MedBot – An NLP based Chat-Bot for Diabetes Prediction using Logistic Regression

Sreekar K, Aswin V, Vishnu Narayanan. Amrita Vishwa Vidyapeetham

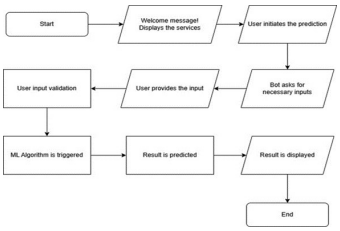
(A) GOAL & CONTRIBUTIONS

Our **goal** is to streamline the process of diabetes prediction:

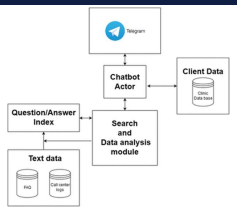
1. Introducing a Chat-Bot that can give preliminary prediction of diabetes given user data.
2. We have explored a number of **Machine Learning techniques** for classification/prediction, and, implemented using **Logistic Regression**.
3. The Chat-Bot interaction data-stream has been implemented using **NLP mechanisms**, facilitating a coherent conversation involving user queries regrading diabetes.

Keywords: ***Chat-Bot, Logistic Regression, Natural Language Processing, Diabetes, Telegram, Health Care.***

(B) Flow Diagram of Algorithm



(C) Chat-Bot API Flow and Interaction



(D) Dataset Description

The dataset is divided into TRAIN,TEST sub directories for easy access. The train-test split after extraction was done was 70-30.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Dataset Link: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

(E) SMOTE Analysis

SMOTE class from the imbalanced-learn library to oversample the minority class in the dataset. The fit_resample method is used to apply SMOTE and obtain the resampled feature matrix (X_resampled) and target variable (y_resampled).

```
33]: import pandas as pd
     from imblearn.over_sampling import SMOTE

     data = pd.read_csv(r'C:\Users\Darksoul\Desktop\PIMA Indian Dataset\archive (2)\diabetes.csv')

     X = data.drop(['Outcome', 'BloodPressure', 'Age'], axis=1)
     y = data['Outcome']

     smote = SMOTE(random_state=42)
     X_resampled, y_resampled = smote.fit_resample(X, y)

     print("Original dataset shape:", X.shape, y.shape)
     print("Resampled dataset shape:", X_resampled.shape, y_resampled.shape)

Original dataset shape: (768, 6) (768,)
Resampled dataset shape: (1000, 6) (1000,)
```

(F) Pre-Processing

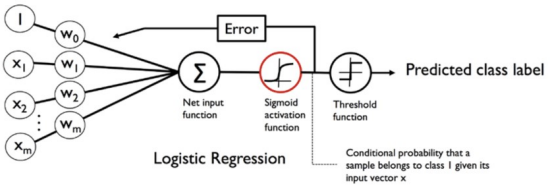
We handle the dataset by first identifying the columns that help in diabetes prediction and eliminate the rest. Then, we impute the missing values by replacing the missing data points by arithmetic mean (or) mode, depending on the type of data.

```
[11] df.head()
```

	Pregnancies	Glucose	BloodPressure	Insulin	DiabetesPedigreeFunction	Outcome
0	6	148.0	72.0	79.799479	0.627	1
1	1	85.0	66.0	79.799479	0.351	0
2	8	183.0	64.0	79.799479	0.672	1
3	1	89.0	66.0	94.000000	0.167	0
4	0	137.0	40.0	168.000000	2.288	1

```
[12] df.shape
     (768, 6)
```

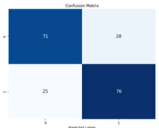
(G) Logistic Regression Architecture



(H) Hyper Parameters Logistic Regression

Parameter Name	Purpose	Value
Penality	Specifies the type of regularization to be applied. 'l2' refers to L2 regularization	12
C	Inverse of regularization strength. Smaller values of C result in stronger regularization.	1.0
Random State	Provides seed for random number generation.	None
max_iter	Maximum number of iterations taken for the solvers to converge.	100

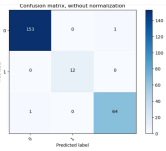
```
Accuracy: 0.73
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.67      0.71        99
           1       0.71      0.80      0.75       101

    accuracy                           0.73       200
   macro avg       0.74      0.73      0.73       200
weighted avg       0.74      0.73      0.73       200
```



(I) Hyper Parameters KMeans with KNN

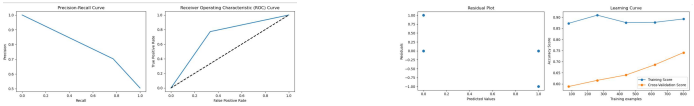
Parameter Name	Purpose	Value
Test size	To split the dataset into training and testing in a ratio	0.3
Random State	This parameter sets the seed for the random number generator used by the data splitter. By using a fixed seed, the random splitting process becomes deterministic, allowing for result reproducibility.	42
No. of Clusters	Number of the centroid which will be formed	3
No. of Neighbours	Its specifies the number of neighbors to consider when making predictions for a data point	Sqrt (no of clusters)

```
KMeans & KNN Classifier Performance Metrics
              precision    recall  f1-score   support

     class 0       0.99      0.99      0.99       154
     class 1       1.00      1.00      1.00        12
     class 2       0.98      0.98      0.98        65

    accuracy                           0.99       231
   macro avg       0.99      0.99      0.99       231
weighted avg       0.99      0.99      0.99       231

accuracy: 0.9913419913419913
```



(J) Hyper Parameters Decision Tree

Parameter Name	Purpose	Value
Test size	To split the dataset into training and testing in a ratio	0.2
Random State	This parameter sets the seed for the random number generator used by the data splitter. By using a fixed seed, the random splitting process becomes deterministic, allowing for result reproducibility.	42
Grid Search CV	The grid search explores various alpha values using 5-fold cross-validation.	5-fold cross-validation

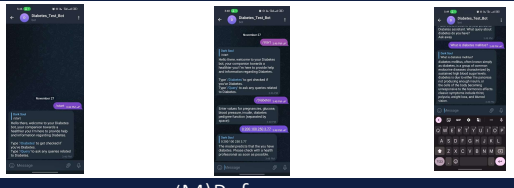



```
Best Hyperparameters: {'criterion': 'gini', 'max_depth': 10
t': 5}
Accuracy: 0.72
Precision: 0.7027027027027027
Recall (Sensitivity): 0.7722772277227723
F1 Score: 0.7358490566037735
Sensitivity (True Positive Rate): 0.7722772277227723
Specificity (True Negative Rate): 0.6666666666666666
```

(K) Accuracy

S.No.	Model	Precision	Accuracy
1.	**Logistic Regression**	**0.74**	**0.74**
2.	**KMeans with KNN**	**0.99**	**0.99**
3.	**Decision Tree**	**0.70**	**0.72**
4.	**Fuzzy CMeans with KNN**	**0.74**	**0.76**

(L) App Demo



(M)References

[1] J.-A. Moldt, T. Festl-Wietek, A. Madany Mamlouk, K. Nieselt, W. Fuhl, and A. Herrmann-Werner, "Chatbots for future docs: exploring medical students' attitudes and knowledge towards artificial intelligence and medical chatbots," Med. Educ. Online, vol. 28, no. 1, p. 2182659, 2023.

[2] D. Choubey, S. Paul, S. Kumar, and S. Kumar, "Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection," in Communication and Computing Systems, 2016.

[3] M. Abedini, A. Bijari, and T. Banirostam, "Classification of Pima Indian diabetes dataset using ensemble of decision tree, logistic regression and neural network," Nternational J. Adv. Res. Comput. Commun. Eng., vol. 9, no. 7, pp. 1–4, 2020.