

INSY-5339

PRINCIPLE OF BUSINESS DATA MINING

SPRING 2021

FINAL REPORT





TEAM 15

Deepakvanshikumar Tuniya

ID:  1001894483

Komal Pimpariya

ID: 100195192

TABLE OF CONTENTS

CHAPTER PAGE

SUMMARY 3

PROJECT BACKGROUND 3

TACKLING BUSINESS IDEA 3

DATASET DESCRIPTION 4

DATA VISUALIZATION TECHNIQUES 5

MODEL SELECTION 8

MODEL DESCRIPTION 8

RESULTS 9

CONCLUSION 14

MANAGERIAL INSIGHTS 15

1. **Summary:**

* Healthcare company going through CVD reports of patient to check who is at risk of diseases and if they can apply control measures to prevent those who are at borderline as the disease is number 1 cause of death globally.

1. **Project Background:**

* Healthcare consultant looking into Cardiovascular Disease (CVD) Patient examination results to predict if they have CVD or not by model comparison technique.
* Focusing on the patients that don’t have CVD and to check on risk factors and aspects like alcohol consumption and smoking if done in excess can affect them with CVD.

1. **Tackling Business Idea:**

* Patient profile would be different based on health and lifestyle patterns.
* To achieve the goal patients will be segregated on their habits such as smoker or non-smoker and consumption of alcohol.
* Paying more attention to people who don’t have CVD but can have the diseases if frequency of their habits increase.
* The company has factual information, information given by the patient and results of medical examination to support business problem.

1. **Dataset Description: Cardiovascular Disease (CVD)**

* The project is based on patient examination of Cardiovascular Disease records where the data has 13 columns and 70000 rows, 4 binary variables, 1 categorical variable, 5 continuous variables and is well balanced.
* Source of the dataset is Kaggle:

<https://www.kaggle.com/sulianova/cardiovascular-disease-dataset>

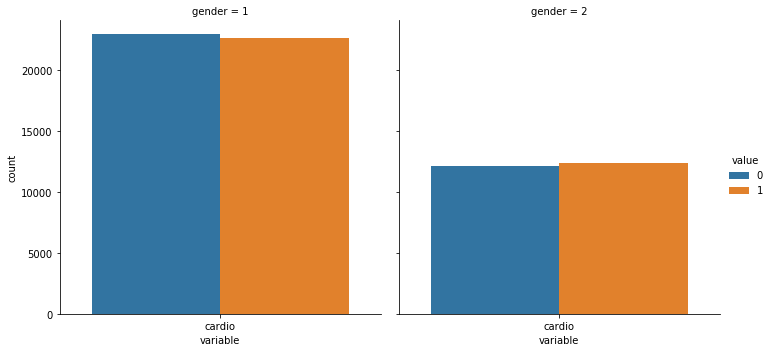
* Data consist of 3 types of information. Objective, examination and subjective.
* Features of the data:

1. Age | Objective Feature | age | int (days)
2. Height | Objective Feature | height | int (cm) |
3. Weight | Objective Feature | weight | float (kg) |
4. Gender | Objective Feature | gender | categorical code |
5. Systolic blood pressure | Examination Feature | ap\_hi | int |
6. Diastolic blood pressure | Examination Feature | ap\_lo | int |
7. Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
8. Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
9. Smoking | Subjective Feature | smoke | binary |
10. Alcohol intake | Subjective Feature | alco | binary |
11. Physical activity | Subjective Feature | active | binary |
12. Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

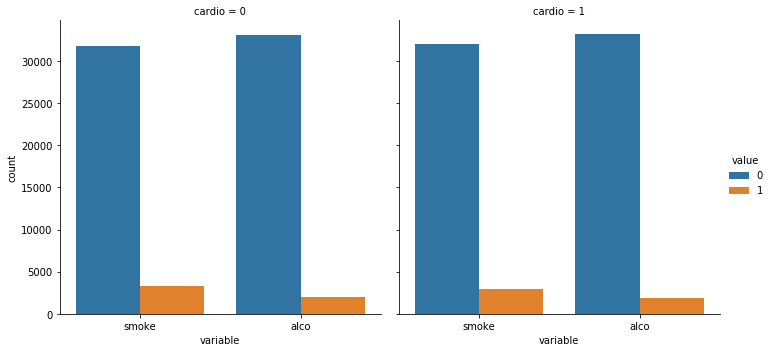
* All of the dataset values were collected at the moment of medical examination.
* The data of target variable is a balanced data with 49.3% not having CVD and 50.3% having CVD.
* As the target variable is binary so we will use classification algorithms for our business problem for prediction and model comparison technique.

1. **Data Visualization Techniques:**

* The visualization technique below is a categorized bar plot where patients are divide with gender and target variable cardio.
* Gender 1 displays female data and gender 2 displays male data.
* Whereas in cardio variable with 0 and 1 states patients having absence and presence of disease based on gender.



* Bar plot of variable smoke and alcohol which is segregated on the basis of patients with cardiovascular diseases and patient without it.
* The plot shows that only few Smoke and Alcohol consuming patients are affected with CVD.



* The heatmap displays the relationship of variables in data. Good relationship shows the usefulness of data.
* As we can see cholesterol has high correlation with glucose.
* Height has high correlation with gender which helped us define gender with checking mean height in the category.
* We can’t drop these variable with high correlation.



1. **Model Selection:**

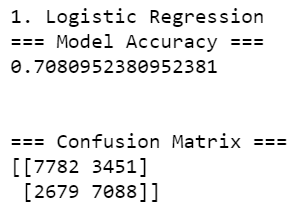
* Problem is executed in Python programming.
* The target variable cardio is used for prediction of our problem.
* We have applied four classification algorithms to discuss our results.
* One best model out of four is selected based on model performance.
* Four models are selected based on accuracies compared to rest of the models.
* The four models implemented are:

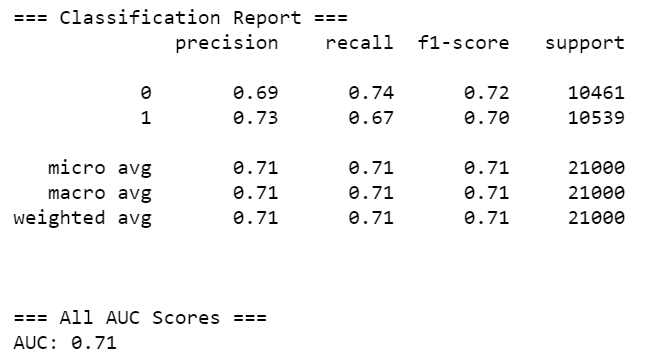
1. Logistic Regression
2. Decision Tree
3. eXtreme Gradient Boosting
4. Random Forest
5. **Model Description:**

* Logistic Regression: The regression model is used to examine relationship between binary target variable ‘cardio’ and independent variables which will have only two outcomes.
* Decision Trees: Decision Tree is applied to implement training model which predicts target variable ‘cardio’ by learning simple decision rules inferred from prior training data.
* eXtreme Gradient Boosting: This model suits well as it uses many trees where one tries to predict the target variable ‘cardio’ and another tries to predict the residual of first tree and so on.
* Random Forest: As this is a classification problem, decision trees are generated from the training data for those who don’t have CVD. Votes are collected from each tree and best are selected by votes. Most voted tree is final prediction.

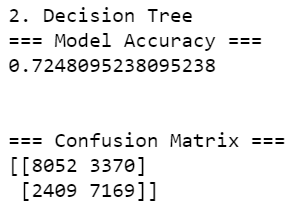
1. **Results:**

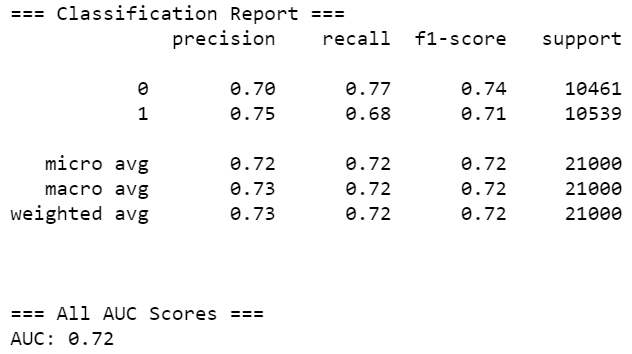
Results of the for four data mining models are based on seven measures such as model accuracy, confusion matrix, precision, recall, f1-score,support and auc scores.



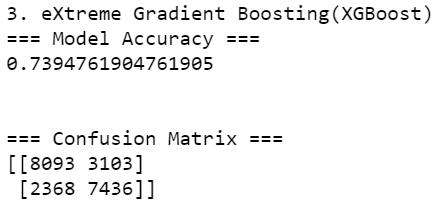


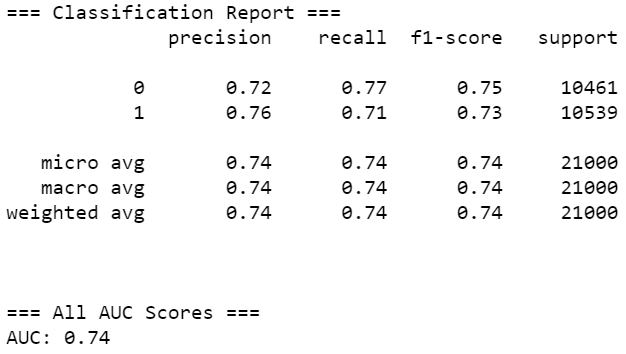
Fitting Logistic Regression as the first model for our data with the above mentioned result. The results say that Logistic Regression is the worst performing model of the four selected models, even though there is not enough difference in the results of all models. The accuracy of the model is 70.8%. The rest of the measures such as confusion matrix and classification report with precision, recall, f1-score and support all have similar poor performance as the accuracy of the model.



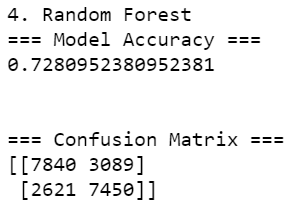


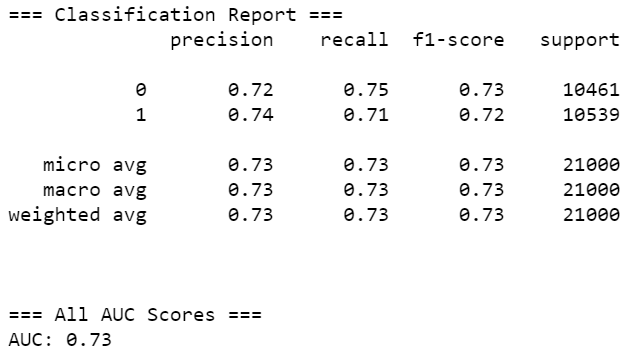
This model is fit with Decision Tree. It is a classification tree. The accuracy of this model is 72.4% one of the top performing if considered overall with confusion matrix. The precision of predicting presence CVD is high among the compared model.



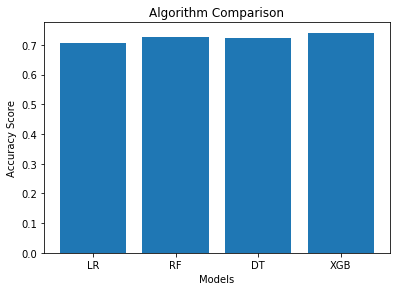


XGBoost is the best fit model for the prediction of our data. It has the top accuracy with 73.9%. It is the outperforming model compared to all performing models. Performance states that XGBoost is superior in all the compared measures. The confusion matrix model has least number of invalid data for people having cardiovascular diseases which also makes the model with most precise data.





The above model is fitted by applying Random Forest algorithm. This is one of the finest model with accuracy almost similar to top performing XGBoost which is 72.8%. The remaining measures of this models are close to Decision Tree.



The bar plot displays accuracy comparison of various algorithms we have applied in our business problem. The bar is built from matplotlibrary of python package. There are four models applied in the proposed work. Random Forest and Decision Tree have same accuracy in the range of 72% and Logistic Regression being the less better model having 71% accuracy. We can say that XGBoost model has the best accuracy but there is a negligible difference in the result of the four models.

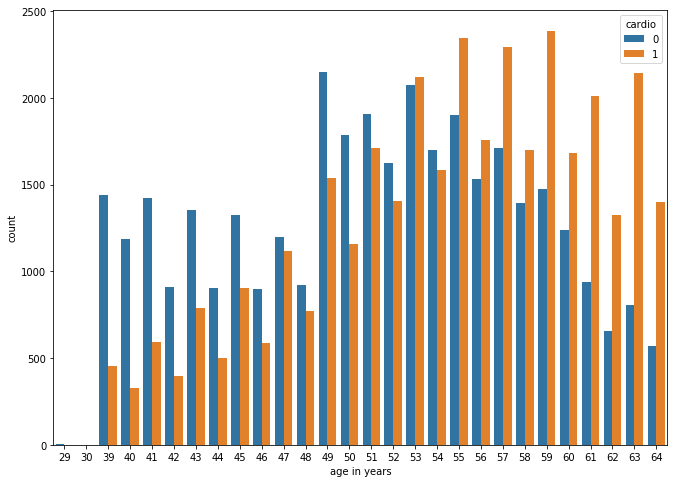
Hence we can say that XGBoost is the best prediction model to predict if a patient has presence or absence of Cardiovascular Diseases(CVD).

1. **Conclusion:**

* Count plot of age in years of the patient where 0 represents patients not having Cardiovascular diseases and 1 shows patients with

Cardiovascular diseases.

* The plot shows that people with age 53 and above have higher chances of having presence of CVD.
* This states that patients with age 53 not having CVD should be more focused by the company to look for risk of getting diseases, their habits and lifestyle patterns need to be taken care of.



**10)Managerial Insights:**

* Extra attention provided to patients with age 53 and above.
* Promote good lifestyle and spread awareness about smoking can be harmful causing CVD.
* Imply a limit in consumption of alcohol.
* Keep a regular tab on health attributes like blood pressure, cholesterol level and sugar levels.
* Inculcate physical fitness for the patients according to their BMI as we have gender, height and weight stats of patient.