Text

Description automatically generated

**UNIVERSITI TEKNOLOGI MARA**

**ASSIGNMENT 1A**

**DESIGNING A PREDICTIVE MODEL FOR CATEGORICAL/BINARY DATA**

1. Find a predictive problem with categorical data.
2. Search for a dataset whose outcome is either binary or categorical data.
3. Train and test the data with **two ML algorithms (classifier)**.
4. Fine-tuning at least one hyper-parameter of the two ML models to analyze its effect to the performance.
5. Write a technical report that discusses the experimental analysis.

**RAMADAN YACIN IBRAHIM**

(SID- 2021935561) (GRP- CS7073A)

**FOR:**

**PROFESOR MADYA ZAIDAH BINTI IBRAHIM**

**2022**

1. **Introduction**

In general, Machine learning is a branch of artificial intelligence and computer science which focuses on capability of a machine to imitate intelligent human behaviour., gradually improving its accuracy, however collecting a dataset for your AI project might seem like an easy task that can be done in the background while you pour most of your time and resources into building the machine learning model. A dataset in machine learning is, quite simply, a collection of data pieces that can be treated by a computer as a single unit for analytic and prediction purposes. This means that the data collected should be made uniform and understandable for a machine that doesn't see data the same way as humans do (According to South Africa).

In this Assignment, we are going to train and test on a dataset from Kaggle dataset using the following Machine Learning algorithms :

* Support Vector Machines
* random forest classifier

then several fine-tuning will be done on the hyper-parameter to evaluate their performance and accuracy. From the dataset used, features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.

1. **Materials and Methods**
2. ***Predive problem***

Each year number of deaths is increasing extremely because of breast cancer. It is the most frequent type of all cancers and the major cause of death in women worldwide. Any development for prediction and diagnosis of cancer disease is capital important for a healthy life. Consequently, high accuracy in cancer prediction is important to update the treatment aspect and the survivability standard of patients. Machine learning techniques can bring a large contribute on the process of prediction and early diagnosis of breast cancer.

1. ***Dataset (Categorical)***

Description of the Dataset

*Graphical user interface, text, application

Description automatically generated*

In this Experiment, we will be using the dataset of Breast Cancer by Dr. William H. Wolberg, W. Nick Street and Olvi L. Mangasarian.

**Attribute Information:**

1) ID number

2) **Diagnosis** (M = malignant, B = benign)

**Ten real-valued features are computed for each cell nucleus:**

1) radius (mean of distances from center to points on the perimeter)

2) texture (standard deviation of gray-scale values)

3) perimeter

4) area

5) smoothness (local variation in radius lengths)

6) compactness (perimeter^2 / area - 1.0)

7) concavity (severity of concave portions of the contour)

8) concave points (number of concave portions of the contour)

9) symmetry

10) fractal dimension ("coastline approximation" - 1)

In this dataset the variable we are going to predict is the **diagnosis** on the second columns.

Table

Description automatically generated

Fig 1 : data csv file

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. In addition, all feature values are recoded with four significant digits. Missing attribute values: none and Class distribution: 357 benign, 212 malignant

Link : <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>

1. ***IDE and Programming language***

In this assignment we are using Python IDLE and programming language. Python is a programming language that supports the creation of a wide range of applications. Developers regard it as a great choice for Artificial Intelligence (AI), Machine Learning, and Deep Learning projects.

1. ***Machine learning Algorithm***

* Support Vector Machines

The SVM [1] was designed to solve large margin classification problems as an implementation of statistical learning theory. It establishes a separating hyperplane and a maximal margin free of training data by choosing a subset SV ⊂ X called support vectors.

*Python Code :*

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.svm import SVC

from sklearn.metrics import classification\_report,confusion\_matrix

data = pd.read\_csv('breast-cancer.csv')

X = data.drop('diagnosis',axis=1)

y = data['diagnosis']

X\_train , X\_test , y\_train, y\_test = train\_test\_split(X, y, test\_size = 0.10)

svclassifier = SVC(kernel='linear')

svclassifier.fit(X\_train,y\_train)

y\_pred = svclassifier.predict(X\_test)

print(confusion\_matrix(y\_test,y\_pred))

print(classification\_report(y\_test,y\_pred))

y\_pred = svclassifier.predict(X\_test)

print(confusion\_matrix(y\_test,y\_pred))

print(classification\_report(y\_test,y\_pred))

* random forest classifier

RF algorithm (Random Forest, RF-algorithm) is actively applied for the RF classifier development in many scientific spheres for solving different classification problems. Usually, it is assumed that the objects of the dataset U which is used for the RF classifier development, are deleted into L classes with the labels from the set Y L}...,,...,,2,1{ = ηl (η l is the label of the l -th class; 1 ηl ≤≤ L ).

*Python Code :*

from sklearn import datasets

import pandas as pd

data = pd.read\_csv('breast-cancer.csv')

print(data)

# Import train\_test\_split function

from sklearn.model\_selection import train\_test\_split

X = data [{'radius\_mean', 'texture\_mean', 'perimeter\_mean', 'area\_mean', 'smoothness\_mean', 'compactness\_mean', 'concavity\_mean', 'concave points\_mean', 'symmetry\_mean', 'fractal\_dimension\_mean', 'radius\_se', 'texture\_se', 'perimeter\_se', 'area\_se', 'smoothness\_se', 'compactness\_se', 'concavity\_se', 'concave points\_se', 'symmetry\_se', 'fractal\_dimension\_se', 'radius\_worst', 'texture\_worst', 'perimeter\_worst', 'area\_worst', 'smoothness\_worst', 'compactness\_worst', 'concavity\_worst', 'concave points\_worst', 'symmetry\_worst', 'fractal\_dimension\_worst'}]

y = data['diagnosis']

# Split dataset into training set and test set

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.60) # 80% training and 20% test

#Import Random Forest Model

from sklearn.ensemble import RandomForestClassifier

clf=RandomForestClassifier(n\_estimators=100)

#Train the model using the training sets y\_pred=clf.predict(X\_test)

clf.fit(X\_train,y\_train) #train model on training set

y\_pred=clf.predict(X\_test) #predict model on test set

#Import scikit-learn metrics module for accuracy calculation

from sklearn import metrics

# Model Accuracy, how often is the classifier correct?

print("Accuracy:",metrics.accuracy\_score(y\_test, y\_pred))

feature\_imp = pd.Series(clf.feature\_importances\_,index=X\_train.columns).sort\_values(ascending=False)

feature\_imp

import matplotlib.pyplot as plt

import seaborn as sns

# Creating a bar plot

sns.barplot(x=feature\_imp, y=feature\_imp.index)

# Add labels to graph

plt.xlabel('Feature Importance Score')

plt.ylabel('Features')

plt.title("Visualizing Important Features")

#plt.legend()

plt.show()

***Diagram

Description automatically generated***

Fig 2: workflow

1. **Experimentation (Train and testing )**

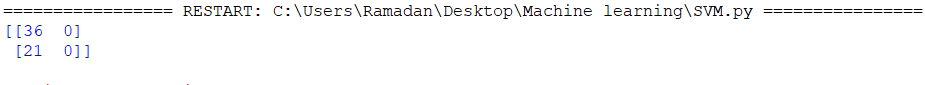
In this experiment we will test the mentioned algorithms (SVM and RF) on the dataset to obtain the accuracy of each of the ML algorithm performs.

Attention : All experiments on the machine learning algorithms described during this paper were conducted using Scikit-learn library and Python programming language. Scikit-learn also known as sklearn is a free software machine learning library for the Python programming language

Graphical user interface, text, application, email

Description automatically generated

The first test was performed with Support Vector Machine with hyper-plane (Kernel) of Gaussian (RBF). After running the algorithm, the performance was obtained. The following image is the result.



A picture containing table

Description automatically generated

Fig 3: Support Vector Machine (rbf) **Accuracy : 0.63**

The accuracy obtained with Support Vector Machine (rbf) is 0.63. secondly, we have performed the test on with the same algorithm by different kernel. This time we used polynomial. After running the algorithm, the performance was obtained. The following image is the result.

Table

Description automatically generated

Fig 4: Support Vector Machine (Poly) **Accuracy : 0.72**

The accuracy obtained with Support Vector Machine (Poly) is 0.72.

Text, table

Description automatically generated

Fig 5: Support Vector Machine (sigmoid) **Accuracy : 0.65**

Lastly with Support Vector Machine, we performed the sigmoid kernel which result 0.65 on the accuracy.After performing several time, the accuracy of the Support Vector Machine on the dataset does not go above 80.

After testing and train on Support Vector Machine, we switched to the Random Forest algorithms.

Graphical user interface, text, application, email

Description automatically generated

After multiple train and test, Random Forest algorithms, the result was obtained.

Table

Description automatically generated

****

Fig 6: Random Forest **Accuracy : 0.956140350877193**

the Accuracy obtained with this classifier has reached 0.95.

Chart, bar chart

Description automatically generated

Fig 7: bar plot

1. **Result and Discussion**

|  |  |
| --- | --- |
| **Support Vector Machine** | **Random Forest** |
| **rbf : 0.63** | **0.96** |
| **Poly : 0.72** |
| **Sigmoid : 0.65** |

Table 1 : Accuracy of ML algorithms

After applying Machine Learning Algorithms on Breast Cancer Wisconsin Diagnostic dataset. We used Confusion Matrix, Accuracy and F1 Score as performance metrics to evaluate and compare the two models and identify the best algorithm for the breast cancer Prediction.

From table 1 , we first notice that the Support Vector Machine’s accuracy can be affected by fine-turning the kernels, not in significant way but can make difference on the performance and accuracy. The polynomial kernel has got the higher accuracy among the other kernels with 0.72 accuracy. Secondly and the most important one is that the difference of accuracy of Support Vector Machine and Random Forest is significantly high with 0.96 for Random Forest which make it the best ML algorithms classifier for the breast cancer dataset.

1. **Conclusion**

On the Breast Cancer dataset, we applied two main algorithms which are: SVM, Random Forests and calculate, compare, and evaluate different results obtained based on confusion matrix, accuracy to identify the best machine learningalgorithm that are precise, reliable and find the higher accuracy.

All algorithms have been programmed in Pythonusing scikit-learn library in IDLE environment. After an accurate comparison between our models, we foundthat Random Forest achieved a higher efficiency of 0.96, andoutperformed Support Vector Machine. Random Forest has demonstrated its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of accuracy.