FINAL PROJECT

Concepts & Technologies of Artificial Intelligence

Submission Date: 02/05/2023

Project Start Date and End Date: March-May

Submitted by: Mehmet Omer DEMIR

Faculty of Science and Engineering

University of Wolverhampton
7CS070/UZ1: Concepts & Technologies of Artificial Intelligence

Instructor: Ahmed Khubaib

Email: M.O.Demir@wlv.ac.uk

Table of Contents

1.Regression Task- House price data from King County

1.1.Introduction

For the purpose of predicting house sale prices in King County United States, this project aims at developing prediction algorithms based on machine learning. The data is download in Kaggle, and you can download it from the following link. The information on the sale of houses between May 2014-May 2015 is included in this report. This page provides a complete description of this dataset with the meanings for all variables. As this is a classic regression problem, it's very likely we will be able to apply many methods of achieving our objective.

1.2.Problem Statement

The real estate industry not only constitutes a crucial component of the national economy, but also commands significant attention among citizens as a major area of interest.
The escalating demand for housing has resulted in a commensurate rise in house prices.
The provision of precise prognostications concerning the pricing of residences is of utmost significance.
The establishment of appropriate and justified property prices has the potential to engender a significant degree of transparency and credibility within the real estate industry, a crucial factor for a vast majority of its consumers.

Despite the potential benefits of using AI to predict housing prices, some challenges are worth considering. One of these challenges is the availability and quality of data. Predicting house prices acurately requires large amounts of clean data about many variables that affect price, including location, number of room, size, age, condition and local economic conditions. Collecting and analyzing all this data can be costly.

1.3.Dataset

.The correlation heatmap shows how different features are related to the log price in US dollars. The features are arranged based on how much they affect the price.

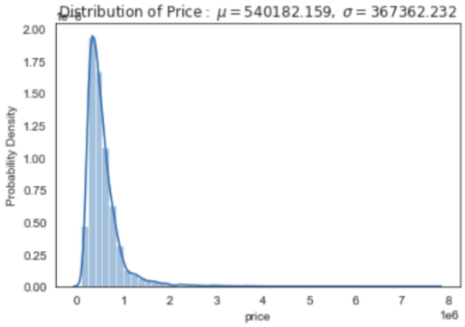


Figure: Price and Probability Density.

We can see from the correlation heatmap how important sqft_living and grade are in describing housing values.

Strong correlations between predictors can also be observed by examining the correlation heatmap more closely.

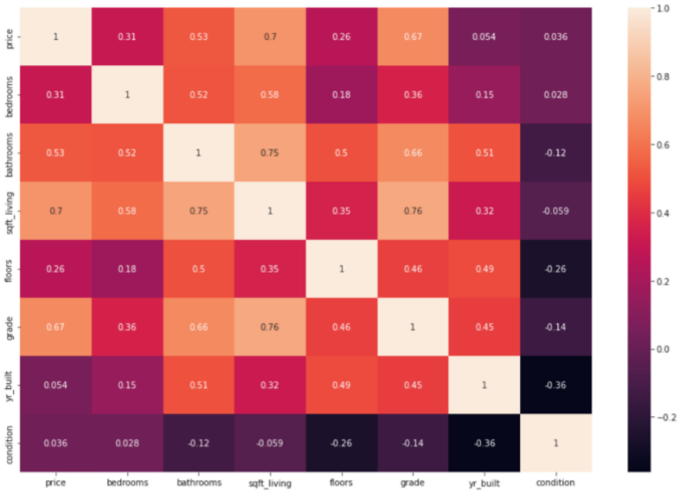


Figure: Heat map.

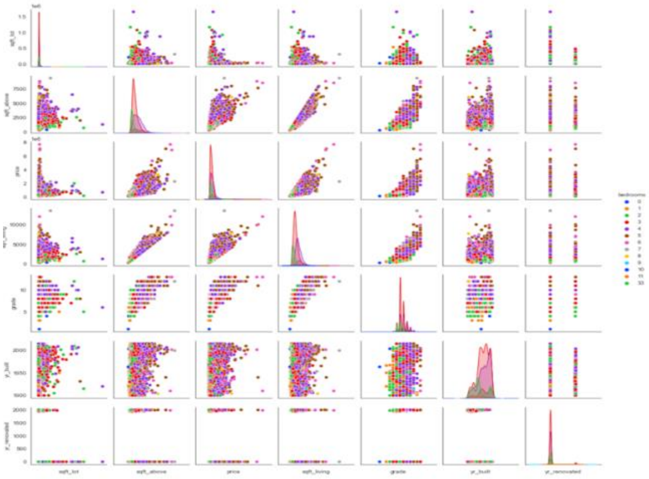


Figure: Scatter plot.

1.4.Model

- ### Decision Tree Regression(DTR)

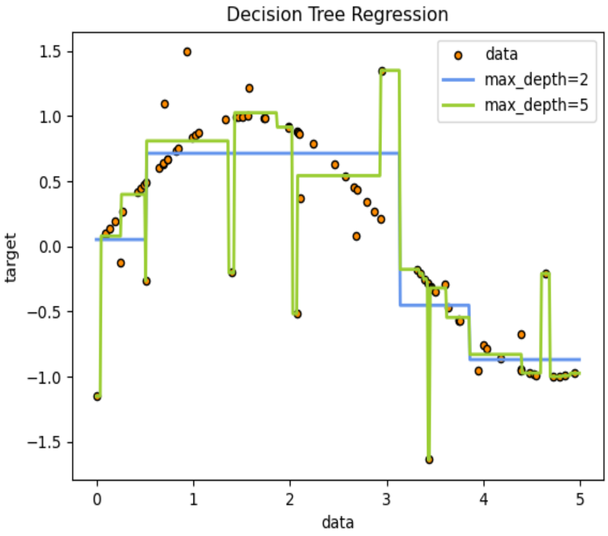


Figure 1: DTR

DTR is a machine learning technique used to discover relationships between input variables and make predictions. This method creates a series of decision trees to understand and summarize the structure of the data. Decision trees are created by dividing the input features using a decision rule at each node. Each node allocates an input feature based on a certain threshold value and thus creates new child nodes**[7]**. This process continues depending on the number and characteristics of child nodes, and eventually a leaf node is reached. The leaef nodes represent the outcome of the decision tree and contain the predicted output values.

Decision tree regression is especially effective when the data has a nonlinear structure. For example, when you want to build a salary forecasting model, you can use decision tree regression to understand how different characteristics (age, gender, education, work experience, etc.) affect salary. This technique creates a model that can predict any value of any independent variable (input property) in a data set.

Decision tree regression provides high accuracy and intelligibility. Its outputs are easily interpretable and provide information on how the model predicts. However, it is a technique that can have overfitting problems. Overfitting means that the model fits the training data too tightly and loses its generalizability. Therefore, it is important to set up the model correctly and check its accuracy using test data.

- Random Forest Regression(RFR)

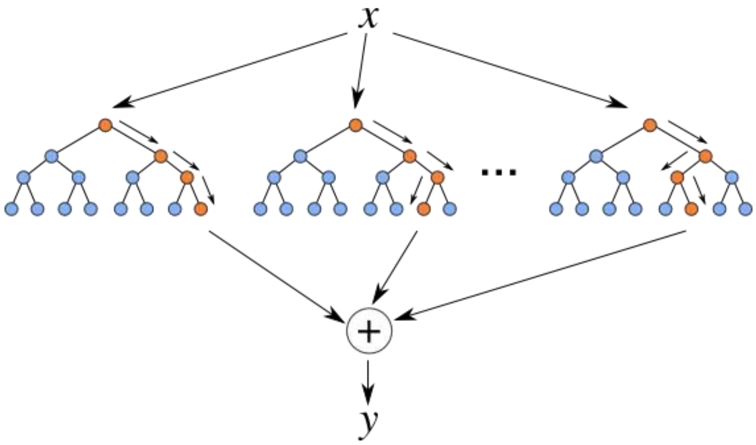


Figure 2:RFR Work Map

RFR is a machine learning algorithm used to predict a numerical target variable, such as house prices. This algorithm is used to combine many decision trees to make a more accurate and consistent forecast.

The Random Forest Regression algorithm first splits the data set into random pieces and constructs a decision tree in each piece. Although each tree has the same variables, it learns using a different subset of data. This allows each tree to learn from a different perspective.

For example, to predict the price of a house, the algorithm first randomly splits the dataset and builds a decision tree in each piece. The nodes of each tree are divided according to the different properties of the dataset**[1]**. These features include the layout of the house, its size, the number of rooms, the size of the garden, etc. it could be. Each tree learns using a different subset of the dataset. This allows each tree to learn uniquely and focus on different features.

Then, to make the prediction, the algorithm feeds data to all the trees, and each tree makes its own prediction. Finally, the final estimate is obtained by averaging the estimates of all trees.

RFR combines many trees, giving it the ability to make a more accurate prediction. Also, learning each tree using a different subset of data reduces the problem of overfitting and helps to obtain a more generalized model.

As a result RFR is a very effective algorithm for estimating numerical target variables such as house price estimation.

1.5.Algorithm Comparison and Conclusion

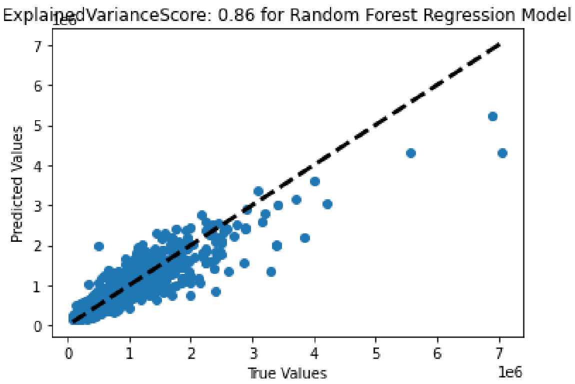


Figure 3: RFR Graph of House price

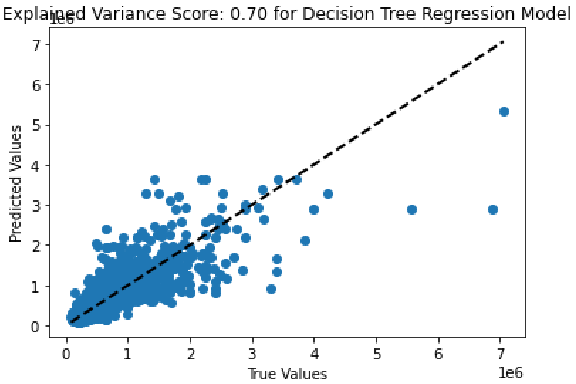
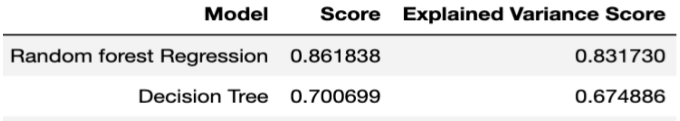


Figure 4: DTR Graph of House Price



Model	Score	Explained Variance Score
Random forest Regression	0.861838	0.831730
Decision Tree	0.700699	0.674886

Figure 5: Model Comparison

Random Forest Regression(RFM)(**Figure 3**) is more accurate than Decision Tree Regression(DTR)(**Figure 4)** in house price prediction because it is an ensemble method that uses multiple decision tree to make prediction.

In Decision Tree Regression, a single tree is built to make predictions based on the features of the dataset. This can lead to overfitting or underfitting of the data, which can result in inaccurate predictions.

RFM, on the other hand, creates multiple decision trees on different random subsets of the dataset and then combines the predictions of those trees to make the final prediction**[10]**.This reduces the overfitting and underfitting of the data and improves the accuracy of the model. Additionally, RFR also reduces the variance of the model and provides a better generalization ability.

Therefore, RFM (**Figure 3**) is more accurate than DTR (**Figure 4)** in house price prediction because it uses an ensemble of decision trees to make predictions, which results in improved accuracy and better generalization ability.

1.6.Adding New Relevant Features

We could try a number of strategies to potentially enhance the models for forecasting house prices. In order to remove any features with low correlation, we could first examine the correlation between each feature and the target variable. We might also try including new elements, like crime rates, school rankings, or accessibility to public transport, that may have a significant effect on the target variable. Additionally, we might think about gathering more information to expand our dataset, which might enhance the precision of our models. Last but not leest, we could test various algorithmas or hyperparameters to see if we can get better outcomes. To better understand the connections between the features and the target, we should carefully analyse the data and use visualisations before making any changes.

2.Classification Task - Heart Condition

2.1 .Introduction

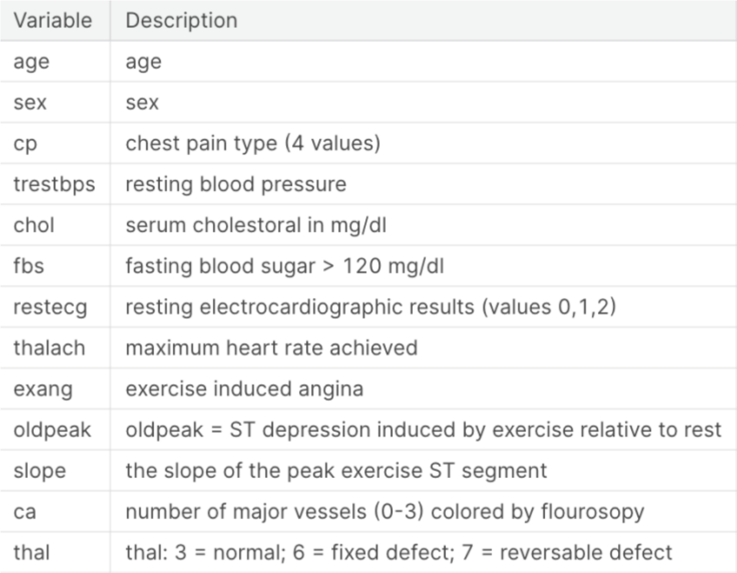
A major factor in the death of many adults in recent years is heart disease. Our project can combine a number of criteria and offer help to predict individuals who are most likely to analyze with heart disease by offering the help of their medical history. It quickly recognizes anyone who triggers any heart disease, such as high blood pressure or chest pain heaviness, and can offer a high degree of assistance in diagnosing the disease with fewer medical tests and effective accurate medications so that they can be optimally treated.**[7].**

2.2 .Problem Statement

This makes heart disease a major concern to be managed with. But it is difficult to identify heart illness since of a few contributory chance components such as diabetes, high blood weight, high cholesterol, irregular beat rate, and numerous other variables**(figure 2.1)**. Because of such imperatives, researchers have turned to today's technological approaches such as Knowledge Mining and Machine Learning to predict disease quickly.In this project, I will apply Machine Learning approaches KNN and SVM to classify whether a person is suffering from heart disease or not. I got the dataset I use for this project from the homework project page.

2.3 .Dataset

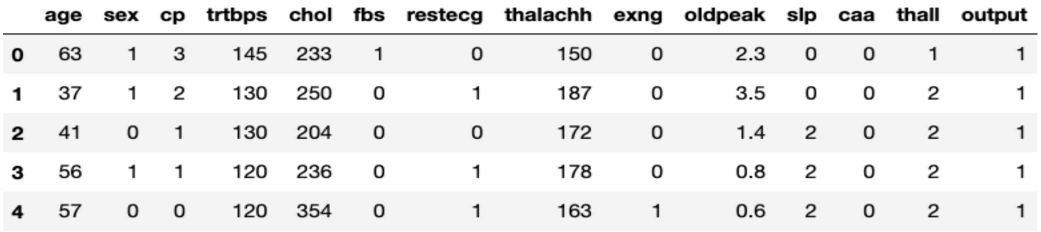
This information is related to people's health and indicates whether they are going to have a heart attack or not. We'll look at the information, examine the dataset, and use different Machine Learning models to sort the target variable. At the end of the project we will discover which algorithm works best for this dataset.



Variable	Description
age	age
sex	sex
cp	chest pain type (4 values)
trestbps	resting blood pressure
chol	serum cholestoral in mg/dl
fbs	fasting blood sugar > 120 mg/dl
restecg	resting electrocardiographic results (values 0,1,2)
thalach	maximum heart rate achieved
exang	exercise induced angina
oldpeak	oldpeak = ST depression induced by exercise relative to rest
slope	the slope of the peak exercise ST segment
ca	number of major vessels (0-3) colored by flourosopy
thal	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

Figure 2.1: Descripton of Dataset.

This data contains information about 303 people. The dataset has 14 parts called columns, and they are showed below(**Figure 2.2)**.



	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figure 2.2: Contains of Dataset.

First, let's look at a graph of the ages of people who may or may not suffer from the disease. Here, output = 1 indicates that the person is suffering from heart disease and output = 0 indicates that the person is not suffering **(Figure 2.3)**.

Then, to better interpret our data set after disease distribution, we should look at the age range **(Figure 2.4)**.

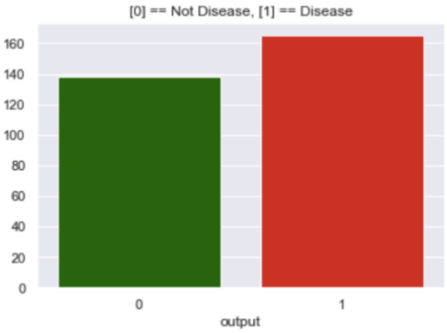


Figure 2.3: Number of diseases.

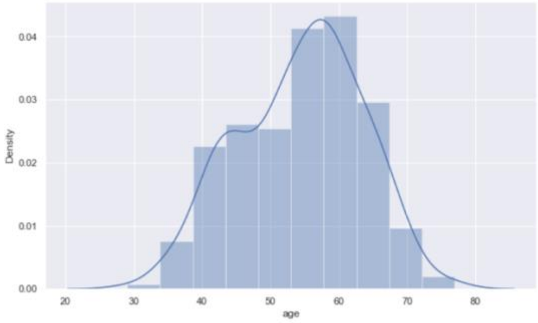


Figure 2.4: Age distribution of the data set.

Since the metabolism of elderly people is weaker than the young, they are more prone to heart attack(**Figure 2.4)**. Therefore, I will divide our dataset into 3 groups, these are young middle elderly.

After dividing the data set into 3 groups, the number of young people is 16, the number of middle-aged people is 128, the number of seniors is 151. We can see the distribution as a percentage using the peiplot method **(Figure 2.5)**.

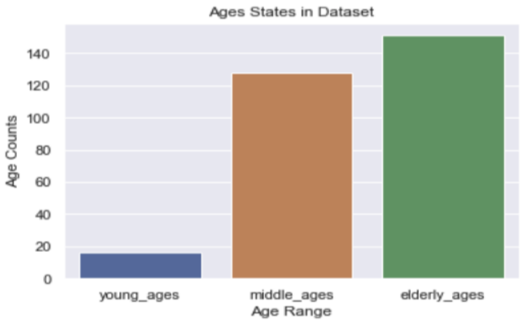


Figure 2.4: Age states.

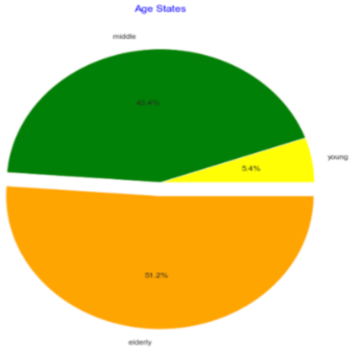


Figure 2.5: Age states in percent.

Then we look at the distribution of the results of the fluoroscopy device by gender, this value varies between 0-3 **(Figure 2.6)**.

We look at the distribution of the disease by sex and then we can examine the heart attack graph by age range**(Figure 2.7)(Figure 2.8)**.

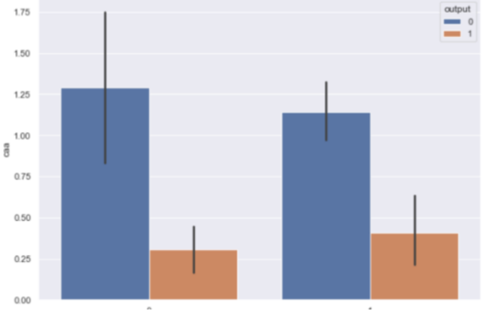


Figure 2.6: fluoroscopy rates by gender.

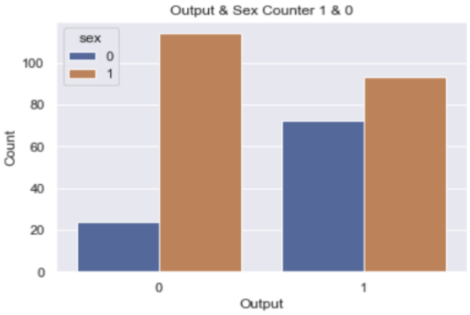


Figure 2.7: Disease by gender.

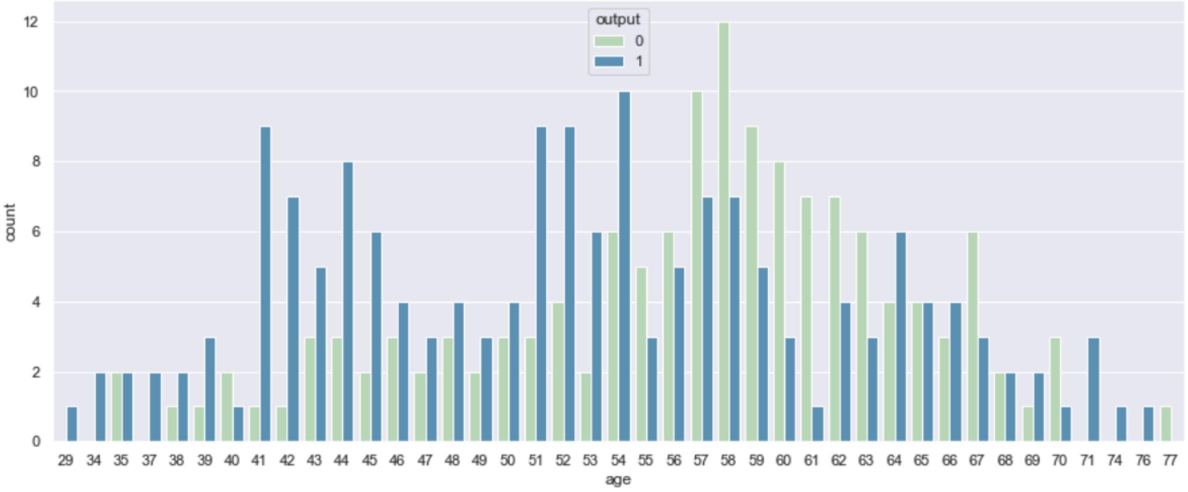


Figure 2.8: Age-related heart attack.

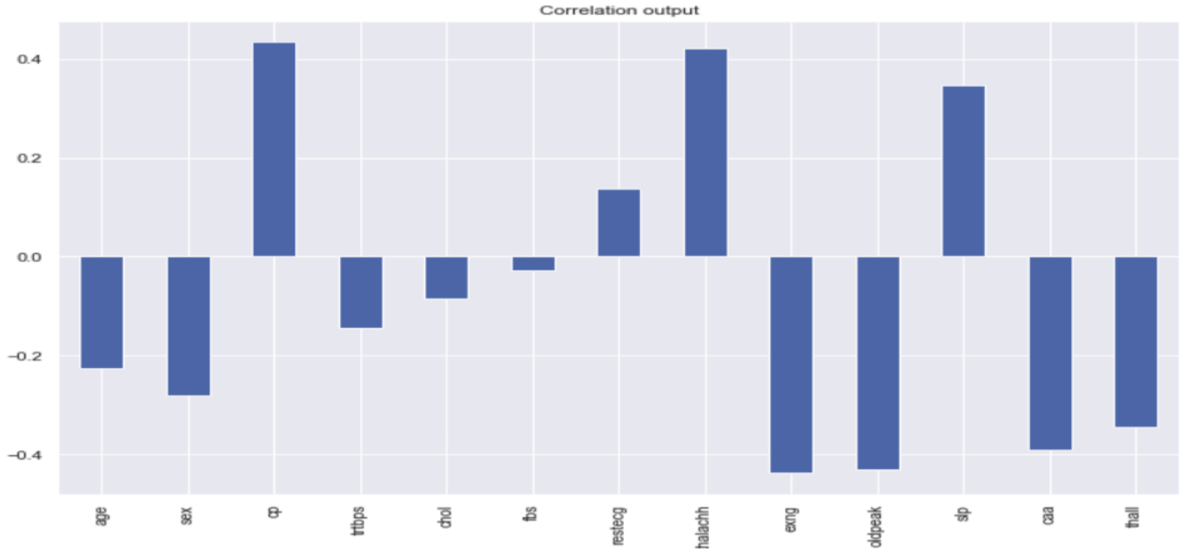


Figure 2.9: Correlation only with output and other variables.

Explanation of heatmap;

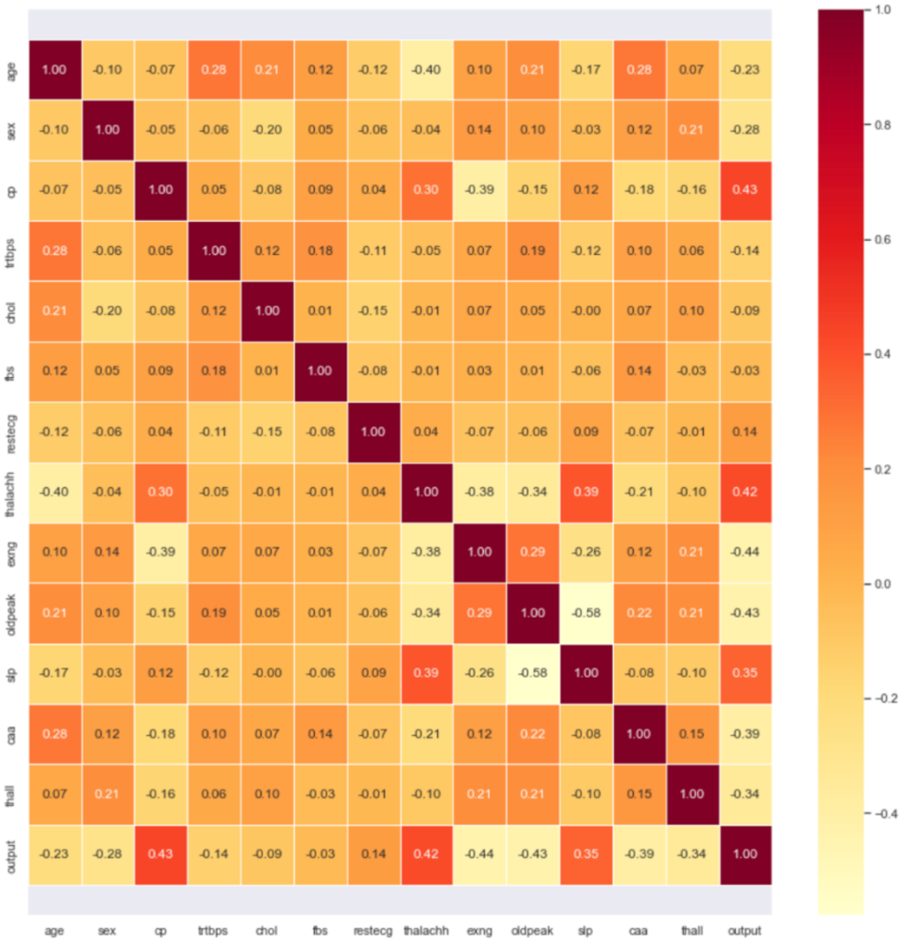


Figure 2.10: Correlation matrix-heatmap

We can figure out things from the map with colors that shows how things are connected;

The output and exang variables have a weak negative connection (-0.44 correlation coefficient).

The output and cp variables are somewhat related, with a correlation of 0.43.

The output and thalach variable have a slight positive relationship, with a correlation coefficient of 0.42.

The output and slope have a small positive connection. The correlation coefficient is 0.35

The output and ca variable are related, but in a weakly negative way. This negative correlation is 0.39.

The output and thall variable have a weak negative relationship (scored at -0.34)

We are able see that those individuals enduring from heart disease (output = 1) have generally higher heart rate (thalach) as compared to individuals who are not enduring from heart disease (output = 0) **(Figure 2.11) (Figure 2.12)**.

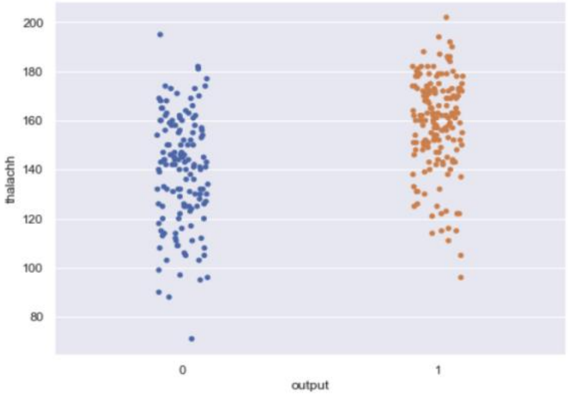


Figure 2.11: Thalach- output graph

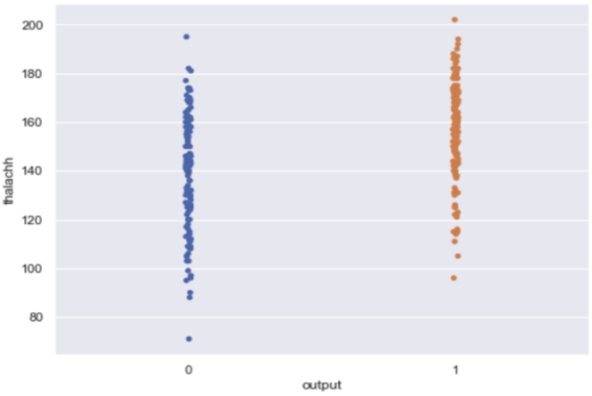


Figure 2.12: thalach- output graph with jitter

2.4.Data Visilazation Results

The results of the analyzes I have made based on two variables are as follows;

- There's no variable which has solid positive relationship with output variable.
- There's no variable which has solid negative relationship with output variable.
- There's no relationship between output and fbs.
- The cp and thalach factors are mildly positively related with output variable.
- We can clearly see that the Talach variable is slightly negatively skewed..
- The individuals enduring from heart illness (output = 1) have moderately higher heart rate (thalach) as compared to individuals who are not enduring from heart illness (output = 0).

2.5.Model

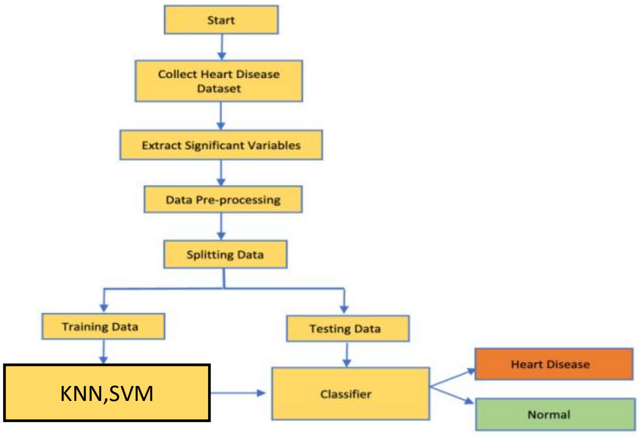
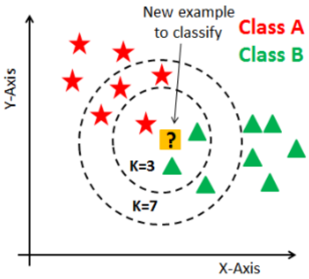


Figure:2.1: Model of Project.

I. K-Nearest Neighbors (KNN)



The K nearest neighbor(KNN) algorithm, as the name suggests, is an algorithm that makes predictions by looking at its neighbors. In the KNN algorithm, the assumption that similar things are near to each other is valid.

When the picture above is examined, it is observed that, in general, similar classes are close to each other. Therefore, the KNN algorithm is also based on this observation, and new future predictions are estimated according to the proximity to these points**[4].**

We may have a question here. What is the measured distance? Yes, different distance measurements are used in machine learning algorithms. Although the Euclidean distance is used mainly, it is available in different measurements. An example table is presented below. In this section, distances that are frequently used in KNN algorithms are mentioned.

Advantages of the KNN algorithm;

- The algorithm in KNN is simple and easy to implement[3].
- There isn’t need to build a new model, adjust a few parameters, or make additional assumptions.
- The algorithm is versatile and can be used for regression, classification, search.

Disadvantages of KNN algorithm;

- The algorithm slows down significantly as the number of samples and/or predictors/arguments increases. Computationally expensive.
- High memory requirement. Because it keeps all train data in memory.
- As the sample size increases, the estimation may take longer.

II. Support Vector Machines(SVM)

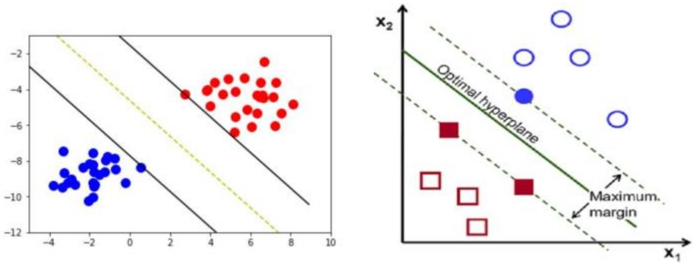


Figure 2.13: Support Vector Machines(SVM)

SVM separates the data into two separate classes, ensuring that the dividing line between the two classes is maximally marginal. This dividing line is a hyperplane that seeks to maximize the gap between classes. To find this hyperplane, it first represents the data points in multidimensional space. It then creates the optimal hyperplane for separating data points from different classes **[9] (Figure 2.13)**.

SVM is more successful when we can linearly separate the data points. However, we can also use the SVM method for data that cannot be fully linearly separated. SVM uses a different method called kernel method to solve nonlinear classification problems. It moves data into high-dimensional space using the kernel method, making it linearly separable.

In short, SVM is an algorithm that saves data compared to other classification algorithms that work well with high-dimensional data. SVM is widely used in many different fields, especially in image processing, text classification, informatics and financial analysis. Because it is a proven classification method, SVM is a method frequently used in industrial and academic research.

Advantages of SVM;

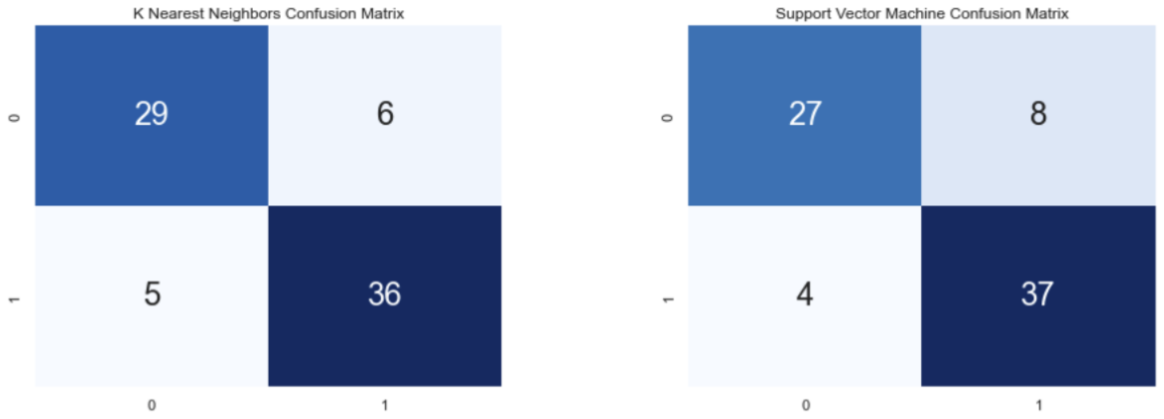
- Since it uses hyperplane while separating the data, it separates the data with the lowest error margin.
- SVM works much more effectively on high-dimensional data. We generally prefer it for text classification and image processing.
- SVM works regardless of the linearity of the data, but it uses the kernel method when classifying nonlinear data.
- SVM can work very effectively even if we have a limited dataset.

Disadvantages of SVM;

- Because the computational load is very high, it consumes a lot of time and processing power for large data.
- In cases where there are many hyperparameters, the accuracy of the SVM classification may decrease.
- Interpreting the result can be difficult in some cases because it works in a multidimensional space.

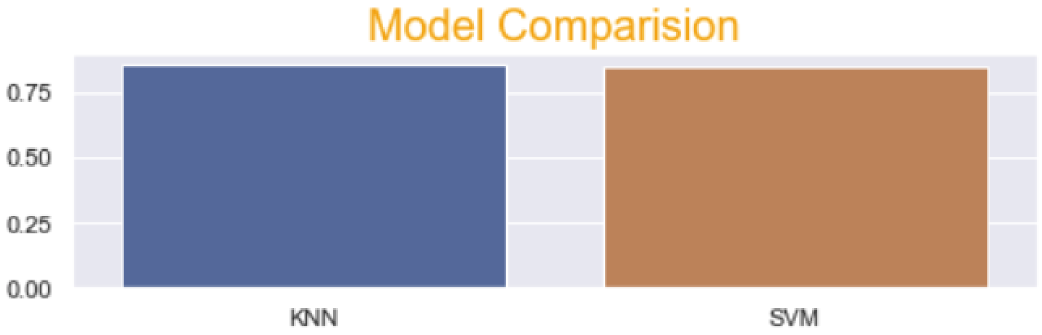
2.6.Results

Confusion Matrixes



K-NN
((124+100)/(5+13+124+100))*100=85.5263

SVM
((124+100)/(5+13+124+100))*100=84.2105



We tried a lot of parameters and settings for both models and maximized the accuracy rate for both models.
As a result, we can say that both are suitable for this model, which is very close to each other in two values.

2.7.Conclution

It is challenging to manually calculate the likelihood of developing heart disease based on risk factors. To predict the outcome from the available data, however, machine learning techniques are helpful.

It is challenging to manually calculate the likelihood of developing heart disease based on risk factors. To predict the outcome from the available data, however, machine learning techniques are helpful.

References

[1] AKCA, M.F. (2020). Nedir Bu Destek Vektör Makineleri? (Makine Öğrenmesi Serisi-2). [online] Deep Learning Türkiye. Available at: https://medium.com/deep-learning-turkiye/nedir-bu-destek-vekt%C3%B6r-makineleri-makine-%C3%B6%C4%9Frenmesi-serisi-2-94e576e4223e.

[2] Analytics Vidhya. (2021, May 11). Random Forest Regression: A Complete Guide with Python Scikit-Learn. Retrieved April 5, 2023, from https://www.analyticsvidhya.com/blog/2021/05/random-forest-regression-a-complete-guide-with-python-scikit-learn/

[3] ATCILI, A. (2022). KNN (K-En Yakin Komsu). [online] Machine Learning Turkiye. Available at: https://medium.com/machine-learning-t%C3%BCrkiye/knn-k-en-yak%C4%B1n-kom%C5%9Fu-7a037f056116 [Accessed 1 May 2023].

[4] Dangare Chaitrali S and Sulabha S Apte. "Improved study of heart disease prediction system using data mining classification techniques."International Journal of Computer Applications 47.10(2012): 44-8

[5] Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd ed.). O'Reilly Media. (pp. 159-188)

[6] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer. (pp. 307-346)

[7] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ open, 4(5), e005025.

[8] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[9] Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics, 21(3), 660-674.

[10] Scikit-learn. (n.d.). Sklearn.ensemble.RandomForestRegressor. Retrieved April 6, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[11] Wolgast G, Ehrenborg C, Israelsson A, Helander J, Johansson E & Manefjord H (2016). Wirelessbody area network for heart attack detection [Education Corner]. IEEE antennas and propagation magazine, 58(5), 84-92.