TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
Отделение школы (НОЦ) Информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Titanic passenger data analysis using cluster analysis **(Анализ данных пассажиров Титаника с использованием кластерного анализа)**

УДК 004.65:004.451:519.23:656.085.3

Студент

Группа	ФИО	Подпись	Дата
8ПМ0И	Чжан Цзифэн		20.06.2022 г.

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф.-м.н.		20.06.2022 г.

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОСГН ШБИП	Меньшикова Е. В.	к.ф.н.		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ООД ШБИП	Антоневич О. А.	к.б.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Савельев А.О.	к.т.н.		

Томск – 2022

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОСВОЕНИЯ ООП
по направлению 09.04.04 «Программная инженерия»

Код компетенции	Наименование компетенции
Универсальные компетенции	
УК(У)-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий
УК(У)-2	Способен управлять проектом на всех этапах его жизненного цикла
УК(У)-3	Способен организовывать и руководить работой команды, вырабатывая командную стратегию для достижения поставленной цели
УК(У)-4	Способен применять современные коммуникативные технологии, в том числе на иностранном (-ых) языке (-ах), для академического и профессионального взаимодействия
УК(У)-5	Способен анализировать и учитывать разнообразие культур в процессе межкультурного взаимодействия
УК(У)-6	Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки
Общепрофессиональные компетенции	
ОПК(У)-1	Способен самостоятельно приобретать, развивать и применять математические, естественно-научные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте
ОПК(У)-2	Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач
ОПК(У)-3	Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями
ОПК(У)-4	Способен применять на практике новые научные принципы и методы исследований

ОПК(У)-5	Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем
ОПК(У)-6	Способен самостоятельно приобретать с помощью информационных технологий и использовать в практической деятельности новые знания и умения, в том числе в новых областях знаний, непосредственно не связанных со сферой деятельности
ОПК(У)-7	Способен применять при решении профессиональных задач методы и средства получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе, в глобальных компьютерных сетях
ОПК(У)-8	Способен осуществлять эффективное управление разработкой программных средств и проектов
Профессиональные компетенции	
ПК(У)-1	Способен к созданию вариантов архитектуры программного средства
ПК(У)-2	Способен разрабатывать и администрировать системы управления базам данных
ПК(У)-3	Способен управлять процессами и проектами по созданию (модификации) информационных ресурсов
ПК(У)-4	Способен проектировать и организовывать учебный процесс по образовательным программам с использованием современных образовательных технологий
ПК(У)-5	Способен осуществлять руководство разработкой комплексных проектов на всех стадиях и этапах выполнения работ

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки (специальность) 09.04.04 Программная инженерия
Отделение школы (НОЦ) Информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Савельев А.О.
(подпись) (дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Магистерской диссертации

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

Группа	ФИО
8ПМ0И	Чжан Цзифэн

Тема работы:

Titanic passenger data analysis using cluster analysis (Анализ данных пассажиров Титаника с использованием кластерного анализа)	
Утверждена приказом директора (дата, номер)	№ 145-46/с от 25.05.2022

Срок сдачи студентом выполненной работы:	15.06.2022

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе *(наименование объекта исследования или проектирования; производительность или нагрузка; режим работы (непрерывный, периодический, циклический и т. д.); вид сырья или материал изделия; требования к продукту, изделию или процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.).*	Разработка метода кластерного анализа для набора данных о пассажирах Титаника с использованием программирования Python и создайте программу для предоставления случаев и методов анализа данных для аналитиков рисков или данных.

Перечень подлежащих исследованию, проектированию и разработке вопросов *(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования, конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов, подлежащих разработке; заключение по работе).*	1. Аналитический обзор литературных источников. 2. постановка задачи исследования. 3. разработка методики. 4. реализация методики. 5. выбор программного обеспечения. 6. обсуждение результатов выполненной работы. 7. финансовый менеджмент. 8. социальная ответственность. 9. заключение
Перечень графического материала *(с точным указанием обязательных чертежей)*	1. Скриншот программы. 2. UML диаграммы.

Консультанты по разделам выпускной квалификационной работы

(с указанием разделов)

Раздел	Консультант
Основная часть	Доцент ОИТ ИШИТР, к.ф.-м.н., доцент Губин Е. И.
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Доцент ОСГН ШБИП, к.ф.н., доцент Меньшикова Е. В.
Социальная ответственность	Доцент ООД ШБИП, к.б.н., доцент Антоневич О. А.

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	1.03.2022

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф.-м.н., доцент		1.03.2022

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ0И	Чжан Цзифэн		1.03.2022

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки (специальность) 09.04.04 Программная инженерия
Уровень образования магистратура
Отделение школы (НОЦ) Информационных технологий
Период выполнения весенний семестр 2021 /2022 учебного года

Форма представления работы:

Магистерская диссертация

(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	15.06.2022

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
10.06.2022	Основная часть	70
10.06.2022	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	10
10.06.2022	Социальная ответственность	10
10.06.2022	Английский язык	10

СОСТАВИЛ:
Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф.-м.н.		

СОГЛАСОВАНО:
Руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Савельев А. О.	к.т.н.		

TASK FOR SECTION
«FINANCIAL MANAGEMENT, RESOURCE EFFICIENCY AND RESOURCE SAVING»

To the student:

Group	Full name
8PM0I	Zhang Jifeng

School	ИШИТР	**Division**	Big Data Solutions
Degree	Master	**Educational Program**	09.04.04 Software Engineering

Input data to the section «Financial management, resource efficiency and resource saving»:	
1. *Resource cost of scientific and technical research (STR): material and technical, energetic, financial and human*	– Salary costs – 95145.4 – STR budget – .172788.5
2. *Expenditure rates and expenditure standards for resources*	– Electricity costs – 5,8 rub per 1 kW
3. *Current tax system, tax rates, charges rates, discounting rates and interest rates*	– Labor tax – 27,1 %; – Overhead costs – 30%;
The list of subjects to study, design and develop:	
1. *Assessment of commercial and innovative potential of STR*	– comparative analysis with other researches in this field;
2. *Development of charter for scientific-research project*	– SWOT-analysis;
3. *Scheduling of STR management process: structure and timeline, budget, risk management*	– calculation of working hours for project; – creation of the time schedule of the project; – calculation of scientific and technical research budget;
4. *Resource efficiency*	– integral indicator of resource efficiency for the developed project.
A list of graphic material *(with list of mandatory blueprints):*	
1. *Competitiveness analysis* 2. *SWOT- analysis* 3. *Gantt chart and budget of scientific research* 4. *Assessment of resource, financial and economic efficiency of STR* 5. *Potential risks*	

Date of issue of the task for the section according to the schedule	06.05.2022

Task issued by adviser:

Position	Full name	Scientific degree, rank	Signature	Date
Associate professor	E.V. Menshikova	PhD		

The task was accepted by the student:

Group	Full name	Signature	Date
8PM0I	Zhang Jifeng		

TASK FOR CHAPTER «SOCIAL RESPONSIBILITY»

Student:

Group	Name		
8PM0I	Zhang Jifeng		
School	ИШИТР	**Division**	Big Data Solutions
Educational level	Magistracy	**Course/Specialty**	09.04.04 Software Engineering

Topic of FQW:

Titanic passenger data analysis using cluster analysis	
Initial data for the chapter «social responsibility»:	
1. Characteristics of the researched object (substance, material, device, algorithm, technique, working area)	– Base EDA to get result of effects of individual or composite features on survival rate. – Machine learning algorithm to predict survival in Titanic disaster. – Working area: desktop of TPU dormitory and PC.
List of questions to be researched, designed and developed:	
1. Legal and organizational issues of occupational safety – consider special (specific to the projected work area) law norms of labor legislation. – indicate the features of the labor legislation in relation to the specific conditions of the project.	– SP 2.4.3648-20. Sanitary and Epidemiological Requirements for Organizations of Education and Training, Recreation and Recreation of Children and Youth. – Labor Code of the Russian Federation Article 349.1. Features of labor regulation of employees of state corporations, public companies, state companies.
2. Occupational safety: 2.1. Analysis of the identified harmful and dangerous factors: the sourse of factor, the impact on human's body 2.2 Suggest measures to reduce the impact of identified harmful and dangerous factors	– Increased levels of electromagnetic radiation. – Insufficient illumination of workplace. – Excessive noise. – Increased / decreased air humidity in the workplace. – physical overload (static - long-term preservation of a certain posture). – overstrain of analyzers (vision). – Increased voltage in an electrical circuit, the closure of which can pass through the human body.
3. Environmental Safety: Influence on the atmosphere, hydrosphere, lithosphere	– Impact of the object on the lithosphere: disposal of electrical components and waste (failed PC, keyboard, mouse and luminescent lamps).
4. Emergency Safety: describe the most likely emergency situation	– Fire.
Date issue of the task for the chapter	22.02.2022

Consultant:

Post	Name	Academic degree	Date	Signature
Docent professor	Antonevich O. A	PhD		

Student:

Group	Name	Date	Signature
8PM0I	Zhang Jifeng		

Abstract

Final qualifying work 70 pages, 27 figures, 16 tables, 28 sources.

Nowadays, people have stepped into the era of big data, and machine learning has become the core technology to deal with the complex information in reality. However, a single machine learning algorithm is limited in handling specific problems. There are still many parts of cluster analysis to be studied in machine learning. Facing the huge amount of information in the risk analysis agency, insurance companies are eager to grasp the first-hand effective information and explore the law of the change of the probability that the basic information of the insured is the subject of insurance. Therefore, scholars have intensified their efforts to apply machine learning to risk analysis. Taking dataset of passengers of Titanic as an example, it is of great significance to systematically study the application of information cluster analysis in machine learning algorithms for the development of machine learning algorithms, as well as the development of risk control and technology.

Key words: data analysis, feature engineering, data cleaning, logistic regression, decision trees, random forest, K-Nearest Neighbor, cross validation, machine learning..

Publications:

Zhang Jifeng. Data preparation of the titanic dataset for training a random forest model for the purpose of survival rate prediction /E. I. Gubin, Zhang Jifeng // Молодежь и современные информационные технологии сборник трудов XIX Международной научно-практической конференции студентов, аспирантов и молодых учёных, 21-25 марта 2022 г., г. Томск: / Национальный исследовательский Томский политехнический университет, Инженерная школа информационных технологий и робототехники; ред. кол. А. Ю. Дёмин, Н. Г. Марков, В. Г. Спицын [и др.] . — Томск : Изд-во ТПУ , 2022 . — [С. 255-256]

CONTENT

List of terms and abbreviations

S/D – Survive or Death

EDA – Exploratory Data Analysis

LR – Logistic Regression

DT – Decision Trees

RF – Random Forest

KNN - K-Nearest Neighbor

AUC – Area Under the ROC curve

ROC - Receiver Operating Characteristic Curve

CV – Cross Validation

Introduction

The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew [1]. That's why the name “DieTanic”, it is a very unforgettable disaster that no one in the world can ignore. Although there were some reasons, like element of luck, it seems some kinds of people with individual features have more possibility to survive than other passengers. It took about $7.5 million to build the Titanic and it sunk under the ocean due to collision. The Titanic Dataset is a very good dataset for us to gain data analysis workflow and experience.

Classification can be seen as a process of identifying, understanding, and grouping ideas and objects into predetermined categories or "subgroups." Machine learning programs use pre-classified training data sets to classify future data sets through various algorithms. Classification algorithms in machine learning use training data to predict the likelihood that subsequent data will be classified into a category [2].

In this work, I made a literature overview to describe the step to create the survival prediction Model using python programming, progress including:

- Exploratory Data Analysis(EDA):
    1. Analysis of the features.
    2. Finding any relations or trends considering multiple features.
- Feature Engineering and Data Cleaning:
    1. Adding any few features.
    2. Removing redundant features.
    3. Converting features into suitable form for modeling.
- Predictive Modeling:
    1. Running Basic Algorithms.
    2. Cross Validation.
    3. Hyper-parameters tuning.

4. Important Features Extraction.

Dataset introduction:

There are 2 datasets in original data:

— Training data, includes whole features of dataset.
— Test data, compare and verify the final accuracy of the created model.

Table 1. Dataset basic info

Variable Name	Description	Type
PassengerID	ID	integer
Survived	Survive/Death	S/D(1/0)
Pclass	Ticket Class	interger
Name	Full Name of Passengers	string
Sex	Sex	string
Age	Age	integer
SibSp	# of siblings / spouses aboard the Titanic	integer
Parch	# of parents / children aboard the Titanic	integer
Ticket	Ticket number	string
Fare	Passenger fare	float
Cabin	Cabin number	string
Embarked	Port of Embarkation	string

In the base info of the training set we can know the meanings of each column, as well as the target variable (Survival/Die, 0/1).1 and 0 represent the state of alive and death, the ticket is divided into three grades according to the price, the sum of the siblings (including step-siblings) and spouse owned by the passenger is represented by SipSp. Similarly, parch represents the number of parents and children (including stepchildren) the passenger has. Some children travel with a babysitter, so its number can be 0. There are three boarding ports, they are: Cherbourg，Queenstown and Southampton.

1. Exploratory Data Analysis

Exploratory data analysis (EDA) is a well-established statistical tradition that provides conceptual and computational tools for discovering patterns to foster hypothesis development and refinement [3]. By understanding the dataset and the relationship between variables and predicted values, it considered an important step in the process of data mining and analysis to help us better perform feature engineering and build models later. Required tools: data science libraries (pandas, numpy, scipy), visualization libraries (matplotlib, seabon).

The difference between Exploratory Data Analysis (EDA) and Traditional Statistical Analysis (Classical Analysis): The traditional statistical analysis method usually assumes that the sample obeys a certain distribution, and then sets the data into the hypothesis model for analysis. However, because most data do not meet the assumed distribution, the results of traditional statistical analysis are often unsatisfactory. The exploratory data analysis method pays attention to the real distribution of the data and emphasizes the visualization of the data, so that the analyst can see the hidden laws in the data at a glance, so as to be inspired, so as to help the analyst find a model suitable for the data. "Exploratory" means that the analyst's understanding of the problem to be solved changes continuously as the research progresses.

After the emergence of EDA, the process of data analysis is divided into two steps: the exploration phase and the verification phase. The exploration phase focuses on discovering patterns or models contained in the data, and the validation phase focuses on evaluating the discovered patterns or models. Many machine learning algorithms (divided into training and testing) follow this idea. In data analysis, statistics could be used to observe more deeply and in detail how the data is accurately organized, and to determine the method of data analysis based on this organizational structure to obtain more information.

The goal of EDA is to discover patterns in data. The role of the data analyst is to listen to the data in as many ways as possible until a plausible "story" of the data is apparent, even if such a description would not be borne out in subsequent samples. Finch (1979) asserted that "we claim for exploratory investigation no more than that it is an activity directed toward the formation of analogy. The end of it is simply a statement that the data look as if they could reasonably be thought of in such and such a way"[4].

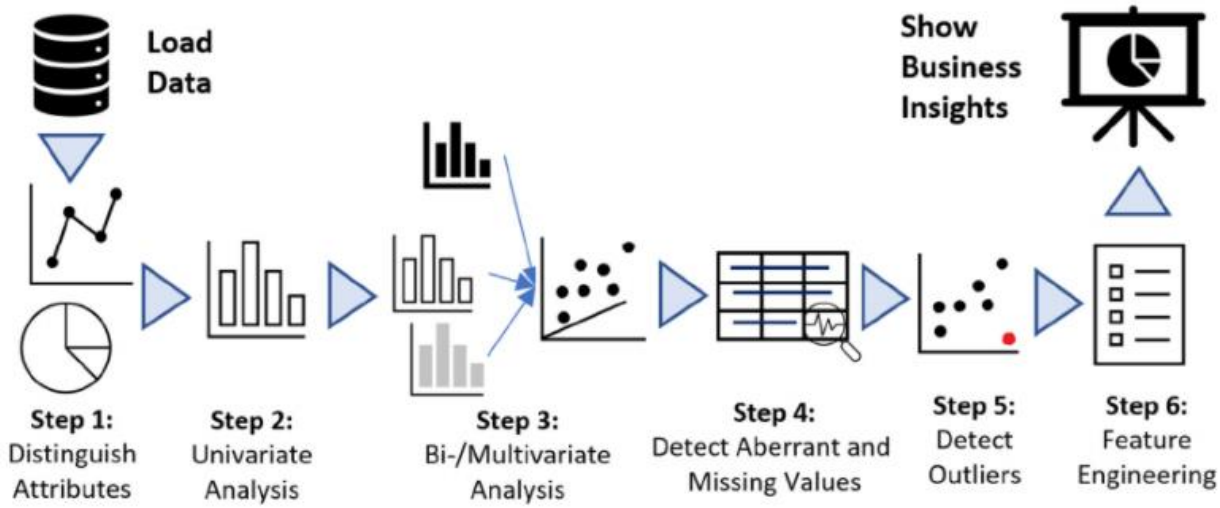


Figure 1.1. Exploratory Data Analysis (EDA) steps

EDA usually consists of six steps (see Figure 1.1) namely: (i) distinguish/identify attributes; (ii) univariate data analysis to characterize the data in the dataset; (iii) detect interactions among attributes by performing bivariate and multivariate analysis; (iv) detect and minimize impact of missing and aberrant values; (v) detect outliers (further analysis or errors); and, finally, (vi) feature engineering, where features are transformed or combined to generate new features. In our mission, EDA progress focuses on step 3 and 6.

1.1 Analysis of features

Firstly, we check out original datasets and import them.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Figure 1.1.1. Data frame of training dataset

Then, we can check the total null values.

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

Figure 1.1.2. Missing value of training dataset

The Age, Cabin and Embarked have null values. I will try to fix them. Missing data mechanisms can be divided into three categories: 1) missing completely at random (MCAR), 2) missing at random (MAR), 3) Missing not at random (MNAR) [5], for our task, the type of missing data is MNAR, we will fix them step by step.

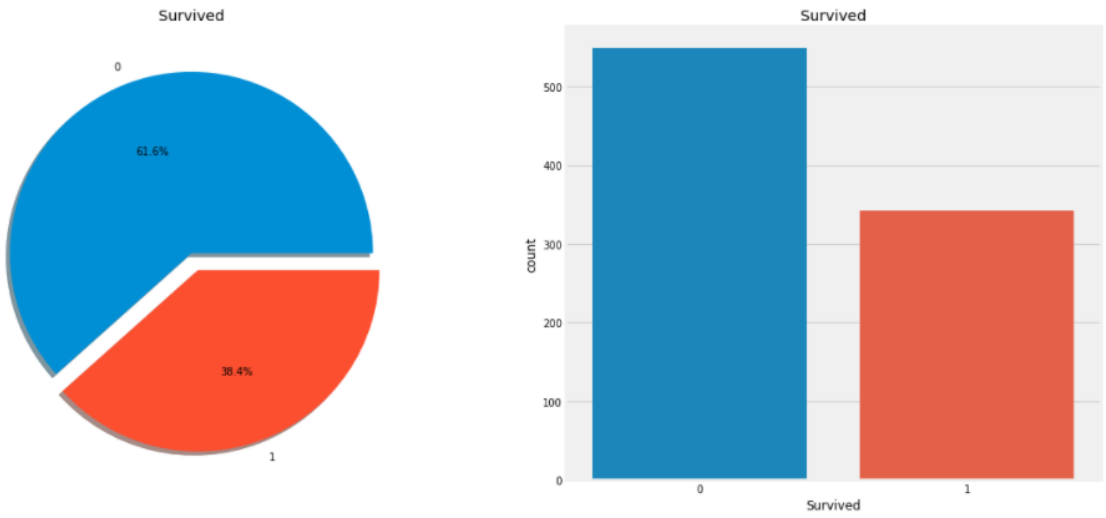


Figure 1.1.3. Pie and bar chart on S/D of training dataset

From fig 1.1.3, it is evident that not many passengers survived in the accident. There are 350 passengers survived out of whole amounts of passengers in training set, which means the survival rate is less than 40% from the crash. Then we should dig deeper and gain better insights from the dataset to see which categories of passengers survived and which did not. The method is to use the different features from dataset like Sex, Embarked, Age, etc. See and check whether the survival rate change according to different variables.

Before start analyzing the features, we should introduce the concept of type of features and assign different feature values to the correct type.

Categorical features are used to represent classification. Unlike numerical features, which are continuous, classification features are discrete. Some classification features are also numerical values, such as account IDs and IP addresses. But these values are not continuous. This feature has no concept of order or distance.

Ordinal features: categorical features that have some sort of meaning, but are different from numerical features. For example: academic qualifications (bachelor/master/doctorate), train ticket level (first-class/second-class), although they are categorical features, they are also ordered or distanced in a certain sense.

Continuous features: A feature that can be obtained arbitrarily within a length of time, and its value is uninterrupted. For example, the number between [0,1] can take n numbers. So, according to these concepts, we can categorize features, here the results:

- Categorical features: Sex, Embarked.
- Ordinal features: PClass.
- Continuous features: Age.

Sex--> Categorical Feature

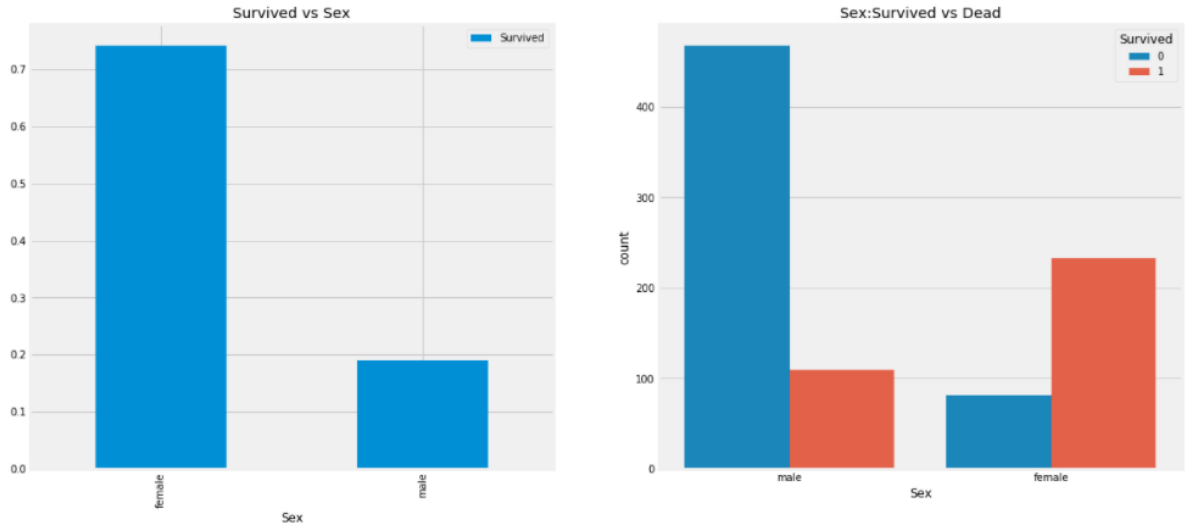


Figure 1.1.4. Bar charts of distribution on S/D and Sex

By fig 1.1.4, although there were far more men than women aboard the Titanic, women had a 75 percent survival rate and far more women were rescued than men. At the same time, the survival rate for men is less than 20 percent, so the Sex seems to be a very important feature.

Pclass --> Ordinal Feature

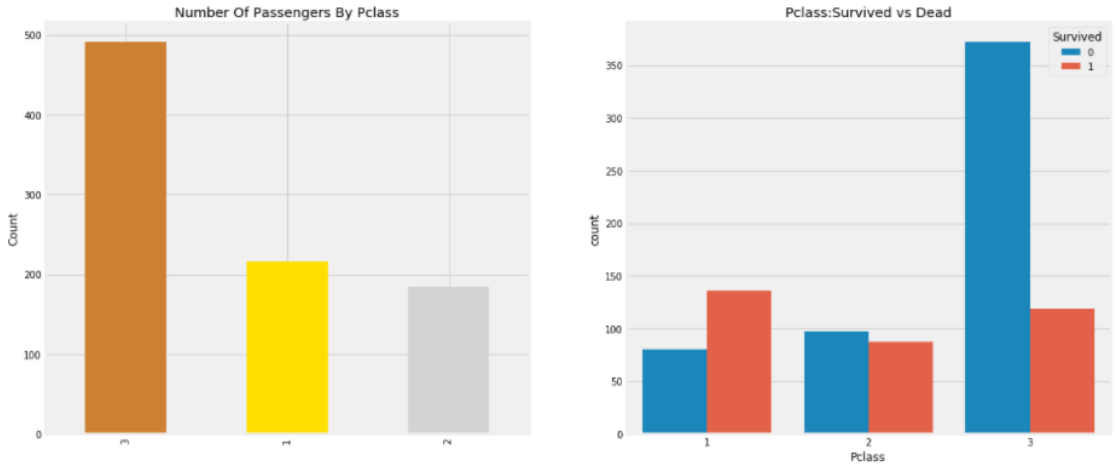


Figure 1.1.5. Bar charts of distribution on passengers in each Pclass and S/D

People say “Money Can't Buy Everything”, however it is clearly that passengers of Pclass1 had a very high possibility to get rescue. Although the number of passengers in Pclass3 were more, the number of survival rate from them was lowest, somewhere just around 25%. For Pclass1 survived is around 63% while for Pclass2 is around 48%. Passengers who buy premium tickets have higher social status and higher wealth than other passengers. From this we can realize that material wealth may not solve all problems, but in the Titanic incident, a high-priced ticket may be able to save a life. Let’s dive in little bit more and check for other interesting observations-check survival rate with Sex and Pclass together.

Sex	Survived	Pclass 1	2	3	All
female	0	3	6	72	81
	1	91	70	72	233
male	0	77	91	300	468
	1	45	17	47	109
All		216	184	491	891

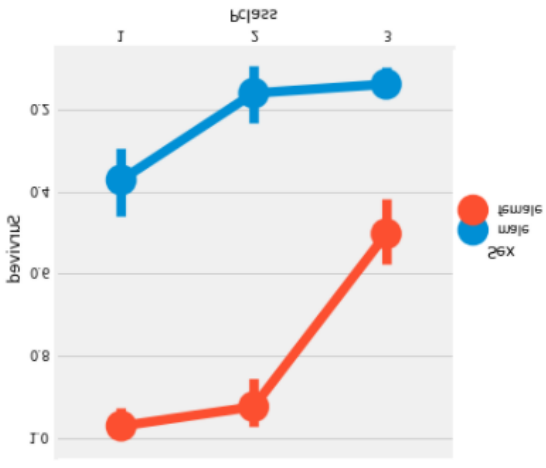


Figure 1.1.6. Cross tap and factor chart of two factors(features)

The reason why use factor plot in this case is that it makes the separation of categorical values easy. From the cross tab and the factor plot, it is clearly see that

survival rate for female from Pclass1 is about 95-96%, there are 3 out of 94 females from Pclass1 were died. It is evident that no matter which ticket they bought, females were given first priority to get rescue. At the same time, males from Pclass1 have such a low survival rate. So Pclass also has an important influence in this event.

Age--> Continuous Feature

Age of oldest passenger: 80.0 years

Age of youngest Passenger: 0.42 years

Average age: 29.7 years

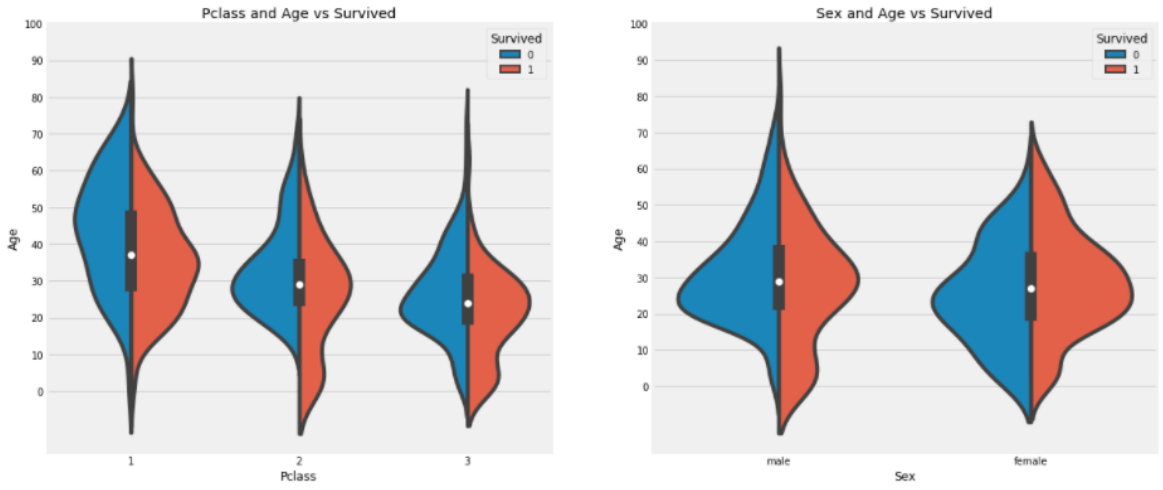


Figure 1.1.7 Violin chart of Pclass and Sex with distribution of Age by S/D

From the violin chart, we can see:1) The number of children is positively correlated with Pclass, and the survival rate of children under 10 were better than adults and regardless of which kind of tickets. 2) Survival rate for range of age for passengers 20-50 years old from Pclass1 is high and it is even better for females. 3) But males, survival was negatively correlated with age.

There are 177 null values in the Age feature. To replace these Null values, we can easily fill them up with mean value. But here is the problem, just cannot simply take the average of the existing ages to fill in the null values, it would be incorrect to classify a little child under five as a 30-year-old adult. Instead of this method, from Name column, there are many salutations like Mr. or Mrs. in the part of names.

It is a useful information so that we could distribute the mean values of passengers whose name has Mr. or Mrs. to the groups respective.

We can also use the Regex: [A-Za-z] +) to extract the Initials from the name.

Initial	Capt	Col	Countess	Don	Dr	Jonkheer	Lady	Major	Master	Miss	Mlle	Mme	Mr	Mrs	Ms	Rev	Sir
Sex																	
female	0	0	1	0	1	0	1	0	0	182	2	1	0	125	1	0	0
male	1	2	0	1	6	1	0	2	40	0	0	0	517	0	0	6	1

Figure 1.1.8. Cross tap of initials from the Name

There are some mistakes of spelling like Mme or Mlle that should be Miss. we can revise them correctly so as any other wrong values. And then fill null value of ages with mean value by the initial.

```
Initial
Master     4.574167
Miss      21.860000
Mr        32.739609
Mrs       35.981818
Other     45.888889
Name: Age, dtype: float64
```

And then we call fill ages with average values to each group according to the salutations (keep one significant digit).

Code:

```
: ## Assigning the NaN Values with the Ceil values of the mean ages
data.loc[(data.Age.isnull())&(data.Initial=='Mr'),'Age']=33
data.loc[(data.Age.isnull())&(data.Initial=='Mrs'),'Age']=36
data.loc[(data.Age.isnull())&(data.Initial=='Master'),'Age']=5
data.loc[(data.Age.isnull())&(data.Initial=='Miss'),'Age']=22
data.loc[(data.Age.isnull())&(data.Initial=='Other'),'Age']=46
```

```
data.Age.isnull().any() #So no null values left finally
```

```
False
```

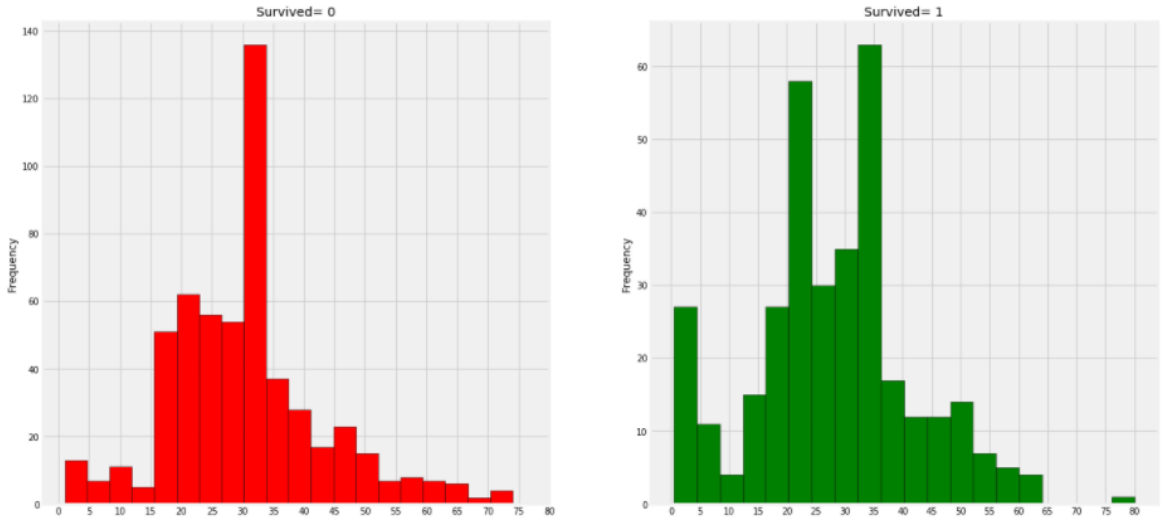


Figure 1.1.9. Bar chart of S/D by distribution of age

As we can see: 1) The children which age<5 were almost saved thanks to the principle (the women and children first).2) The oldest passenger got rescue.3) Passengers in the age group of 30-40 has the highest survival rate.

Embarked--> Categorical Value

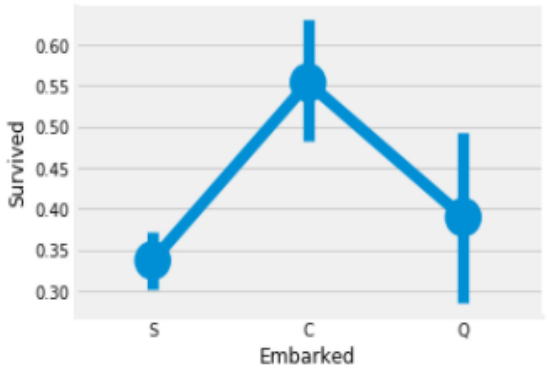


Figure 1.1.10. Factor chart of Embarked and survival rate

The chances for survival for Port C is highest around 0.55 while it is lowest for S, Q is middle.

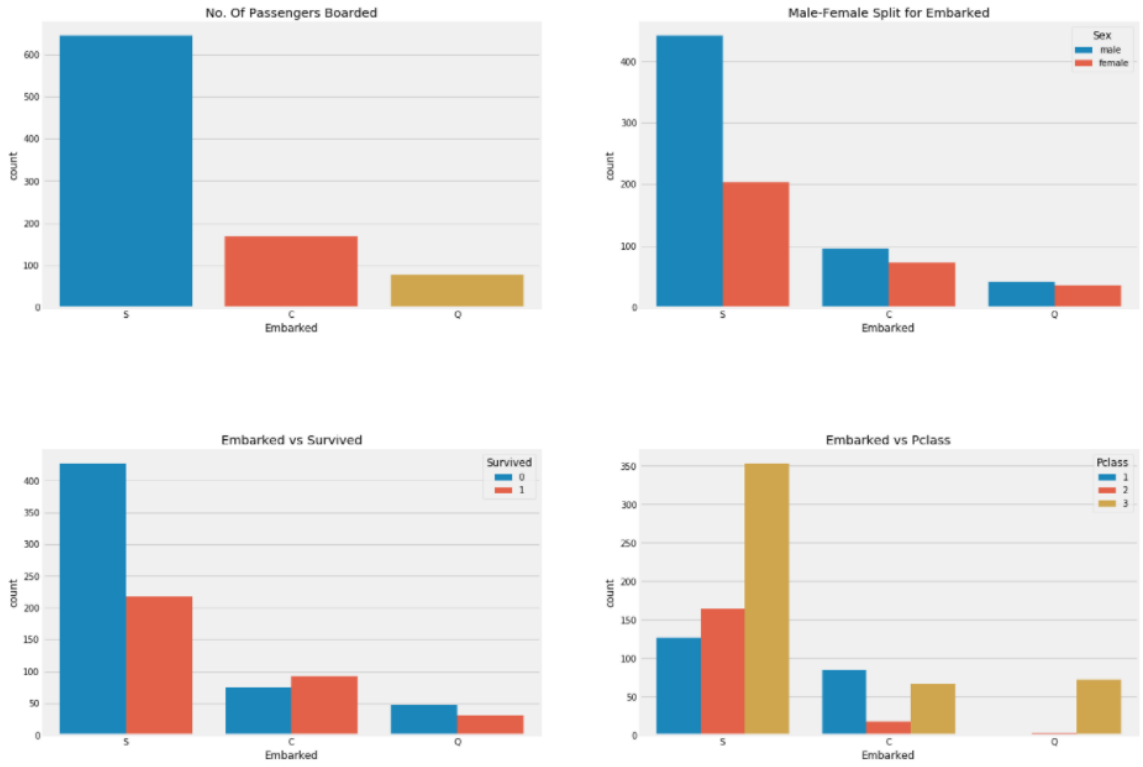


Figure 1.1.10. Bar charts of Embarked with other features

We can get observations :1) Maximum passengers boarded from S. Majority of them being from Pclass3. 2) The Passengers from C look to be lucky as a good proportion of them survived. The reason for this maybe the rescue of all the Pclass1 and Pclass2 Passengers. 3) The Embark S looks to the port from where majority of the rich people boarded. Still the chances for survival is low here, that is because many passengers from Pclass3 around 81% didn't survive. 4) Port Q had almost 95% of the passengers were from Pclass3.

SibSp, Parch-->Discrete Feature

This feature represents whether a person is alone or with his family members. Sibling = brother, sister, stepbrother, stepsister, Spouse = husband, wife, Parch = parents, children.

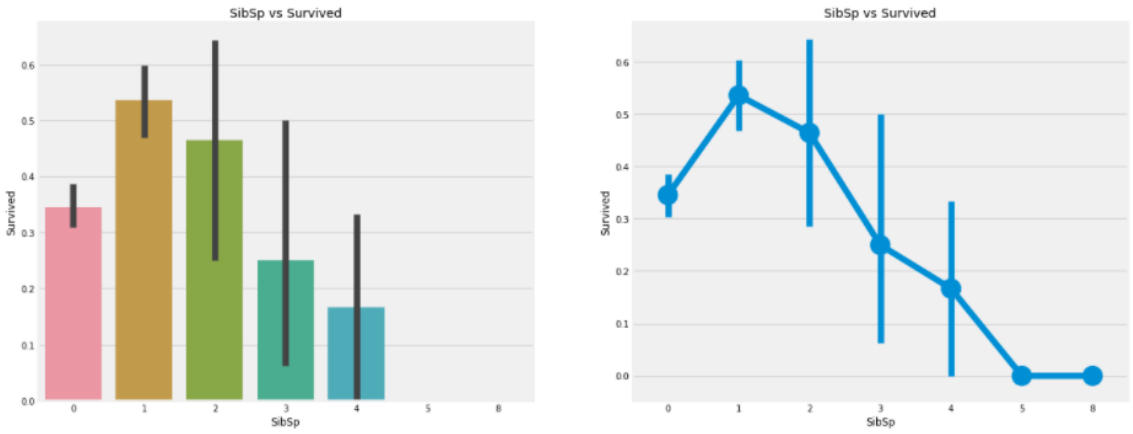


Figure 1.1.11. Bar and factor chart of SipSp and Survival rate

The bar chart and factor chart shows that if a passenger is alone onboard with no siblings, he has 34.5% survival rate. The graph roughly decreases if the number of sibling increase. This makes sense. That is, if I have a family on board, I will try to save them instead of saving myself first. Surprisingly the survival for families with 5-8 members is 0%. The reason is Pclass. The crosstab (fig 1.1.12) shows that Person with SibSp>3 were all in Pclass3. It is imminent that all the large families in Pclass3(>3) died. Here too the results are quite similar as Parch.

Pclass	1	2	3
SibSp			
0	137	120	351
1	71	55	83
2	5	8	15
3	3	1	12
4	0	0	18
5	0	0	5
8	0	0	7

Figure 1.1.12. Crosstab of SipSp with distribution of Pclass

Fare--> Continuous Feature

Highest Fare was: 512.3292

Lowest Fare was: 0.0

Average Fare was: 32.2042079685746

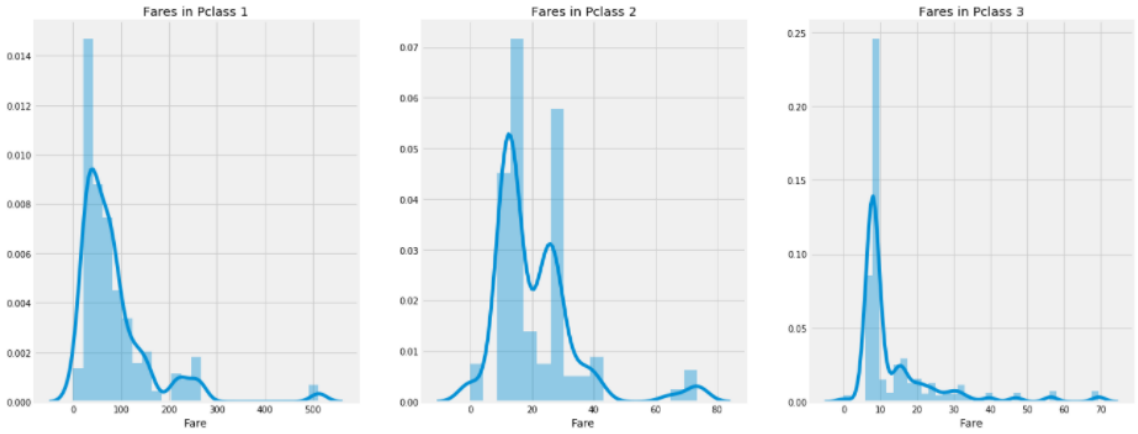


Figure 1.1.13. Distplot of Fares with Pclass

There looks to be a large distribution in the fares of Passengers in Pclass1 and this distribution goes on decreasing as the standards reduces. As this is also continuous, we can convert into discrete values by using binning.

Observations in a Nutshell for all features:

Sex: The chance of survival for women is high as compared to men.

Pclass: There is a visible trend that being a 1st class passenger gives you better chances of survival. The survival rate for Pclass3 is very low. For women, the chance of survival from Pclass1 is almost 1 and is high too for those from Pclass2.

Age: Children less than 5-10 years do have a high chance of survival. Passengers between age group 15 to 35 died a lot.

Embarked: This is a very interesting feature. The chances of survival at C looks to be better than even though the majority of Pclass1 passengers got up at S. Passengers at Q were all from Pclass3.

Parch+SibSp: Having 1-2 siblings, spouse on board or 1-3 Parents shows a greater chance of probability rather than being alone or having a large family travelling with.

1.2 Correlation between the features

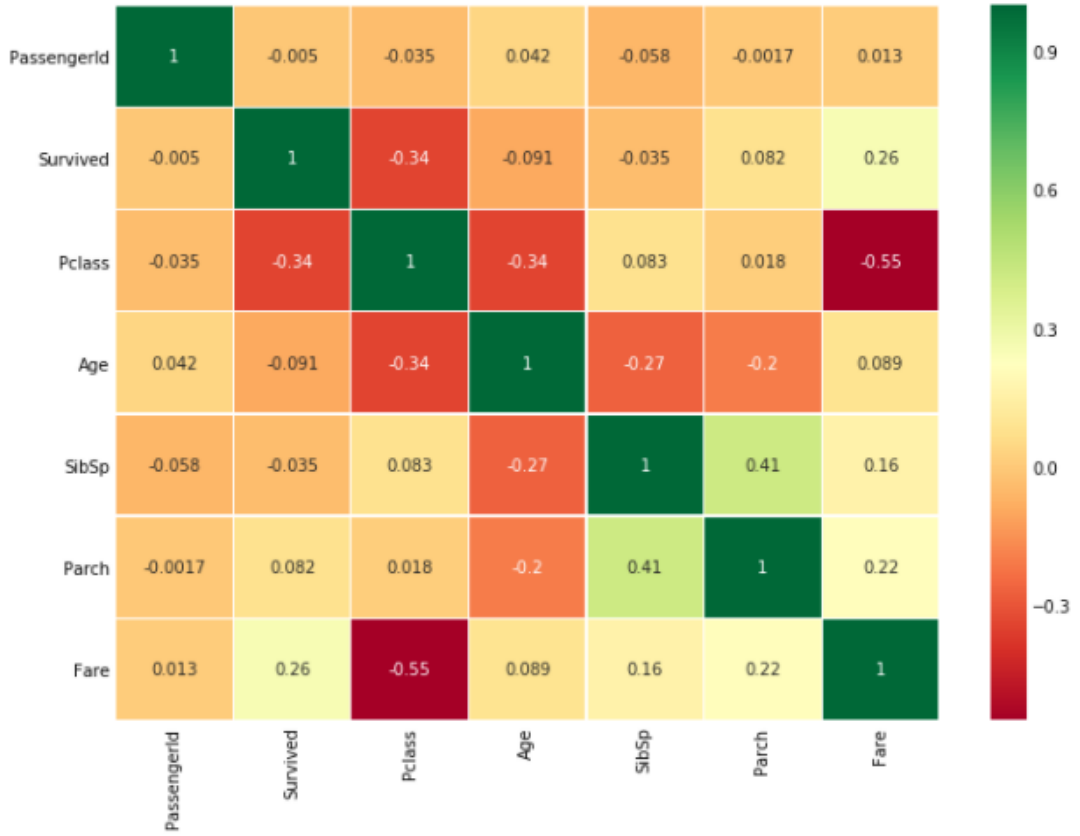


Figure 1.2.1. Heat map of original dataset

Interpreting the Heat map:

The first thing to notice is that only the numeric features are compared as it is obvious that we cannot correlate between alphabets or strings. Before understanding the plot, let us see what exactly correlation is (POSITIVE CORRELATION or NEGATIVE CORRELATION).

Now let's say that two features are highly or perfectly correlated, so the increase in one leads to increase in the other. This means that both the features are containing highly similar information and there is very little or no variance in information. This is known as Multi-Collinearity as both of them contains almost the same information.

While making or training models, we should try to eliminate redundant features as it reduces training time and many such advantages. Now let's look at this heat map, we can see that the features are not much correlated. The highest correlation is between SibSp and Parch i.e. 0.41. So we can carry on with all features.

2. Feature engineering and data cleaning

Whenever we are given a dataset with features, it is not necessary that all the features will be important. There maybe be many redundant features which should be eliminated. Also we can get or add new features by observing or extracting information from other features.

An example would be getting the Initials feature using the Name Feature. Let's see if we can get any new features and eliminate a few. Also we will transform the existing relevant features to suitable form for Predictive Modeling.

2.1 Adding any few features

We add some new features, first named Age_band, because there is a problem with it, As I have mentioned earlier that age is a continuous feature, there is a problem with continuous variables in ML Models.

We need to convert these continuous values into categorical values by either binning or normalization. I will be using binning i.e. group a range of ages into a single bin or put them a single value in, so the maximum value of age of the passenger was 80, we can divide the range of age into 5 bins. 80/5=16, so size of bins is 16.

Code:

```
data['Age_band']=0
data.loc[data['Age']<=16,'Age_band']=0
data.loc[(data['Age']>16)&(data['Age']<=32),'Age_band']=1
data.loc[(data['Age']>32)&(data['Age']<=48),'Age_band']=2
data.loc[(data['Age']>48)&(data['Age']<=64),'Age_band']=3
data.loc[data['Age']>64,'Age_band']=4
data.head(2)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Initial	Age_band
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	Mr	1
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	Mrs	2

Figure 2.1.1. Dataset with new feature(Age_band)

At this point, we can create a new feature called "Family_size" and "Alone" and analysis it. This feature is the summation of Parch and SibSp. It gives us a combined data so that we can check if survival rate has anything to do with family size of the passengers. Alone will denote whether a passenger is alone or not. Family_size=0 means that the passenger is alone. Clearly, if you are alone or Family_size=0, then chances for survival is very low. For family size > 4, the chances decrease too. This also looks to be an important feature for the model. Let's examine this further.

Since fare is also a continuous feature, we need to convert it into ordinal value. For this we will use pandas.qcut. So what qcut does is it splits or arranges the values according the number of bins we have passed. So if we pass for 5 bins, it will arrange the values equally spaced into 5 separate bins or value ranges.

Code:

```
data['Fare_Range']=pd.qcut(data['Fare'],4)
data.groupby(['Fare_Range'])['Survived'].mean().to_frame().style.background_gradient(cmap='summer_r')
```

Fare_Range	Survived
(-0.001, 7.91]	0.197309
(7.91, 14.454]	0.303571
(14.454, 31.0]	0.454955
(31.0, 512.329]	0.581081

Figure 2.1.1. example new feature(Fare_Range)

As discussed above, we can clearly see that as the fare_range increases, the chances of survival increases. Now we cannot pass the Fare_Range values as it is. We should convert it into singleton values same as we did in Age_Band, named Fare_cat.

Code:

```
data['Fare_cat']=0
data.loc[data['Fare']<=7.91,'Fare_cat']=0
data.loc[(data['Fare']>7.91)&(data['Fare']<=14.454),'Fare_cat']=1
data.loc[(data['Fare']>14.454)&(data['Fare']<=31),'Fare_cat']=2
data.loc[(data['Fare']>31)&(data['Fare']<=513),'Fare_cat']=3
```

Clearly, as the Fare_cat increases, the survival chances increase. This feature may become an important feature during modeling along with the Sex.

2.2 Removing redundant features

Name--> We don't need name feature as it cannot be converted into any categorical value.

Age--> We have the Age_band feature, so no need of this.

Ticket--> It is any random string that cannot be categorized.

Fare--> We have the Fare_cat feature, so unneeded

Cabin--> A lot of Null values and also many passengers have multiple cabins. So this is a useless feature.

Fare_Range--> We have the fare_cat feature.

PassengerId--> Cannot be categorized.

2.3 Converting features into suitable form for modeling

Since we cannot pass strings to a machine learning model, we need to convert features like Sex, Embarked, etc. into numeric values.

Code:

```
data['Sex'].replace(['male','female'],[0,1],inplace=True)
data['Embarked'].replace(['S','C','Q'],[0,1,2],inplace=True)
data['Initial'].replace(['Mr','Mrs','Miss','Master','Other'],[0,1,2,3,4],inplace=True)
```

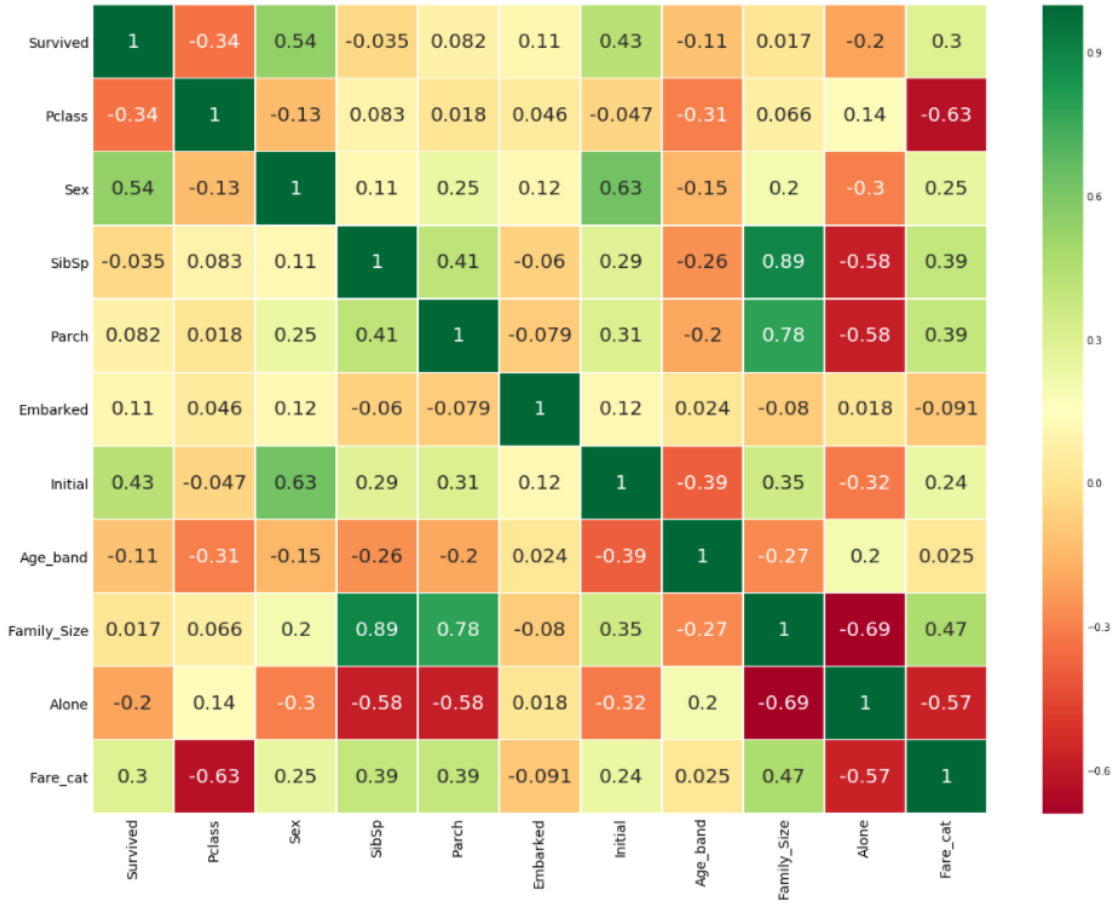


Figure 2.3.1. Heat map of processed dataset

Above correlation plot, we can see some positively related features. Some of them being SibSp and Family Size and Patch and Family Size and some negative ones like Alone and Family_Size.

	Survived	Pclass	Sex	SibSp	Parch	Embarked	Initial	Age_band	Family_Size	Alone	Fare_cat
0	0	3	0	1	0	0	0	1	1	0	0
1	1	1	1	1	0	1	1	2	1	0	3
2	1	3	1	0	0	0	2	1	0	1	1
3	1	1	1	1	0	0	1	2	1	0	3
4	0	3	0	0	0	0	0	2	0	1	1

Figure 2.3.2. Processed dataset

Now that we are finished all processed work with the dataset, next step is to start modeling and making predictions.

3. Predictive modeling

We have gained some insights from the EDA part. But with that, we can't accurately predict or tell whether a passenger survived or died. So now we will predict the whether the passengers' S/D using some classification algorithms. Here are the algorithms I will use to make the model:

1)Logistic Regression

2)Decision Tree

3)Random Forest

4)K-Nearest Neighbors

3.1 Running Basic Algorithms

Logistic Regression

LR is a transformation of a linear regression using the sigmoid function. The vertical axis stands for the probability for a given classification and the horizontal axis is the value of x. It assumes that the distribution of y|x is Bernoulli distribution. The formula of LR is as follows [6]:

$$\mathrm{F}(x) = \frac{1}{1 + e^{-(\beta 0 + \beta 1 x)}}$$

Here is $\beta 0 + \beta 1 x$ similar to the simple linear function y = ax + b. The logistic model applies a sigmoid function to limit the y value from a large measure to the range 0–1.

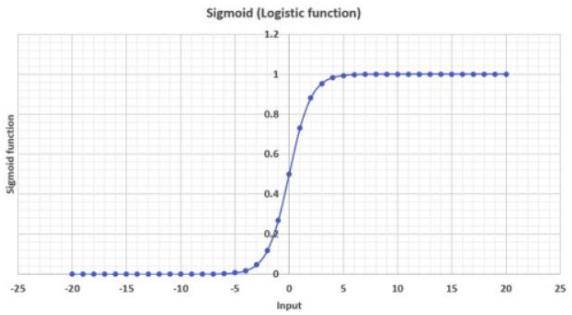


Figure 3.1.1. Sigmoid function of LR

Accuracy of LR:

```
The accuracy of the Logistic Regression is 0.817
```

Figure 3.1.2. The ROC curve of LR

The ROC curve is a common tool used with binary classifiers. The dashed line represents the ROC curve for a purely random classifier, a good classifier is as far away from this line as possible (towards the upper left corner).

Decision Tree

Decision tree is a flow-chart-like tree structure, where each internal node is denoted by rectangles and the leaf nodes are denoted by ovals. It the most commonly used algorithm because of its ease of implementation and easier to understand compared to other classification algorithms, decision tree classifiers obtain similar and sometimes better accuracy when compared with other classification methods [7]. Decision tree algorithm can be implemented in a serial or parallel fashion based on the volume of data, memory space available on the computer resource and scalability of the algorithm [8].

Accuracy of DT:

```
The accuracy of the Decision Tree is 0.81
```

Random Forest

Random forest is composed of of many individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. new input samples enter, and each decision tree in the forest will be judged and classified separately. In data science speak, the reason that the random forest model works so well is: A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models [9]. Each decision tree will get its own classification result, which one of the classification results of decision tree model is classified. At most, then the random forest will treat this result as the final result.

Accuracy of RF:

```
The accuracy of the Random Forests is 0.821
```

K-Nearest Neighbors

KNN classifier is to classify unlabeled observations by assigning them to the class of the most similar labeled examples. The KNN algorithm assigns a category to observations in the test dataset by comparing them to the observations in the training dataset. Because we know the actual category of observations in the test dataset, the performance of the KNN model can be evaluated. One of the most commonly used parameter is the average accuracy that is defined by the following equation [10]:

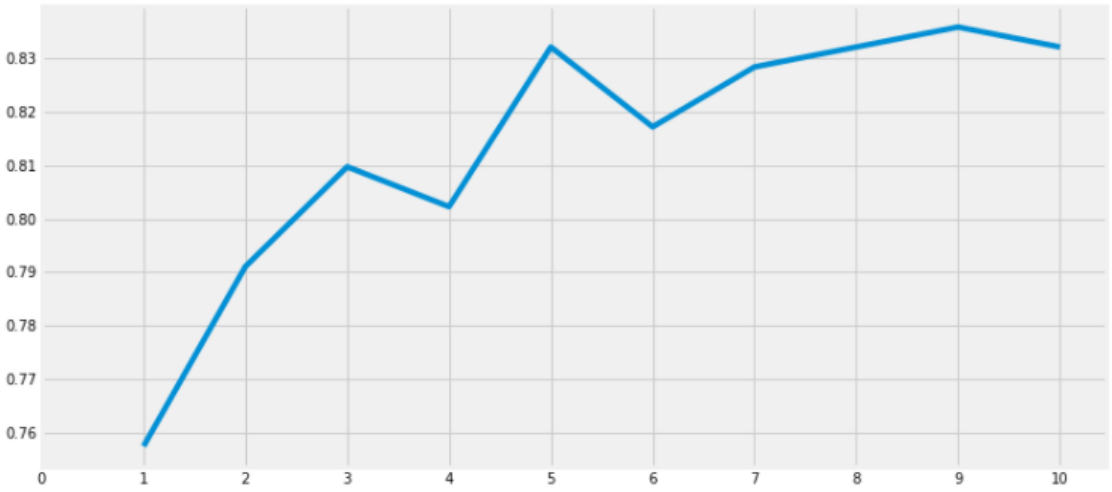
$$Average\ Accuracy = \sum_{i=1}^{l} \frac{TP_i + TN_i}{TP_i + FN_i + FP_i + TN_i} / l$$

where TP represents true positive, TN represents true negative, FP represents false positive and FN represents false negative. The subscript i indicates category, and l refers to the total category.

Accuracy of KNN:

```
The accuracy of the KNN is 0.832
```

Now the accuracy for the KNN model changes as we change the values for n_neighbors attribute. The default value is 5. Let's check the accuracies over various values of n_neighbors.




```
Accuracies for different values of n are: [0.75746269 0.79104478 0.80970149 0.80223881 0.83208955 0.81716418
 0.82835821 0.83208955 0.8358209  0.83208955] with the max value as  0.835820895522388
```

Figure 3.1.3. The change of Accuracy with different n_neighbors of KNN

The accuracy of a model is not the only factor that determines the robustness of the classifier. Let's say that a classifier is trained over a training data and tested over the test data and it scores an accuracy of 90%. Now this seems to be very good accuracy for a classifier, but can we confirm that it will be 90% for all the new test sets that come over? The answer is No, because we can't determine which all instances will the classifier will use to train itself. As the training and testing data changes, the accuracy will also change. It may increase or decrease. This is known as model variance.

To overcome this and get a generalized model, we use Cross Validation.

3.2 Cross Validation

Many times, the data is imbalanced, i.e. there may be a high number of class1 instances but less number of other class instances. Thus we should train and test our algorithm on each and every instance of the dataset. Then we can take an average of all the noted accuracies over the dataset.

1)The K-Fold Cross Validation works by first dividing the dataset into k-subsets.2) Let's say we divide the dataset into (k=5) parts. We reserve 1 part for testing and train the algorithm over the 4 parts. 3)We continue the process by changing the testing part in each iteration and training the algorithm over the other parts. The accuracies and errors are then averaged to get an average accuracy of the algorithm. This is called K-Fold Cross Validation.4) An algorithm may under-fit over a dataset for some training data and sometimes also over-fit the data for other training set. Thus with cross-validation, we can achieve a generalized model.

	CV Mean	Std
Logistic Regression	0.805843	0.021861
KNN	0.813783	0.041210
Decision Tree	0.805880	0.032258
Random Forest	0.815968	0.031248

Figure 3.2.1. The mean values of CV and standard deviation of models

The RF has highest accuracy of CV.

	predict	real
0	1	1
1	0	0
2	0	0
3	1	1
4	0	1
5	1	1
6	1	1
7	0	0
8	1	1
9	1	1
10	0	0

Figure 3.2.2. Comparison of RF model prediction and real S/D

As we can see, there are some non-right prediction, but we still have more than 80% accuracy.

3.3 Hyper-parameters tuning

The machine learning models are like a Black-Box. There are some default parameter values for this Black-Box, which we can tune or change to get a better model. Similar different parameters for different classifiers, are called the hyper-parameters, which we can tune to change the learning rate of the algorithm and get a better model, this is known as hyper-parameter tuning.

We will tune the hyper-parameters for the best classifier – RF.

```
Fitting 3 folds for each of 9 candidates, totalling 27 fits

[Parallel(n_jobs=1)]: Done  27 out of  27 | elapsed:   18.4s finished

0.817059483726
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=900, n_jobs=1,
            oob_score=False, random_state=0, verbose=0, warm_start=False)
```

Figure 3.3.1. The result of hyper-parameters tuning of RF

The best score for RF, score is about 81.8% with n_estimators=900.

3.4 Feature importance of best model

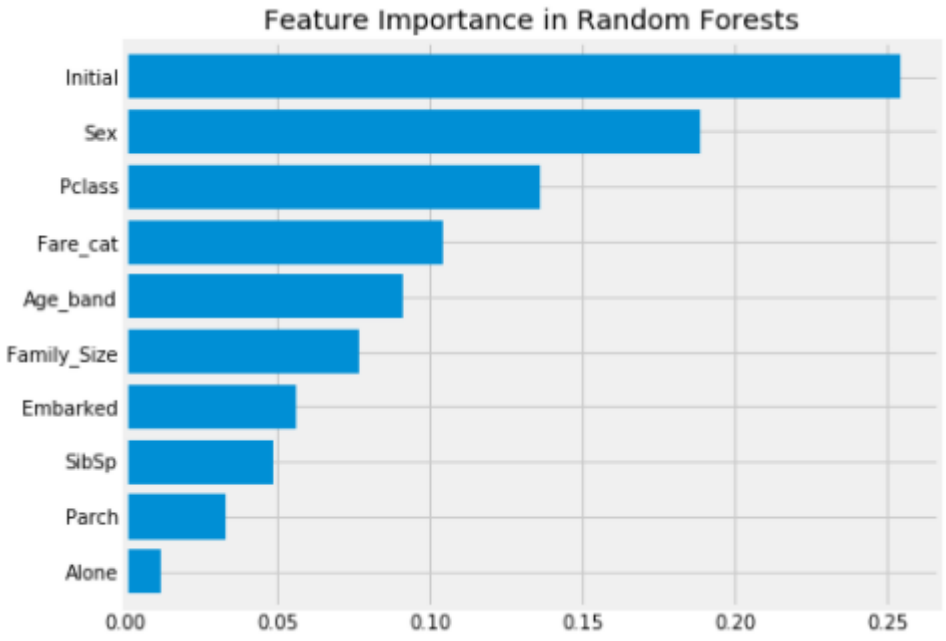


Figure 3.4.1. The bar chart of feature importance in RF

Observations:

1) Some of the common important features are Initial, Fare_cat, Pclass, Family_Size.

2) The Sex feature doesn't seem to give any importance, which is shocking as we had seen earlier that Sex combined with Pclass was giving a very good differentiating factor. Sex looks to be important only in RF. However, we can see the feature Initial, which is at the top in many classifiers. We had already seen the positive correlation between Sex and Initial, so they both refer to the gender.

3) Similarly the Pclass and Fare_cat refer to the status of the passengers and Family_Size with Alone, Parch and SibSp.

4. Financial management, resource efficiency and resource saving

The purpose of this section discusses the issues of competitiveness, resource efficiency and resource saving, as well as financial costs regarding the object of study of Master's thesis. Competitiveness analysis is carried out for this purpose. SWOT analysis helps to identify strengths, weaknesses, opportunities and threats associated with the project, and give an idea of working with them in each particular case. For the development of the project requires funds that go to the salaries of project participants and the necessary equipment, a complete list is given in the relevant section. The calculation of the resource efficiency indicator helps to make a final assessment of the technical decision on individual criteria and in general.

The object of research and development is to create the survival prediction model by using python programming, the process includes three parts: 1. Exploratory data analysis. 2.Feature engineering and data cleaning. 3.Predictive Modeling. The resulting machine learning models provides predicted results and shows accuracy, after cross-validation, the best model is selected and display the importance of each feature variable.

Potential consumers of the developed solution are enterprises or person involved in the data analysis. In order to effectively use the scientific potential of the project, it is necessary to pay attention both to the development and to its analysis with in terms of its relevance, as well as resource efficiency and resource saving.

The purpose of this section is to design and create competitive developments, technologies that meet modern requirements in the field of resource efficiency and resource saving.

4.1 Competitiveness analysis of technical solutions

In order to find sources of financing for the project, it is necessary, first, to determine the commercial value of the work. Analysis of competitive technical solutions in terms of resource efficiency and resource saving allows to evaluate the comparative effectiveness of scientific development. This analysis is advisable to carry out using an evaluation card.

First of all, it is necessary to analyze possible technical solutions and choose the best one based on the considered technical and economic criteria.

Evaluation map analysis presented in Table 1. The position of your research and competitors is evaluated for each indicator by you on a five-point scale, where 1 is the weakest position and 5 is the strongest. The weights of indicators determined

by you in the amount should be 1. Analysis of competitive technical solutions is determined by the formula:

$$C = \sum W_i \cdot P_i,$$

C - the competitiveness of research or a competitor;

Wi– criterion weight;
Pi – point of i-th criteria.

When i=1 – Machine learning method in this thesis -random forest (It has highest accuracy). As a result of data preparation, we gained some insights from the Exploratory Data Analysis. It should be noted that random forest is a kind of bagging algorithm, but random forest uses CART decision tree as a weak learner, and the feature selection of decision tree is also random. Due to the randomness, it is very useful to reduce the variance of the model, so the random forest generally does not need additional pruning, that is, it can achieve better generalization ability and anti-overfitting ability (Low Variance).
When i=2 – Classical method (K-Nearest Neighbors). The amount of calculation is large, especially when the number of features is very large.
When i=3 – Classical method (Decision Trees). Unable to handle high-latitude data, the risk of overfitting is higher.

Table 1. Evaluation card for comparison of competitive technical solutions

Evaluation criteria	**Criterion weight**	**Points**			**Competitiveness**		
		P_f	P_{i1}	P_{i2}	C_f	C_{i1}	C_{i2}
1	2	3	4	5	6	7	8
Technical criteria for evaluating resource efficiency							
1. Energy efficiency	0.1	4	3	4	0.4	0.3	0.4
2. Reliability	0.15	5	4	3	0.75	0.6	0.45
3. Safety	0.2	5	5	5	1	1	1
4. Ease of operation	0.1	4	3	3	0.4	0.3	0.3
5. Ability to connect to PC	0.15	5	4	4	0.75	0.6	0.6
Economic criteria for performance evaluation							
1. Development cost	0.1	5	3	4	0.5	0.3	0.4
2. Market penetration rate	0.1	4	3	4	0.4	0.3	0.4
3. Expected lifecycle	0.1	5	3	4	0.5	0.3	0.4

Total	**1**	**37**	**28**	**31**	**4.5**	**3.7**	**3.95**

From the scorecard, we can conclude that the best algorithm is random forest. Also, the most important criteria for the developed solution are high scores, and the total score is 4.5. Thus, this development, due to its safety, reliability and ability to connect to PC can be competitive in the market.

4.2 SWOT analysis

Complex analysis solution with the greatest competitiveness is carried out with the method of the SWOT analysis: Strengths, Weaknesses, Opportunities and Threats. The analysis has several stages. The first stage consists of describing the strengths and weaknesses of the project, identifying opportunities and threats to the project that have emerged or may appear in its external environment. The second stage consists of identifying the compatibility of the strengths and weaknesses of the project with the external environmental conditions. This compatibility or incompatibility should help to identify what strategic changes are needed.

Table 2. SWOT analysis

	Strengths: S1. Use machine learning algorithm to predict target feature S2. Ability to auto-calculate by using dependency grammar. S3. Expandable functional. S4. Low cost	**Weaknesses:** W1. This dataset is not the final solution actual tasks, but only constructor of this solutions. W2. The binary classifier lacks scalability. W3. Lack of positive reviews.
Opportunities: O1. Cluster of classification elements makes the progress of analysis more comprehensive O2. Market Demand problem solving extraction information O3. The data preprocessing process is professional and market attractive	– Adding new constructions for rules with taking into account the needs users will make more configuration flexible; – Attracting new customers thanks to product configuration for their task.	– Development of various training materials: detailed documentation with examples, development courses, seminars; – Get various datasets to create model and use them to coordinate analysis.
Threats: T1. Appearance on the market new players from similar development.	– Regular research of new solutions and models with subsequent implementation in the solution; – Work on optimizing the performance of the solution.	– Provision of configuration services for specific task; – Maintaining a stable pricing policy.

T2. The emergence of new algorithm different methods solutions.		

4.3 Project Initiation

The initiation process group consists of processes that are performed to define a new project or a new phase of an existing one. In the initiation processes, the initial purpose and content are determined and the initial financial resources are fixed. The internal and external stakeholders of the project who will interact and influence the overall result of the research project are determined.

Table 3. Stakeholders of the project

Project stakeholders	**Stakeholder expectations**
Company or individual wants to conduct a data analysis business	Use machine learning model to predict the target feature.

Table 4. Purpose and results of the project

Purpose of project:	Creating survival prediction models in different algorithm using Python programming, use program to completing the data cleaning, feature analysis and engineering. Finally, evaluate, validate and optimize model.
Expected results of the project:	ML Models cluster of classification elements to predict target feature.
Criteria for acceptance of the project result:	Model work correctly and get high accuracy
Requirements for the project result:	Python 3
	Jupiter notebooks
	Scikit-learn
	Anaconda 3

The organizational structure of the project

It is necessary to solve the any questions: who will be part of the working group of this project, determine the role of each participant in this project, and prescribe the functions of the participants and their number of labor hours in the project.

Table 5. Structure of the project

№	Participant	Role in the project	Functions	Labor time, hours (working days (from table 7) ×6 hours)
1	Supervisor: Е.И. Губин	Supervisor of project	Set goals, purpose and objectives, references give advices, review master's thesis	60
2	Student: Zhang Jifeng	Executor	Exploratory Data Analysis, Feature Engineering and Data Cleaning, Predictive Modeling.	384

Project limitations

Project limitations are all factors that can be as a restriction on the degree of freedom of the project team members.

Table 6. Project limitations

Factors	**Limitations / Assumptions**
3.1. Project's budget	172,788.5
3.1.1. Source of financing	TPU
3.2. Project timeline:	10 Feb 2022-10 Jun 2022
3.2.1. Date of approval of plan of project	10 Feb 2022
3.2.2. Completion date	25 May 2022

Project Schedule

As part of planning a science project, you need to build a project timeline and a Gantt Chart.

Table 7. Project Schedule

Job title	Duration, working days	Start date	Date of completion	Participants

Set goal, purpose and objectives	7	10 Feb 2022	18 Feb 2022	Supervisor
Analysis of the features	14	19 Feb 2022	10 Mar 2022	Student
Finding any relations or trends considering multiple features	14	11 Mar 2022	30 Mar 2022	Student
Feature engineering	12	31 Mar 2022	15 Apr 2022	Student
Data cleaning	5	16 Apr 2022	22 Apr 2022	Student
Predictive Modeling	14	23 Apr 2022	13 May 2022	Student
Prepare thesis	5	14 May 2022	20 May 2022	Student
Check thesis	3	21 May 2022	25 May 2022	Supervisor

A Gantt chart, or harmonogram, is a type of bar chart that illustrates a project schedule. This chart lists the tasks to be performed on the vertical axis, and time intervals on the horizontal axis. The width of the horizontal bars in the graph shows the duration of each activity.

Table 8. A Gantt chart

№	Activities	Participants	T_c, days	Duration of the project											
				February			March			April			May		
				1	2	3	1	2	3	1	2	3	1	2	3
1	Set goal, purpose and objectives	Supervisor	7												
2	Analysis of the features	Student	14												
3	Finding any relations or trends considering multiple features	Student	14												
4	Feature engineering	Student	12												
5	Data cleaning	Student	5												
6	Predictive Modeling	Student	14												
7	Prepare thesis	Student	5												
8	Check thesis	Supervisor	3												

4.4 Scientific and technical research budget

The amount of costs associated with the implementation of this work is the basis for the formation of the project budget. This budget will be presented as the lower limit of project costs when forming a contract with the customer.

To form the final cost value, all calculated costs for individual items related to the manager and the student are summed.

In the process of budgeting, the following grouping of costs by items is used:

- Material costs of scientific and technical research;
- costs of special equipment for scientific work (Depreciation of equipment used for design);
- basic salary;
- additional salary;
- labor tax;
- overhead.

Calculation of material costs

The calculation of material costs is carried out according to the formula:

$$C_m = (1 + k_T) \cdot \sum_{i=1}^{m} P_i \cdot N_{consi}$$

where m – the number of types of material resources consumed in the performance of scientific research;

N_{consi} – the amount of material resources of the i-th species planned to be used when performing scientific research (units, kg, m, m^2, etc.);

P_i – the acquisition price of a unit of the i-th type of material resources consumed (rub./units, rub./kg, rub./m, rub./m^2, etc.);

k_T – coefficient taking into account transportation costs.

Prices for material resources can be set according to data posted on relevant websites on the Internet by manufacturers (or supplier organizations).

Table 9. Material costs

Name	Amount	Price per unit, rub.	Material costs, rub.

Paper SvetoCopy Classic A4, 80g/m2, 500 sheets.	1	481	481
Laser cartridge Cactus CS-TK1170 (TK-1170) black for Kyocera Mita Ecosys	2	670	1340
Total			1821

Costs of special equipment

During whole work, we used our own PC. Supervisor's computer was provided by Tomsk polytechnic university, so there is no extra cost on hardware or any other special equipment.

In the progress of complete project, we used Python to clean the dataset and create model. Python3 is free for users, the same as Anaconda3, Jupiter Notebook and Scikit-learn.

Thus, we don't have costs for special equipment no matter in hardware but also software.

Basic salary

This point includes the basic salary of participants directly involved in the implementation of work on this research. The value of salary costs is determined based on the labor intensity of the work performed and the current salary system

The basic salary (S_b) is calculated according to the formula:

$$S_b = S_a \cdot T_w ,$$

where S_b – basic salary per participant;

T_w – the duration of the work performed by the scientific and technical worker, working days;

S_d - the average daily salary of an participant, rub.

The average daily salary is calculated by the formula:

$$S_d = \frac{S_m \cdot M}{F_v},$$

где S_m – monthly salary of an participant, rub .;

M – the number of months of work without leave during the year:

at holiday in 48 days, M = 11.2 months, 6 day per week;

F_v – valid annual fund of working time of scientific and technical personnel (251 days).

Table 10. The valid annual fund of working time

Working time indicators	
Calendar number of days	365
The number of non-working days - weekend - holidays	 52 14
Loss of working time - vacation - isolation period - sick absence	 48
The valid annual fund of working time	251

Monthly salary is calculated by formula:

$$S_{month} = S_{base} \cdot (k_{premium} + k_{bonus}) \cdot k_{reg}, \quad (x)$$

where S_{base} – base salary, rubles;

$k_{premium}$ – premium rate;

k_{bonus} – bonus rate;

k_{reg} – regional rate.

Table 11. Calculation of the base salaries

Performers	S_{base}, rubles	k_{reg}	S_{month}, rub.	W_d, rub.	T_p, work days (from	W_{base}, rub.

					table 7)	
Supervisor	37700	1,3	49010	2030.7	10	20307
Student	19200		24960	1,034.2	64	66,188.8

Additional salary

This point includes the amounts of payments stipulated by the legislation on labor, for example, payment of regular and additional holidays; payment of time associated with state and public duties; payment for work experience, etc.

Additional salaries are calculated on the basis of 10-15% of the base salary of workers:

$$W_{add} = k_{extra} \cdot W_{base}, \quad \text{(x)}$$

where W_{add} – additional salary, rubles;

k_{extra} – additional salary coefficient (10%);

W_{base} – base salary, rubles.

Total salary:

Supervisor: 20307*1.1=22,337.7

Sudent: 66,188.8*1.1=72,807.7

Total additional salary:

(20307+66,188.8) *0.1=8649.6

Labor tax

Tax to extra-budgetary funds are compulsory according to the norms established by the legislation of the Russian Federation to the state social insurance (SIF), pension fund (PF) and medical insurance (FCMIF) from the costs of workers.

Payment to extra-budgetary funds is determined of the formula:

$$P_{social} = k_b \cdot (W_{base} + W_{add}) \quad \text{(x)}$$

where k_b – coefficient of deductions for labor tax.

In accordance with the Federal law of July 24, 2009 No. 212-FL, the amount of insurance contributions is set at 30%. Institutions conducting educational and scientific activities have rate - 27.1%.

Table 12. Labor tax

	Supervisor	**Student**
Coefficient of deductions	27.1%	
Salary (basic and additional), rubles	22,337.7	72,807.7
Labor tax, rubles	6,053.5	19,730.9

Overhead costs

Overhead costs include other management and maintenance costs that can be allocated directly to the project. In addition, this includes expenses for the maintenance, operation and repair of equipment, production tools and equipment, buildings, structures, etc.

Overhead costs account from 30% to 90% of the amount of base and additional salary of employees.

Overhead is calculated according to the formula:

$$C_{ov} = k_{ov} \cdot (W_{base} + W_{add})$$

where k_{ov} – overhead rate.

Table 13. Overhead

	Project leader	**Engineer**
Overhead rate	50%	
Salary, rubles	22,337.7	72,807.7
Overhead, rubles	11,168.9	36,403.9

Other direct costs

Energy costs for equipment are calculated by the formula:

$$C = P_{el} \cdot P \cdot F_{eq},$$

where P_{el} − power rates (5.8 rubles per 1 kWh);

P − power of equipment, kW;

F_{eq} − equipment usage time, hours.

C=5.8*0.06*384+5.8*0.15*60=1254.9 rub

Internet (my PC): 350*4=1400 rub

Total:1254.9+1400=2654.9

Formation of budget costs

The calculated cost of research is the basis for budgeting project costs.

Determining the budget for the scientific research is given in the table.

Table 14. Budget for the scientific and technical research

Name	Cost, rubles
1. Material costs	1821
2. Equipment costs	0
3. Basic salary	86,495.8
4. Additional salary	8649.6
5. Labor tax	25,784.4
6. Overhead	47,572.8
7. Other direct costs	2654.9
Total planned costs	172,788.5

4.5 Conclusion of financial management

Thus, in this section was developed stages for design and create competitive development that meet the requirements in the field of resource efficiency and resource saving.

These stages includes :

- development of a common economic project idea, formation of a project concept;
- organization of work on a research project;
- identification of possible research alternatives;
- research planning;

- assessing the commercial potential and prospects of scientific research from the standpoint of resource efficiency and resource saving;
- determination of resource (resource saving), financial, budget, social and economic efficiency of the project.

5. Social responsibility

The object of research and development is to create the survival prediction model by using python programming, the process includes three parts: 1. Exploratory data analysis. 2.Feature engineering and data cleaning. 3.Predictive Modeling. The resulting machine learning models provides predicted results and shows accuracy, after cross-validation, the best model is selected and display the importance of each feature variable. Theoretical the significance of this work lies in the development of algorithms and methods for predicting target variable from processed dataset with machine learning methods.

Potential readers of whole works are enterprises or person involved in the data analysis.

The development of whole algorithm was carried out in PC, which means, it requires a high-performance CPU.

5.1 Legal and organizational issues of occupational safety

5.1.1 The projected working area, law norms of labor legislation

According to SP 2.4.3648-20 “Sanitary and epidemiological requirements for organizations of education and training, recreation and rehabilitation of children and youth”, since 2021 the issue of establishing breaks while working at a computer has not been regulated by law [1].

The employer may independently establish the procedure for granting breaks from work for computer for recreation in the rules of internal labor regulations. These breaks are included in working hours. That is, they do not extend the duration of the employee's working day. According to the Labor Code 62 of the Russian Federation dated 30.12. 2001 No. 197-FZ (as amended on 04/01/2019) during these breaks, the employee should not perform other work [2].

Since working with this library in an enterprise implies collection and analysis of personal data. To restrict access to medical data and ensure their security, data processing should carry out in accordance with federal law on the protection personal data [3]:

1. The processing of personal data must be carried out on legal and fair basis.
2. The processing of personal data should be limited achieving specific, predetermined and legitimate goals. Not processing of personal data incompatible with the purposes of collection is allowed personal data.
3. It is not allowed to combine databases containing personal data processed for purposes incompatible between yourself.
4. Processing is subject only to personal data that meet the purposes of their processing.
5. The content and scope of the processed personal data must comply with the stated purposes of processing. processed personal data should not be redundant in relation to the stated purposes of their processing.

Basic ergonomic requirements for the correct location and arrangement of researcher's workplace

According to SanPiN 2.2.2 / 2.4.1340-03 [4], the workplace when working with a PC should be at least 6 square meters. The legroom should correspond to the following parameters: the legroom height is at least 600 mm, the seat distance to the lower edge of the working surface is at least 150 mm, and the seat height is 420 mm. It is worth noting that the height of the table should depend on the growth of the operator.

The following requirements are also provided for the organization of the workplace of the PC user: The design of the working chair should ensure the maintenance of a rational working posture while working on the PC and allow the posture to be changed in order to reduce the static tension of the neck and shoulder muscles and back to prevent the development of fatigue.

The type of working chair should be selected taking into account the growth of the user, the nature and duration of work with the PC. The working chair should be lifting and swivel, adjustable in height and angle of inclination of the seat and back, as well as the distance of the back from the front edge of the seat, while the adjustment of each parameter should be independent, easy to carry out and have a secure fit.

5.1.2 Features of the labor legislation in relation to the specific conditions of the project.

An employee of a state corporation, a public company or a state company, in the cases and in the manner established by the Government of the Russian Federation, is obliged to [5]:

1. Provide information on their income, expenses, property and property obligations and on income, expenses, property and property obligations of his spouse and minor children.
2. Inform the employer about personal interest in the performance of labor duties, which may lead to a conflict of interest, take measures to prevent such a conflict.

An employee of a state corporation, state company, public company, in cases established by the Government of the Russian Federation, is prohibited from:

1. Participate in the activities of the management and control bodies of a commercial organization, with the exception of participation with the consent of the supreme management body of a state corporation, state company or public company.

2. Carry out entrepreneurial activities.

3. Be an attorney or third party representative in a public corporation, public company or public company, unless such activity is carried out with the consent of the supreme governing body of a public corporation, state company or public company.

4. Receive in connection with the performance of labor duties remuneration from other legal entities, individuals (gifts, monetary remuneration, loans, services, payment for entertainment, recreation and other remuneration), with the exception of remuneration for performance in the case provided for in clause 1 of this part, functions of members of the management and control bodies of a commercial organization and compensation for travel expenses related to the performance of such functions.

5. Use for purposes not related to the performance of labor duties, the property of a state corporation, state company or public company, as well as transfer it to other persons.

6. Disclose or use information classified by the legislation of the Russian Federation as confidential information, or proprietary information, as well as information that became known to him in connection with the performance of his labor duties.

7. Accept awards, honorary and special titles (except for scientific titles) from foreign states, international organizations without the written permission of the employer's representative.

8. Use official powers in the interests of political parties, other public associations, religious associations and other organizations that are not the object of the activities of a state corporation, state company or public company.

9. Create structures of political parties, other public associations (with the exception of trade unions, veterans and other bodies of public amateur performance) and religious associations in a state corporation, state company or public company or facilitate the creation of these structures.

10. Be a member of management bodies, boards of trustees or supervisory boards, other bodies of foreign non-profit non-governmental organizations and their structural divisions operating in the Russian Federation.

11. Be engaged without the written permission of the employer in paid activities financed exclusively at the expense of foreign states, international and foreign organizations, foreign citizens, stateless persons, unless otherwise provided by an international treaty of the Russian Federation or the legislation of the Russian Federation.

5.2 Occupational safety

Occupational safety is understood as a system of organizational measures and technical means that prevent or reduce the possibility of exposure of workers to

dangerous harmful production factors arising at nuclear power plants during work activity. In our work, it is necessary to detect dangerous and harmful factors that may arise when working with an information system. Subsequent selection is carried out using GOST 12.0.003–2015 "Hazardous and harmful production factors. Classification". The results of the selection are shown in the table below. [6]

5.2.1 Analysis of harmful and dangerous factors that can be created

All dangers which are seated and faced with PC are divided into several categories, the basis of their classification lies in determining their effect on certain organs of the human body, as well as the method of affecting them.

Table 5.1 - Potential hazardous and harmful production factors

Factors (GOST 12.0.003-2015)	The Type	Impact on the human body	Legislation documents
Increased levels of electromagnetic radiation	Physical	Harmful	GOST 12.1.030-81 Electric safety. Protective Conductive earth, neutralling.
Insufficient illumination of workplace	Physical	Harmful	SanPiN2.2.1/2.1.1.1278-03 Hygienic requirements for natural, artificial and mixed lighting of residential and public buildings.
Excessive noise	Physical	Harmful	GOST 12.1.003-2014 Occupational safety standards system. Noise. General safety requirements.
Increased / decreased air humidity in the workplace	Physical	Harmful	GOST 12.1.005-88 General sanitary requirements for working zone air
physical overload (static - long-term preservation of a certain posture)	Physical	Harmful	GOST 9241-4-2009 Ergonomic requirements for office work with visual display terminals (VDT). Part 4. Keyboard requirements.
overstrain of analyzers (vision)	Physical	Harmful	GOST 9241-4-2009 Ergonomic requirements

			for office work with visual display terminals (VDTs). Part 5. Workstation layout and postural requirements.
Increased voltage in an electrical circuit, the closure of which can pass through the human body.	Physical	Dangerous	GOST 12.1.019-2017 Electrical safety. General requirements and nomenclature of types of protection.

- **Increased levels of electromagnetic radiation:**

Electromagnetic radiation is an oscillation of electric and magnetic fields that propagates through space at the speed of light. A person does not see or feel it, therefore, is not able to assess how it affects health. Meanwhile, doctors all over the world are sounding the alarm that EMR acts on the body like radiation. Let's figure out how electromagnetic waves affect a person, whether there are ways to protect against adverse effects. [7]

To minimize the impact:

Use special dosimeters to find out the level of danger from various sources of electromagnetic radiation at home and at work;

Arrange electrical appliances according to the indicators, try to stay away from play areas and dining tables (at least 2 meters);

The distance from the CRT monitor or TV should be at least 30 cm;

If possible, remove all electrical appliances from bedrooms and children's rooms;

Place the electronic clock with the alarm clock no more than 10 cm away from the pillow;

Do not stay near a working microwave, microwave or heater;

It is not recommended that the mobile phone be within 2.5 cm of the head. It is best to make a hands-free call and keep the mobile phone away from you as much as possible;

Always turn off electronic devices that are not in use, as they produce a certain dose of radiation even in sleep mode;

Using a hair dryer before bed is harmful: EMR slows melatonin production and disrupts sleep cycles. Do not use a computer or tablet within 2 hours of bedtime;

In sockets where electrical appliances are connected, it is necessary to check for grounding.

- **Insufficient illumination of workplace:**

The harmful effect of lighting parameters is manifested in the absence or lack of natural light, as well as insufficient illumination of the working area. Properly designed and rationally executed lighting of production facilities has a positive impact on workers, improves efficiency and safety, reduces fatigue and injuries, and also maintains high performance. Visual discomfort and physiological strain such as

anxiety, fatigue, lethargy, headaches, eyestrain, migraine, nausea, back pain, neck pain, shoulder pain, poor concentration or lack of mental alertness, and daytime sleepiness among video display terminal (VDT) workers are primarily connected with inadequate lighting in the working place and in most cases decrease work performance and efficiency.

To minimize the impact:

The illumination on the table surface in the area of the working document should be 300-500 lux. [8] Lighting should not create glare on the surface of the monitor. Illumination of the monitor surface should not be more than 300-lux;

The brightness of the lamps of common light in the area with radiation angles from 50 to 90 °should be no more than 200 cd/m, the protective angle of the lamps should be at least 40 °. The ripple coefficient should not exceed 5%.

- **Excessive noise:**

Air-condition and equipment cooling fans necessary for the proper operation of IT equipment run continuously and create excessive noise that influence comfort, poses risk to hearing and impairs communication and concentration. Noise worsens working conditions; have a harmful effect on the human body, namely, the organs of hearing and the whole body through the central nervous system. It results in weakened attention, deteriorated memory, decreased response, and increased number of errors in work.

To minimize the impact:

When working faced with a PC, the noise level in the workplace should not exceed 65 dB; [9]

In order to study in a quiet environment, irrelevant applications of the computer should be closed to reduce computer power consumption, thereby reducing computer noise, and windows should also be closed to reduce environmental noise;

Numerous studies have shown that with an equal integral noise level, the development of occupational hearing loss will be observed more often and with less work experience, if the noise at the workplace is predominantly impulsive.

- **Increased / decreased air humidity in the workplace:**

Computer workstations, lighting and electronic devices generate heat constantly, causing variations in the microclimate in the work environment. Increased evidence shows that indoor environmental conditions substantially influence health and productivity. Relative humidity less than 30% can lead to skin and throat irritation, producing high static.

To minimize the impact:

According to the General sanitary requirements for working zone air (GOST 12.1.005-88) [10] a computer workstation worker belongs to Category-I work, which implies light physical work. Parameters such as: humidity is the one of determine the thermal comfort. In any seasons: it should be around 50%.

- **physical overload (static - long-term preservation of a certain posture):**

Individuals who use computers and do monotonous repetitive manipulations with or without objects over long periods of time may experience discomfort or pain as a result of poor posture, improper adjustment or use of workstation components or other factors which may lead to musculoskeletal disorders (MSD). In most cases,

there are relatively simple and inexpensive corrective measures which can be employed to reduce the likelihood of discomfort or injury.

Musculoskeletal conditions are typically characterized by pain (often persistent) and limitations in mobility, dexterity and overall level of functioning, reducing people's ability to work. Low back pain is the main contributor to the overall burden of musculoskeletal conditions which are also the highest contributor to the global need for rehabilitation. Regardless of the working position, sitting for long periods of time is unhealthy.

The design of the workplace and the relative position of all its elements (seats, keyboard, ways of displaying information, etc.) must conform to anthropometric, physiological and psychological requirements and the nature of the work.

To minimize the impact:

Use an ergonomic keyboard, which enable users to comfortably, quickly and accurately identify and use the keys they need. Keyboard characteristics that affect performance include: alpha and numeric key layout, language differences (country variants), the physical characteristics of individual keys, and the overall configuration of the keyboard housing; [11]

prepare a comfortable small pillow at ordinary times and place the pillow on the chair, which will help relieve the pressure on the lumbar spine of the human body, and turn the neck more in leisure time, which can move the muscles and bones well.

- **overstrain of analyzers (vision):**

The main factors that determine the configuration of a workstation are the seat, work surface, viewing angle, keyboard height, knee clearance, forearm tilt, and armrest height.

Configure the workstation correctly to ensure comfortable and productive work. Determine the parameters agreed with the individual user in terms of specific requirements in operating the computer, body size, allowable and recommended work postures, and work comfort.

To minimize the impact:

In order to be able to formulate admissible requirements in a qualified manner that ensure efficient and comfortable work, taking into account the dimensions of the human body, to specify the right working postures;

a well-designed seat is to provide stable, comfortable body support that does not interfere with body movement while working and helps to complete the job;

The user must be able to tilt or rotate the video display in such a way as to maintain a relaxed working posture, regardless of eye height, with minimal effort, while avoiding annoying reflections and glare on the screen. Having the ability to adjust the monitor height setting is also useful. Adaptability is provided by adjusting mechanisms built into the video display, or by special devices that are part of the office equipment or the display itself. When making adjustments, the user must not lift the blocks with objects placed on them, such as books or manuscripts. Adjustment mechanisms should be clear, unambiguous, and adjustment should be easy to perform. [12]

- **Increased voltage in an electrical circuit, the closure of which can pass through the human body:**

Electric current is able to create severe burns in the body. The reason is hidden in the power dissipation across the body ́s electrical resistance. Shock can cause: cardiac arrest, burns to tissues and organs, muscle spasms, serious effects to the nervous system and other unexpected consequences. [13]

To minimize the impact:

Providing insulating materials and providing insulating clothing. Keep hands dry when touching sockets.

5.3 Environmental Safety

The subsection considers the nature of the impact of the projected environmental solutions. Alleged sources identified environmental pollution arising from the development and implementation of proposed solutions.

5.3.1 Analysis of the influence of the object and the research process on environment

The object of study does not affect the environment, so how the computer does not emit harmful substances into the atmosphere and hydrosphere.

At the end of the life of the PC, they can be classified as waste electronic industry. The processing of such waste is carried out separation into homogeneous components, chemical isolation of suitable for further use of the components and sending them for further use according to GOST R 55102-2012"Resource saving. Waste management. Safety guide collection, storage, transportation and disassembly of spent electrical and electronic equipment, except mercury-containing devices and instruments" [14]. List of elements and waste electrical and electronic equipment containing them equipment that must be collected separately at the time of withdrawal waste electrical and electronic equipment from operation:

- Capacitors - contain polychlorinated biphenyls;
- printed circuit boards and other devices with a surface area more than 10 cm2 - contain lead, mercury, cadmium;
- cartridges - contain lead, cadmium, benzene, toluene, phenol;
- plastic;
- cathode ray tubes - contain lead glass, barium compounds, phosphors;
- elements of used electrical and electronic equipment - contain lead, cadmium, tin;
- fluorescent lamps - contain mercury.

5.3.2 Rationale for measures to protect the environment

Protection of soil cover and subsoil from solid waste is implemented for collection, sorting and disposal of waste and its organized burial. The main regulations governing the issue disposal of personal computers are federal laws of the Russian Federation"On Environmental Protection" and "On Production and Consumption Wastes". According to these laws, all office equipment must be

disposed of in compliance with certain rules: dismantling of spare parts, sorting of waste and disposal.

Fluorescent lamps are classified as mercury-containing waste, and for their disposal, the Decree of the Government of the Russian Federation [15] applies. Establishes the procedure for handling production and consumption waste in terms of lighting devices, electric lamps, improper collection, accumulation, use, neutralization, transportation and placement of which may cause harm to life, health citizens, harm to animals, plants and the environment.

Self-disposal, use, transportation and disposal of spent mercury-containing lamps consumers of waste mercury-containing lamps, as well as their accumulation in places that are the common property of the owners of the premises apartment building, with the exception of placement in places of primary collection and placement and transportation to them. Collection of waste mercury-containing lamps at consumers carry out specialized organizations. Waste that cannot be recycled or recycled use, be disposed of in landfills.

5.3 Emergency Safety

The object of research can initiate the emergence of such emergency like a fire. The cause of the fire may be power supply or computer failure.

When conducting research in the laboratory, it may also occur fire. The causes of the fire can be: ignoring the basic rules fire safety, electrical wiring failure, fire artificial lighting devices, ignition of computing devices equipment due to insulation failure or malfunction of the equipment.

According to NPB 105-03 “Definition of categories of premises, buildings and outdoor installations for explosion and fire hazards”
the room in which the system was developed belongs to category B3 according to fire hazard, contains substances and materials that can burn at interaction with water, air oxygen or with each other [16].

The room contains a computer, therefore, according to SP 9.13130.2009“Fire equipment. FIRE EXTINGUISHERS. Operating requirements" for elimination of fires caused by the ignition of electrical equipment, carbon dioxide fire extinguishers are used [17].

To protect against fires, it is necessary to have such a fire extinguisher available, equipment such as fire cabinets, fire shields and fire extinguishers. Employees must be familiar with the use of such equipment. Employees must know the evacuation plan from the premises, the location of exits from the building. It is also necessary to carry out planned evacuation from the building in order to prepare employees for actions in emergency. To prevent a fire in the production area, it is necessary:

- work should only be carried out if the electrical equipment;
- the power grid should not be overloaded simultaneously by several powerful consumers of electricity;
- the last person to leave the premises must check heaters, electrical appliances, equipment, etc.

In the event of a fire, it is advisable to extinguish it independently only at its early stage when a fire is detected.

According to the Decree of the Government of the Russian Federation of September 16, 2020 N1479 "On the approval of the Rules for the fire regime in the Russian Federation" If a fire or signs of burning (smoke, smell of burning, temperature increase) in the production room or on the territory of the enterprise, the employee is obliged to immediately report this to fire protection. The fire brigade is informed of the object's address and location, the occurrence of a fire. Notify the fire brigade even if the fire is extinguished on its own. fire can go unnoticed in hidden places (in the voids of wooden floors and baffles, etc.), and subsequently combustion may resume. Further it is necessary to take, if possible, measures to evacuate people, extinguish fire and safety of material assets [18].

5.4 Conclusion of social responsibility

Thus, in this section, the issues observance of the rights of personnel to work, compliance with the requirements for occupational safety and health, industrial safety, environmental protection and resource conservation.

It was found that the developed system does not affect the atmosphere and hydrosphere, but in order to prevent pollution of the lithosphere, it is necessary to dispose of waste during the study. Also, during the work, it was found that the object of research or the performance of research in the laboratory can initiate the occurrence of such an emergency as a fire.

In addition, dangerous and harmful factors were identified that may arise at different stages of work, namely: Increased levels of electromagnetic radiation, Insufficient illumination of workplace, Excessive noise, Increased / decreased air humidity in the workplace, physical overload (static - long-term preservation of a certain posture), overstrain of analyzers (vision), Increased voltage in an electrical circuit, the closure of which can pass through the human body.

The results obtained are also recommendations for elimination and prevention of fires, waste disposal and elimination of dangerous and harmful factors considered in this paper. Calculations were made to create illumination E = 300 lux for the laboratory room in which the library was made. There were also requirements to the organization of a workplace and the organization of work are formulated. These recommendations and requirements can be implemented at enterprises where the use of the considered library is planned.

5.5 Reference of social responsibility

1. СП 2.4.3648-20 Санитарно-эпидемиологические требования к организациям воспитания и обучения, отдыха и оздоровления детей и молодежи.
2. Трудовой кодекс Российской Федерации от 30.12. 2001 г. № 197– ФЗ (ред. от 01.04.2019 г.). – М., 2015. – 123 с.
3. Федеральный закон от 27.07.2006 N 152-ФЗ (ред. от 30.12.2020) “О персональных данных.”
4. SanPiN 2.2.2 / 2.4.1340-03 Гигиена труда, технологические процессы, сырье, материалы, оборудование, рабочий инструмент.
5. Labor Code of the Russian Federation Article 349.1. Features of labor regulation of employees of state corporations, public companies, state companies.
6. GOST 12.0.003–2015 "Hazardous and harmful production factors. Classification".
7. GOST 12.1.030-81 Electric safety. Protective Conductive earth, neutralling.
8. SanPiN2.2.1/2.1.1.1278-03 Hygienic requirements for natural, artificial and mixed lighting ofresidential and public buildings.
9. GOST 12.1.003-2014 Occupational safety standards system. Noise. General safety requirements.
10. GOST 12.1.005-88 General sanitary requirements for working zone air.
11. GOST 9241-4-2009 Ergonomic requirements for office work with visual display terminals (VDT). Part 4. Keyboard requirements.
12. GOST 9241-4-2009 Ergonomic requirements for office work with visual display terminals (VDTs). Part 5. Workstation layout and postural requirements.
13. GOST 12.1.019-2017 Electrical safety. General requirements and nomenclature of types of protection.
14. ГОСТ 55102-2012 Ресурсосбережение. Обращение с отходами. Руководство по безопасному сбору, хранению, транспортированию и разборке 85 отработавшего электротехнического и электронного оборудования, за исключением ртутьсодержащих устройств и приборов.
15. Постановление Правительства РФ от 03.09.2010 № 681 (ред. от 01.10.2013) Об утверждении Правил обращения с отходами производства и потребления в части осветительных устройств, электрических ламп, ненадлежащие сбор, накопление, использование, обезвреживание.
16. НПБ 105-03 Определение категорий помещений, зданий и наружных установок по взрывопожарной и пожарной опасности.
17. СП 9.13130.2009 Техника пожарная. ОГНЕТУШИТЕЛИ. Требования к эксплуатации.
18. Постановление Правительства РФ от 16 сентября 2020 г. № 1479 Об утверждении Правил противопожарного режима в Российской Федерации.

Conclusion

In this thesis have created the prediction of S/D model by using python programming including the following step: EDA, feature engineering and data cleaning, predictive modeling. After step 1 and 2, we got the processed dataset with 10 useful features to start modeling. We selected 4 machine learning algorithm (LG, DT, RF, KNN) to create models, then used cross validation to find best model - random forest because it has highest score of CV accuracy. Last extracted the importance of each feature and finished project of data analysis of Titanic dataset.

Reference

1. Amy Tikkanen. 2020.“Millionaire’s Special”, “RMS Titanic”, “Royal Mail Ship Titanic”.
2. Rachel Wolff. 2020. 5 Types of Classification Algorithms in Machine Learning, UML: https://monkeylearn.com/blog/classification-algorithms/
3. John T. Behrens. 1997. Principles and procedures of exploratory data analysis.
4. Guillermo L. Taboada, Isabel Seruca, Cristina Sousa, Ángeles Pereira. 2020. Exploratory Data Analysis and Data Envelopment Analysis of Construction and Demolition Waste Management in the European Economic Area.
5. Huang Shan, Gubin E.I. Data cleaning for data analysis // Молодежь и современные информационные технологии: Труды XVI Междунар. научно - практической конференции студентов, аспирантов и молодых ученых.
6. R.O.Sinnott, H.Duan, Y.Sun. 2016. A Case Study in Big Data Analytics: Exploring Twitter Sentiment Analysis and the Weather // Big Data Principles and Paradigms - C. 387-389.
7. Kamal Bunkar, Umesh Kumar Singh, Bhupendra Pandya, Rajesh Bunkar. 2012. Data mining: Prediction for performance improvement of graduate students using classification, UML: https://ieeexplore.ieee.org/document/6335530
8. J Gehrke, R Ramakrishnan, V Ganti. 2000. Rainforest - a framework for fast decision tree construction of large datasets, UML: http://citeseer.ist.psu.edu/showciting?cid=55374&sort=cite&start=20
9. Kadyr Arailym. 2020. Kadyr_Arailym_report.docx.
10. Zhongheng Zhang. 2016. Introduction to machine learning: k-nearest neighbors.

Appendix A. Program code of predictive modeling

```
#importing all the required ML packages

from sklearn.linear_model import LogisticRegression #logistic
regression

from sklearn.ensemble import RandomForestClassifier #Random Forest

from sklearn.neighbors import KNeighborsClassifier #KNN

from sklearn.tree import DecisionTreeClassifier #Decision Tree

from sklearn.model_selection import train_test_split #training and
testing data split

from sklearn import metrics #accuracy measure

from sklearn.metrics import confusion_matrix #for confusion matrix


train,test=train_test_split(data,test_size=0.3,random_state=0,stratify
=data['Survived'])

train_X=train[train.columns[1:]]

train_Y=train[train.columns[:1]]

test_X=test[test.columns[1:]]

test_Y=test[test.columns[:1]]

X=data[data.columns[1:]]

Y=data['Survived']


model = LogisticRegression()

model.fit(train_X,train_Y)

prediction3=model.predict(test_X)

print('The accuracy of the Logistic Regression
is',round(metrics.accuracy_score(prediction3,test_Y),3))
```

```
from sklearn.metrics import roc_auc_score

from sklearn.metrics import roc_curve


logit_roc_auc = roc_auc_score(test_Y, model.predict(test_X))

fpr, tpr, thresholds = roc_curve(test_Y,
model.predict_proba(test_X)[:,1])

plt.figure()

plt.plot(fpr, tpr, label='Logistic Regression (area = %0.2f)' %
logit_roc_auc)

plt.plot([0, 1], [0, 1],'r--')

plt.xlim([0.0, 1.0])

plt.ylim([0.0, 1.05])

plt.xlabel('False Positive Rate')

plt.ylabel('True Positive Rate')

plt.title('Receiver operating characteristic')

plt.legend(loc="lower right")

#plt.savefig('Log_ROC')

plt.show()

model=DecisionTreeClassifier()

model.fit(train_X,train_Y)

prediction4=model.predict(test_X)

print('The accuracy of the Decision Tree
is',round(metrics.accuracy_score(prediction4,test_Y),3))


model=RandomForestClassifier(n_estimators=100)
```

```
model.fit(train_X,train_Y)

prediction7=model.predict(test_X)

print('The accuracy of the Random Forests
is',round(metrics.accuracy_score(prediction7,test_Y),3))


model=KNeighborsClassifier()

model.fit(train_X,train_Y)

prediction5=model.predict(test_X)

print('The accuracy of the KNN
is',round(metrics.accuracy_score(prediction5,test_Y),5))


a_index=list(range(1,11))

a=pd.Series()

x=[0,1,2,3,4,5,6,7,8,9,10]

for i in list(range(1,11)):

    model=KNeighborsClassifier(n_neighbors=i)

    model.fit(train_X,train_Y)

    prediction=model.predict(test_X)

    a=a.append(pd.Series(metrics.accuracy_score(prediction,test_Y)))

plt.plot(a_index, a)

plt.xticks(x)

fig=plt.gcf()

fig.set_size_inches(12,6)

plt.show()

print('Accuracies for different values of n are:',a.values,'with the
max value as ',a.values.max())
```

```
from sklearn.model_selection import KFold #for K-fold cross validation

from sklearn.model_selection import cross_val_score #score evaluation

from sklearn.model_selection import cross_val_predict #prediction

kfold = KFold(n_splits=10, random_state=22) # k=10, split the data
into 10 equal parts

xyz=[]

accuracy=[]

std=[]

classifiers=['Logistic Regression','KNN','Decision Tree','Random
Forest']

models=[LogisticRegression(),KNeighborsClassifier(n_neighbors=9),Decis
ionTreeClassifier(),RandomForestClassifier(n_estimators=100)]

for i in models:

    model = i

    cv_result = cross_val_score(model,X,Y, cv = kfold,scoring =
"accuracy")

    cv_result=cv_result

    xyz.append(cv_result.mean())

    std.append(cv_result.std())

    accuracy.append(cv_result)

new_models_dataframe2=pd.DataFrame({'CV
Mean':xyz,'Std':std},index=classifiers)

new_models_dataframe2


from sklearn.model_selection import GridSearchCV
```

```
n_estimators=range(100,1000,100)

hyper={'n_estimators':n_estimators}

gd=GridSearchCV(estimator=RandomForestClassifier(random_state=0),param
_grid=hyper,verbose=True)

gd.fit(X,Y)

print(gd.best_score_)

print(gd.best_estimator_)


f,ax=plt.subplots(2,2,figsize=(15,12))

model=RandomForestClassifier(n_estimators=900,random_state=0)

model.fit(X,Y)

pd.Series(model.feature_importances_,X.columns).sort_values(ascending=
True).plot.barh(width=0.8,ax=ax[0,0])

ax[0,0].set_title('Feature Importance in Random Forests')
```