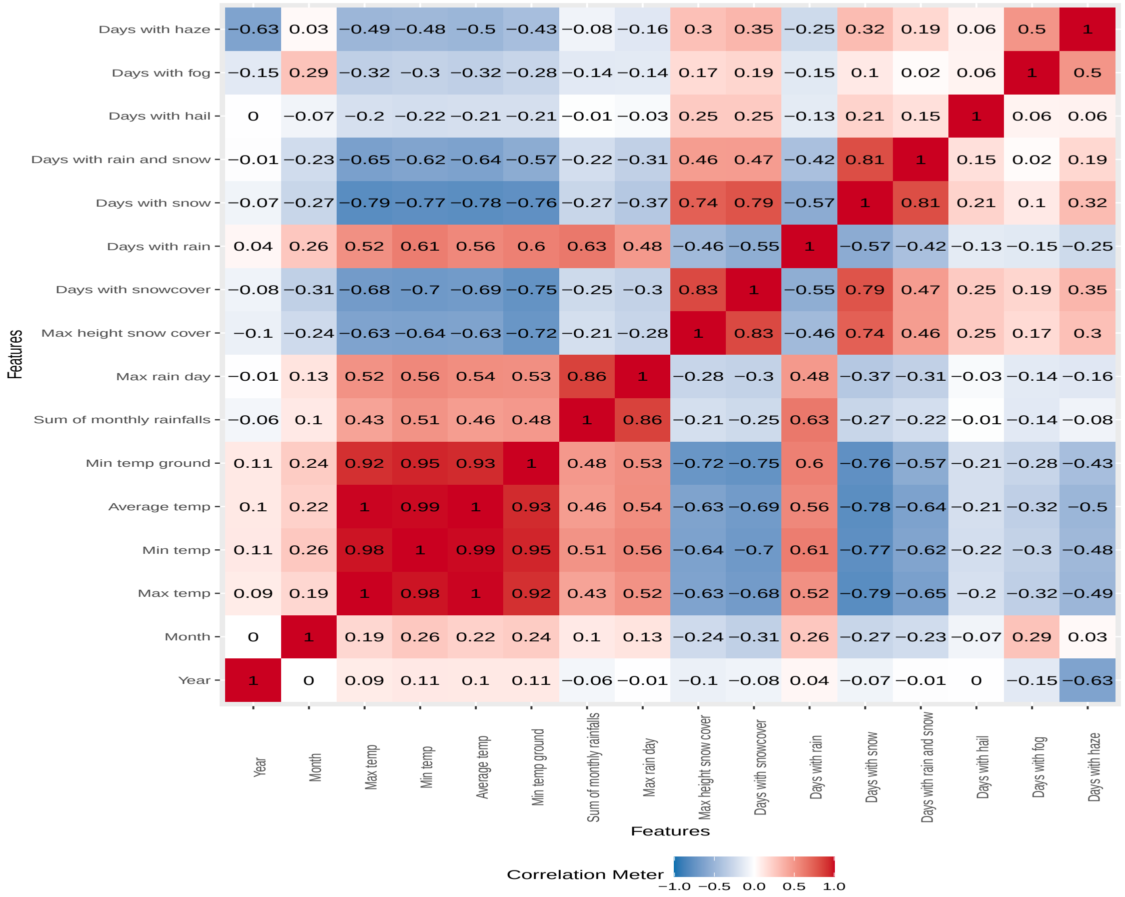


Figure 4.1: Correlation table for first part of variables. Prediction for rainfall for present time.

In the first figure we can see two groups of variables which are correlated with each other. Variables *Min Temperature*, *Max Temperature*, *Avg Temperature* and *Min Temp near ground* have correlations between 0.92-0.98. This area has dark red colour on the chart. These variables are correlated also with *Max height snow cover*, *Days with snow* etc at the level of 0.57-0.79- dark blue area.

Concentrating on the dependent variable *Sum of monthly rainfall* we can observe correlations with *Max rain day*- 0.86 and *Days with rain*- 0.63. These dependencies are quite obvious.

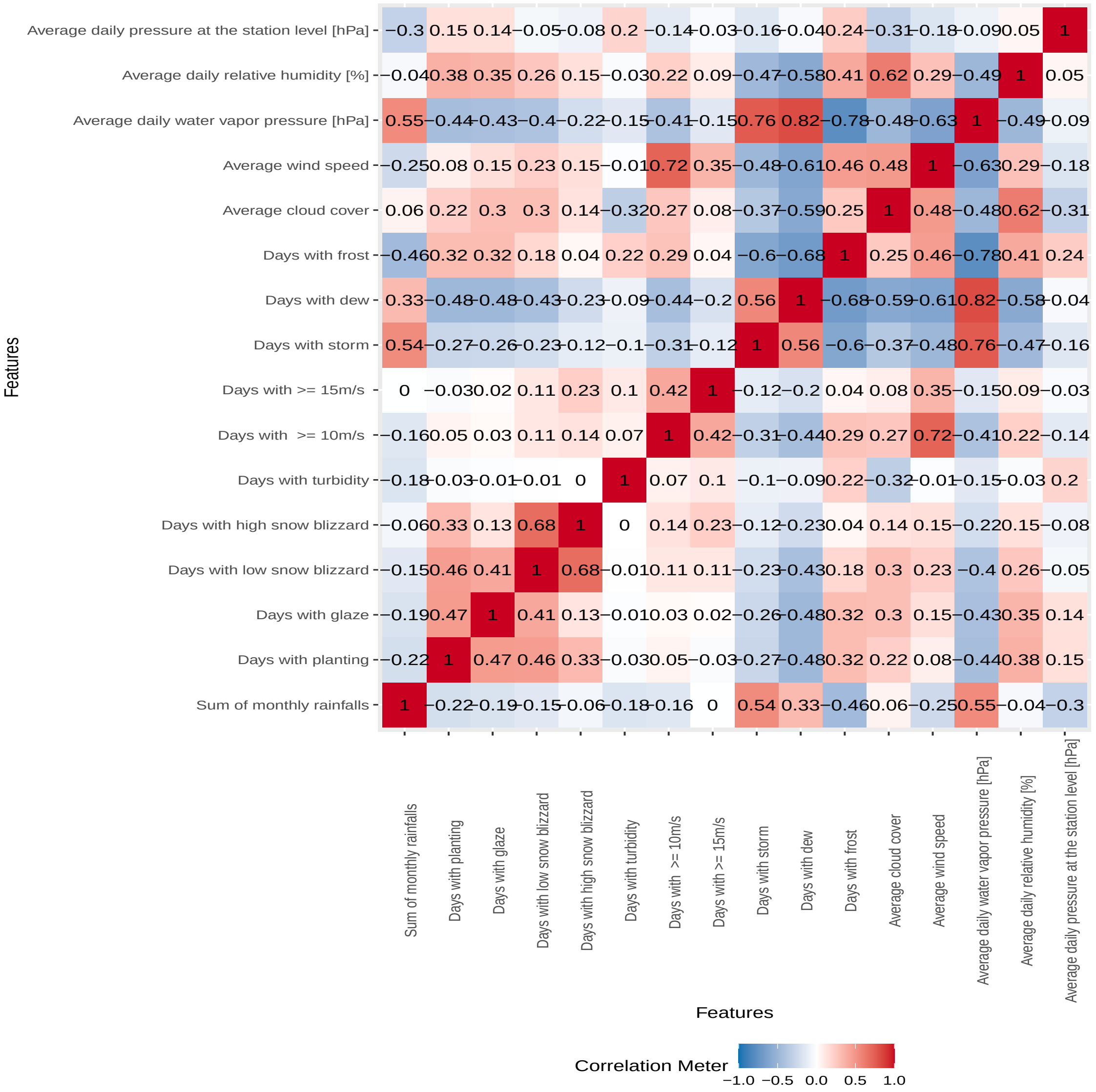


Figure 4.2: Correlation table for second part of variables. Prediction for rainfall for present time.

The highest values are between *Average daily water vapor pressure* and variables like: *Days with dew*, *Days with storm* and *Days with frost*. Here *Sum of monthly rainfall* has only weak correlations (0.54-0.55) with *Days with storm* and *Average daily water vapor pressure*.

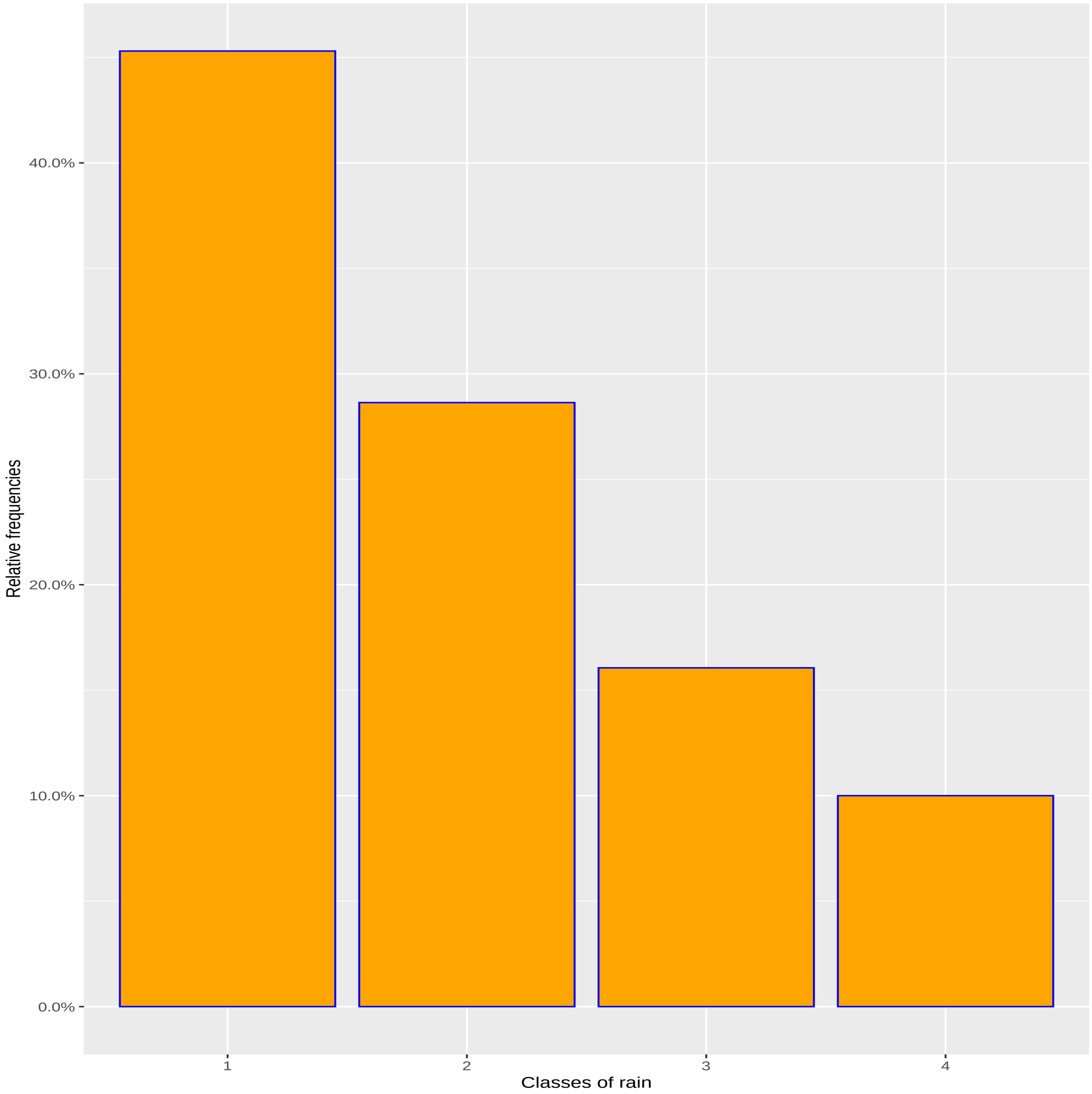


Figure 4.3: Rain classes

The chart demonstrates four classes of rain. Over 45% of 660 observations are in the most common class **1**. Rains at higher levels are less likely to occur.

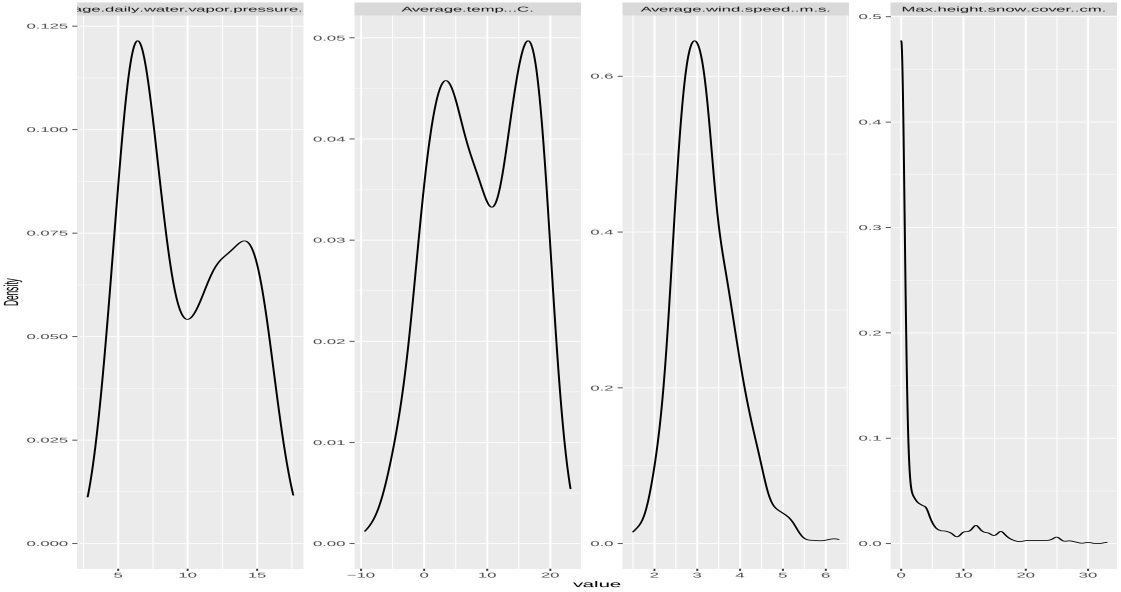


Figure 4.4: Presentation of some variable densities

The main *Sum of monthly rainfall* variable takes values close to zero with high probability. Values of *Average Temperature* and *Water Vapor Pressure* have class isolation properties into 2 groups. The *Average wind speed* density resembles Normal distribution. The mean is shifted to the left so it may be right-skewed distribution.

Chapter 5

Experimental part

In this chapter, we want to present the method of operation of individual machine learning methods and carry out rainfall predictions in various scenarios.

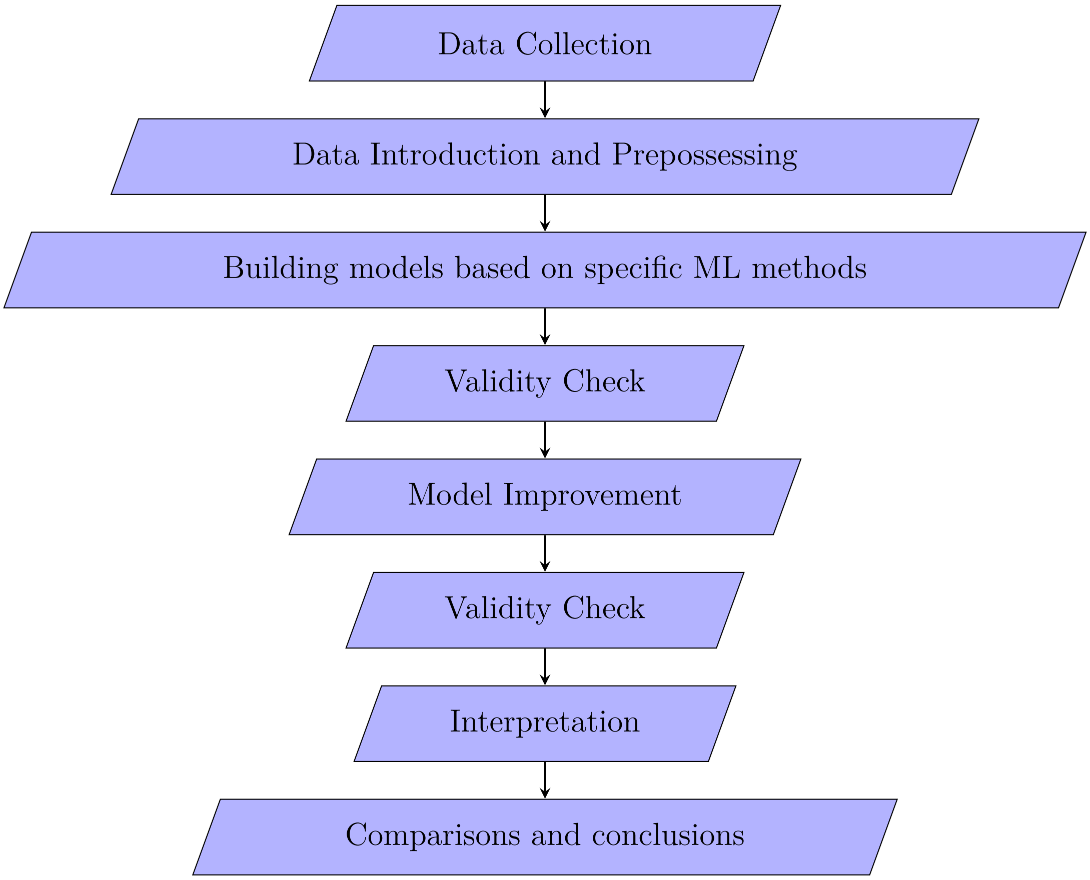

Data Collection
Data Introduction and Prepossessing
Building models based on specific ML methods
Validity Check
Model Improvement
Validity Check
Interpretation
Comparisons and conclusions

Figure 5.1: Diagram of individual stages in the analysis

The diagram above shows the scheme of proceeding during the prediction problem in our case.

In the experimental part, all the simulations are running on the same data set. For the purposes of greater reliability of the fit of the predictive models to the data in each method, simulations are performed separately for the learning and test parts.

The data is split randomly. The training sample contains 70%, and the test sample 30% of observations from the data set.

5.1 Regression models

5.1.1 Simple Model 1.

At the beginning we want to check the correctness of the model using all independent variables for the present time. In other words we are trying to predict: **How many litres of rain will fall in a proper month using all the data** (also variables that relate to water).

The forward regression is used to build the **Model 1** - all variables, prediction for present time.

Table 5.1: Table describing features of the Model 1

Variable	**Coefficient**	**Std. Error**	**Test value**	**p-value**
Intercept	212.002	190.367	1.114	0.26602
Maximum daily rainfall	2.11061	0.07702	27.403	< 2e-16
Days with rain	2.55217	0.20410	12.504	< 2e-16
Days with snow	1.60260	0.27053	5.924	6.24e-09
Days with storm	1.15066	0.38132	3.018	0.00269
Days with haze	0.28065	0.12002	2.338	0.01981
Month	-0.58612	0.22496	-2.605	0.00948
Average water vapor pressure [hPa]	0.74970	0.37618	1.993	0.04687
Average relative humidity [%]	0.41369	0.17060	2.425	0.01570
Days with fog	-0.38440	0.24202	-1.588	0.11292
Average pressure at the station level [hPa]	-0.26770	0.18936	-1.414	0.15813

Results of the simulation were obtained via R.

It should be quite logical for us that the parameters directly related to rainfall have the greatest impact on the total rainfall in a month.

From the table definitely it is known, that variables *Maximum daily rainfall* and *Days with rain* have the greatest influence on dependent variable- *Sum of monthly rainfall.* Also the Pearson's Correlation figure shows that dependence. All three parameters-**Standard Error**, **Test Value**, **p-value** are responsible for that fact.

The *p-value* for both variables has the lowest value- <2e-16. If the *p-value* is lower than significance level (in our case 0.05), then for sure the given independent variable should appear in the model (H_0 is rejected).

The *Test value* for both variables is 27.403 and 12.504, the next one for variable *Days with snow* equals 5.924. this again demonstrates the importance of the variables *Maximum daily rainfall* and *Days with rain.*

Variables *Days with fog* and *Average pressure at the station level [hPa]* have the highest *p-value* parameter, over 0.05. The *Test value* also for these variables is close to zero comparing on the rest. It means that these variables have the weakest influence for dependent variable (rainfall). It is also shown on the *Coefficient* values. The reason why these variables appear in the model is the increase of the R^2 factor in the model.

It would seem that since the *p-value* for those variables is so high, removing it from the model should improve the goodness of fit. Unfortunately, this intuition is wrong, because a **high *p-value* is not a premise for a strong insignificance of the variable**. It can appears for two reasons:

- the null hypothesis (of the variable's insignificance) is true.
- the null hypothesis is false but the test is of low power.

The coefficient Standard Error measures the average distance that the observed values fall from the regression line. If the number is close to zero it indicates that the observations are closer to the fitted line.

Assuming that, all model assumptions are satisfied, we can say that with 95% confidence the true parameter $\hat{\beta}_i$ lies in such proper confidance interval.

$$\hat{\beta} \in \left[\hat{\beta}_i - 1.96 s.e(\hat{\beta}_i), \hat{\beta}_i + 1.96 s.e(\hat{\beta}_i)\right]$$

, where $s.e(\hat{\beta}_i$ means Standard Error of $\hat{\beta}_i$.

In the table below we have computed the specific intervals for each variable. There is a 2.5% chance to get a value lower than the value of second column and 2.5% to get a value higher than the value of third column.

Table 5.2: Confidence Intervals

Variable	**2.5 %**	**97.5 %**
Intercept	-162.1143	586.119
Maximum daily rainfall	1.9592	2.662
Days with rain	2.151	2.9533
Days with snow	1.0709	2.1342
Days with storm	0.4013	1.9
Days with haze	0.0448	0.5165
Month	-1.028	-0.144
Average water vapor pressure [hPa]	0.0104	1.489
Average relative humidity [%]	0.078	0.7489
Days with fog	-0.86	0.091
Average pressure at the station level [hPa]	-0.6398	0.104

Table 5.3: Comparison of predicted and real rainfall data from test sample

Index	**Real rainfall value**	**Predicted rain value**
5	58.4	42.466
7	123.4	91.36
8	95.2	81.62
9	9.3	16.46
11	38.6	38.89
15	35.3	41.34

After the demonstration of the computed rainfalls values it can be noticed that the relevant values do not differ from each other - they may belong to the same rainfall classes.

5.1.2 Model 2- deleting two variables

The next step is to build similar regression model (like *Model 1*), but now the most significance variables, which are directly related to rainfall (*Maximum daily rainfall* and *Days with rain*) will be not used.

Table 5.4: Table describing features of the Model 2

Variable	**Coefficient**	**Std. Error**	**Test value**	**p-value**
Intercept	937.27811	370.52554	2.530	0.011760
Average water vapor pressure [hPa]	3.52453	0.88610	3.978	8.11e-05
Average cloud cover	9.06329	2.29502	3.949	9.10e-05
Days with storm	3.02792	0.67807	4.465	1.01e-05
Year	0.05411	0.10955	0.494	0.621618
Average pressure at the station level [hPa]	-1.18108	0.35664	-3.312	0.001002
Average relative humidity [%]	1.34117	0.35402	3.788	0.000172
Days with fog	-1.05545	0.43986	-2.400	0.016822
Min temp ground	0.98077	0.33643	2.915	0.003732
Month	-0.91607	0.39923	-2.295	0.022214
Days with wind speed >= 15m/s	6.78972	2.93624	2.312	0.021207
Days with haze	0.38672	0.27449	1.409	0.159567

Results of the simulation were obtained via R.

The **Model 2** contains 11 independent variables. In that case based on *Test value* and *p-value* three variables: *Average water vapor pressure [hPa]*, *Average cloud cover* and *Days with storm* have the highest level of significance.

We can not observe as low *p-values* as in the *Model 1.* The fact that the model contains *Year* variable is quite undesirable, because it is known that it carries a big randomness.

The model fitting comparison will take place in the **Summary** subsection.

5.1.3 Deleting given observations- Cook's Distance

At this stage we would like to check how the model will behave after removal some outliers. We are rejecting all influential observations, using **Thumb Rule**. For the *Model 1* the limit equals 0.009. During that process 23 observations have been rejected.

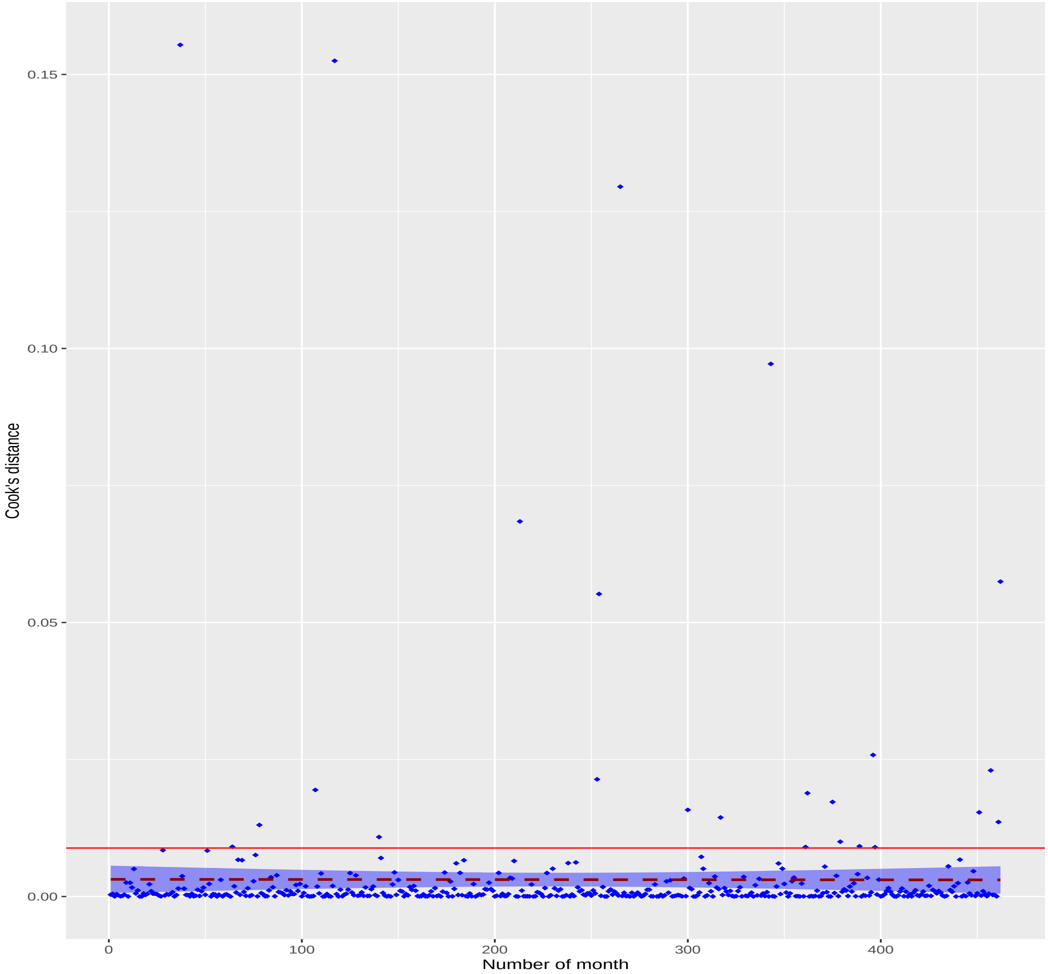


Figure 5.2: Graph showing the Cook's distances for individual observations together with the limit corresponding to the Rule of Thumb

At the boxplots below are shown the residuals for *Model 1* and *Model 1 after Cook*. The mean values and quantiles in both charts are similar. Definitely outliers scatter is observed. After removal the influential data the sample is more clustered.

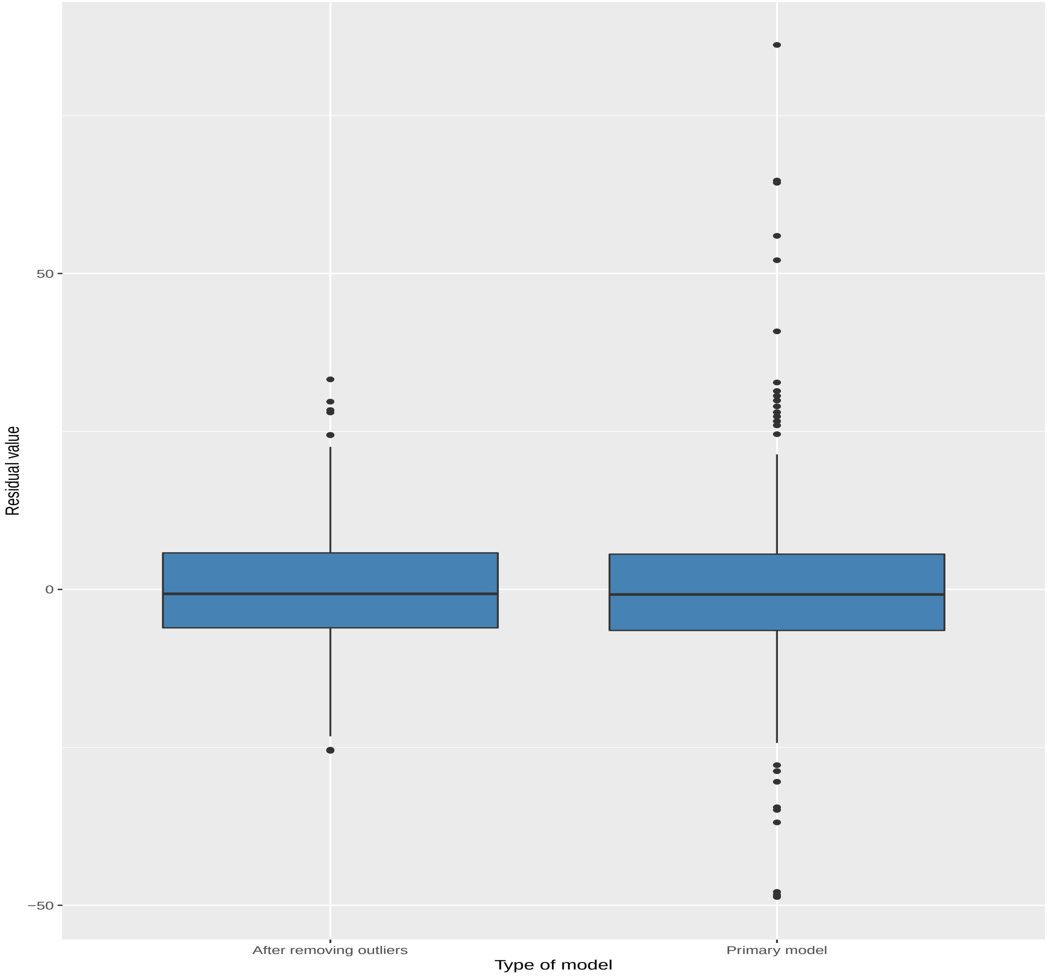


Figure 5.3: Comparison of the residuals from the Model 1 and the Model 1 after removal the Cook's influance observations

Table 5.5: Comparison of Residual Summary Statistics

Type of model	**Min**	**First Quartile**	**Median**	**Third Quartile**	**Max**
Model 1	-48.681	-6.493	-0.803	5.58	86.18
Model 1 after Cook	-25.546	-6.08	-0.7	5.781	33.242

The outliers removal successfully has imporved the *Model 1*. The **median** is closer to zero, which means better model fitting.

5.1.4 Future data prediction

One of the main goals was an attempt to predict rainfall in the future time period. Due to the prediction's problem it is needed to join a proper column of Sum of rainfalls in the next 1 and 3 months- respectively **Model 3** and **Model 4**. Thanks to that we will be able to build models where the last column of data set- **Rainfall in 1 month** or **Rainfall in 3 month** will be our new dependent variable.

At first we would like to know how will look like the Correlation Table now.

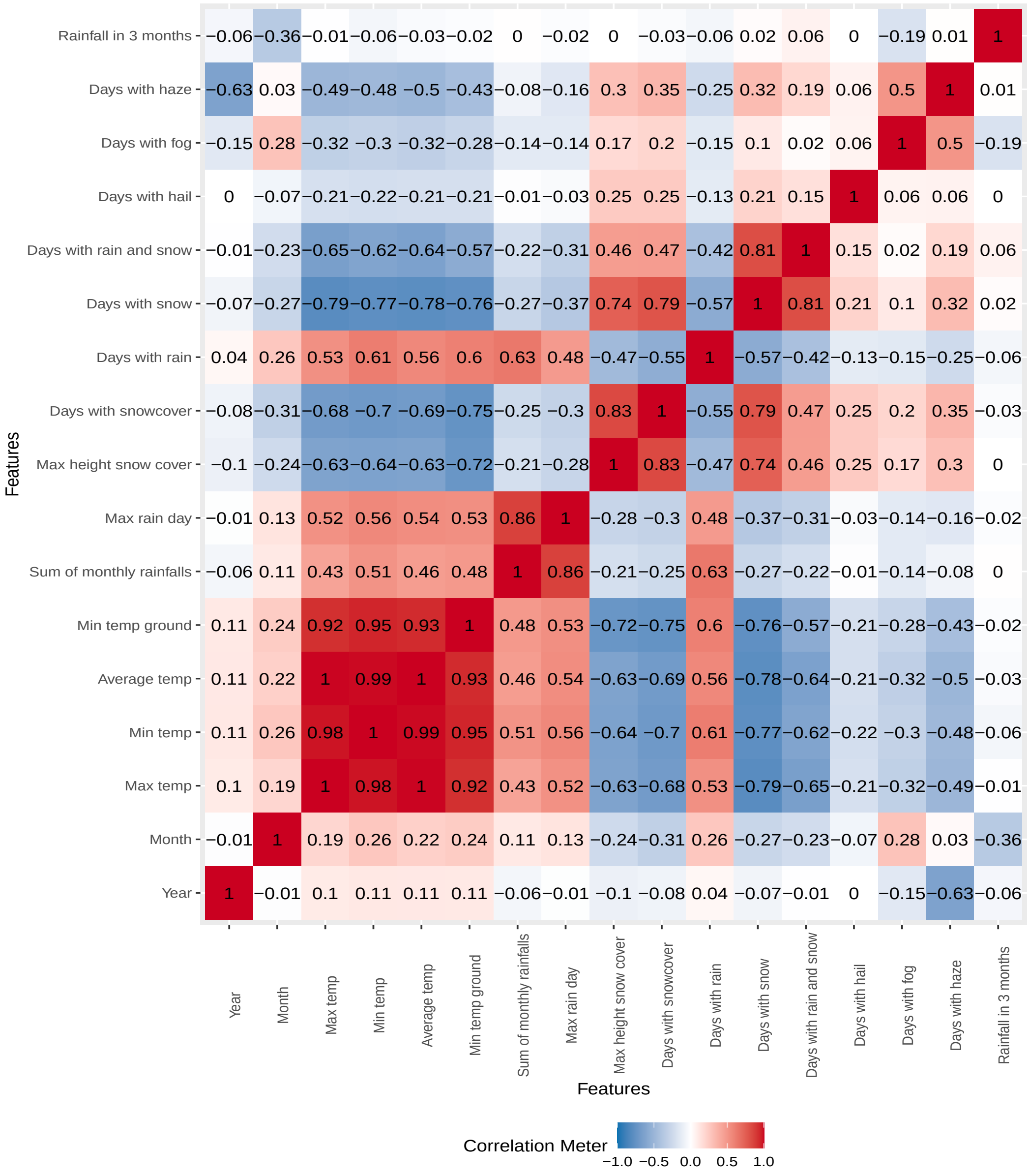


Figure 5.4: Correlation table for first part of variables. Prediction for rainfall in 3 months.

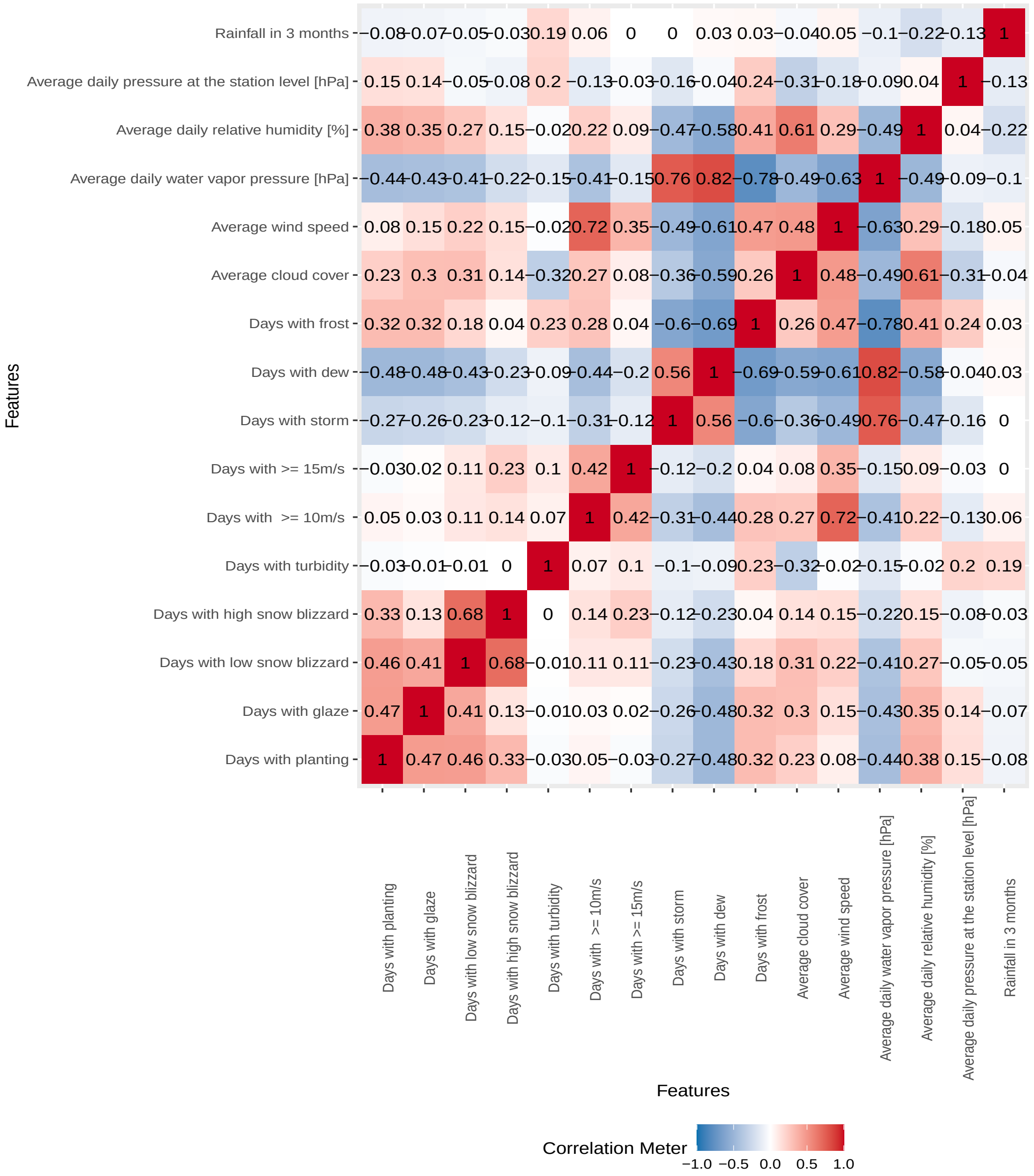


Figure 5.5: Correlation table for second part of variables. Prediction for rainfall in 3 months.

As we see the correlations between *Rainfall in 3 months* and all dependent variable are very weak. Actually it can be said, that no correlations are here observed.

Table 5.6: Table describing features of the Model 4

Variable	**Coefficient**	**Std. Error**	**Test value**	**p-value**
Intercept	124.58988	51.41567	2.423	0.015779
Month	-1.81796	0.59679	-3.046	0.002454
Average relative humidity [%]	-2.21836	0.67504	-3.286	0.001095
Days with haze	1.45387	0.33933	4.285	2.24e-05
Average cloud cover	12.93757	3.65072	3.544	0.000436
Days with low snow blizzard	-4.42766	1.70331	-2.599	0.009645
Average water vapor pressure [hPa]	-4.86565	3.13191	-1.554	0.120991
Sum of monthly rainfalls	0.17027	0.05886	2.893	0.004005
Max temperature	11.11933	3.82510	2.907	0.003831
Average temperature	-10.85719	4.98134	-2.180	0.029808

Results of the simulation were obtained via R.

Almost every *p-value* is in a lover level than 0.05. The |Test Value| are in range $(1.5, 4.5)$- close to zero.

In that subsection, due to the similarity of models only prediction of *Rainfall in 3 months* is presented. The summary of fitting models factor will be described in the next subsection.

5.1.5 Summary of regression models

In the end of Regression Models section it is needed to compare the constructed model based on specific factors. We will use **Multiple R-squared**, **Adjusted R-squared**, **F-statistic and Degrees of Freedom**, **Pearson Correlation** and **Mean Square Error**.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

, where n represents the number of data points in the specific sample.

Table 5.7: Confusion Matrix of Naive Bayes Classifier

Type of model	R^2	**Adj**R^2	**F-statistic and DT**	**Pearson Correlation**	**MSE**
Model 1	0.8399	0.8363	236.6 on 10 and 451	0.9376	202.47
Model 1 after Cook	0.87	0.8675	319 on 9 and 428	0.9378	220.45
Model 2	0.4903	0.4778	39.35 on 11 and 450	0.768	725.89
Model 3	0.2506	0.2356	16.76 on 9 and 451	0.5132	934.42
Model 4	0.2568	0.2419	17.24 on 9 and 449	0.4302	733.14

Control Model in the Training Sample

Only first two models have sense to being used in prictical tasks. There is a huge differance between these two and the rest of models in every single parameter.
We can observe, that the *Model 1 after Cook* has better fitting indicators- R^2 and $AdjR^2$ are higher. The *F-Statistic* has also raised.

The *Model 2* tells actually nothing, it can not be used because of high MSE factor. It means, that if we delete two most important water variables from the model it will be very difficult to predict the dependent variable.

Shifting the Rainfall variable and prediction for the next one or three months is impossible in that case. Both scenarios where we use *Model 3* and *Model 4* have very close to each other indicators. The conclusion is quite clear, it does not matter how far we want to predict the rainfall. There don't exist any linear correlation between given variables.

5.2 Classification Methods

As it was written before, for the next methods of classification the variable prediction approach will change. The problem of classification will be more general. In that stage will try to predict the proper class of rainfall.
1 class- the lowest rainfall level, 4 class- the highest rainfall level (the rarest type of rain).

In section for Naive Bayes and Decision Trees the algorithms the effectiveness of a particular model will be checked by examining the classification error and presenting by the Confusion Matrix.

In each simulation we are using two models.
Control model- all the independent variables are being used in the model
Regression model- all the independent variables from *Model 1* are being used.

Confusion Matrix, also known as Table Errors. It is used to present the exact way in which the classes of a given variable have been classified by the proper, used model. Thanks to it, we are able to check the number of actual classes that correspond to the classes provided by the model. The rows of the matrix represent the classes provided by the model. The columns refer to the real classes. The marked **elements on the matrix diagonal** show the **number of correctly classified classes**.

5.2.1 Decision Trees

Table 5.8: Confussion Matrixes using Decision Trees

Control model for Present Time using Training sample

	1	**2**	**3**	**4**
1	**177**	22	0	0
2	28	**97**	26	4
3	0	13	**44**	9
4	0	0	7	**35**

Control model for present Time using Test sample

	1	**2**	**3**	**4**
1	**76**	15	0	0
2	18	**36**	10	0
3	0	5	**12**	5
4	0	1	7	**13**

Regression model for present Time using Training sample

	1	**2**	**3**	**4**
1	**179**	21	0	0
2	25	**106**	25	5
3	1	5	**45**	9
4	0	0	7	**34**

Regression model for present Time using Test sample

	1	**2**	**3**	**4**
1	**76**	15	0	0
2	18	**36**	10	0
3	0	5	**12**	5
4	0	1	7	**13**

Control model in 3 months using Training sample

	1	**2**	**3**	**4**
1	**191**	36	2	0
2	15	**90**	18	1
3	2	7	**34**	7
4	1	2	12	**42**

Control model in 3 months using Test sample

	1	**2**	**3**	**4**
1	**75**	19	3	0
2	13	**25**	13	1
3	0	7	**13**	3
4	0	3	11	**11**

Regression model in 3 months using Training sample

	1	**2**	**3**	**4**
1	**189**	34	2	0
2	19	**89**	17	1
3	0	10	**35**	7
4	1	2	12	**42**

Regression model in in 3 months using Test sample

	1	**2**	**3**	**4**
1	**75**	19	3	0
2	13	**25**	13	1
3	0	7	**13**	3
4	0	3	11	**11**

In the table containing Confusing matrices above we can say that the *Regression model* definitely works better in the **Training samples**- has higher values on the diagonals comparing with the *Control model.* Comparing the matrices where models are based on *Training samples* it is hard to choose which type of model and which prediction works better only during checking the diagonals.

We can see good separation of class **4** in *Training sample* during *Prediction for present.* Matrices from *Test samples* are for each prediction time identically. We can actually find the elements on the diagonals which are quite similar between both prediction times.

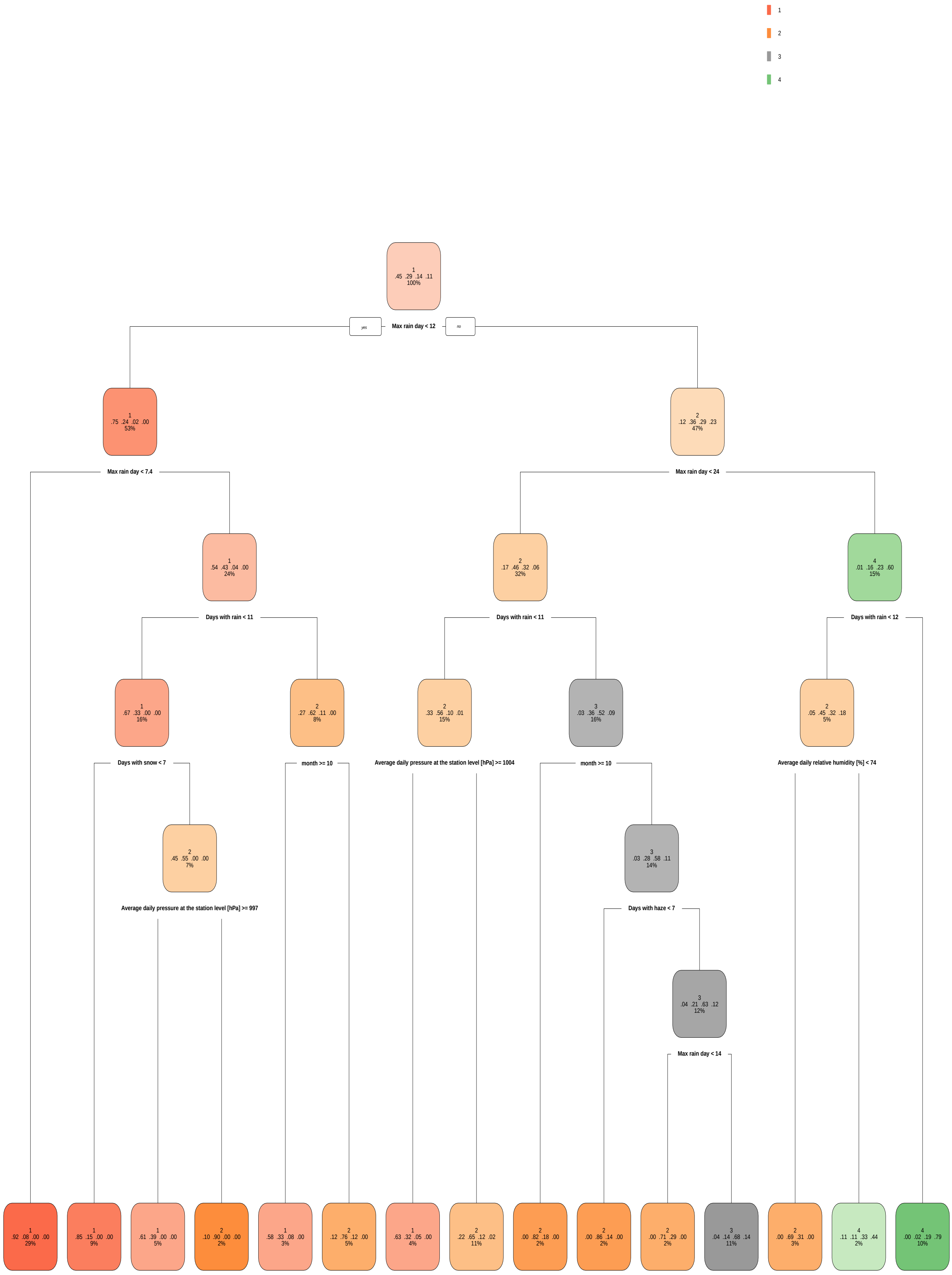


Figure 5.6: Generated chart of CRAM Classification Tree on predicting rainfall class for 3 months.

The Tree above shows the final results of classification on the *Test sample.* It contains 15 leafs- 5 represents the **1** class, 7 class **2** , only 1 class **3** and 2 class **4**. The darker the color shades of the leaves, the greater the purity of the individual leaves of the classification tree.

5.2.2 Naive Bayes

Table 5.9: Confusion Matrices using Naive Bayes Classifier

Control model for Present Time using Training sample

	1	**2**	**3**	**4**
1	**172**	54	4	0
2	28	**31**	18	6
3	20	42	**37**	11
4	2	0	8	**29**

Control model for Present Time using Test sample

	1	**2**	**3**	**4**
1	**52**	28	1	1
2	14	**12**	9	2
3	9	19	**23**	8
4	2	3	6	**9**

Regression model for present Time using Training sample

	1	**2**	**3**	**4**
1	**186**	53	2	0
2	22	**32**	14	3
3	14	42	**45**	10
4	0	0	6	**33**

Regression model for present Time using Test sample

	1	**2**	**3**	**4**
1	**57**	24	0	0
2	14	**18**	9	0
3	6	19	**25**	7
4	0	1	5	**13**

Control model in 3 months using Training sample

	1	**2**	**3**	**4**
1	**158**	50	6	0
2	37	**33**	15	0
3	16	44	**50**	20
4	2	0	6	**25**

Control model in 3 months using Test sample

	1	**2**	**3**	**4**
1	**67**	28	1	2
2	11	**12**	11	2
3	6	19	**12**	8
4	2	3	5	**9**

Regression model in 3 months using Training sample

	1	**2**	**3**	**4**
1	**171**	51	3	1
2	32	**34**	13	2
3	10	42	**54**	14
4	0	0	7	**28**

Regression model in in 3 months using Test sample

	1	**2**	**3**	**4**
1	**69**	28	2	0
2	10	**14**	6	0
3	7	20	**19**	5
4	0	0	2	**16**

In the table containing Confusing matrices above we can say that the *Regression model* definitely works better- has higher values on the diagonals comparing with the *Control model.* It is not obvious if the Model, which predict present rainfall has better accuracy.

We can see good separation of class **4** in *Regression Model* during *Prediction in 3 months.* It fits class **4** only to the real class **3**, but most of examples it classify correctly.

5.3 Methods Comparison and Summary

To summarise all Confusion Matrices in simple and clear way we can use the **Classification Error**.

$$\text{Classification Error} = 1 - \sum_{i=1}^{C} \frac{m_{ii}}{n}$$

, where C means number of classes. m_{ii} refers to elements on the matrix's diagonal, and n is the number of all observations.

For the final comparison of the all three algorithms we have created Confusion Matrix for the *Model 4* in regression models. The predicted values have been transformed to classes.

Table 5.10: Confusion Matrix of Regression Model 4

	1	**2**	**3**	**4**
1	**31**	16	2	1
2	42	**26**	15	5
3	11	17	**17**	7
4	1	3	4	1

Table 5.11: Table of comparing Classification Errors of all algorithms

Proper model and sample	**NV**	**Shift NV**	**DT**	**Shift DT**	**Shift Regr**
Control model, training sample	0.5195	0.5390	0.2359	0.2239	-
Control model, test sample	0.6111	0.5657	0.3081	0.3706	-
Regression model, training sample	0.3593	0.3788	0.2121	0.2283	-
Regression model, test sample	**0.4293**	**0.4040**	**0.3081**	**0.3706**	**0.6231**

NV-Naive Bayes Classifier prediction for present time;
Shift NV-Naive Bayes Classifier prediction for 3 months;
DT-Decision Tree prediction for present time;
Shift DT-Decision Tree prediction for 3 months;
Shift Reg-Regression Model 4 prediction for 3 months

Conclusions regarding the algorithms presented in the experimental part:

- Definitely *Regression Models* in Naive Bayes Classifier were more effective, what is shown in the table above. It means, that often the maximum number of independent variables in model does not give the highest accuracy.

- The Classification Errors in Decision Trees in *Regression Models* in most cases are also lower or equal to errors in *Control Model* . Only comparing the *Training samples* in **Shift DT** we are getting a minimal worse result in the *Regression Models* (0.2239 vs 0.2283).

- In each method the Classification Errors are higher for test samples (normally it happens quite often)
- We can observe that in Naive Bayes Classifier lower errors occur during prediction of the future rainfalls (in our case in 3 months) on test samples (0.4293 vs 0.4040 and 0.611 vs 0.5657).
  However, on the other hand, the opposite situation occurs on the training set- the Classification Errors is lower in that case.(0.5195 vs 0.5390 and 0.3593 vs 0.3788)
- Only during comparison the *Control Models* in *Training Samples* the lowest Classification Error occurs in the **Shift DT**. Normally the **DT** has the lowest errors.
- The Decision Trees present the best accuracy of all algorithms (0.3081 and 0.3706).
- The Model of Linear Regression has by far the highest error (0.6231).- this may be due to a different prediction concept.

Chapter 6

Summary and conclusion

Unfortunately, the Long term forecast in the context of rainfall is too complex for the demonstrated methods. Various improvements could be applied to the methods presented, such as *bagging, boosting* or *Random Forest.* There are also many other conceptions and methods of classification and prediction.

In the thesis was shown, that weather forecasting is a long and complicated process, that is why one of main priorities should be water independence. The agricultural sector should use water more efficiently. Technologies of irrigation systems are becoming more and more common in the world, but unfortunately for economic reasons they are not yet found in Poland. The construction of appropriate solutions in this context could be the key of importance for the development of Polish agriculture on the global market.

Nowadays, using global weather engines, the forecast for 3-7 days is accurate to over 80% [14]. They are created thanks to subseasonal models simulating the evolution of the atmosphere in the following days. Weather models are "fed" with millions of data each day on a variety of parameters. Current factors such as stratospheric winds and cloud oscillations are also taken into account. These are very complex processes that require specialists from various fields, using also machines with very high computing power. The goal for the coming years is to achieve similar effectiveness with monthly forecasts. Predicting the occurrence of a given weather event can bring diametrical benefits, making people's lives better.

Bibliography

[1] Agresti, A. *Foundations of linear and generalized linear models.* John Wiley & Sons, 2015.

[2] Ajitesh, K. Introduction to overfitting and underfitting. *Lecture# 14* (2020).

[3] Breiman, L., Friedman, J., Olshen, R., Stone, C. Classification and regression trees. wadsworth int. *Group 37*, 15 (1984), 237–251.

[4] Cure, J. D., Acock, B. Crop responses to carbon dioxide doubling: a literature survey. *Agricultural and forest meteorology 38*, 1-3 (1986), 127–145.

[5] Johnson, R. A., Wichern, D. W., et al. *Applied multivariate statistical analysis*, vol. 6. Pearson London, UK:, 2014.

[6] Kazdin, A. A primer on statistical significance. math vault.

[7] Kus, J. Gospodarowanie wodą w rolnictwie. *Studia I Raporty IUNG-PIB [In Polish: Water Management in Agriculture. IUNG-PIB Studies and Reports] 47*, 1 (2016), 83–104.

[8] Majewski, J. Rola owadów zapylających w zapewnieniu bezpieczeństwa żywnościowego polski. *Roczniki Naukowe Stowarzyszenia Ekonomistów Rolnictwa i Agrobiznesu 19*, 3 (2017).

[9] Mazur, B. O testowaniu.

[10] Mbow, C., Rosenzweig, C. E., Barioni, L. G., Benton, T. G., Herrero, M., Krishnapillai, M., Ruane, A. C., Liwenga, E., Pradhan, P., Rivera-Ferre, M. G., et al. Food security.

[11] of Meteorology, I., Water Management, N. R. I. Imwm-pib public data, 2021.

[12] Ortiz-Bobea, A., Ault, T. R., Carrillo, C. M., Chambers, R. G., Lobell, D. B. Anthropogenic climate change has slowed global agricultural productivity growth. *Nature Climate Change 11*, 4 (2021), 306–312.

[13] Ortiz-Bobea, A., Ault, T. R., Carrillo, C. M., Chambers, R. G., Lobell, D. B. Anthropogenic climate change has slowed global agricultural productivity growth. *Nature Climate Change 11*, 4 (2021), 306–312.

[14] Pegion, K., Kirtman, B. P., Becker, E., Collins, D. C., LaJoie, E., Burgman, R., Bell, R., DelSole, T., Min, D., Zhu, Y., et al. The subseasonal experiment (subx): A multimodel subseasonal prediction experiment. *Bulletin of the American Meteorological Society 100*, 10 (2019), 2043–2060.

[15] Rolnictwa, R. S. Główny urząd statystyczny, 2015.

[16] Van den Hove, S., McGlade, J., Mottet, P., Depledge, M. H. The innovation union: a perfect means to confused ends? *Environmental science & policy 16* (2012), 73–80.