Ministry of Education and Science of the Republic of Kazakhstan
Suleyman Demirel University

Aigerim Temirali

Statistical inference and machine learning in big data

THESIS

Presented in Partial Fulfillment for the
(degree code: 6M060100)
Department of Computer Sciences
Faculty of Engineering and Natural Sciences

Supervisor: **Zhaparov Meirambek**

Kaskelen, 2019

Аңдатпа

Үлкен деректермен жұмыс істеудің жаңа әдістерінің қажеттілігі көптеген ғылыми салалардағы жалпы тақырып болып табылады. бірақ оның анықтамасы контекстке байланысты өзгереді. Статистикалық идеялар мұның ажырамас бөлігі болып табылады және ішінара жауап ретінде статистикалық көшіру. оқыту бойынша тақырыптық бағдарлама. Бұл мақалада қамтылған тақырыптар туралы шолу келтіріледі. түрлі қосымшаларға ортақ болып көрінетін проблемалар мен стратегияларды сипаттайды және осы проблемаларды және стратегияларды нақты анықтау үшін бірнеше мысал қосымшаларын қамтиды.

Аннотация

Потребность в новых методах для работы с большими данными является общей темой в большинстве научных областей, хотя ее определение имеет тенденцию варьироваться в зависимости от контекста. Статистические идеи являются неотъемлемой частью этого и, как частичный ответ, тематической программы по статистическому выводу, обучению. В этом тезисе дается обзор затронутых тем, описываются проблемы и стратегии, которые кажутся общими для многих различных областей применения, и включаются некоторые примеры приложений, чтобы сделать эти проблемы и стратегии более конкретными.

Contents

1. Introduction

1.1 Motivation

In my practice, I met with different definitions: · Big Data is when data is more than 100GB (500GB, 1TB, who likes it)
· Big Data is data that cannot be processed in Excel

· Big Data is data that cannot be processed on a single computer.
And even such:
· Big Data is generally any data.
· Big Data does not exist, marketers have invented it.
Thus, under Big Data I will understand not some specific amount of data or even the data itself, but their processing methods, which allow distributed information to be processed. These methods can be applied both to huge data arrays (such as the content of all pages on the Internet) and to small ones (such as the content of this thesis).

1.2 Aims and Objectives

Principles of working with big data

Based on the definition of Big Data, you can formulate the basic principles of working with such data:
1. Horizontal scalability. Since there can be as much data as possible - any system that involves processing big data should be expandable. The data volume increased by 2 times - the amount of iron in the cluster increased by 2 times and everything continued to work.
2. Fault tolerance. The principle of horizontal scalability implies that there can

be many machines in a cluster. For example, Yahoo's Hadoop Cluster has more than 42.000 machines (this link can be used to look at cluster sizes in different organizations). This means that some of these machines will be guaranteed to fail. Methods of working with big data should take into account the possibility of such failures and survive them without any significant consequences.

3. Local data. In large distributed systems, data is distributed across a large number of machines. If the data is physically located on the same server, and processed on the other - the cost of data transfer may exceed the cost of processing itself. Therefore, one of the most important design principles for BigData solutions is the principle of data locality - if possible, we process data on the same machine on which we store them.

All modern means of working with big data somehow follow these three principles. In order to follow them - it is necessary to invent some methods, ways and paradigms of developing data development tools. I will analyze one of the most classical methods in my thesis.

2. What is Machine Learning, Statistical inference?

2.1 Machine learning

Machine Learning is an enormous and developing zone. In this part, we can't in any way, shape or form even study this region, however we can give some unique situation and a few associations with likelihood and insights that should make it simpler to consider AI and how to apply these techniques to certifiable issues. The key issue of insights is fundamentally equivalent to AI: given a few information, how to make it noteworthy? For measurements, the appropriate response is to develop expository estimators utilizing amazing hypothesis. For AI, the appropriate response is algorithmic forecast. Given an informational collection, what forward-looking surmisings would we be able to draw? There is an inconspicuous piece in this portrayal: how might we know the future if the sum total of what we have is information about the past? This is the essence of the issue for AI, as we will investigate in the section.

Machine learning is a data analysis method that automates building an analytical model. It is a branch of artificial intelligence based on the idea that machines should be able to learn and adapt through experience. It is closely related to computational statistics, which makes predictions based on statistical data, computer generated. It is sometimes confused with data mining, but it's more focused on analyzing exploration data, while machine learning includes complex algorithms that are mainly used to predict when a machine training concentrates on prediction based on already known attributes derived through the training data, then the data search focuses more on the search for unknowns attributes in any data.

Machine learning arose because of the desire for artificial intelligence. In the

early days artificial intelligence already as an intellectual field the researchers were very interested in machines learning from data. Therefore, they tried to approach the problem with the help of various symbolic methods, as well as of the methods that at that time were called neural networks, usually these were only models that were later discovered for repacking general linear probability models and statistics.

With a focus on logical and knowledge-based approaches, the gap between artificial intelligence and machine learning. Probabilistic systems were infected with both theoretical and practical issues of data collection and presentation. By 1980 expert systems emerged in the year to dominate AI, but statistics were unsuccessful. Work continued on symbolic and knowledge-based systems that lead to inductive logic programming but the statistical research area is currently time is outside the AI area, in the pattern recognition lines and data retrieval.

Machine learning became a separate field and began to expand in the 1990s. Line changed Its goal is to achieve AI, trying to solve more practical solvable problems. Then the field pushed his attention away from the symbolic methodologies that it inherited from artificial intelligence, and instead passed on to methods and models taken from probability and statistics.[1]

These days, data is too large for people to process and analyze by themselves. Machine learning mainly uses a range or spectrum based on an optimization method, a large number of parameters. For people it is inappropriate to find such an optimal manual setup. For example, recognition of dynamics from tone, tone and amplitude. There is no guarantee that machine learning will work in every case. Sometimes machine learning fails, requiring an understanding of the problem that needs to be solved in order to apply the correct algorithm, very large data requirements. These learning algorithms require large amounts of data, learning. It would be very difficult to work with such large amounts of data or to collect such data. But things like increasing the amount and variation of available data, the variety processing, which is cheaper and more powerful, and more accessible data storage, in our days we can quickly and automatically create models and algorithms that can analyze larger and more complex data providing faster and more accurate results in large scale. Therefore, machine learning is quickly becoming very important and widespread, an embedded part of our daily life.[2]

Machine learning applications may be related to spam filtering, optical character recognition and search engines. Machine learning is used data for determine which algorithm is best for generating results based on quantity of data, quality and nature of data. This data is then used to mining in various ways, for example, recommendation systems such as similar products on eBay, personalized content on google plus pages, video ads on sites like YouTube and the latest but not least offers friends on facebook. Also used for intelligent search in Google search engines and Bing.

2.2 Statistical Inference

Factual deduction is characterized as the procedure deriving the properties of the given dissemination dependent on the information. At the end of the day, it concludes the properties of the populace by leading theory testing and getting gauges. Here, the information utilized in the examination are gotten from the bigger populace.

Measurable surmising varies from the illustrative measurements as the clear insights is relies on the watched information. Likewise, in the illustrative measurements, there is no supposition that the information is gotten from the greater populace. However, in the factual surmising, the outcomes about the populace can be made, by acquiring the information from the populace and by applying inspecting strategies.

Steps for making a statistical inference :

1. The population is modelled by a probability distribution for which the parameter is known.

2. Generalisation about the population can be made by selecting the sample.

That is, when the hypothesis about the population is given, the first step is the select a statistical model that gives data and the second step is to reduce proportions from the model so that the statistical inferences can be drawn.

Also, the statistic obtained from the sample gives the information about the parameter. The below points are important while making the statistical inference.

- The statistic obtained from the sample is not same as the value of the

parameter as the sample is subset of the population

- Observed value depends upon the sample selected.
- Variability in the values of the statistic is unavoidable. Everyone has known about enormous information. Numerous individuals have enormous information. In any case, just a few people realize how to manage huge information when they have it.

I research genuine issues with enormous information. These techniques will be portrayed in the event that reviews that clarify how each is utilized to take care of certifiable issues. What's more, we can likewise build up our programming aptitudes utilizing the techniques we simply figured out how to perform useful undertakings and get results

Similarly as there are numerous factual and AI strategies for breaking down enormous information.

The dangerous development of Big Data carries critical difficulties and possibilities to a wide scope of fields, from genomics to customized medication to data innovation. It likewise makes huge open doors for measurements. The rise of information science guarantees to alter ventures from business to human services to government, and to change how we work, live and impart.

Measurable surmising varies from the illustrative measurements as the clear insights is relies on the watched information. Likewise, in the illustrative measurements, there is no supposition that the information is gotten from the greater populace. However, in the factual surmising, the outcomes about the populace can be made, by acquiring the information from the populace and by applying inspecting strategies.

3. Using methods for ML

3.1 About company of "EduCon"

The company of EduCon is a set of training centers. there are 26 branches in Kazakhstan. The company Edukon gave us the data of its students. which was present in the winter training camp. And below is the data. and I took that I need the data of the students and their test result each month and the last as they evaluated part of the organization of the winter camp. In this training camp there were 627 students from different regions. from different cities of Kazakhstan.

Figure 3.1: Data of students

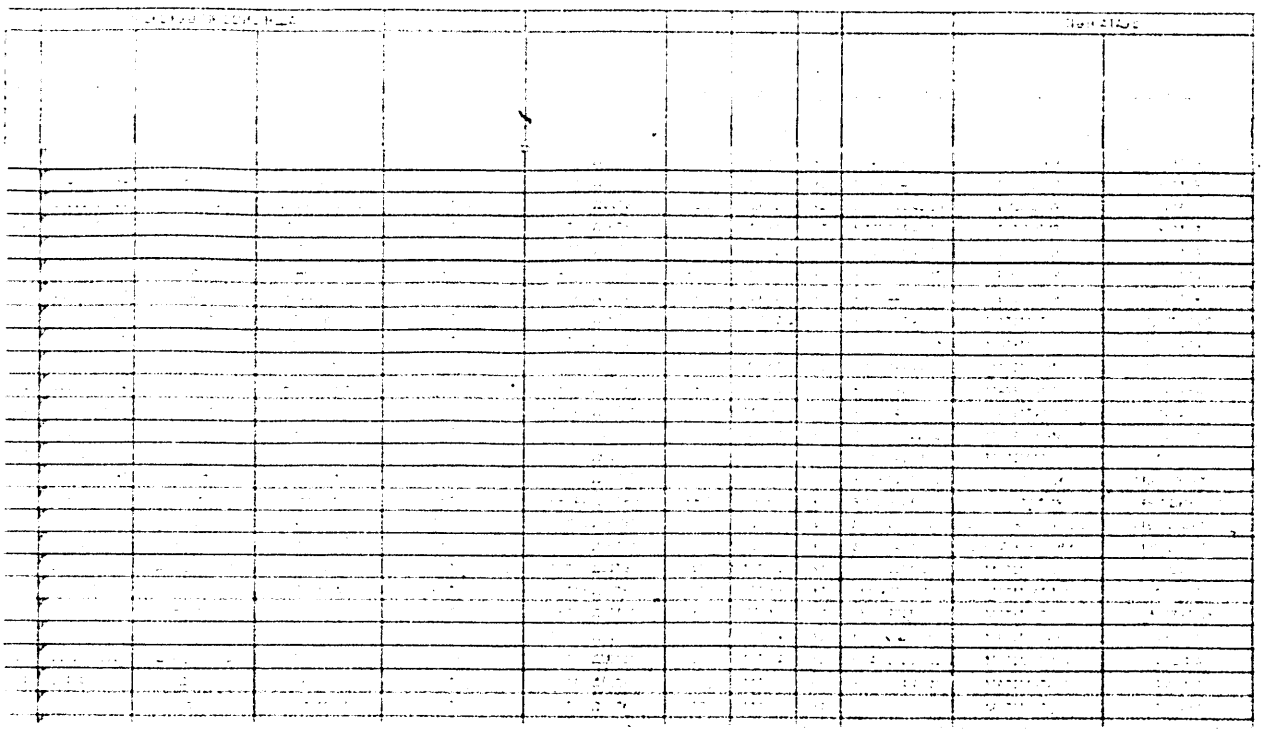
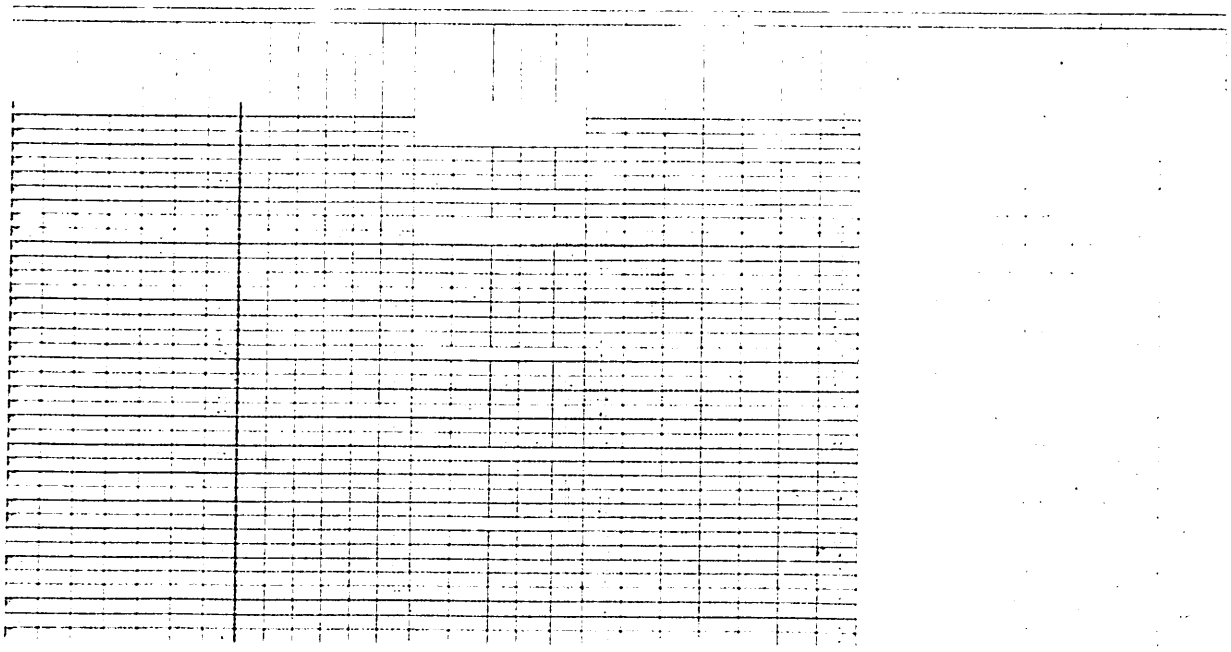


Figure 3.2: Data of students



"EduCon" LLP is a single national for additional education centers Long-term experience with complex maintenance of preparation for testing education and consulting company.

Mission of EduCon Company is knowledge and human intelligence the value of science and technology is growing rapidly taking into consideration the education policy adopted by the state learn critical thinking. analyze technology. combining knowledge with science to introduce innovations into science and society receiving psychological stability. high moral quality and high quality education involvement in the upbringing of young people.

Purpose of "EduCon" Company:

- Our state offers educational centers and educational institutions based on the best practice in the world resulting in an effective program of training and preparation:

- To be able to think critically and apply knowledge in life preparation of substantial teaching aids and assignments;

- Formation of the student's qualitative knowledge and moral consciousness;

- Personality. which can work with student psychology who can master the best techniques, search engineers. qualified specialists Conducting workshops for the formation:

- Implementation of training and retraining programs:

- Organization and implementation of educational programs methodological support:

- Promotes the development of the personality and abilities of the student preparation of complex methodical instructions:

- Promoting professions in accordance with our state, education orientation and specialty of recipients compilation help manuals:

- Formation of psychological stability of students organization of training workshops on the way.

EduCon Company contracts with additional education centers and works closely with general education schools and provides services to educational centers. Currently in additional education centers in all regions of the country Special attention is paid to the preparation for the Unified National Testing.

EduCon has been teaching at additional education centers teachers' professional development, special curriculum and supplement training in the field of textbooks.

3.2 Using methods

By Decision Tree Method

Decision trees are a way of representing classification rules in a hierarchical, consistent structure. Typically, each node includes checking one independent variable. Sometimes in a tree node two independent variables are compared with each other or some function is determined from one or several variables.

The decision tree should be applied in the following cases:

- when it is necessary to study all possible elements of the topic under consideration (problems);
- when it is necessary to reveal hidden patterns in the data;

- when the achievement of short-term goals must be obtained earlier than the results of the entire work

Decision tree training refers to a class with a teacher. that is. a training and test sample contains a classified set of examples. The evaluation function used by the CART algorithm is based on the intuitive idea of reducing sewage (uncertainty) in a node [3].
Consider the problem with two classes and a node with 50 examples of one class. Node has maximum uncleanness. If a partition is found that breaks the data into two subgroups of 40: 5 examples in one and 10:45 in another. then intuitively "impurity" will decrease. It will completely disappear when it is found. the split that will create subgroups of 50: 0 and 0:50. In the CART algorithm. the idea of "sewage" is formalized in Gini index. If the data set T contains data of n classes. then the index of G is defined as [4]:

$$G(T) = 1 - \sum_{i=1}^{n} p_2^i \ (1)$$

$$G_split = N - (\frac{1}{L}\sum_{i=1}^{n} i^2 + \frac{1}{R}\sum_{i=1}^{n} r^2) \rightarrow min \ (2)$$

$$G_split = \frac{1}{L}\sum_{i=1}^{n} i^2 + \frac{1}{R}\sum_{i=1}^{n} r^2 \rightarrow max \ (3)$$

In accordance with the tasks of machine learning that have A data table is not considered to be a sample of some general population to myself. To do this. you need to build a decision function for this sample. not caring how it will work for new ones objects. This formulation is often used in cluster problems. analysis.

In these cases. an empirical (selective) function is specified. national quality. Often it is chosen heuristically - for example measures as a reflection of ideas about "good" clustering.

The mathematical problem is to minimize the functional quality.

Such a statement of the problem is not statistical (the probability but can be associated with it. Indeed. to minimize By measuring the sample analogue of the criterion. we to some extent Nimiziruem criterion itself. For example. the method that minimizes empirical risk significantly minimizes the risk. However. this relationship is not straightforward. At a minimum. you need to take into

account var complexity of the solution.

Since the method (algorithm) of constructing the decision function there is a function (mapping of multiple samples into multiple decision-making functions), then methods, like functions, can be classified cite explicit and implicit.

Such a classification makes sense, but it is more common important and important is the division of methods into methods nive distributions, and methods of direct construction decision functions.

The methods are based on first estimating the distribution and then build a crucial function, substituting the estimate instead of distribution.

Traditionally, in applied statistics it is customary to divide rametric and nonparametric methods.

In parametric methods, it is assumed that the distribution belongs to a given parametric family, for example, the class of normal distributions. In that In case of restoring distribution, it suffices to estimate options.

A special case of the parametric approach is the bayer sovskiy approach. In this case, the parameters postulate non- which is the prior distribution and by the Bayes formula a posteriori (subject to sampling) distribution, according to which can build a solution.

This class of methods is based on a non-parametric estimate distribution. Example - parson density estimates, which can be considered as a generalization of one of the basic density estimation methods - using histogram

An example of a method from this class can serve as minant Fisher, which is analytically expressed as explicit sampling function. This category also includes metric methods the structure of the crucial function, in particular, considered earlier precedent method.

A large class of methods involves optimizing some quality functional (as a rule, empirical risk) in some class of decision functions, for example, the support method vectors, decision trees, neural networks.

The idea of the support vector method is to construct separating surface (in the simplest case - linear), located at the maximum distance from the nearest quotients of shared classes.

Neural functions are used to build decision functions, direct distribution networks that constitute the perposition of the basic functions (neurons). Base func-

tions are usually selected as superposition, linear combination of arguments and the so-called transfer function (nonlinear). The resulting value of the direct distribution neural network Sagging is a unique function of input values. Another class of neural networks is "memory" networks.

Currently, there are a lot of studies conducted in the field of educational camps as what influences to effective conveying of knowledge to participants. In one of those studies it was investigated that asking reflective questions during preparation sessions in camps is very important in order to enhance the level of understanding of materials across participants of camps. [5]

Also, constructive approach to studies between participants have been considered as an important part in organization of camps, and can be manifested in the principles as teacher should be a facilitator of learners, not instructor; materials should be relevant to learning (they should be aligned with the purpose of learning of students); teaching should be conducted according to explanation of multiple perspectives on the provided material, whereas in order to understand some theory or term, instructor should show the explanation from different perspectives in order to convey to others multidimensional view of the taught subject. [6]

Camps have also been conducted in the field of robotics. In one of studies, critical design issues have been deeply explored for educational robotics camps and some observations have been generated on the basis of the study as instruction strategies should be implemented from simple to complex, project studies should be highly encouraged due to enjoyability across the students of the camp, group size should be suitable for fair distribution of work between group members. [7]

In the study about the impact of cheering in classrooms on the camp indicated that it has positive influence on perception of technology across participants. [8]

Machine learning seems to be indispensable tool in solving variety of tasks. In this paper we used one of its algorithms - decision trees in order to predict the feedback grade of a participant to the camp on the basis of input data about each student of EduCon camp regarding his personal information, favourite activities, and results of exams that have been conducted for 4 months. So each row corresponds to a participant, which has some information as his/her city, results of exams, favourite activities in a textual form.

Decision trees, for the last times, have been applied in variety of contexts and situations as in one of those studies test and training data sets have been made from two types of geographical areas and two types of sensors - multispectral Landsat ETM - and hyperspectral DAIS, which were used to measure the effectiveness of univariable and multivariable decision trees in the task of classification of land cover. [9]

Also, decision trees have been vastly used in the field of natural language processing systems. In one of such studies part-of-speech tagging have been implemented using decision trees given small training data, which achieved remarkable accuracy. [10] As to large datasets, decision trees also proved that they can deal with large amounts of data with right methodology. In one of such studies, decision trees have been trained on the basis of 1 terabyte in size labeled dataset by building them in parallel on tractable data chunks, which were subsets of original dataset, and it achieved good results according to cross-validation experiments on the dataset. [11]

Objectives There are the following objectives in this work: ∘Investigating the data by visualizing it to see some patterns and general trends useful for the study: ∘Preprocessing the data (removing outliers, changing textual representation of data into numeric, so that decision tree models will be able to work with them):
∘Dividing the data into training and testing parts:
∘Building effective decision tree model using hyperparameter optimization in order to achieve high optimal results in classification metrics such as precision, recall, f1 score, support using classification report:

∘Generate recommendation points of improvement for the camp according to analysis.

Exploratory Data Analysis

Before applying decision tree model we decided to perform exploratory data analysis to understand the general patterns and trends of data.

Firstly, we visually represented distribution of feedback grade for teaching organization part using histogram. Here, we can clearly see that there is no grade

less than 3. so generally everyone liked this organizational part. Majority of people voted 5. and tentatively of them voted 4. and only minority voted 3. which are less than 5 people according to the y axis.

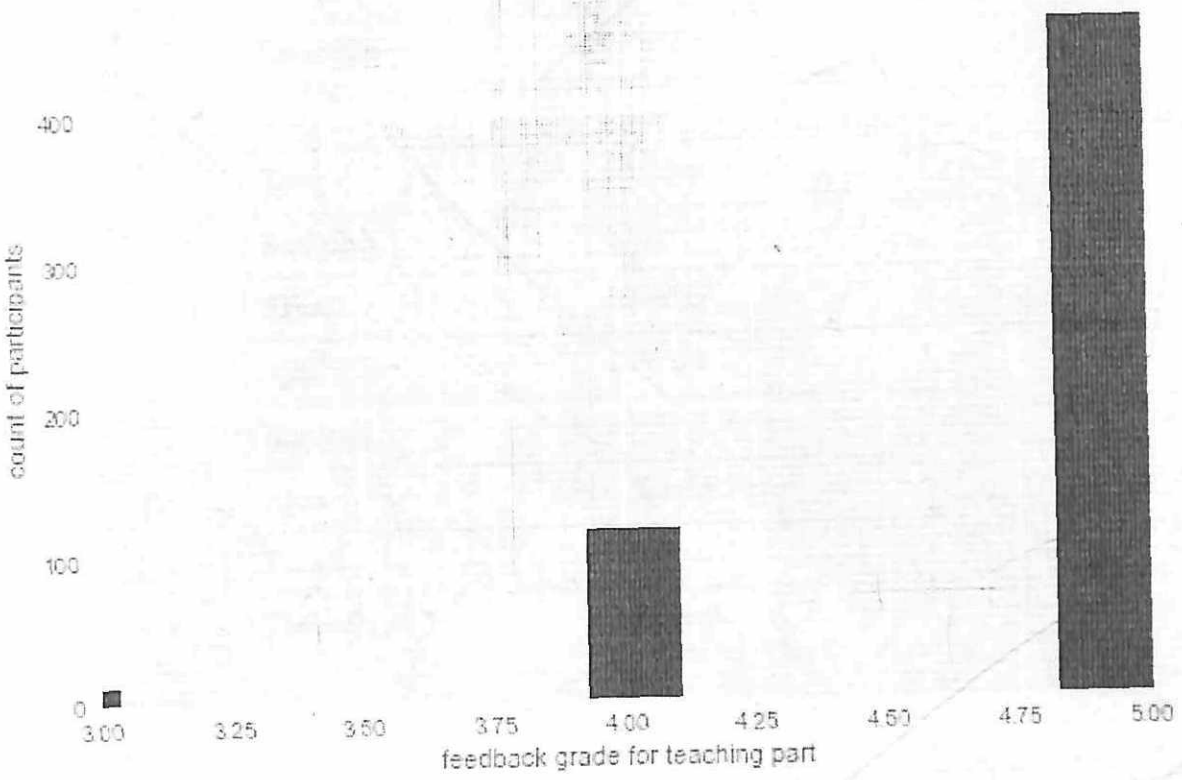


Figure 3.3: Feedback for grade teaching part

Then we decided to look at the relationship between target variable (feedback grade about teaching system) and cities. Given that majority voted 5. we can see that the outlier here is Shymkent city with 1.61 mean of voting.

Here, the mean of Semey city is 5, which is absolute maximum, that also seems to be strange and somehow "weird". Then, we decided to visually represent this data to see the general trend (we decided to exclude some cities, because their values are close to cities after Shymkent city).

Figure 3.6: Feedback of cities teaching grade

City	Mean of Feedback grade for Teaching part
Shymkent	4.614286
Atyrau	4.642857
Nursultan	4.650602
Kokshetau	4.700000
Oral	4.714286
Taraz	4.732143
Karaganda	4.733333
Aktau	4.740741
Aktobe	4.780488
Kostanay	4.785714
Almaty	4.789474
Pavlodar	4.800000
Turkestan	4.800000

Figure 3.4: Mean of Feedback grade for teaching part

So as we see, other remaining cities are very similar regarding the feedback grade.

Kyzylorda	4.809524
Taldykorgan	4.833333
Oskemen	4.852941
Semey	5.000000

Figure 3.5:

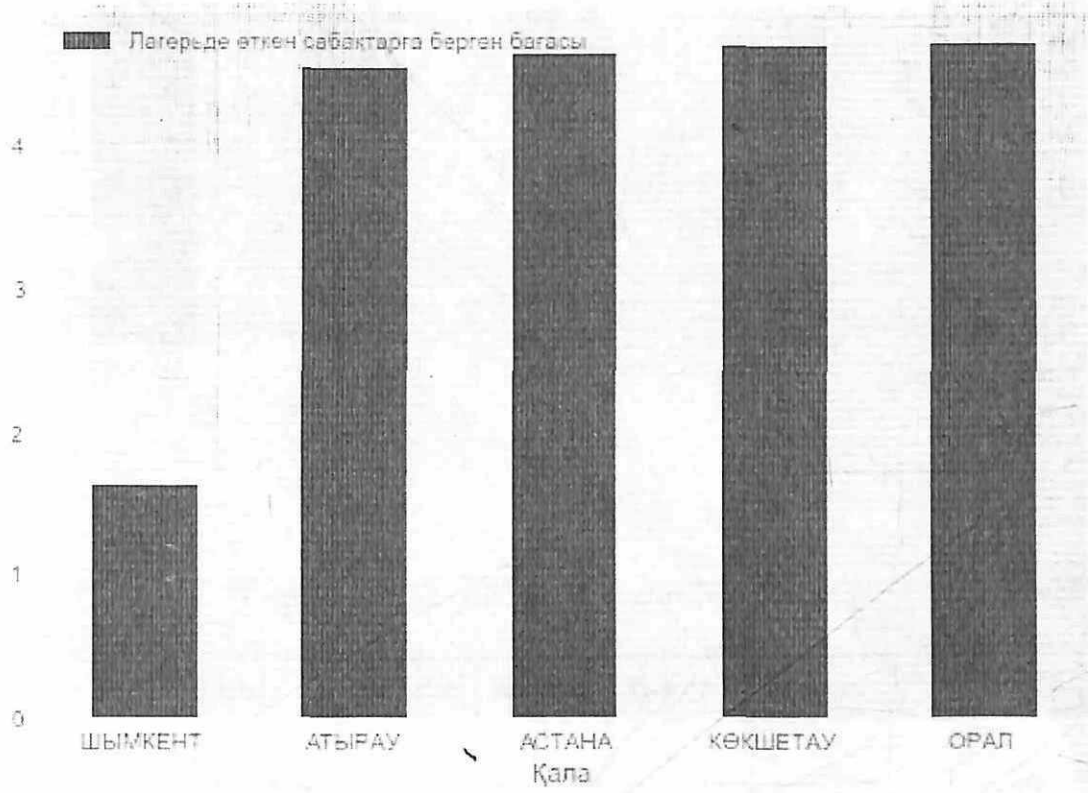
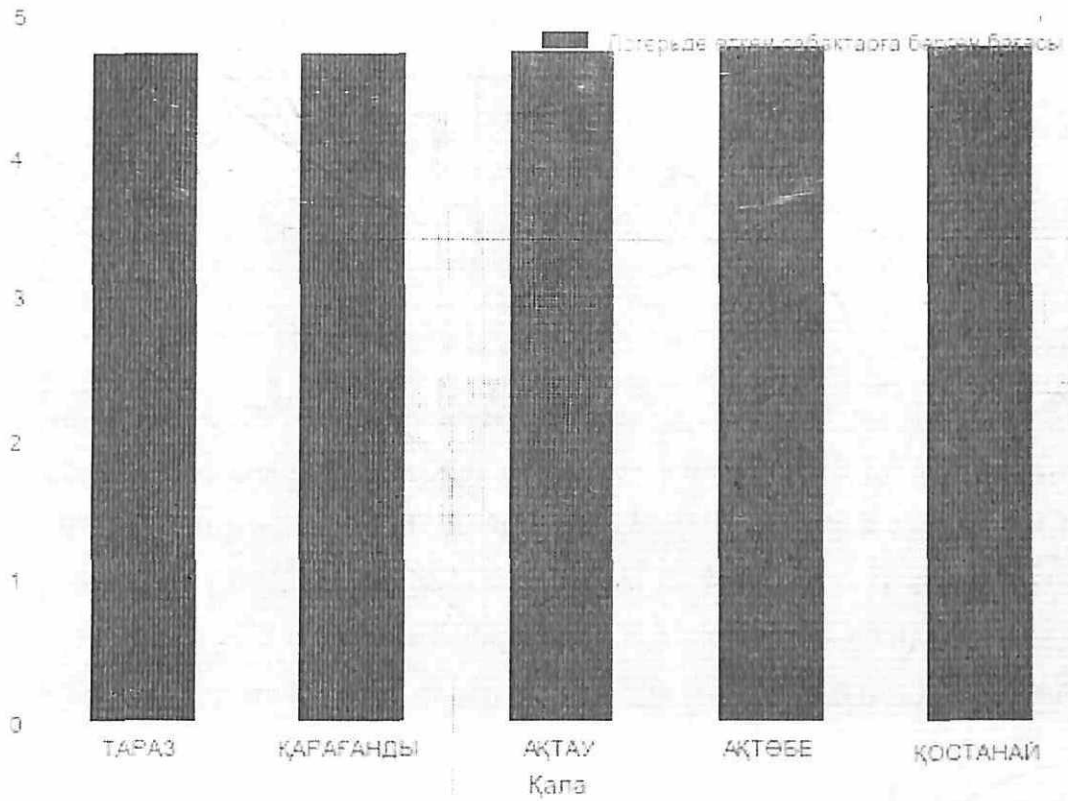


Figure 3.7: Feedback of cities teaching grade

I used variety of decision tree models in terms of changing its parameters to train and test the models in order to predict the feedback grade of a student about the camp teaching organizational part. It varies between 0 to 5.

Decision Tree is a machine learning algorithm, which works by building a tree, where the root corresponds to the feature with the highest information gain, and leaves correspond to the target value, whereas the path from the leaf till the root can be considered as questions that the model asks in order to predict target variable for some provided input data. I divided the data into training and testing part, where 67% of dataset have been allocated to training part, whereas remaining 33% to the testing phase.

In our first attempt, these are the results we achieved according to classification



report:

Figure 3.8: Decision tree models results

Class	Precision	Recall	f1-score	Support
3	1.00	0.25	0.40	4
4	0.28	0.20	0.23	35
5	0.82	0.89	0.85	155

I decided to exclude 0 value, because actually no one gave 0 for the camp, but it was just replacement variable for nulls. There was also almost no 1 and 2 grades, so that we did not consider them in the classification report. In order to deal with the features, which values have been represented in a textual manner, we used pandas dummy variables. After that, we decided to start hyperparameter optimization phase. The first parameter which we optimized is maximum depth of our decision tree model. Depth term refers to number of edges from the root node (which was splitted according to methodology of selection the feature with the highest information gain) till the leaf, and these are the results which we achieved in this case :

Figure 3.9: Decision tree models results

Class	Precision	Recall	f1-score	Support
3	0.00	0.00	0.00	4
4	0.22	0.23	0.22	35
5	0.82	0.83	0.82	155

In comparison with the previous results, we can see that for the class 3 results did not change, because the number of people who voted 3 for the organization of study in the camp is very low in comparison with classes 4 and 5. As to class 4 we can see that precision decreased a little bit, recall and f1 score increased a little bit. As to class 5 only recall decreased a little bit. It clearly shows that in this case optimizing maximum depth of the tree is not essential, because it did not lead to some extreme changes.
As to the best value for maximum depth, we found that it is 7 using grid search methodology.

The next interesting hyperparameter which we wanted to optimize was the minimum number of splits, which refers to the minimum number of samples of data which are needed to split an internal node of the decision tree model. These are the results, which we achieved :

Figure 3.10: Decision tree models results

Class	Precision	Recall	f1-score	Support
3	0.00	0.00	0.00	4
4	0.25	0.03	0.05	35
5	0.80	0.98	0.88	155

I can see that recall of class 5 have been drastically increased from 0.83 to 0.98, as to class 3 and 4 these are of less importance, because the specificity of the dataset is that almost all participants gave 5 vote, which diminishes the importance of other grades due to their very few number in the data. We found that 192 is the most optimum number of splits for this dataset. Now, we thought what

if we perform grid search on the intersection of maximum depth and minimum samples for splitting. These are the results we achieved :

Figure 3.11: Decision tree models results

Class	Precision	Recall	f1-score	Support
3	0.00	0.00	0.00	4
4	0.25	0.03	0.05	35
5	0.80	0.98	0.88	155

and see that precision, recall, and f1-score for grades less than 5 have been equalized to zero due to their unimportance, which was described in the previous analysis. As to metrics for class 5, some of them have been increased, some of them have been decreased. These are the most optimal parameters for maximum depth and number of samples for splitting according to intersectional grid search : max depth: 3, min samples split: 192

Class	Precision	Recall	f1-score	Support
3	0.00	0.00	0.00	4
4	0.00	0.00	0.00	35
5	0.79	0.99	0.88	155

Figure 3.12: Decision tree models results

3.3 By linear regression

Direct relapse gets to the core of measurements: Given a lot of information focuses, what is the relationship of the information close by to information yet observed? By what means should data from one informational index engender to other information? Direct relapse offers the accompanying model to address this inquiry.

$$E(Y|X = x) \approx ax + b \ (4)$$

That is, given explicit qualities for X, accept that the restrictive desire is a direct capacity of those particular qualities. In any case, in light of the fact that

the watched qualities are not simply the desires. the model suits this with an added substance clamor term. At the end of the day. the watched variable (a.k.a. reaction. target. subordinate variable) is demonstrated as.

$$E(Y|X = x_i) + \epsilon_i \approx ax + b + \epsilon_i = y \ (5)$$

where $E(\epsilon_i) = 0$ and the ϵ_i are iid and where the distribution function of ϵ_i relies upon the issue, despite the fact that it is frequently accepted Gaussian. The $X = x$ qualities are known as free factors. covariates. or regressors. How about we check whether we can utilize the majority of the strategies we have grown so far to comprehend this type of relapse. The principal assignment is to decide how to assess the obscure linear parameters. a and b. To make this concrete. let's assume that $\epsilon \sim \mathbb{N}(0, \sigma^2)$. Bear in mind that

$$E(Y|X = x) \ (6)$$

is a deterministic function of x. In other words. the variable x changes with each draw. but after the data have been collected these are no longer random quantities.Thus. for fixed x. y is a random variable generated by ϵ . Perhaps we should denote ϵ as ϵ_x to emphasize this. but because ϵ is an independent. identically-distributed (iid) random variable at each fixed x. this would be excessive. Because of Gaussian additive noise. the distribution of y is completely characterized by its mean and variance.

$$E = ax + y \ (7)$$
$$V = \sigma^2$$

Utilizing the greatest probability methodology, we work out the log-probability work as

$$\mathcal{L}(a, b) = \sum_{i=1}^{n} \log \mathcal{N}(ax_i + b, \sigma^2) \propto \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Figure 3.13:

Note that we smothered the terms that are irrelevent to themaximum-finding. Taking

$$\frac{\partial \mathcal{L}(a,b)}{\partial a} = 2\sum_{i=1}^{n} x_i(b + ax_i - y_i) = 0$$

Figure 3.14:

the subsidiary of this concerning a gives the accompanying condition: The going with code replicates a couple of data and usages Numpy devices to enroll the parameters as showed up.

```
>>> import numpy as np
>>> a = 6;b = 1 # parameters to estimate
>>> x = np.linspace(0,1,100)
>>> y = a*x + np.random.randn(len(x))+b
>>> p,var_=np.polyfit(x,y,1,cov=True) # fit data to line
>>> y_ = np.polyval(p,x) # estimated by linear regression
```

Figure 3.15:

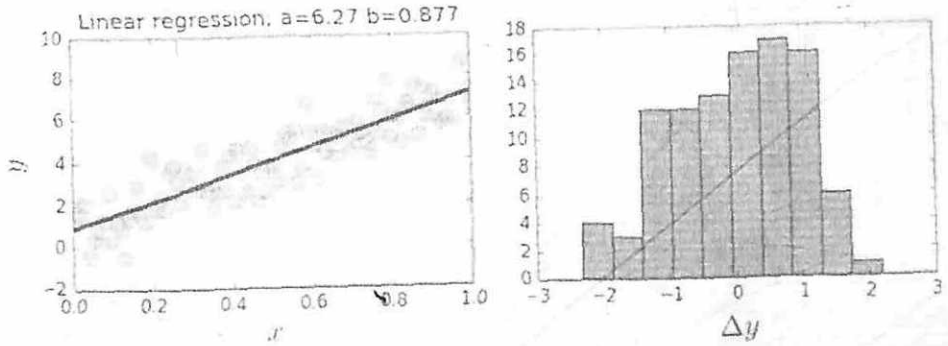


Figure 3.16:

The chart on the left of Fig.1 demonstrates the relapse line plotted against the information.

The evaluated parameters are noted in the title. The histogram on the privilege of Fig. 3.12

demonstrates the lingering blunders in the model. It is dependably a smart thought to examine the residuals

of any relapse for ordinariness. These are the differences between the fitted line for each x_i value and the corresponding y_i value in the data. Note that the x term does

not need to be consistently monotone.

To decouple the deterministic variety from the irregular variety, we can fix the

file and compose separate issues of the structure

$$y_i = ax_i + b + \epsilon_i \ (7)$$

where $E(\epsilon_i) = 0$. What could we do with just this one component of the problem? In other words, suppose we had m-samples of this component as in

$$y_i.\, k_m^{k=1} \ (8)$$

Following the usual procedure. we could obtain estimates of the mean of y_i as

$$y_i = \frac{1}{m} \sum_m^{k=1} y_{i,k} \ (9)$$

In any case, this discloses to us nothing roughly the individual parameters an and b since they are not recognizable inside the terms that are processed, to be explicit, we may have

$$E(y_i) = ax_i + b \ (10)$$

yet, we still just have one condition and the two questions, an and b. What about in the event that we

consider and fix another part j as in

$$y_j = ax_j + b + \epsilon_i \ (11)$$

Then, we have

$$E(y_j) = ax_j + b \ (12)$$

so at smallest directly we have two conditions and two inquiries and we realize how to assess the got out hand sides of these conditions from the data using the estimators y_i and y_j . We should perceive how this functions inside the code test underneath. **Programming Tip**

```
>>> x0, xn =x[0],x[50]
>>> # generate synthetic data
>>> y_0 = a*x0 + np.random.randn(20)+b
>>> y_1 = a*xn + np.random.randn(20)+b
>>> # mean along sample dimension
>>> yhat = np.array([y_0,y_1]).mean(axis=1)
>>> a_,b_=np.linalg.solve(np.array([[x0,1],
...                                 [xn,1]]),yhat)
```

Figure 3.17:

The earlier code utilizes the illuminate work in the Numpy linalg module,

which contains the center direct variable based math codes in Numpy that join the

fight tried LAPACK library.

I can work out the answer for the assessed parameters for this situation where $x_0 = 0$

$$a = \frac{y_i - y_0}{x_i} \quad (13)$$

$$b = y_0$$

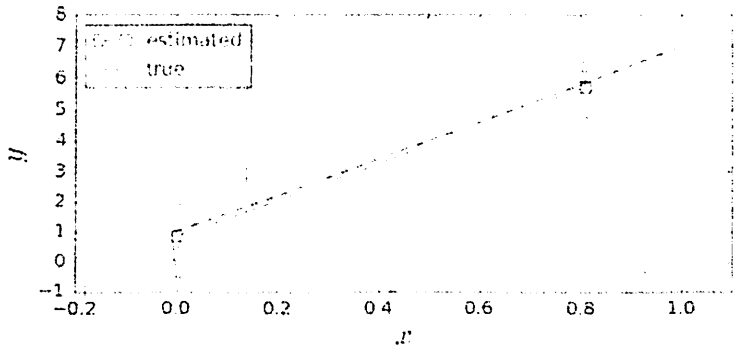


Figure 3.18:

The expectations and variances of these estimators are the following,

$$E(a) = \frac{ax_i}{x_i} = a \quad (14)$$

$$E(b) = b$$

$$V(a) = \frac{2\sigma^2}{x_2^i} \quad (15)$$

$$V(b) = \sigma^2 \quad (16)$$

he desires demonstrate that the estimators are fair-minded. The estimator a has a difference

that diminishes as bigger focuses x_i are chosen. That is. it is smarter to have tests

farther along the even hub for fitting the line. This change evaluates the influence of those far off focuses.

Relapse From Projection Methods. We should check whether we can apply our insight into

projection techniques to the general case. In vector documentation, we can compose the accompanying:

$$y = ax + b1 + \epsilon \quad (17)$$

where 1 is the vector of all ones. Let's use the inner-product notation.

$$< x, y >= E(x^T y)$$

Then, by taking the inner-product with some x1 1 we obtain. $x_1 \in 1^{\perp}$ then we get the MMSE approximation to x in the $1^{\perp}$ space. Thus, we take

$$x_1 = P_1 \perp (x)$$

Remember that $P_1 \perp$ is a projection matrix so the length of x_1 is at most x. This means that the denominator in the a equation above is really just the length of the x vector in the coordinate system of $P_1 \perp$. Because the projection is orthogonal (namely, of minimum length), the Pythagorean theorem gives this length as the following:

$$< x, x_1 >^2 =< x, x > - < 1, x >^2$$

The first term on the right is the length of the x vector and last term is the length of x in the coordinate system orthogonal to $P_1 \perp$, namely that of 1. We can use this geometric interpretation to understand what is going on in typical linear regression in much more detail. The fact that the denominator is the orthogonal projection of x tells us that the choice of x_1 has the strongest effect (i.e., largest value) on reducing the variance of a. That is, the more x is aligned with 1, the worse the variance of a. This makes intuitive sense because the closer x is to 1, the more constant it is, and we have already seen from our one-dimensional example that distance between the x terms pays off in reduced variance. We already know that a is an unbiased estimator and because we chose x_1 deliberately as a projection, we know that it is also of minimum variance. Such estimators are known as Minimum-Variance Unbiased Estimators (MVUE). I can write $x - 1$ as the following:

$$x_1 = x - P_1 x$$

where P_1 is projection matrix of x onto the 1 vector. Using this, the numerator of a becomes

$$< y, x_1 >=< y, x > - < y, P_1 x >$$

Note that,

$$P_1 = 11^T \frac{1}{n}$$

so that writing this out explicitly gives

$$< y, P_1 x >= (y^T 1)(1^T x)/n = (\sum y_i)(\sum x_i)/n$$

$$< x, P_1 x >= (x^T 1)(1^T x)/n = (\sum x_i)(\sum x_i)/n$$

So, stopping all of this together gives the taking after,

$$a = \frac{x^T y - (\sum x_i)(\sum y_i)/n}{x^T - (\sum x_i)^2 \ n}$$

with corresponding variance,

$$V(a) = \sigma^2 \frac{\|x_1\|}{<x, x_1>^2} = \frac{(\sigma)^2}{(\|x_i\|)^2 - n(x^2)}$$

Using the same approach with b gives,

$$b = \frac{<y, x->}{1.x-} =$$
$$= \frac{<y, 1 - P_x(1)>}{<1, 1 - P_i(1)>} =$$
$$= \frac{x^T x(\sum y_i)/n \quad x^T y(\sum x_i) \ n}{x^T x - (\sum x_i)^2/n}$$

Python Machine Learning Modules

Python provides many bindings for machine learning libraries, some specialized for technologies such as neural networks, and others geared towards novice users. For our discussion, we focus on the powerful and popular Scikit-learn module. Scikit-learn is distinguished by its consistent and sensible API, its wealth of machine learning algorithms, its clear documentation, and its readily available datasets that make it easy to follow along with the online documentation. Like Pandas, Scikit-learn relies on Numpy for numerical arrays. Since its release in 2007, Scikit-learn has become the most widely-used, general-purpose, open-source machine learning modules that is popular in both industry and academia.

Figure 3.19:

As with all of the Python modules we use, Scikit-learn is available on all the major platforms.

To get started, let's revisit the familiar ground of linear regression using Scikit-learn. First, let's create some data. I next import and create an instance of the Linear Regression class from Scikit-learn. Scikit-learn has a wonderfully consistent

Figure 3.20:

API. All Scikit-learn objects use the fit method to compute model parameters and the predict method to evaluate the model. For the Linear Regression instance, the fit method computes the coefficients of the linear fit. This method requires a matrix of inputs where the rows are the samples and the columns are the features. The target of the regression are the Y values, which must be correspondingly shaped, as in the following.

Figure 3.21:

The coef property of the linear regression object shows the estimated parameters for the fit. The convention is to denote estimated parameters with a trailing underscore. The model has a score method that computes the R^2 value for the regression. Recall from our statistics part that the R^2 value is an indicator of the quality of the fit and varies between zero (bad fit) and one (perfect fit). Now, that we have this fitted, I can evaluate the fit using the predict method

Figure 3.23:

Scikit-learn module can easily perform basic linear regression. The circles show the training data and the fitted line is shown in black

Figure 3.22:

Linear regression is used for data analysis and machine learning. Build your model in Python and get the first results!

I take the students of EduCon company and understand how the value of points depends on other data, such as monthly points, level of education, role, city in which they study, and so on.
Regression solves the problem of a single presentation of analysis data for each student. Moreover, points, education, role and city are independent variables with a total average score depending on them.
Regression considers a certain phenomenon and a series of observations. Each observation has two or more variables. Assuming that one variable is dependent on others, you are trying to build relationships between them.

In other words, you need to find a function that displays the dependence of some variables or data on others.

Dependent data is called dependent variables, outputs, or responses.

Independent data is called independent variables, inputs or predictors.

Usually there is one continuous and unlimited dependent variable in the regression. Input variables can be unlimited, discrete or categorical data, such as gender, nationality, brand, etc.

The linear relationship may be direct or inverse. If you think that with an increase in air temperature, the number of accidents increases, then this is a positive

relationship and a direct correlation. Another way to describe this same direct relationship is to say that the number of accidents decreases with decreasing air temperature.

Conversely, if you believe that the number of crimes decreases with an increase in the number of police officers who patrol this territory, the relationship is reversed. You can describe this inverse relationship and so - the number of crimes increases with a decrease in the number of police officers patrolling this territory. In the illustration, that, in addition to the direct and inverse relationship, there may not be any significant correlation at all.

The correlation analysis and graphs of the interrelation of phenomena show how strongly two phenomena depend on each other. Regression analysis, in turn, allows you to get even more information about the relationship of phenomena. This analysis allows you to show the degree (influence) with which one or more variables can potentially cause a positive or negative change in another variable.

It is common practice to designate output data - y, input data - x. In the case of two or more independent variables, they can be represented as a vector $x = (x_1, \ldots . x_r)$, where r is the number of input variables.

One-step regression is one of the most important and widely used regression techniques. This is the easiest regression method. One of its advantages is the ease of interpretation of the results.

$$E(Y|X = x) \approx ax + b$$

$$E(Y|X = x_i) + \epsilon_i \approx ax + b + \epsilon_i = y$$

$$E(Y|X = x_i) + \epsilon_i \approx ax + b + \epsilon_i = y$$

$$E(Y|X = x_i) + \epsilon_i \approx ax + b + \epsilon_i = y$$

$$y_i = ax_i + b + \epsilon_i$$

$$y_i, k_m^{k-1}$$

Following the usual procedure, we could obtain estimates of the mean of y_i as

$$y_i = \frac{1}{m} \sum_{m}^{k-1} y_{i \cdot k}$$

And by using linear regression methods in Machine learning:

```
data = data.fillna(0)


mean = 0

mean_list = []
counter_list = []

for j in range(744):
    n_list = []
    counter = 0
    for i in range(1,3):
        if int(data['res'+str(i)][j]) !=0:
            counter += 1
        n_list.append(int(data['res'+str(i)][j]))
    try:
        mean = sum(n_list)/counter
        counter_list.append(counter)
        mean_list.append(mean)
        del counter
        del n_list
    except:
        mean = 0
        mean_list.append(mean)
        counter_list.append(counter)
        del counter
        del n_list
```

Figure 3.24: Linear regression results

Huge information gives enormous chances to factual deduction, however maybe much greater difficulties, particularly when contrasted with the examination of painstakingly gathered, normally littler, arrangements of information. Logically, the program underlined the jobs of measurements, software engineering, and arithmetic in getting logical knowledge from huge information. Two correlative strands were presented: cross-cutting, or primary, inquire about that supports examination, and space explicit research concentrated on specific application regions. The previous class included AI, factual induction, enhancement, arrange investigation, and representation. Theme explicit workshops tended to issues in wellbeing strategy, social approach, ecological science, digital security and interpersonal organizations. These divisions are not inflexible, obviously, as central and application regions are a piece of an input cycle where each motivates advancements

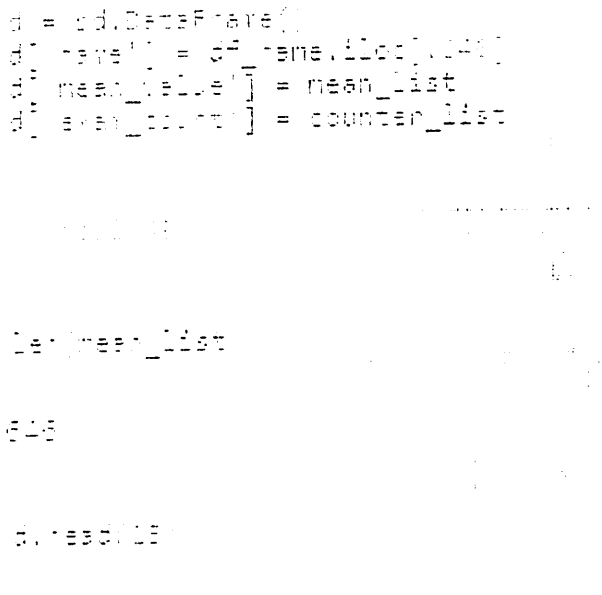


Figure 3.25: Linear regression results

in the other. Some very importantapplication zones huge information is crucial were not ready to be the subject of centered workshops. yet a significant number of these applications featured in individual introductions. Currently, there are a lot of studies conducted in the field of educational camps as what influences to effective conveying of knowledge to participants. In one of those studies it was investigated that asking reflective questions during preparation sessions in camps is very important in order to enhance the level of understanding of materials across participants of camps. Also, constructive approach to studies between participants have been considered as an important part in organization of camps, and can be manifested in the principles as teacher should be a facilitator of learners, not instructor; materials should be relevant to learning (they should be aligned with the purpose of learning of students); teaching should be conducted according to explanation of multiple perspectives on the provided material, whereas in order to understand some theory or term, instructor should show the explanation from different perspectives in order to convey to others multidimensional view of the taught subject. Camps have also been conducted in the field of robotics. In one of studies, critical design issues have been deeply explored for educational robotics camps and some observations have been generated on the basis of the study as instruction strategies should be implemented from simple to complex, project studies should be highly encouraged due to enjoyability across the students of the camp, group size should be suitable for fair distribution of work between

	name	mean_value	exam_count
0	Лаззат Абайқызы	72.000000	3
1	Қалдыбек Абдуллаев	86.750000	4
2	Шереке Амантаев	71.000000	1
3	Жасын Арапбаев	96.000000	1
4	Дана Ахметова	99.000000	4
5	Дінмұхаммед Бақтыбай	81.000000	2
6	Салтанат Бексарлеева	78.000000	3
7	Лаззат Бисенова	68.000000	1
8	Анар Газизова	85.000000	2
9	Айбота Ерлбаева	68.000000	1
10	Амина Ермакова	82.000000	3
11	Медина Әбдіхалық	67.000000	3
12	Дінмұхаммед Жеңгірхан	86.250000	1
13	Лаура Жомартова	44.750000	4
14	Диана Жұманқанова	64.250000	1

Figure 3.26: Linear regression models results

group members. In the study about the impact of cheering in classrooms on the camp indicated that it has positive influence on perception of technology across participants. Machine learning seems to be indispensable tool in solving variety of tasks. In this paper we used one of its algorithms - decision trees in order to predict the feedback grade of a participant to the camp on the basis of input data about each student of EduCon camp regarding his personal information, favourite activities, and results of exams that have been conducted for 4 months. So each row corresponds to a participant, which has some information as his/her city, results of exams, favourite activities in a textual form. Decision trees, for the last times, have been applied in variety of contexts and situations as in one of those studies test and training data sets have been made from two types of geographical

```
[illegible]
    [illegible]
```

```
0 0.0
1 78.26517571884985
2 75.55833333333334
3 81.59481037924151
4 82.26190476190476
```

Figure 3.27: Linear regression models results

```
data.head()
```

	res1	res2	res3	res4	res5
0	0.0	47.0	83.0	82.0	93.0
1	33.0	33.0	48.0	33.0	85.0
2	0.0	0.0	71.0	0.0	84.0
3	0.0	0.0	96.0	0.0	84.0
4	0.0	0.0	98.0	0.0	78.0

Figure 3.28: Linear regression models results

areas and two types of sensors - multispectral Landsat ETM+ and hyperspectral DAIS, which were used to measure the effectiveness of univariable and multivariable decision trees in the task of classification of land cover. Also, decision trees have been vastly used in the field of natural language processing systems. In one of such studies part-of-speech tagging have been implemented using decision trees given small training data, which achieved remarkable accuracy. [6] As to large datasets, decision trees also proved that they can deal with large amounts of data with right methodology. In one of such studies, decision trees have been trained on the basis of 1 terabyte in size labeled dataset by building them in parallel on tractable data chunks, which were subsets of original dataset, and it achieved good results according to cross-validation experiments on the dataset.

4. Conclusion

I investigated the dataset of educational camp across participants and highlighted the necessary points of improvement according to exploratory data analysis as checking the state of some cities whether they have some problems or not. and developing more sophisticated measures for capturing the feedback grade about teaching part. Also. we trained successfully decision tree model to predict feedback grade of participants about teaching organization part and achieved high optimal results.

5. Acknowledgements

ACKNOWLEDGMENT I acknowledge EduCon and Dostyk educational organizations for providing us with data to perform the analysis and apply decision tree models on it. Thanks to my thesis supervisor M.Zhapparov for constant support and useful discussion.

6. References

[1]Aivazian S.A..Enuykov I.S.. Meshalkin L.D. Applied statistics: basics of modeling and primary data processing. M.: Finance and statistic. 1983. 471p.

[2]Flakh P. Machine learning. M.:DMK press. 2015. 400p

[3]Mikheev M. Yu. Situational-structural approach to the analysis of information objects , M. Yu. Mikheev. A. B. Scherban News of higher educational institutions. Volga region. -2006. -№ 6. -C. 128-135. -(Technical science).

[4]Mikheev M.Yu. Using decision support systems for modeling professional activities of a specialist in the process of practice-oriented learning Mikheev M.Yu., Novikov AV. Symochkina I.Yu. Proceedings of the International Symposium "Reliability and Quality", Penza State University. - 2009. - p. 174-175

[5] Bischoff bischopj@oneonta.edu. P. J. ., Castendyk. D., Gallagher. H.. Schaumloffel. J.. Labroo. S.. Bischoff, P. J. (2008). A science summer camp as an effective way to recruit high school students to major in the physical sciences and science education. International Journal of Environmental Science Education. 3(3), 131–141. Retrieved from http: /search.ebscohost.com/ login.aspx?direct=truedb= 05392-005site ehost-live

[6] Doolittle. P. E.. Camp. W. G. (1999). Electronic Journals Journal of Vocational and Technical Education CONSTRUCTIVISM: THE CAREER AND TECHNICAL EDUCATION PERSPECTIVE. Journal of Vocational and Technical Education. 16(1). 23–16. https: /doi.org/0021-9150(74)90027-6 [pii]

[7] Ucgul. M.. Cagiltay. K. (2014). Design and development issues for educational robotics training camps. International Journal of Technology and Design

Education. 24(2). 203–222. https://doi.org 10.1007 s10798-013-9253-9

[8] Jacobs-Rose. C.. Harris. K. (2010). Educational camps and their effects on female perceptions of technology programs. Journal of Industrial Teacher Education; 47(1), 11–41. Retrieved from http: /scholar.lib.vt.edu ejournals JITE v47n1 ros

[9] Pal. M.. Mather. P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sensing of Environment, 86(4), 554–565. https: doi.org/10.1016/S0034-4257(03)00132-9

[10] Màrquez. L.. Rodríguez, H. (2005). Part-of-speech tagging using decision trees. 25–36. https: doi.org/10.1007/bfb0026668

[11] Hall, L. O., Chawla, N., Bowyer, K. W., Ave, E. F. (n.d.). Decision Tree Learning on Very Large Data Sets