1DL034
Introduction to Machine Learning
2021 Spring
Project Report

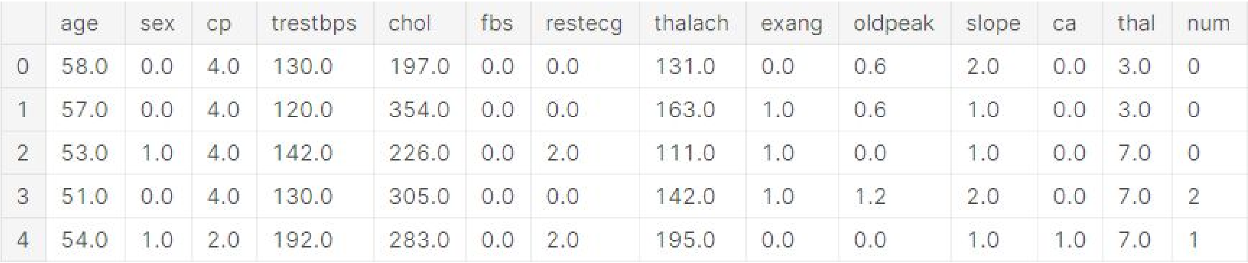
UPPSALA
UNIVERSITET

Dora Akbulut
Polina Gorodnichina
M. Tarik Demir

1. The Dataset

Our dataset has 153 rows and 14 rows, including the target row.

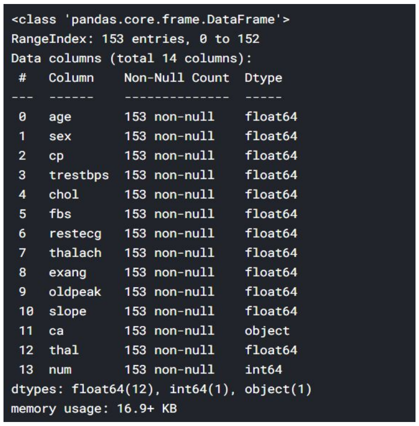
This is what our dataset initially looks like:



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	58.0	0.0	4.0	130.0	197.0	0.0	0.0	131.0	0.0	0.6	2.0	0.0	3.0	0
1	57.0	0.0	4.0	120.0	354.0	0.0	0.0	163.0	1.0	0.6	1.0	0.0	3.0	0
2	53.0	1.0	4.0	142.0	226.0	0.0	2.0	111.0	1.0	0.0	1.0	0.0	7.0	0
3	51.0	0.0	4.0	130.0	305.0	0.0	0.0	142.0	1.0	1.2	2.0	0.0	7.0	2
4	54.0	1.0	2.0	192.0	283.0	0.0	2.0	195.0	0.0	0.0	1.0	1.0	7.0	1

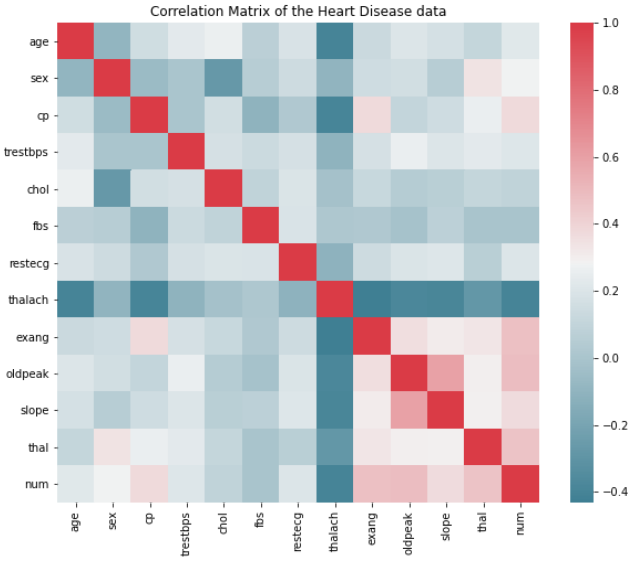
Continuous features: age, trestbps, chol, thalach, oldpeak.
Categorical features: sex, cp,fbs, restecg, exang, slope, ca, thal
Our target is also categorical.

Then, we checked for null values. We have no null values as NaN. We decided to dig a bit deeper to see if the null values existed in another format. They did indeed. The *"ca"* feature had a "?" value in three rows. Unique values that *"ca"* had: ['0.0' '1.0' '?' '3.0' '2.0']. We imputed the missing values later on, during our preprocessing steps.



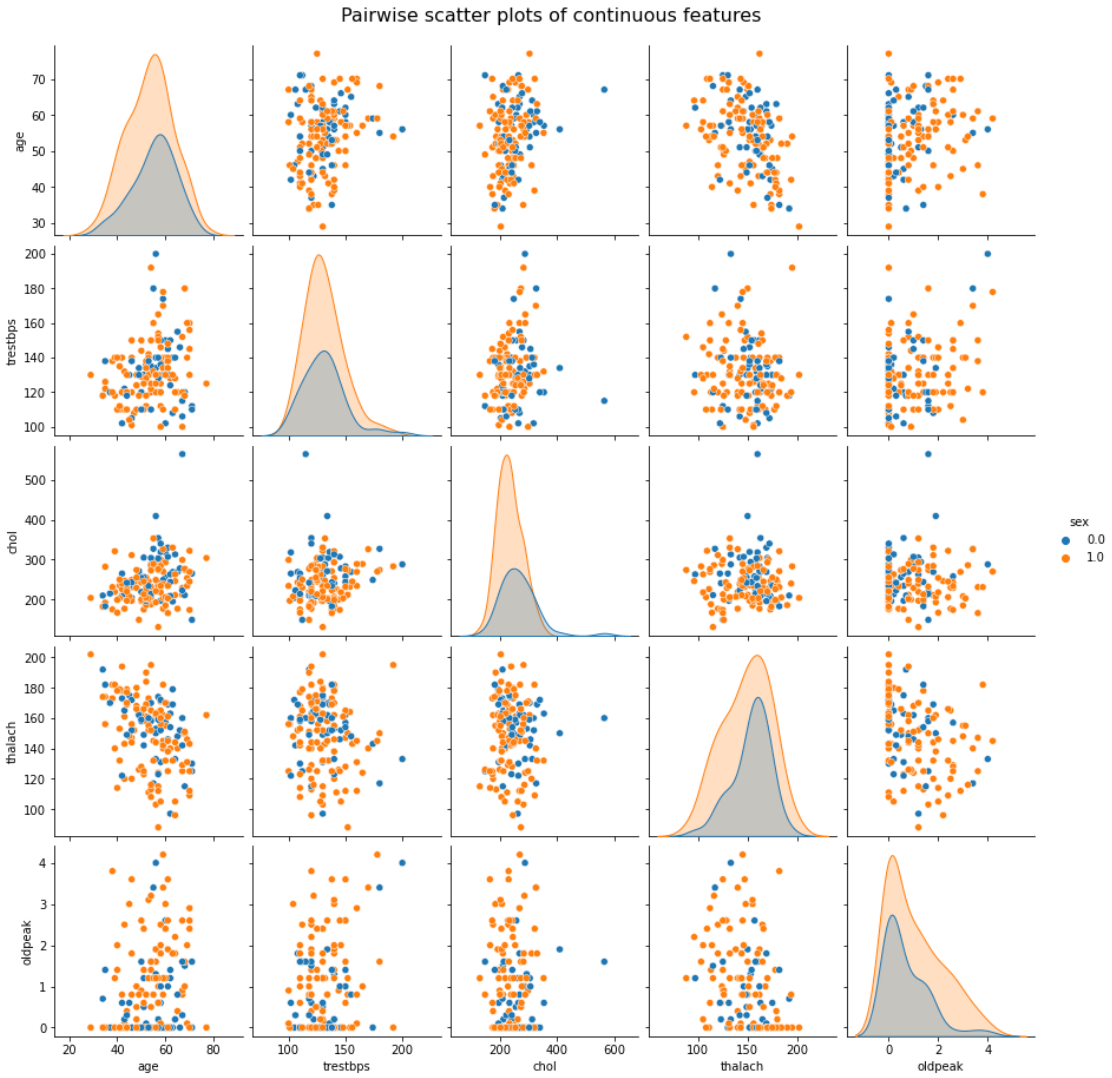
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 153 entries, 0 to 152
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       153 non-null    float64
 1   sex       153 non-null    float64
 2   cp        153 non-null    float64
 3   trestbps  153 non-null    float64
 4   chol      153 non-null    float64
 5   fbs       153 non-null    float64
 6   restecg   153 non-null    float64
 7   thalach   153 non-null    float64
 8   exang     153 non-null    float64
 9   oldpeak   153 non-null    float64
 10  slope     153 non-null    float64
 11  ca        153 non-null    object
 12  thal      153 non-null    float64
 13  num       153 non-null    int64
dtypes: float64(12), int64(1), object(1)
memory usage: 16.9+ KB
```

2. Exploratory Analysis



The Correlation Matrix on the left shows the correlation between features. So, the most negatively correlated features are "max heart rate achieved (thalach)" and "age", which suggests that the maximum pressure indicators are highly dependent on a person's age.

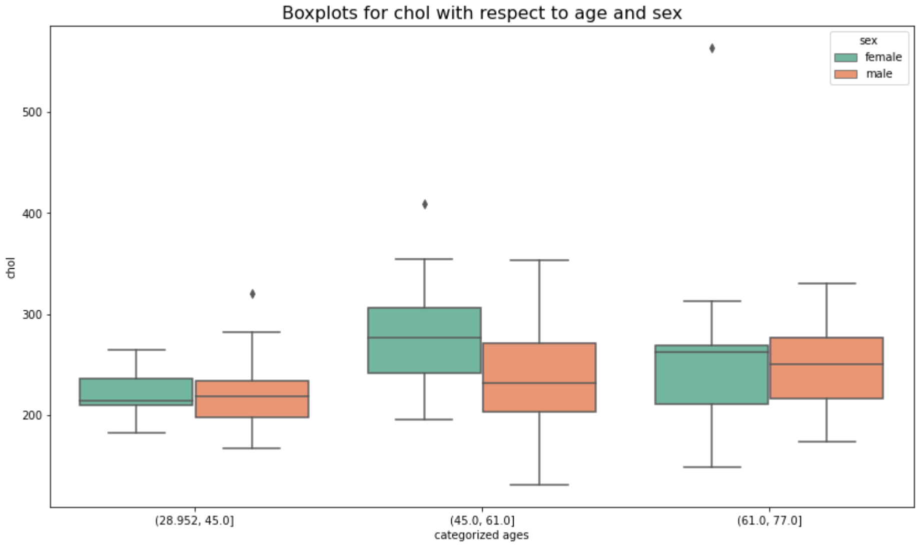
The most meaningful connection between diagnosis of heart disease is observed in the presence or absence of detected angina (exang), angina ST indicators (oldpeak), as well as thalassemia (thal), and peak exercise ST segment (slope).



These plots show the correlation of continuous features to each other. So, for example, deviating indicators for "Serum Cholesterol" are observed in people over 60. Also, the "resting blood pressure" is fairly uniform with an increase in the "max heart rate".

The most frequent indicators of "resting blood pressure" are observed at 3-4 ST depression types induced by exercise relative to rest, which suggests a general increase in angina pectoris in these types of diagnostics.

This graph also shows the distribution of all indicators for heart disease relative to comparison by gender (1 (orange) - men, 0 (blue) - women). For example, the "Serum Cholesterol" score is generally lower in men than in women. Interesting that only the second 'num' type has an ST-T wave abnormality in a woman because the rest of the types are manifested in men. Also interesting that any kind of sloping with resting blood pressure is observed in men on average, and in deviating (low and high) values are more typical for women

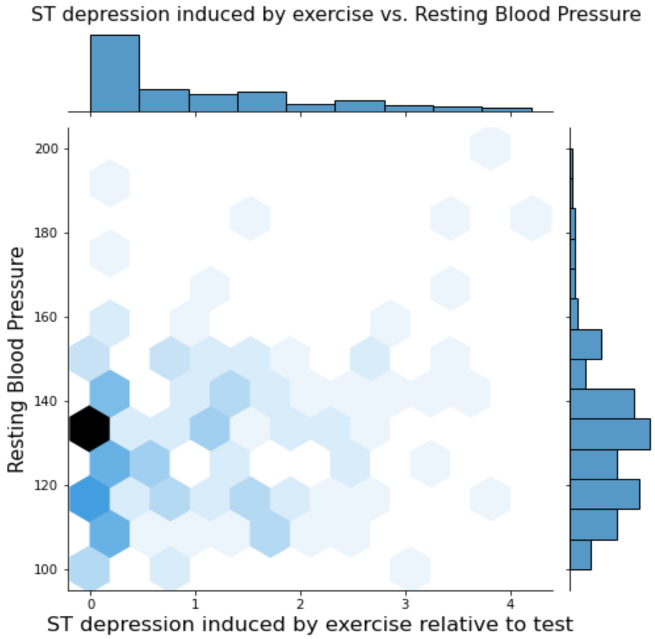


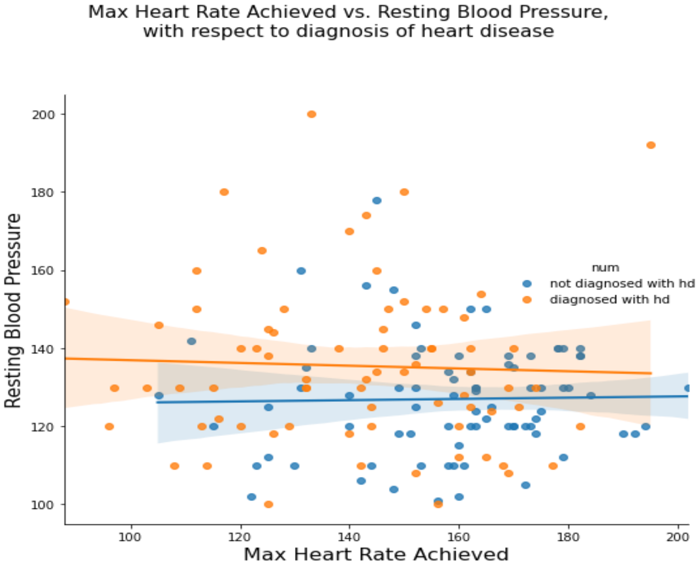
This plot shows the distributions of cholesterol concentrations, compared between groups of age and sex. It can be seen that men and women have similar levels of cholesterol, except for the age group of [45, 61].

The most common value that appeared for people with low angina at rest is 130-135 mmHg (interestingly the normal pressure is supposed to be much lower, around 120.)

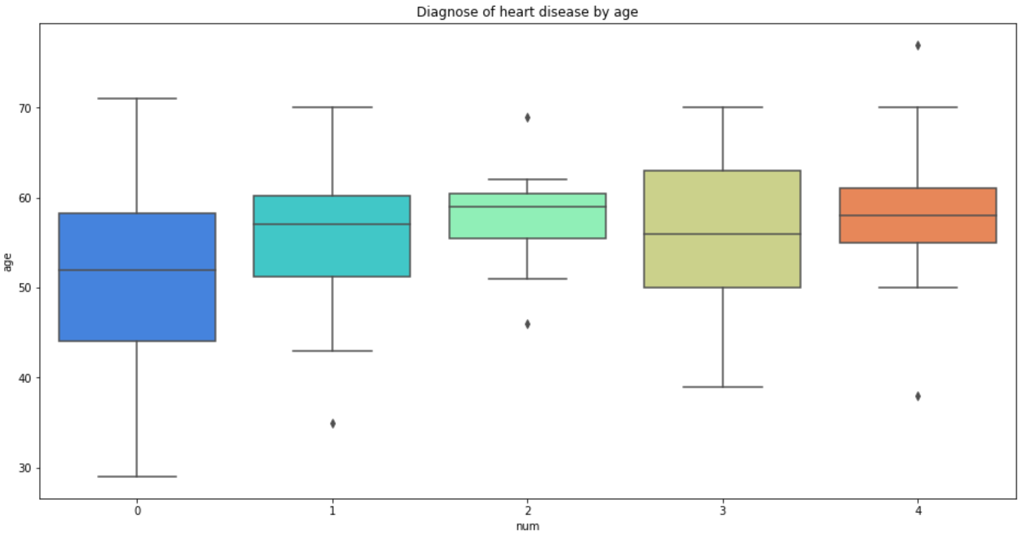
As the risk of high resting blood pressure increases, the number of people who had these symptoms decreased.

For example, there were cases of high (180-200 mmHg) blood pressure, but the number of people with these indicators was very few.
It's in general indicates a low risk of oldpeak and ST depression with a normal average of blood pressure



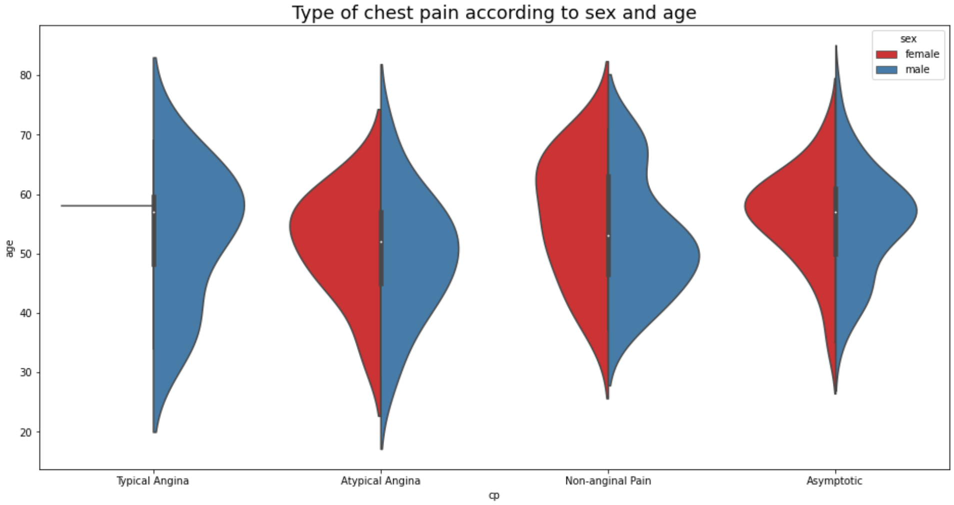


This plot shows that healthy people have an increase in “resting blood pressure”, that provokes “max heart rate”, while diagnostic unhealthy patients have this value decreased, which indicates abnormalities in the cardiovascular system.



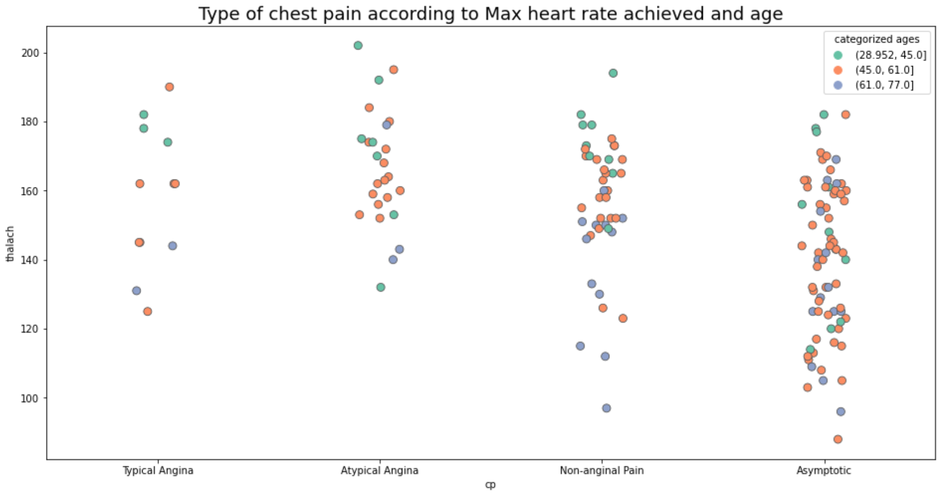
This boxplot represents the category of heart disease with respect to age. The first scatter of values is observed in representatives with no diseases and an average age of about 52. The age cohort spread presented on the data and ranges from 45 to 59 years old

The most common average age for heart disease in any of the diagnoses is 55-57 years. In most diagnostics, The 25% interval has a greater age range compared to 75%. This shows that people older than 55 have a much larger possibility of having heart disease.

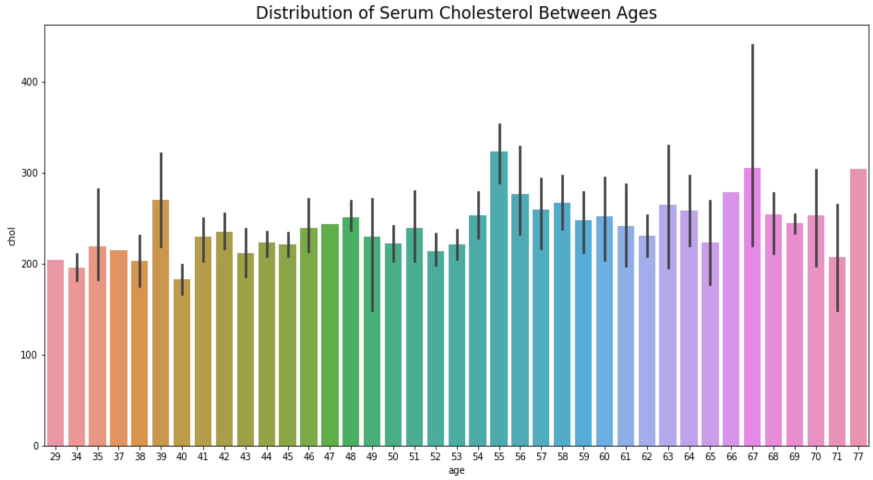


According to our training data, women practically do not get typical angina, but they are more susceptible to serious diseases compared to men (55-60 year old people are most susceptible to experiencing chest pain).

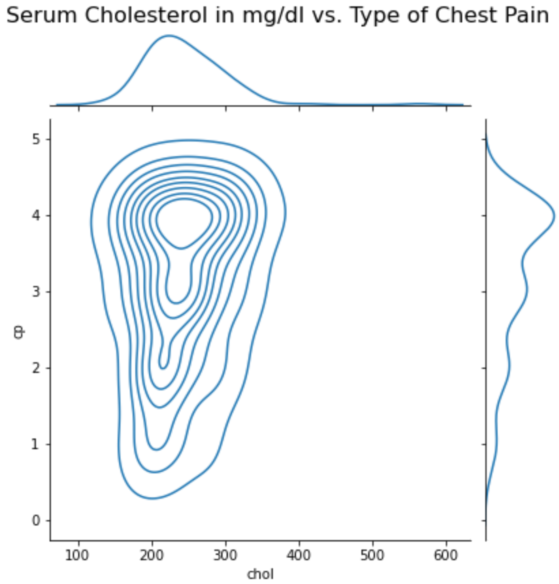
Around 45 years old men are highly prone to anginal pain.



This is the plot of the connection between maximum heart rate and age. It can be observed that people with asymptotic chest pain, on average, have a lower “max heart rate”. The highest rates are shown by patients with atypical angina.



This graph shows the patient's age and the amount of Serum Cholesterol. The highest indicators of “chol” are observed at 55 years old patients, then, there is a gradual decrease until reaching 67-77 years interval, where the indicators start to increase again. The general trend shows that the number of Serum Cholesterol on average increases with age, despite deviating cases when this indicator reaches the norm (or even below the norm) at the age of over 50.



This plot shows the relationship between the distribution of cholesterol among different types of chest pains. The concentration of cholesterol with 200 mg/dl is grouped around asymptotic angina, compared to others.

Deviations in the direction of decreasing and increasing Serum Cholesterol are most often observed in asymptotic and non-anginal pain. It indicates a distribution of data, where are the most frequent cases with asymptotic disease and least with common typical angina.

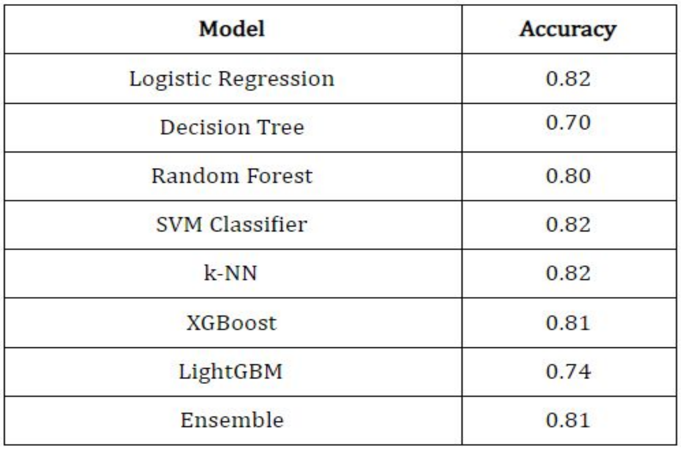
3. Preprocessing

To preprocess our data, we implemented a Preprocessing Pipeline which handled One-Hot encoding and data scaling. We could not find a way to impute the missing values within our pipeline, so we did it by hand for training and test data. Our preprocessing steps were as follows:

1. Imputation of the missing categorical values with the mode, and continuous values with the median.
2. Fixed the target value *“num”* from [0, 1, 2, 3, 4] to [0, 1].
3. We applied One-Hot Encoding to our categorical data.
4. We applied to scale to our continuous data.

4. Machine Learning Models

To evaluate the performance of several machine learning models on the training data, we performed 5 fold cross-validation and checked the average accuracy of the models. The next table shows the average accuracy for each model.



Model	Accuracy
Logistic Regression	0.82
Decision Tree	0.70
Random Forest	0.80
SVM Classifier	0.82
k-NN	0.82
XGBoost	0.81
LightGBM	0.74
Ensemble	0.81

With these results, we chose the models with the highest scores and did a Grid Search to get the best results. The only exception is k-NN, which we did not use on the test data. This is because k-NN does not output probabilities in the desired format.

Here, LightGBM and XGBoost models are gradient boosting algorithms, which use an ensemble of decision trees with different approaches.

We used all of the features that are initially given since we did not observe any improvement in accuracy when we discarded the least fitting features from training data.

5. Choosing the Best Hyperparameters

We ran Grid Search to find the best hyperparameters for our high performance. We could not run the Grid Search on XGBoost and Ensemble Learning because we received many errors during the hyperparameter testing. The following table shows the accuracy we received with the models we used Grid Search with.

Model with GridSearchCV Parameters	**Accuracy**
Logistic Regression	0.83
Random Forest	0.82
SVM Classifier	0.83

There is not much of a big difference between the performance of our models. The best model on the test data was the Logistic Regression Model.

In the course of this work, we mastered and improved our skills of working with training and test models, mastered the system of working with visual components and learned more about machine learning models for supervised machine learning.