ARTIFICIAL INTELLIGENCE

NAME	P.GOWTHAM
TEAM ID	344
PROJECT NAME	DIABETES PREDICTION
DATE	01 NOV 2023

INTRODUCTION

- *Diabetes is a common chronic disease and poses a great threat to human health. The characteristic of diabetes is that the blood glucose is higher than the normal level, which is caused by defective insulin secretion or its impaired biological effects, or both. Diabetes can lead to chronic damage and dysfunction of various tissues, especially eyes, kidneys, heart, blood vessels and nerves.*

- *Diabetes can be divided into two categories, type 1 diabetes (T1D) and type 2 diabetes (T2D). Patients with type 1 diabetes are normally younger, mostly less than 30 years old. The typical clinical symptoms are increased thirst and frequent urination, high blood glucose levels.*

- *This type of diabetes cannot be cured effectively with oral medications alone and the patients are required insulin therapy. Type 2 diabetes occurs more commonly in middleaged and elderly people, which is often associated with the occurrence of obesity, hypertension, dyslipidemia, arteriosclerosis, and other diseases.*

- *With the development of living standards, diabetes is increasingly common in people's daily life. Therefore, how to quickly and accurately diagnose and analyze diabetes is a topic worthy studying. In medicine, the diagnosis of diabetes is according to fasting blood glucose, glucose tolerance, and random blood glucose levels.*

- *Moreover, more and more studies used ensemble methods to improve the accuracy proposed a newly ensemble approach, namely rotation forest, which combines 30 machine learning methods. Proposed a machine learning method, which changed the SVM prediction rules.*

CLASSIFICATION

- *In this section, we used decision tree, RF and neural network as the classifiers. Decision tree and RF can implement in WEKA, which is a free, non-commercial, Source machine learning and data mining software based on JAVA environment.*

- *Neural network can be implemented in MATLAB, which is a commercial mathematics software exploited by MathWorks, Inc. It is used for algorithmic development, data visualization, data analysis and provides advanced computational language, and interactive environment for numerical calculation.*

RELATED WORKS

- *Sajida et al. in [20] discusses the role of Adaboost and Bagging ensemble machine learning methods [18] using J48 decision tree as the* *basis for classifying the Diabetes Mellitus and patients as diabetic or non diabetic, based on diabetes risk factors.*

- *Results achieved after the experiment proves that, Adaboost machine learning ensemble technique outperforms well comparatively bagging as well as a J48 decision tree. Orabi et al. in [19] designed a system for diabetes prediction,*

- *whose main aim is the prediction of diabetes a candidate is suffering at a particular age. The proposed system is designed based on the concept of machine learning, by applying decision tree.*

MODEL VALIDATION

- *In many studies, authors often used two validation methods, namely hold-out method and kfold cross validation method, to evaluate the capability of the model .*

- *According to the goal of each problem and the size of data, we can choose different methods to solve the problem. In hold-out method, the dataset is divided two parts, training set and test set.*

- *The training set is used to train the machine learning algorithm and the test set is used to evaluate the model . The training set is different from test set. In this study, we used this method to verity the universal applicability of the methods.*

- *In k-fold cross validation method, the whole dataset is used to train and test the classifier . First, the dataset is average divided into k sections, which called folds.*

- *In training process, the method uses the k-1 folds to training the model and onefold is used to test. This process will be repeat k times, and each fold has the chance to be the test set. The final result is the average of all the tests performance of all folds .*

- *The advantage of this method is the whole samples in the dataset are trained and tested, which can avoid the higher variance . In this study, we used the five-fold cross validation method.*

IMPORT REQUIRED LIBRARIES

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O
import matplotlib.pyplot as plt #to plot charts
import seaborn as sns #used for data visualization
import warnings #avoid warning flash
warnings.filterwarnings('ignore')
import os
for Dir name, _, filenames in os.walk('/Kaggle/input'):
for filename in filenames:
print(os.path.join(drimane, filename))
```

OUTPUT

```
Pregnancies                0
Glucose                    0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
ditype: int64
```

CONCLUTION

- *Diabetes mellitus is a disease, which can cause many complications. How to exactly predict and diagnose this disease by using machine learning is worth studying.*

- *According to all above experiments, we found the accuracy of using PCA is not good, and the results of using all features and using mRMR have better results.*

- *The result, which only used fasting glucose, has a better performance especially in Luzhou dataset.*

- *It means that the fasting glucose is the most important index for prediction, but only using fasting glucose cannot achieve the best result, so if want to predict accurately, we need more indexes.*

- *In addition, by comparing the results of three classifications, we can find there is not much difference among random forest, decision tree and neural network, but random forests are obviously better than the another classifiers in some methods.*

- *The best result for Luzhou dataset is 0.8084, and the best performance for Pima Indians is 0.7721, which can indicate machine learning can be used for prediction diabetes, but finding suitable attributes, classifier and data mining method are very important.*

- *Due to the data, we cannot predict the type of diabetes, so in future we aim to predicting type of diabetes and exploring the proportion of each indicator, which may improve the accuracy of predicting diabetes.*