Diabetes Prediction Based on Health Indicators

Galpayage G.D.T.G.
Department of Computer Engineering
University of Ruhuna
Galle, Sri Lanka
galpayage_gdtg_e22@engug.ruh.ac.lk

Sundarasekara G.O.
Department of Computer Engineering
University of Ruhuna
Galle, Sri Lanka
githmi_oshani_sundarasekara_e22@engug.ruh.ac.lk

***Abstract*— The prevalence of diabetes has been increasing day by day and has urged the necessity of early detection and treatment. In this study, we use machine learning to predict whether a person is having diabetes or not or is in a pre-diabetes condition, based on a set of health indicators using two machine learning algorithms: logistic regression and random forest. Our goal is to determine the best model that would predict the diabetes condition of individuals more accurately. The dataset is preprocessed by incorporating various pre-processing techniques in order to obtain the best accuracy possible. The model performance is measured using performance metrics like F1 score, the area under the ROC curve and also using the confusion matrix. The developed predictive models promise un-biasness and fairness. In conclusion, our work showcases the capability of machine learning in developing predictive models by effectively combining with healthcare and take preventative healthcare measures.**

Keywords—diabetes prediction, machine learning, logistic regression, random forest.

I. INTRODUCTION

Diabetes is a serious chronic disease that is characterized by the elevated levels of blood glucose (or blood sugar), in which individuals lose their ability to regulate levels of glucose in the blood effectively. This can lead to reduced quality of life and life expectancy of humans. In the field of healthcare and medical diagnostics, early identification and prediction of diabetes based on different health indicators is important for its' effective management. As at present, there is a high prevalence of diabetes patients, it is crucial to identify the individuals who are at a higher risk of getting diabetes, and facilitate treatments to them. Hence, machine learning techniques can be used in this project so that it would help the potential diabetic patients to prevent from it and it will also save the healthcare costs. Ultimately, both the patients and the doctor can obtain a second opinion on the diagnosis of diabetes. The primary objective of this project is to develop a robust and accurate model for diabetes prediction, by utilizing a diverse array of health indicators and predict the likelihood of a person getting diabetes beforehand.

For this project, we used the logistic regression and random forest algorithms considering their suitability to our project. Logistic regression is more suitable for our project since the output 'Diabetes_binary' and most of the feature data are binary and due to the simplicity of the model. However, this algorithm struggles to capture the non-linear relationships hence, there is possibility of the data oozing it's value. However, the second algorithm that we have used excels in handling non-linearities and a large no. of feature. But, the computational complexity stands as a disadvantage in using this algorithm in our project. However, the combination of these two algorithms would provide a well-rounded and accurate prediction on the potential risk of having diabetes, meeting our project's objectives.

II. METHODOLOGY

A. Data

The data taken for this project was sources from Kaggle [1] , which provides a diverse collection of datasets.

The dataset comprises 253,680 instances, each with 21 features including both the clinical and lifestyle-related variables, such as some demographics, lab test results, health history, and some personal information.

The 21 features used are: HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education and Income.

The output variable is "Diabetes_Binary". It is coded as 0 for negative cases (non-diabetic individuals) and 1 for positive cases (individuals diagnosed with diabetes). The dataset is balanced and there were no any missing values.

B. Pre-processing

Under pre-processing process, we first down-sampled the majority class of data in order to make the dataset balanced. Then, we first plotted a heat map in order to find the correlation among the features. Then we identified that there was no any relationship between having a health care coverage and getting diabetes, and also that there was no any relationship between seeing or not seeing a doctor due to cost and getting diabetes, all of which had a correlation very much closer to zero. Hence we removed those features. Next, we used Min-Max scaling in order to rescale 03 features named 'BMI', 'MentHlth', 'PhysHlth' into the range of 0 and 1, since all the other features had values in between 0 and 1. Then we did Principal Component Analysis and reduced the dimensions to 10. We randomly splitted the dataset as 70% for training data and 30% for test data.

We randomly splitted the dataset such that 70% is taken as the training data and the remaining 30% (70 – 30 split) is taken as the test data.

C. Algorithm

Logistic Regression : Suppose that x is the independent variable and y is the dependent variable. Then, assuming a linear relationship among the variables, the logistic function or the logit function is used maps y as a sigmoid function of x. The logistic function can thus be represented as:

$$f(x) = \frac{L}{1+e^{-K(x-x_0)}} \quad (1)$$

The standard logistic function is a logistic function with parameters k = 1, x0 = 0, L = 1. This reduces the logistic function as below:

$$f(x) = \frac{1}{1+e^{-x}} \quad (2)$$

As you can see, the logistic function returns only values between 0 and 1 for the dependent variable, irrespective of the values of the independent variable. This is how logistic regression estimates the value of the dependent variable.

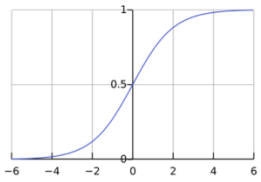

Fig. 1. Logistic regression [2]

Suppose there are multiple independent variables that affect the value of the dependent variable as in our case. To model such input datasets, logistic regression formulas assume a linear relationship between the different independent variables. Here, the logistic function or the logistic function, which is the equation between x(independent variables) and y(dependent variable), maps y(dependent variable) as a sigmoid function of x(independent variables) The logistic function can be represented as:

$$f(x) = \frac{1}{1+e^{-(\theta_0 x_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta n x n)}} \quad (3)$$

The cost function is :

Cost function = -ylog(f(x))-(1-y)log(1-f(x))

-ylog(f(x)): This part is zero when y=0

(1-y)log(1-f(x)) : This part is zero when y = 1

Random Forest: Here, a multitude of decision trees are constructed during training by selecting a sub set of data points and a subset of features. The generated output from each tree are used to find the final output by taking the mode of the classes (classification) or the mean prediction (regression) of the individual trees.

D. Implementation

Both algorithm codes for both the Logistic Regression [3] and Random Forest Classifier [4] models were taken from 'scikit-learn' machine learning library.

The algorithm part of the code was modified when using each classifier when tuning the hyperparameters.

In logistic regression, we adjusted the regularization parameter (C) to control over fitting, penalty to specify the type of regularization to be used, and solver to define the weights that minimize the logistic loss function during model training.

```
parameter_grid = {
   'C': [0.001, 0.01, 0.1, 1, 10, 100], #Inverse regularization strength
   'penalty': [ 'l1', 'l2' ], #Penalty
    'solver': ['liblinear', 'lbfgs' ,  'newton-cg', 'newton-cholesky', 'sag',
'saga'] #Algorithm to use in optimization
}
```

The above parameters were varied in using.

```
grid_search = GridSearchCV(logreg, parameter_grid, cv=3,
scoring='f1', verbose=1)
logreg_best = LogisticRegression(C=10 ,penalty='l2' , solver='sag')
```

In random forest, we adjusted n_estimators

```
parameter_grid_random = {
   'n_estimators': [100, 200],  # Number of trees in the forest
   'max_depth': [ 12, 19],  # Maximum depth of each tree
    'max_features': ['auto', 'sqrt', 'log2'],  # Number of features
considered at each split
   'min_samples_split': [2, 5 ],  # Minimum samples required to split a
node
   'min_samples_leaf': [1, 2, 4],  # Minimum samples required at a leaf
node
   'bootstrap': [True, False]  # Whether to use bootstrap sampling
```

The above parameters were varied in using.

```
grid_search_random_forest = GridSearchCV(randomForest,
parameter_grid_random, cv=10, scoring='f1', verbose=1)
randomForest_best = RandomForestClassifier(n_estimators=100,
max_depth=12, max_features='log2',min_samples_split=5
, min_samples_leaf=4, bootstrap=True)
```

III. RESULTS

Logistic Regression Performance : F1 score = 0.7548039260863554.

The confusion matrix and ROC curve:

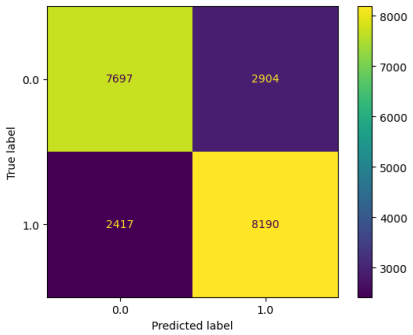

Fig. 2. Confusion matrix obtained for Logistic Regression

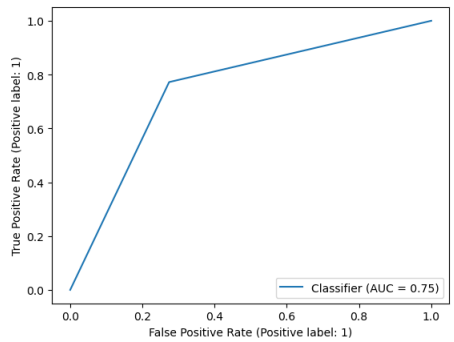

Fig. 3. ROC curve obtained for Logistic Regression

Random Forest Performance : F1 score = 0.7625535754568012.

The confusion matrix ROC curve :

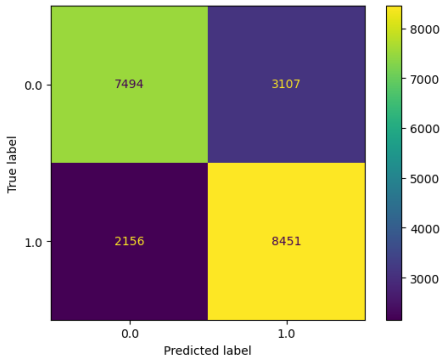

Fig. 4. Confusion matrix obtained for Random Forest Classifier

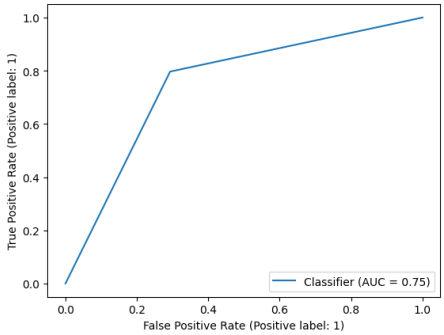

Fig. 5. ROC curve obtained for Logistic Regression

TABLE I. VALUES FOR THE PERFORMANCE METRICS

Performance Metric	**Logistic Regression Model**	**Random Forest Classifier Model**
F1 score	75.50	75.74
Accuracy	74.71	74.35
Precision	74.16	72.75
Recall	76.90	78.99
Specificity	72.46	69.58
AUC of ROC curve	0.75	0.74

Above metrics were found using the confusion matrices and the ROC curves that were obtained.

IV. DISCUSSION

A. *Discussion*

The Logistic regression model exhibited a lower time complexity during training and converging, than Random forest. Almost all the performance metrics found from both the models were nearly closer. The F1 score and recall were greater in the Random Forest Classifier model than those of Logistic Regression model. All the other performance metrics obtained were higher in the Logistic regression model. The ROC curves from both are partially converged, which suggests that the logistic regression and random forest models exhibit similar performance in certain regions of the decision space. Almost all the performance metric values are moderately high, which shows that the models were moderately accurate.

B. *Ethical Aspects*

Possible negative social impact: Since there is a possibility of the model to provide an incorrect output, a person with no diabetes can be labeled as a having diabetes or in a pre-diabetes condition, which might affect him mentally and physically.

Data privacy: Though the data in the dataset are anonymous, there are technologies which can de-identify them today. Hence, it must be ensured that there are no any data privacy issues.

Bias and Fairness: The labels obtained for the data are unbiased since we balanced the dataset. The dataset used is fair since the data obtained from the features, for example the age, represents almost all the possible ages of humans. Similarly, all such data in features are fair with a representation from all categories.

Involvement of healthcare professionals: In order to validate the data, the perspectives and feedback from the healthcare professionals could be incorporated in order to increase data reliability, which is a point for improvement.

C. *Conclusion*

Logistic regression model demonstrates high accuracy, whether the person has diabetes, prediabetes condition or not, than the other model. Random forest classifier will correctly predict the people in prediabetes or diabetes conditions than the other model(recall). Logistic regression model predicts a higher percentage of actual people in prediabetes or diabetes condition out of all positive predictions(precision). But since the f1 score gives a harmonic mean of precision and recall, it provides a better conclusion about performance than all these. As the f1 score is greater in the Random Forest classifier model, and since the true prediction of people in prediabetes and diabetes condition is paramount, we can conclude that Random Forest classifier model is the best for our application. Though it takes a longer time to converge, that has no significant effect for our application. This model can be further improved and accuracy can be increased by accruing more data in the future. Future research could explore

strategies to address potential dataset challenges for improved model outcomes.

V. REFERENCES

[1] CDC, "Diabetes Health Indicators Dataset," kaggle, [Online]. Available: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_binary_health_indicators_BRFSS2015.csv. [Accessed 01 2024].

[2] s.-l. developers, "sklearn.linear_model.LogisticRegression," scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html. [Accessed 01 2024].

[3] s.-l. developers, "sklearn.ensemble.RandomForestClassifier," scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. [Accessed 01 2024].

[4] calbimontedaniel, "Data Mining Introduction Part 10: Microsoft Logistic Regression," sqlservercentral, [Online]. Available: https://www.sqlservercentral.com/steps/data-mining-introduction-part-10-microsoft-logistic-regression. [Accessed 01 2024].