

Abdul Rahman VASSOUF - 2028142016

Regression and Classification models

Karabük Üniversitesi

Mühendislik fakültesi

Biyomedikal mühendisliği

University of Karabuk

Faculty of Engineering

Biomedical Engineering

# **ABSTRACT**

In the field of biomedical engineering and in other fields, some situations require statistical studies in order to predict certain results based on the previous historical data available at hand.

In order to perform these operations it takes a lot of time and effort from the people doing the statistical study. For this reason, software libraries based on software platforms have been developed to perform those statistical operations and predict results based on previous data, and this has saved a lot of time and effort to do statistical and classification studies.

In this research, software models were created to help link data and predict future results based on it, among these models there is a model for predicting the weights of newborns based on the health information of parents, the other model was designed to classify mushrooms into toxic and edible mushrooms based on A set of metadata for mushroom shape, color and size.

# **TABLE OF CONTENTS**

[**ABSTRACT** 1](#_Toc58624244)

[**TABLE OF CONTENTS** 2](#_Toc58624245)

[**TABLE OF FIGURES** 3](#_Toc58624246)

[**CHAPTER 1:** 4](#_Toc58624247)

[**NEWBORN LENGTH PREDICTION MODEL** 4](#_Toc58624248)

[**1.1. INTRODUCTION** 4](#_Toc58624249)

[**1.2. DEVELOPMENT PLATFORM** 4](#_Toc58624250)

[**1.3. DATA COLLECTION STAGE** 4](#_Toc58624251)

[**1.4. DATA CLEANING AND ORGANIZING** 5](#_Toc58624252)

[**1.5.1. Heat map** 6](#_Toc58624253)

[**1.5.2. Columns Selecting** 7](#_Toc58624254)

[**1.5.3. Multiple Linear Regression and Polynomial Linear Regression** 7](#_Toc58624255)

[**1.5.4. Evaluation and Scoring Regression Models** 8](#_Toc58624256)

[**1.5.5. Decision Tree and Random Forest** 9](#_Toc58624257)

[**CHAPTER 2:** 11](#_Toc58624258)

[**TOXIC AND NON-TOXIC (EDIBLE) MUSHROOM CLASSIFIER** 11](#_Toc58624259)

[**2.1. INTRODUCTION** 11](#_Toc58624260)

[**2.2. DATA COLLECTION STAGE** 11](#_Toc58624261)

[**2.3. DATA CLEANING AND ORGANIZING** 11](#_Toc58624262)

[**2.3.1. Looking For Missing and Duplicated Values** 11](#_Toc58624263)

[**2.3. BUILDING THE CLASSIFICATION MODEL** 12](#_Toc58624264)

[**2.3.1. Knn and SVM Classification Models** 12](#_Toc58624265)

[**2.3.3. Evaluation of Knn, SVM, Decision Trees and Random Forests** 13](#_Toc58624266)

[**REFERENCES** 14](#_Toc58624267)

# **TABLE OF FIGURES**

[Figure 1. 1. Dataset Cleaning. 6](#_Toc58624030)

[Figure 1. 2. Heat map, shows correlation between variables. 6](#_Toc58624031)

[Figure 1. 3. Column selecting codes. 7](#_Toc58624032)

[Figure 1. 4. Multiple Linear Regression codes. 7](#_Toc58624033)

[Figure 1. 5. Polynomial Regression Codes. 7](#_Toc58624034)

[Figure 1. 6. Scoring and Evaluating Models. 8](#_Toc58624035)

[Figure 1. 7. Decision Tree Codes. 9](#_Toc58624036)

[Figure 1. 8. Random Forest Codes. 10](#_Toc58624037)

[Figure 2. 1. "Info and "Describe" methods. 12](file:///C:\Users\HP\notebook\Project\Report_Draft01.docx#_Toc58624038)

[Figure 2. 2. Knn, SVM, Decision Tree, and Random Forest Scoring and evaluation. 13](file:///C:\Users\HP\notebook\Project\Report_Draft01.docx#_Toc58624039)

# **CHAPTER 1:**

# **NEWBORN LENGTH PREDICTION MODEL**

## **1.1. INTRODUCTION**

During pregnancy, some environmental factors surrounding the mother affect the safety and health of the fetus, causing changes in the weight and height of the fetus after birth. These factors include smoking of one of the parents, the weight of the mother during and before pregnancy, the height of the father and the height of the mother.

Some health authorities created databases to be studied statistically in order to predict the most influencing factors on the safety of the fetus collected data of some newborns, in this study, a child height prediction model was made based on the parents' data. The height and weight of the newborn child can be an indicator of the influence of external factors.

## **1.2. DEVELOPMENT PLATFORM**

All of coding stages are done using Jupyter-notebook which is a software based on Python programming language, Jupyter is one of the most recommended software to work and deal with data and data science.

Also (Pandas, numpy, matplotlib, Sklearn, Seaborn) python libraries used in this study.

## **1.3. DATA COLLECTION STAGE**

In order to search for suitable dataset to create a model to predict weight and height indicators for a newborn child, the appropriate dataset was searched on many search sites, the most famous of which are Kaggle, Data World, and others. After a long search time, a suitable dataset with important data was found. The search process took many hours, as most of the search results were suffering from loss of some important data.

The data search stage is one of the most time consuming stages of researchers.

The dataset, obtained from the University of Sheffield website, contains more than 200 birth records with details of the health of the parents and the newborn child.

Here are information about the dataset:

|  |  |  |
| --- | --- | --- |
| Column name | Column content | Data Type |
| ID | Baby’s ID | Numeric |
| headcirumference | Head circumference | Numeric |
| length | Baby’s length after birth (inches ) | Numeric |
| Birthweight | Baby’s weight after birth (lbs) | Numeric |
| Gestation | Pregnancy duration (weeks) | Numeric |
| smoker | Does mother smoke? 1=yes , 0= no | Binary |
| motherage | Mother’s age | Numeric |
| mnocig | Mothers number of cigarettes per day | Numeric |
| mheight | Mother’s height (inches) | Numeric |
| mppwt | Mothers pre pregnancy weight (lbs) | Numeric |
| fage | Father’s age | Numeric |
| fedyrs | Father’s education years | Numeric |
| fnocig | Fathers number of cigarettes per day | Numeric |
| fheight | Father’s height (inches) | Numeric |
| lowbwt | Is the birth weight is low? 1=yes , 0=no | Binary |
| mage35 | Is mother’s age more than 32 years? | Binary |
| LowBirthWeight | Is the birth weight is low? , 1=Low, 0=Normal | String |

Table 1. 1. (Birth Dataset Columns)

## **1.4. DATA CLEANING AND ORGANIZING**

After collecting the data, dataset check should be performed to verify that there is no duplicate, missing or corrupt data.

This was done using software commands such as “info, describe, and duplicated” commands, as shown in the image.

After checking the data, the duplicate data was deleted because it is identical to other data in the same dataset.

In addition, some missing data was found in the Weight and Pregnancy Period columns, as shown in the figure,

This was fixed by filling it in with the average value of the sum of the column values because that is the best way to fill in this type of data.

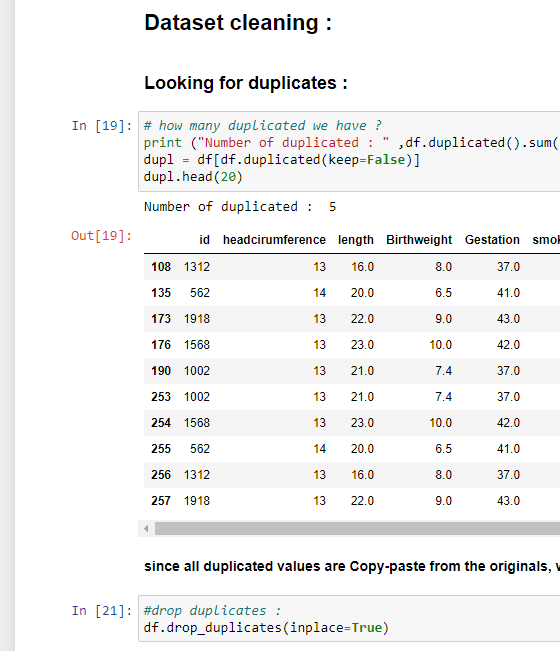
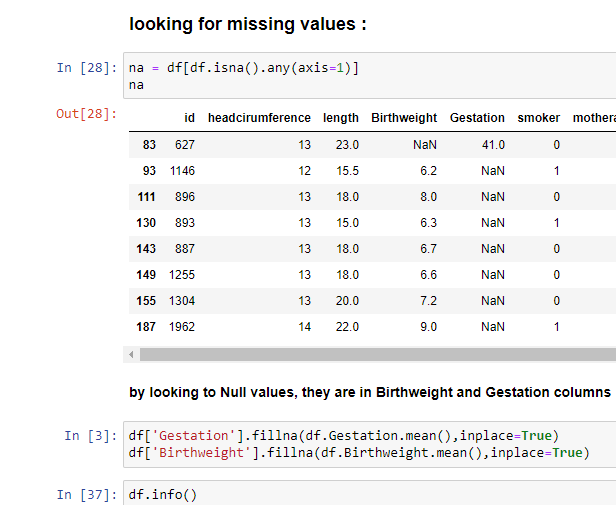


Figure 1. 1. Dataset Cleaning.

**1.5. MODEL CREATION**

### **1.5.1. Heat map**

After performing the data cleaning process, a matrix was created to calculate the correlation between the available variables using a “heat map”.

As we can see in (Figure1.2), the heat map shows the correlation values between each of the two different parameters, in this way we can find the most suitable column to choose in the retraction process.

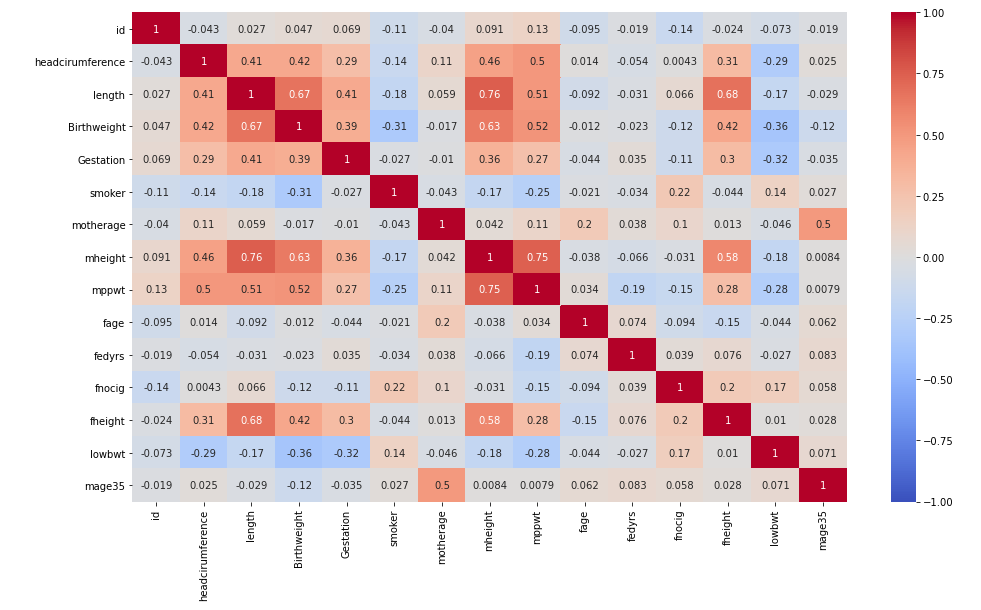


Figure 1. 2. Heat map, shows correlation between variables.

### **1.5.2. Columns Selecting**

Because the best correlation values was between fheight, mheight and length, so we select columns fheight and mheight to use in regression model.

We select length as y parameter because it is the value the model is going to predict.

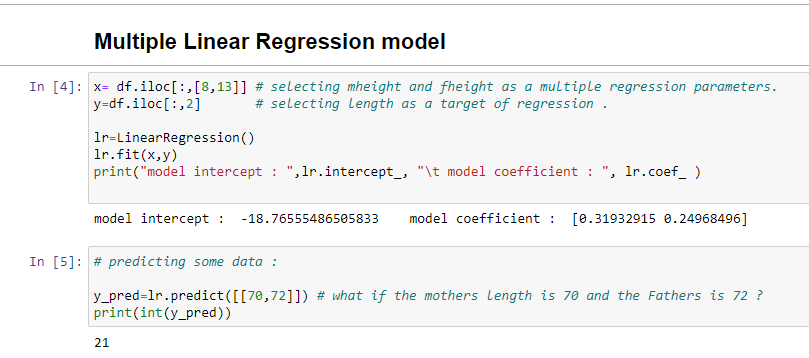


Figure 1. 3. Column selecting codes.

### **1.5.3. Multiple Linear Regression and Polynomial Linear Regression**

After the selection of the right columns, now we are going to create a regression models to train it by using our dataset.

Multiple Linear Regression model will analyze the data considering the liner relationship between them, while Polynomial Linear Regression analyze data according to the curve near to them.

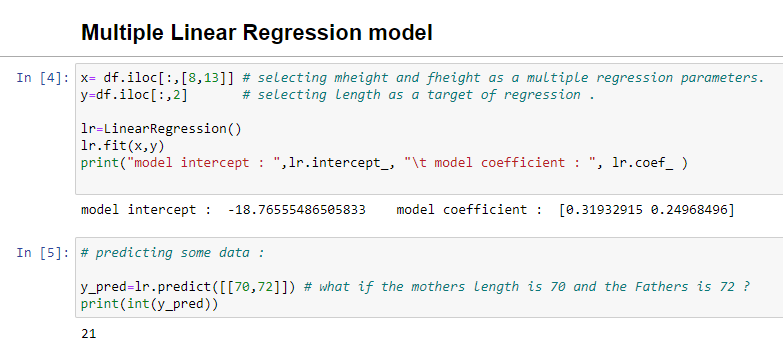


Figure 1. 4. Multiple Linear Regression codes.

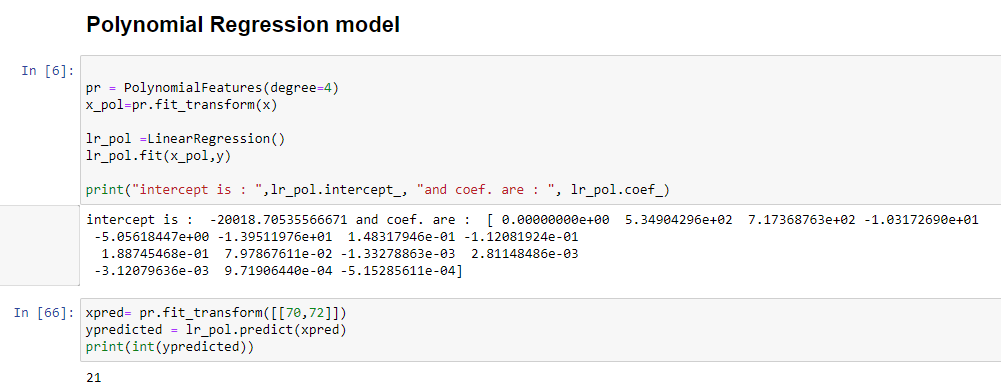


Figure 1. 5. Polynomial Regression Codes.

After the creation of Regression model, we “fit” the data from dataset into it. Then we can use it to predict target data according to training data.

### **1.5.4. Evaluation and Scoring Regression Models**

By using specific functions, we can score the accuracy of the model by comparing the predicted data with the real data. The most important functions are R2 score, Mean absolute error score, and Mean squared error. For Multiple and Polynomial linear Regression models we got R2 score about 0.663 and 0.702, this is a good accuracy according to dataset which used.

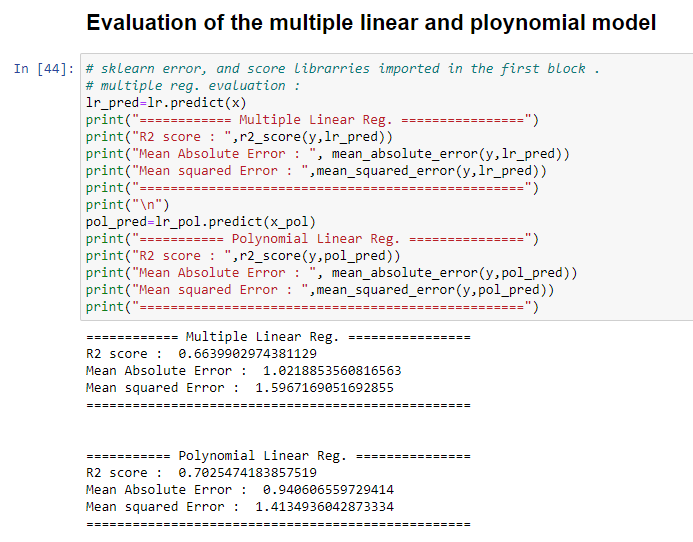


Figure 1. 6. Scoring and Evaluating Models.

### **1.5.5. Decision Tree and Random Forest**

Decision tree is a function used to re-group or to classify data according to their values. So by doing some (true/false) comparisons, decision tree can reach to the best-predicted value by following (True/False) lines until the end of the line which is the best possible prediction.



Figure 1. 7. Decision Tree Codes.

Random forest contain multiple decision trees use them to predict multiple ways to and compare results to reach to the final value.

To use decision trees and random forests, first we have to split the dataset into two datasets, one for training and another one for testing, it is important to do that to avoid getting wrong accuracy numbers.

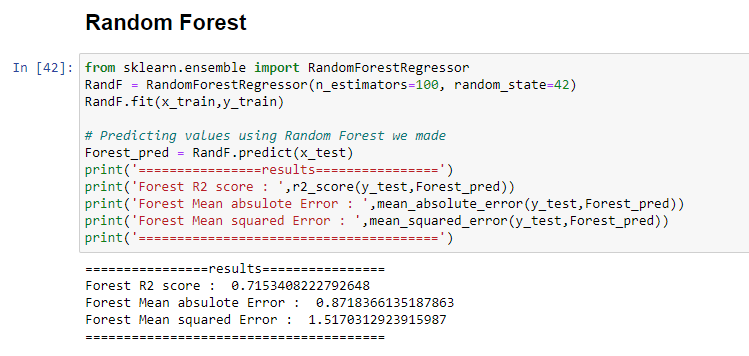


Figure 1. 8. Random Forest Codes.

# **CHAPTER 2:**

# **TOXIC AND NON-TOXIC (EDIBLE) MUSHROOM CLASSIFIER**

## **2.1. INTRODUCTION**

In many cases, researchers need to do a classification process for a specific thing based on its characteristics. A very large number of previous information and results are collected to perform a classification process for the target.

This large amount of information needs a lot of time to be studied manually, so some software libraries have been developed that are able to find appropriate classifications based on previous information provided by the researcher.

In this study, a model was created that classifies some types of mushrooms based on their morphological and color characteristics, and based on that, it determines whether the mushroom is toxic or edible.

## **2.2. DATA COLLECTION STAGE**

After doing some searches for suitable datasets to build classification models, I found PLMB repository on Github, in this repository there are a hundreds of different suitable datasets in different sciences. This dataset of mushrooms was a good choice for classification, since it contain many details about several mushroom types.

The dataset, obtained from the PLMB, contains more than 8000 records with details about morphological and color properties.

The column data of dataset can be found in attached .ipynb files.

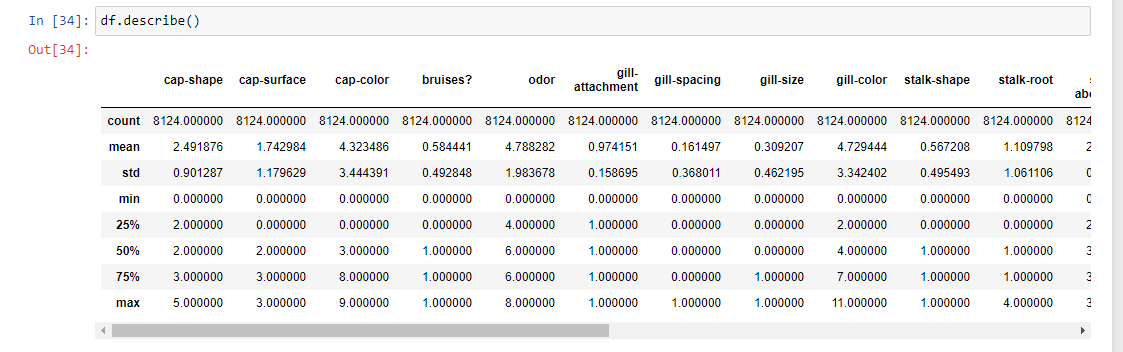
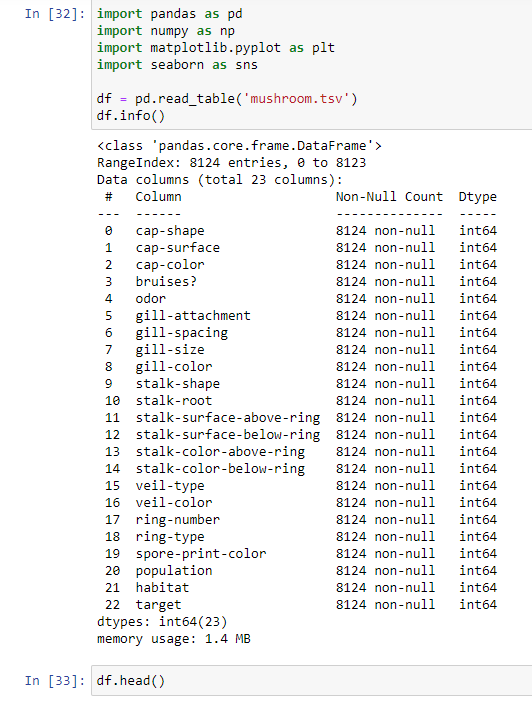
## **2.3. DATA CLEANING AND ORGANIZING**

It is always important to check for the missing data in the dataset we have even if it is checked before.

### **2.3.1. Looking For Missing and Duplicated Values**

By using info, describe, and duplicated methods, there was no any missing or duplicated data in the dataset; all entries were “int64”, and everything was working very well.

Figure 2. 1. "Info and "Describe" methods.



## **2.3. BUILDING THE CLASSIFICATION MODEL**

By using “Sklearn” python library, we can easily create classification models use different methods, such as: Knn, SVM and others.

### **2.3.1. Knn and SVM Classification Models**

In K-Nearest Neighbors (KNN), the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

While SVM algorithm is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.

**2.3.2. Decision Trees and Random Forest**

In classification methods, Decision trees and Random Forests are similar to those in predicting methods, the main Idea is the same, Decision Trees classify object according to (True/False) relations, while random forests combine between multiple decision trees to get a more accurate results by comparing the performances of multiple decision trees.

### **2.3.3. Evaluation of Knn, SVM, Decision Trees and Random Forests**

To make sure that the model doesn’t has any kind of “over fitting” we have to split the dataset into training and testing datasets to avoid getting a wrong accuracy percent.

After that, we should calculate the Score of the model we created to look about the performance and accuracy of the model, and to make sure that the model is working without issues.

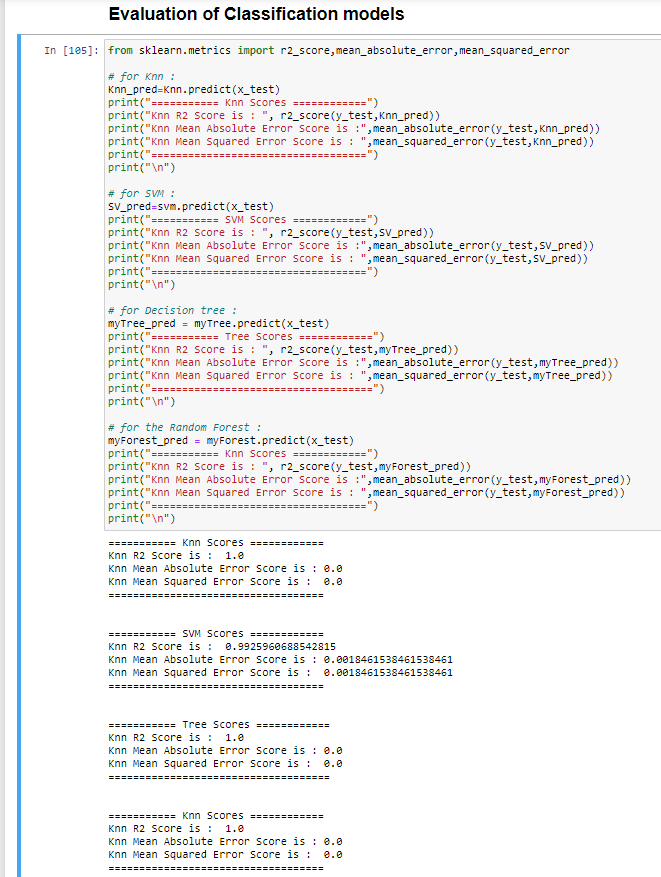


Figure 2. 2. Knn, SVM, Decision Tree, and Random Forest Scoring and evaluation.

# **REFERENCES**

1. “Mathematics and Statistics Help (MASH)”, *The University of Sheffield*, <https://www.sheffield.ac.uk/mash/statistics/datasets>
2. Olson, Randal S. and La Cava, “PMLB”, *GitHub repository*, <https://github.com/EpistasisLab/pmlb/tree/master/datasets/mushroom>
3. Kaggle, Datasets , <https://www.kaggle.com/datasets>
4. Data World, <https://data.world>