






Wholly owned by UTAR Education Foundation  
(Co. No. 578227-M)  
DU012(A)

UCCD2063 Artificial Intelligence Techniques

Group Assignment

June 2024

<b>Student Name:</b>	<b>Ooi Khai Shen</b>	<b>Tan Kia Yee</b>	<b>Tan Qi Yang</b>
<b>Student ID:</b>	<b>2207092</b>	<b>2206932</b>	<b>2207140</b>
<b>Contribution:</b>	<b>33.33%</b>	<b>33.33%</b>	<b>33.33%</b>
<b>Signature:</b>			

## Contents

1.0 Introduction.....	1
1.1 Background .....	1
1.2 Objectives .....	1
2.0 Methods.....	2
2.1 Dataset description.....	2
2.2 Data exploration and visualization.....	2
2.3 Data Pre-processing .....	2
2.4 Model Selection .....	2
2.5 Model Training and Validation .....	3
2.6 Model Tuning and Testing .....	4
3.0 Result and Discussion .....	6
3.1 Summary of Training and Testing Result .....	6
3.2 In-dept Analysis of the Prediction Performance and Errors .....	7
3.3 Performance Comparing between Models.....	8
3.4 Strengths and Weaknesses .....	9
3.5 Feature Importance .....	10
4.0 Conclusion .....	12

## **1.0 Introduction**

### **1.1 Background**

Cardiovascular disease is a kind of heart disease that involve a group of disorders of heart and blood vessels. For example, the blockage of coronary arteries that will lead to heart attacks, irregular heartbeats and high blood pressure that will lead to the risk of stroke. Cardiovascular disease leading to a high percentage of global death. Lack of awareness and personal bad habits is one of the main factors from getting cardiovascular disease. It is important for an individual to identify the risk of getting the disease in the early stage. Machine learning will be one of the solutions to address the problem through early making of prediction on the risk of getting cardiovascular disease. The data can be collected through patient health record including important features such as their habits, lifestyle and their age. Implementing machine learning to train the data and identifying the patterns of the features and predicting the risk of getting cardiovascular.

### **1.2 Objectives**

The aim of the study is to investigate, analyze and compare various machine learning technique to determine the most effective model for cardiovascular risk prediction. Through the model, the machine learning technique able to implement to the real-life lead to earlier detection and intervention. Hence, able to decrease the risk of getting cardiovascular disease by enabling effective preventive measure and treatment.

## **2.0 Methods**

### **2.1 Dataset description**

The dataset contains 2100 entries and 18 columns. There are 10 columns are categorical which are Gender, Family\_history, Alcohol, Junk\_food, Snack, Smoking, Transportation, TV, Discipline and Cardiovascular\_risk(y). The numerical data are including Age, Height, Weight, Vege\_day, Meals\_day, Water\_intake and Income.

### **2.2 Data exploration and visualization**

Analyze the data by understanding the datatype that used by each column. Visualize the dataset by using histogram to show the distribution of each column.

### **2.3 Data Pre-processing**

First, it needs to perform separation between the target variable which is cardiovascular risk that we want to predict from the other columns. Since the cardiovascular disease is categorical data that categories into high, medium and low, due to this the variable need to encode and convert it into numerical values. Besides that, we need to perform preprocessing for each of the categorial and numerical column. We separate the categorial columns into ordinal and nominal which ordinal features have meaningful order among categories while nominal do not have any ranking. One Hot Encoder is used to convert the nominal columns into one hot binary vector while Label Encoder is used to convert ordinal columns into numerical values. For numerical data, Standardization is performed to ensure the features contribute equally to the model's performance. Then, splitting the dataset into 80% of training set and 20% testing set. It is use for the performance evaluation.

### **2.4 Model Selection**

#### **1. Decision Tree**

Decision Tree is a machine learning algorithm that uses a tree-like structure to model the decisions and the outcomes. The internal node of tree represents the decisions based on specific features and the branches show the results of these decisions. The leaf nodes show the final prediction. The simplicity and interpretability are the benefit of decision tree which making us easier to understand and visualize. It can also handle both numerical and categorical data unlike

Linear Regression only primarily handles numerical data. Lastly, decision tree can identify the importance of features in dataset.

## **2. Random Forest**

Random forest is a machine learning algorithm that using multiple decision trees to perform the prediction. The operation is created with a huge bundle of decision trees when training the model to arrive at a final decision. The benefit of random forest is the characteristic of model enable it to become more robust when handling high dimensional data that have many features. It can reduce the risk of overfitting.

## **3. Support Vector Machine (SVM)**

Support Vector Machine which known as SVM is a powerful machine learning algorithm that separate the data and classify them by finding a best boundary. In our work, we employ linear kernel SVM and fine-tuned the performance by optimizing the parameter using cross-validation. This can ensure that the model will be more effective and able to predict more effectively. SVM can handle high-dimensional data and identify the boundary to separate them effectively.

### **2.5 Model Training and Validation**

1. **Decision Tree:** Before training the Decision Tree model, hyperparameters are set to optimize performance. The maximum depth of the tree is limited to 20 to prevent overfitting by reducing the complexity of the model. The minimum samples required to split an internal node and the minimum samples required to be at a leaf node are also specified to further control overfitting. This configuration allows the Decision Tree to generalize better to new data.

**Validation for Decision Tree:** A five-fold cross-validation is applied to assess the model's robustness and effectiveness. The dataset is divided into five parts, and the model is trained and tested on different portions, ensuring comprehensive evaluation and reducing the risk of overfitting.

2. **Random Forest:** Before performing the training of dataset, customized hyperparameters for fine-tuning the model. The number of trees in the forest is set into 200 sets. The depth of the tree is restricted into 20 maximum depths for each tree.

Bootstrap is set into true to allow the model train on a random subset of the training data. This step is to increase the diversity among trees to reduce the risk of overfitting scenarios.

**Validation for Random Forest:** Five-fold cross validation is implemented to ensure the robustness of the model. Five different portions of the whole dataset are separate out for achieve a better testing and training to prevent overfitting issue.

3. **SVM:** The model is initially trained using the linear kernel and with parameter (C) of 1. The parameter is like the margin to separate the data, lower C has the wider margin and the larger C has narrower, we will continue to find out which margin, C parameter will be suitable for the model. Higher C can lead to better performance but may caused overfitting while the lower C allow some misclassification, we must find the best value in the middle. Afterward, whether the linear or RBF model be better and more effectively for our data is trying to find out. Linear kernel is suitable for the linearly separated data while the RBF kernel is for non-linear data relationship. Whether which kernel is suitable for our case could be find out by test and try out the combination of C (0.1, 1, 10) together with kernel ('linear', 'rbf').

**Validation for SVM:** Same method which a five-fold cross-validation is applied to ensure the model's robustness and effectiveness. The dataset is divided into five parts, and each the part are trained and tested on, ensuring comprehensive evaluation and reduce overfitting issues.

## 2.6 Model Tuning and Testing

1. **Decision Tree:** To enhance model performance, hyperparameters are fine-tuned manually. The maximum depth of the tree is restricted to 20 levels to prevent overfitting and reduce complexity. The minimum number of samples required to split a node and the minimum number of samples required at a leaf node are adjusted to improve generalization. The model is trained using these parameters, and predictions are made on both the training and test sets. The model's performance is then evaluated using accuracy scores for both training and testing set, along with analysis of the confusion matrix, precision, recall, and F1-score to understand its effectiveness in classification tasks.

2. **Random Forest:** To improve the overall performance on the classification task, the hyperparameters are fine-tuned manually through specifying the parameter. Specify the model to have 200 decision trees in the ensemble. More trees able to improve the performance and reducing the variance making the model more robust. Then limits the depth of each decision to 20 levels. It is to prevent the model become complex and prevent the pattern that prone to overfitting. Next minimum number of samples split, and minimum number of sample leaf are set manually to avoid overfitting situation. Enable bootstrap sampling for each of the trees to increase the diversity among the decision trees. Make prediction on the training set and the testing set, observe the accuracy score for both set. Analyse the confusion matrix for the prediction. Observe the precision, recall and f1-score values.
3. **SVM:** The initial SVM is further tune by hyperparameter tuning which with the method GridSearchCV to explore that which parameter work the best for the data. The parameter – C which like the margin between the data is found out by test out in between 0.1, 1 and 10. Larger number of C suppose will be more accurate but may lead to overfitting thus we have to find the best immediate value that work the best. Besides, another parameter which the kernel needs to find out, kernel consist of 2 types. First is the linear kernel which is better on deal with the linear data while another rbf kernel is better on deal with the non-linear kernel. After employing the grid search and cross-validation, we found that the best for our data is model with {'C': 10,'kernel': linear}. The model performance is evaluated on both training and testing sets, it should demonstrate high performance, achieving high accuracy. The confusion matrix, precision, recall, and F1-score were also analysed to better understand towards the model's classification capabilities.

### **3.0 Result and Discussion**

#### **3.1 Summary of Training and Testing Result**

##### 1. Result for the Decision Tree

<b>Decision Tree</b>	<b>Accuracy</b>
Cross-validation accuracy score	0.9583, 0.9702, 0.9702, 0.9643, 0.9702
Mean cross-validation accuracy	0.9667
Training set accuracy	0.9929
Testing set accuracy	0.9595

*Table 3.1.1 Result of Decision Tree*

##### 2. Result for the Random Forest

<b>Random Forest</b>	<b>Accuracy</b>
Cross-validation accuracy score	0.9702, 0.9762, 0.9643, 0.9554, 0.9762
Mean cross-validation accuracy	0.9685
Training set accuracy	0.9994
Testing set accuracy	0.9690

*Table 3.1.2 Result of Random Forest*

##### 3. Result for the SVM

<b>SVM</b>	<b>Accuracy</b>
Cross-validation accuracy score	0.9792, 0.9762, 0.9911, 0.9821, 0.9940
Mean cross-validation accuracy	0.9845
Training set accuracy	0.9940
Testing set accuracy	0.9881

*Table 3.1.3 Result of SVM*



### 3.2 In-dept Analysis of the Prediction Performance and Errors

1. **Decision Tree:** From the result the model showed a high accuracy. For the cross-validation accuracy scores, the model presents a consistent accuracy across all the folds with the mean of 0.9697. This indicates that the data have a minimal variance across folds. However, the training set accuracy score achieved a very high accuracy which is 0.9929 indicate that it is slightly overfitting. The confusion matrix shows that there are only 12 error that performing misclassification out of 1680 samples. For testing set has 0.9452 of accuracy which is slightly lower than the training set accuracy. The result show that it has a high accuracy rate which mean that the prediction performance overall is remaining strong. From the confusion matrix, it shows that there are 17 error of the prediction out of 420. Through the overall of the accuracy, we able to observe that the model is well performed.
2. **Random Forest:** From the result the model showed a high accuracy. For the cross-validation accuracy scores, the model presents a consistent accuracy across all the folds with the mean of 0.9685. This indicates that the data have a minimal variance across folds. However, the training set accuracy score achieved a very high accuracy which is 0.9994 indicate that it is slightly overfitting. The confusion matrix shows that there are only 1 error that performing misclassification for training. For testing set has 0.9690 of accuracy which is slightly lower than the training set accuracy. The result show that it has a high accuracy rate which mean that the prediction performance overall is remaining strong. From the confusion matrix, it shows that there are 13 error of the prediction. However, from the overall of the accuracy, it still shows that the model is still robust and well performed.
3. **SVM:** The model particularly after tuning process, has exceptional good performance in predicting cardiovascular risk level. The model shows high cross-validation accuracy, with a mean of 0.9821. The training set accuracy of 0.9946, which slightly higher than the cross-validation scores, indicates minimal overfitting. There is only 10 misclassifications out of 1680 samples, further shows that how strong the model performance is on the training data. In addition, while coming to the testing set, SVM model also performing good, with impressive accuracy of 0.9881. This is slightly lower than the training set accuracy with 5 misclassifications out of 420 samples. From the confusion matrix shows that most errors were in misclassifying 'low' risk individuals as

'medium' risk and few other misclassifications. In a nutshell, SVM model demonstrate robustness and effectiveness in classifying cardiovascular risk levels.

### 3.3 Performance Comparing between Models

<b>Model</b>	<b>Training set accuracy</b>	<b>Testing set accuracy</b>	<b>Cross-validation accuracy</b>
<b>Decision Tree</b>	99.29%	95.95%	96.97%
<b>Random Forest</b>	99.94%	96.90%	96.85%
<b>SVM</b>	99.40%	98.81%	98.45%

*Table 3.3.1 Accuracy Performance of Models*

Through Table 3.0, we able to observe that the Decision Tree Model and Random Forest Model have similar accuracy performance. Both models are slightly overfitting the training data by having 99.29% and 99.94% on training set accuracy. Random Forest is having slightly higher accuracy on testing with 96.90% compared to Decision Tree. There are several factors causing Random Forest to have higher accuracy compared to Decision Tree. Random Forest is an ensemble learning method that combines multiple decision tree to make a final prediction. By combining the prediction of multiple trees, it will reduce the variance among the features and lead to a higher accuracy. Support Vector Machine (SVM) is having the high training accuracy of 99.40% and it does not overfitting as it has the testing accuracy of 98.81% which is highest accuracy. There are several factors resulting SVM has highest accuracy among these three models. SVM use regularization to avoid overfitting, and it is good at generalizing to unseen data if properly regularized.

### 3.4 Strengths and Weaknesses

#### 1. Decision Tree:

**Strength:** Easy to understand and interpret which means that it will provide a clear visual representation of the decision-making process. Besides, it can handle both numerical and categorical without requiring extensive preprocessing or transformation like Linear Regression. Showing high accuracy on this study.

**Weaknesses:** The model initially overfit the training data, which was only 12 errors out of 1680 samples. The testing accuracy is the lowest among the three models.

#### 2. Random Forest:

**Strength:** It performs well when handle a high-dimensional dataset. The structure of random forest ensemble of multiple of decision trees that can reduced the risk of overfitting and improve the accuracy of prediction.

**Weaknesses:** The model is more complex compared to decision tree; the training time is slower. It requires more tuning on the hyperparameter to optimize the performance. It needs a higher memory usage.

#### 3. SVM:

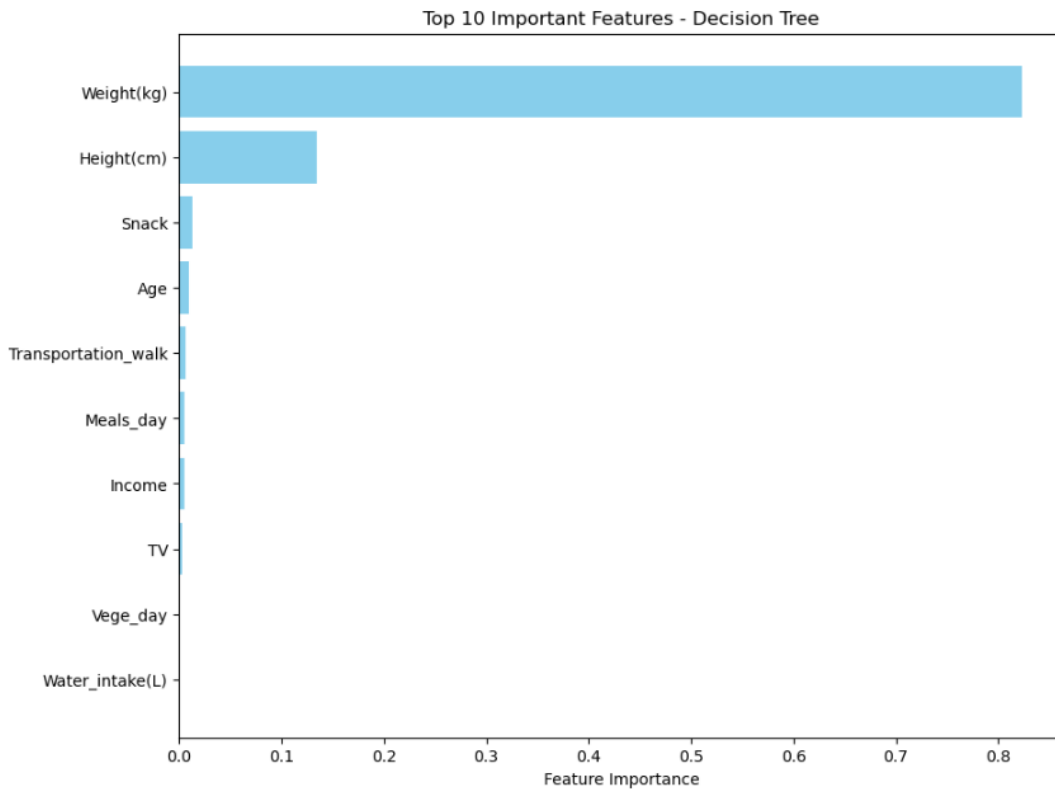
**Strength:** This is best model for identifying cardiovascular diseases, because it achieved the highest accuracy in this study especially after being tuned with optimized parameters. It performed well for handling high dimensional datasets and thus datasets with many features, like this project where cardiovascular risk effected by various factors is suitable for SVM model.

**Weaknesses:** SVM's performance on medium and low-risk classification was slightly weak, and the model was sensitive to the choice of hyperparameters like C, so need to be careful handle which parameter to be used in our case that show the best to the result.

### 3.5 Feature Importance

**Decision Tree** in this study that provided insight into feature importance. The important features in predicting cardiovascular risk based on Decision Tree are sorted ascending:

Weight(kg), Height(cm), Snack, Age, Transportaion\_walk, Meals\_day, Income, TV, Vege\_day, Water\_Intake(L).

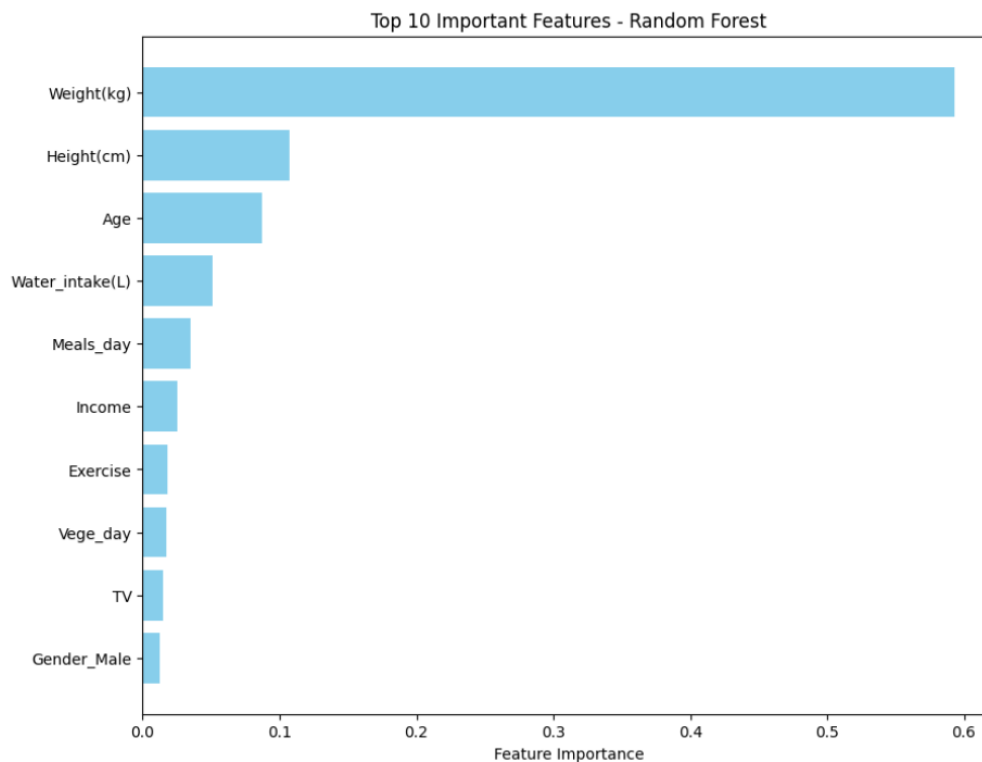


*Figure 3.5.1 The ranking of the important features for Decision Tree*

According to figure, we able to observe that Weight(kg) contributes nearly 80% and Height(cm) contributes nearly 20% to the model while remaining features contributing very little only. This is because the Decision Tree algorithm finds that splits based on Weight(kg) provide the most information in distinguishing the output. Thus, focusing on Weight(kg) and Height(cm) might be beneficial for model simplicity and interpretability.

**Random Forest** in this study that provided into feature importance. The important features in predicting cardiovascular risk based on Random Forest are sorted ascending:

Weight(kg), Height(cm), Age, Water\_intake(L), Meals\_day, Income, Exercise, Vege\_day, TV, Gender\_Male



*Figure 3.5.2 The ranking of the important features for Random Forest*

According to figure, we able to observe that Weight(kg) contributes nearly 60% and Height(cm) and Age contributes nearly 10% to the model while remaining features contributing little to the model. Similar with the Decision Tree, Weight(kg) has highest weightage indicate that the Random Forest relies heavily on it for making prediction. The remaining importance feature is slightly different compared with Decision Tree because Random Forest provides a more robust and stable estimate of feature importance by averaging the importance scores from the trees.

#### **4.0 Conclusion**

This assignment project successfully developed three machine learning model to predict cardiovascular risk which are Decision Tree, Random Forest and Support Vector Machine (SVM). To improve the accuracy and lower the overfitting risk, some efforts have been made to preprocessing the data, including separating feature and the target column, performing label encoding to the target variable, dividing the features into categorical and numerical column then identifying the nominal and ordinal features, encoding categorical variables and standardize numerical column, lastly splitting the data into training set and testing set. Beside that from study we found that tuning process is important for each model to increase the performance and efficiency of prediction.

Among these three models, SVM shown the highest accuracy followed by Random Forest model and Decision Tree. This is because SVM show a highly consistent accuracy when handling complex and high-dimensional data. It able to look for the most optimal decision boundaries and handle non-linear relationships data to make the model become more effective for the prediction. However, Decision Tree provide simpler and more interpretable model but tend to overfit when come into prediction complex datasets. Random Forest is the improve version of Decision Tree that combining multiple decision trees in order to improve the accuracy and reduce the overfitting risk. Hence, based on the comparison of accuracy and the characteristic of models, SVM is recommended for use in the cardiovascular risk prediction model.