CE 802 Machine Learning

Assignment: Design and Application of a Machine Learning System for a Practical Problem

University of Essex

School of Computer Science and Electronic Engineer

Pilot study

Submitted by

Supasun khumpraphan

Registration number:2110366

Supervisor: Dr. Vito De Feo

MSc - artificial intelligence

Word count:747

Executive summary

More and more people in the world are dying from diabetes today, and what is worrying is that the prevalence of diabetes caused by obesity and lifestyle problems among adolescents and working adults is on the rise. This group of people is the main force in driving the country's economy and society. however, these problems will be eliminated if we can predict with machine learning or deep learning whether a patient has diabetes based on the information provided by the National Health Service, which tells the patient characteristics such as age, and blood sugar levels. average, ethnicity To help people know the disease in advance and be able to deal with it in a timely manner. resulting in fewer deaths or less severe illnesses.

1. Solutions

We have used four solutions to address this problem: Type of Predictive Task, feature importance, learning procedures, evaluate the performance.

1.1. Type of Predictive Task

In this solution, we can see that the data type is "supervised learning" because in our dataset there is useful information as a feature variable in multiple columns and has a "label" that labels the answer. Also, the answer given is in Discrete format with only True and False. which is in the form of Binary classification used to classify while regression predicts continuous values in continuous form. Which does not match the label pattern in this problem that only needs True and False, not continuous values.

1.2. feature important

in type predictions supervised learning, data is considered to be very important such as

- **Heredity** is a factor that is difficult to avoid. But the physical health checks controlling weight or avoiding other factors or periodic blood tests will help reduce the risk.
- **Being overweight** can have insulin resistance. When the body lacks insulin or insulin does not function. It causes high sugar levels. There is a risk of various complications that follow.
- **Aging**, therefore, everyone over the age of 35 should have their blood sugar checked regularly.
- **lack of exercise** has the potential for many diseases, including diabetes.
- **eating behavior** not eating healthy food eating only sweets can lead to diabetes.
- **have high blood pressure** People with high blood pressure are more likely to develop diabetes than people with normal blood pressure.
- **Stress** can lead to many diseases that can lead to diabetes.

The information mentioned above will be a feature. Important to help predict the model effectively.

1.3. learning procedures

In the prediction "Supervised Learning" type classification. There are many types. We need to know the advantages of each model for use in different situations. We have done 9 models as follows.

- **Decision Tree Classifier** is a popular, easy-to-use, easy-to-understand, effective algorithm that is based on Random Forest, which is one of the best algorithms today. It is a rule-based model that creates an if-else rule from the values of each feature without an equation to define the relationship between feature & target. It is important to create a Decision Tree to split each feature's value, minimizing the cost function's value to a minimum[2].
- **Random Forest** is a model that trains multiple decision trees together, where each tree receives a feature and data is a random subset of all features and data. When making a prediction, give each Decision Tree makes its predictions and selects the best prediction result from the most voted prediction values! This technique is called bagging or bootstrapping[2].
- **XGBoost** is a model that trains multiple Decision Trees in each tree, where each decision tree learns from the errors of the previous tree, making prediction accuracy more and more accurate. When the learning of the tree continues until it is deep enough, and the model stops learning when there are no more error patterns from the previous tree to learn, both Random Forrest and XGBoost are ensemble models. Using many models comes together to form a complex model[2].
- **Multi-Layer Perceptron** classifier is a type of neural network (ANN). The simplest MLP consists of at least three layers of nodes: the input layer, the input layer, and the hidden layer. hidden layer and the output layer is a model that takes quite a long time because it is deep learning, but it is a model that predicts values very accurately[3].
- There is also a model that is used to make and compare 5 other models, including Logistic Regression, K-nearest Neighbors, Support Vector Classification, Gaussian Naive Bayes, Stochastic Gradient Descent Classifier

1.4. Evaluate the performance

In order to put the model into practice, It is necessary to measure the performance of the Model before the Model is powerful enough to be developed or used in various fields. performance metrics are measured in tables. it is called Confusion Matrix. Confusion Matrix data is an important table for measuring machine learning's ability to solve problems. classification can predict The values are as follows: True Positive (TP) is what the program predicts is true and values are true True Negative (TN) is what the program predicts is not true and values are not True False Positive (FP) is what the program predicts is "true" but is "false". False Negative (FN) is what the program predicts is "not true" but is "true". This is generally a popular measure. Together in research and work, there are 3 values and the equation is Precision is a measure of the accuracy of the data, Recall is a measure of the accuracy of the Model, Accuracy is a

measure of the accuracy of the model.

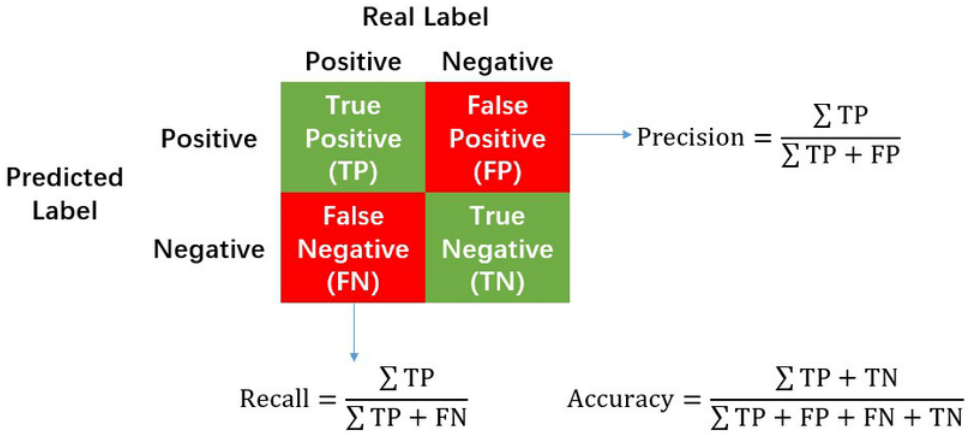

Fig. 1: Confusion Matrix[1].

1.5. Reference

[1]Jun M.,(2019). *Analyzing the Leading Causes of Traffic Fatalities Using XGBoost and Grid-Based Analysis: A City Management Perspective* <https://www.researchgate.net/figure/Calculation-of-Precision-Recall-and-Accuracy-in-the-confusion-matrix_fig3_336402347>

[2]Micheal D.,(2020). *Random Forest vs XGBoost* <https://www.educba.com/random-forest-vs-xgboost/>

[3]Alisneaky Z.,(2011). *Multilayer perceptron* <https://en.wikipedia.org/wiki/Multilayer_perceptron>