GIET UNIVERSITY JOURNAL, 2024, VOL.3, NO.3, PP. -112-120

Manuscript received 28th April, 2024; revised 3rd May, 2024; Accepted 6th May, 2024

COMPARATIVE ANALYSIS ON PREDICTION VIA NOVEL MODEL OF DIABETES DISEASE

Aniket Kumar, Department of CSE, GIET University
Yuktamukhi Mohapatro, Department of CSE, GIET University
Bidush Kumar Sahoo, Department of CSE, GIET University
*bidush.sahoo@gmail.com

Abstract: Diabetes is a chronic disease with the potential to cause a worldwide healthcare crisis. According to the International Diabetes Federation, 382 million people are living with diabetes across the whole world. By 2035, this will be doubled to 592 million. Diabetes Mellitus or simply diabetes is a disease caused due to an increased level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. However, early prediction of diabetes is quite a challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as the kidney, eye, heart, nerves, foot, etc. Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. Our task is to help make predictions on medical data. Machine Learning is an emerging scientific field in data science dealing with how machines learn from experience. This paper aims to develop a system that can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. This paper aims to predict diabetes via three different supervised machine-learning methods including SVM, Logistic regression, and ANN. This paper also aims to propose an effective technique for the earlier detection of diabetes disease. Keywords: Machine Learning, Supervised, SVM, ANN, Logistic Regression. This paper is made for early prediction of diabetes disease, by the prediction we can take some precautions for maintaining the diabetes disease.

***Keywords*—Technology, Optoelectronics, photonics, telecommunications, Circuits, systems, applications**

1. Introduction:

The submitted manuscript should be prepared strictly according to the guidelines presented in this paper. Nowadays people try to lead a luxurious life. Due to this luxurious life, many people are skipping either breakfast, lunch, or dinner, due to hungriness, in the body increases of glucose level and that's why diabetes disease is increasing. The goal of diabetes management is to keep blood glucose levels as close to normal as safely possible. Since diabetes may greatly increase the risk for heart disease and peripheral artery disease, measures to control blood pressure and cholesterol levels are an essential part of diabetes treatment as well.

The main scope of our paper is to create a basic predicting system to predict diabetes of patients by the features of pregnancies, glucose, blood pressure, skin thickness, insulin, body mass index(BMI), diabetes pedigree function, and age. By implementing the predictive system the following profit is getting:

o Protect the patients from different diseases.

GIET UNIVERSITY JOURNAL, 2024, VOL.3, NO.3, PP. -112-120

Manuscript received 28th April, 2024; revised 3rd May, 2024; Accepted 6th May, 2024

o Protect from eye blindness.

o Prevent from kidneys and other organs affects.

o Manage the levels of diabetes.

2. Literature Survey

For the paper, we have used a support vector machine classifier to perform the prediction of diabetes disease, and the same classifier has been used in the following areas:
Support vector machines are used in many tasks when it comes to dealing with images. SVMs are particularly used in one definite application of image processing facial features extraction and recognition. While working with facial features, we need algorithms that can properly classify different features based on very fine-tuned feature extractions. Facial expressions are one of the most versatile features while working with image processing. Like your face can be way different from the face of someone else. But still, the algorithm should be able to classify and acknowledge the facial expressions given by both of you. Even though the model is trained on just your face or the other person's.
Image classification is one of classical problems of concerns in image processing. The goal of image processing. The goal of image classification is to predict the categories of the input image using its features. There are various approaches for solving this problem such as k nearest neighbor (K-NN), Adaptive boost (Adaboosted), Artificial Neural Network (ANN), Support Vector machine (SVM). The k-NN classifier, a conventional non- parametric, calculates the distance between the feature vector of the input image (unknown class image) and the feature vector of training image dataset. Then, it assigns the input image to the class among its k-NN, where k is an integer [1].
SVM is one of the best-known methods in pattern classification and image classification. It is designed to separate of a set of training images two different classes,
(x1, y1), (x2, y2), ..., (xn, yn) where x, in R^d, d-dimensional feature space, and y, in {-1,+1}, the class label, with i=1..n [1]. SVM builds the optimal separating hyper planes based on a kernel function (K). All images, of which feature vector lies on one side of the hyper plane, are belong to class -1 and the others are belong to class +1.
Besides there are some integrated multi techniques model for classifying such as Multi Artificial Neural Network (MANN) applying for facial expression classification and multi classifier scheme applying for adult image classifications.
SVMs work amazingly well because of their ability to create the largest margin possible while dividing different points on the feature maps. The basic algorithm hence can work perfectly when it comes to fitting the model data which is based on the finely accumulated facial features, like expressions.

Manuscript received 28th April, 2024; revised 3rd May, 2024; Accepted 6th May, 2024

One of the fields which has the most uncleaned data is the geoscience field. Geospatial data is one of the noisiest data out there. SVMs still work pretty well, especially when the dataset size is not so high. The most problematic set of geospatial analysis problems is inversion problems. For example, the geo-sounding problem. The thing about geospatial data is that not only that the data is noisy, but it is very delicate too. Because of this, the sample points are very close to each other. If we need to classify those samples properly and train a model which is not going to overfit, then we need to use support vector machines.

The thing about geospatial data is that not only that the data is noisy, but it is very delicate too. Because of this, the sample points are very close to each other. If we need to classify those sample properly and train a model which is not going to overfit, then you need to use support vector machines.

In general, there are three fundamental tasks of statistical learning from data: classification, regression, and probability density modelling. Another two basic problems are of great importance: monitoring networks design/redesign and assimilation/integration of data and science-based models, e.g., physical pollution diffusion models, meteorological models, etc.

Usually learning machines are universal tools, i.e., in principle, they can model any mapping (either categorical or continuous data) with any desired precision. The problem is how to select good structure of the machine (for example, multilayer perceptron) and how to tune its parameters. In machine learning there are several approaches for hyperparameter tuning, e.g. splitting of data, cross-validations, etc. For geospatial data geostatistical tools, for example, variography is a valuable tool to control the quality of machine learning procedures and parameters tuning [Kanevski and Maignan 2004].

The generic methodology of spatial data analysis and modelling is presented in Figure 1. As usually, exploratory spatial data analysis (ESDA) is a first step of the study. Quantitative analysis of monitoring networks using topological, statistical and fractal measures helps to describe data representativity, to remove biases in modelling distributions and to select de-clustering procedures [Kanevski and Maignan 2004].

The variography (well-known geostatistical tool to analyse and to model anisotropic spatial correlations) is proposed to be used both at the phase of exploratory spatial data analysis and at the evaluation of the results. Despite of a variogram is a linear two-point statistics, like auto-covariance function for time series, it characterises the presence of spatial structures, anisotropy and scales [Chiles and Delfiner 1999; Cressie 1993]. Variogram analysis of the residuals, in addition to the traditional statistical analysis, is an important step in understanding the quality of modelling results: variograms of the residuals should demonstrate pure nugget effect (i.e. no spatial structures) on training and validation data sets. Variography can be used as an independent tool during tuning of machine learning hyper-parameters. In this case the cost function can be modified taking into account the difference.

GIET UNIVERSITY JOURNAL, 2024, VOL.3, NO.3, PP. -112-120

Manuscript received 28th April, 2024; revised 3rd May, 2024; Accepted 6th May, 2024

We can use support vector machines to classify the handwriting of two different people. SVMs train better when it comes to applications such as detection of the curves and straights used in typical handwriting. SVMs can also be used in pure computer-based texts. For example, a typical text-based classification task is the email spam classifier. In that, we need to classify an email that is spam from the email which is not a spam. It is one of the most used applications in the email delivery systems provided by platforms like Gmail. SVMs can correctly classify the spams from the pool of emails. Some of the SVMs trained on structured data achieve as high as 97 percent accuracy for this application.

When it comes to text-based applications, we are talking about language. Another part of the language exchange is speech-based applications. Audio-based analysis is also a field in which SVMs offer a solution. We can use many audio-based pre-processing functions and then use the SVMs for classification or just simple speech recognition. Both of these applications are quite useful and widely used.

Text categorization is the task of automatically sorting text documents into a set of predefined classes. Text categorization algorithms usually represent documents as bags of words and consequently have to deal with huge numbers of features. Semantic Analysis will be used for feature extraction, eliminating the text representation errors caused by synonyms and polysemes, and reducing the dimension of text vector.

A text classifier for ci is automatically generated by a general inductive process (the learner) which, by observing the characteristics of a set of documents pre-classified under ci or ¯ci, gleans the characteristics that a new unseen document should have in order to belong to ci. In order to build classifiers for C, one thus needs a set Ω of documents such that the value of Φ (dj, ci) is known for every (dj, ci) ∈Ω × C.

Training efficiency (i.e., average time required to build a classifier ˆΦi from a given corpus Ω), as well as classification efficiency (i.e., average time required to classify a document by means of ˆΦi), and effectiveness (i.e., average correctness of ˆΦi's classification behaviour) are all legitimate measures of success for a learner.

SVM is an effective technique for classifying high-dimensional data. Unlike the nearest neighbour classifier, SVM learns the optimal hyper plane that separates training examples from different classes by maximizing the classification margin. It is also applicable to data sets with nonlinear decision surfaces by employing a technique known as the kernel trick, which papers the input data to a higher dimensional feature space, where a linear separating hyperplane can be found. SVM avoids the costly similarity computation in high-dimensional feature space by using a surrogate kernel function. It is known that support vector machines (SVM) are capable of effectively processing feature vectors of some 10 000 dimensions, given that these are sparse. Several authors have shown, that support vector machines provide a fast and effective means for learning text classifiers from examples. Documents of a given topic could be identified with high accuracy

GIET UNIVERSITY JOURNAL, 2024, VOL.3, NO.3, PP. -112-120

Manuscript received 28th April, 2024; revised 3rd May, 2024; Accepted 6th May, 2024

Well, this may surprise you but there are many applications where SVMs are used for basic encryption as well as complex analysis of different materials to see and even break the encryptions and other security measures. SVMs can also be used to detect the encryption schemas uploaded to the images, to hide them. Yes, images are used to hide the encryption patterns in secretive transmissions. When the resolution of images goes higher, the more difficult it becomes to detect those patterns and crack the schema. The SVMs are hence useful when it comes to analyzing and getting the small and astutely observed changes and modifications in the images. This is how SVMs are used in security-based applications. This field is still under heavy research to get even more power from the SVMs.

When it comes to the medical sector, AI has always tried to give a solution. The first use of SVMs in the medical field was based on cancer recognition, which was an image-based application. But the algorithm took its flight in the field when it was first used in the protein analysis tasks. We all know that human-based proteins are very delicate structures and are prone to too much noise as well as errors while using the algorithms for recognition. Another field there is the remote homology which uses SVMs to the fullest. This is where the analysis is dependent on how the protein sequences are modelled.

Of course, the use of SVMs does spread over to the detection of various diseases, based on either image data or text data (value-based). But as they were discussed earlier, we are not going to repeat the same here again. However, it is important to mention that they are widely used in many fine crafted classification applications, necessary for the medical sector.

3. Implementation

Our analysis will use the PIMA Diabetes Dataset available on the dropbox site (https://www.dropbox.com/s/uh7o7uyeghqkhoy/diabetes.csv?dl=0). We obtained the diabetes disease outcomes of female patients to model to predict whether the patients are diabetes or non-diabetic. This dataset contains a total of 9 features or attributes or variables, which were recorded for 768 observations. This data will allow us to create different regression model to determine how different independent variables help to predict our dependent variable, quality. These 9 features of 768 patients help to predict whether the individual person is diabetic or non-diabetic.

Data Preparation:

Our first step was to clean and prepare the data for analysis. We went through different steps of data cleaning. First, we checked the data types focusing on numerical and categorical to simplify the correlation's computation and visualization. Second, we tried to identify any missing values existing in our data set. Last, we researched each column/feature's statistical summary to detect any problems like outliers and abnormal distributions.

GIET UNIVERSITY JOURNAL, 2024, VOL.3, NO.3, PP. -112-120

Manuscript received 28th April, 2024; revised 3rd May, 2024; Accepted 6th May, 2024

Discussion and Analysis

```
# loading the diabetes dataset to a pandas DataFrame
diabetes_dataset = pd.read_csv('/content/diabetes.csv')
```

```
[4] pd.read_csv?
```

```
[5] # printing the first 5 rows of the dataset
    diabetes_dataset.head()
```

	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

```
[6] # number of rows and columns in this dataset
    diabetes_dataset.shape

    (768, 9)
```

Figure 1: Parameters used for diabetics detection

```
# getting the statistical measures of the data
diabetes_dataset.describe()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

```
[8] diabetes_dataset['Outcome'].value_counts()

    0    500
    1    268
    Name: Outcome, dtype: int64
```

Figure 2: Analysis of data

GIET UNIVERSITY JOURNAL, 2024, VOL.3, NO.3, PP. -112-120

Manuscript received 28th April, 2024; revised 3rd May, 2024; Accepted 6th May, 2024

0-->Non-diabetic 1-->Diabetic

```
[ ] diabetes_dataset.groupby('Outcome').mean()
```

Outcome	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	3.298000	109.980000	68.184000	19.664000	68.792000	30.304200	0.429734	31.190000
1	4.865672	141.257463	70.824627	22.164179	100.335821	35.142537	0.550500	37.067164

```
[ ] # separating the data and labels
    X = diabetes_dataset.drop(columns = 'Outcome', axis=1)
    Y = diabetes_dataset['Outcome']
```

```
[ ] print(X)
```

```
     Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI
0              6      148             72             35        0  33.6
1              1       85             66             29        0  26.6
2              8      183             64              0        0  23.3
3              1       89             66             23       94  28.1
4              0      137             40             35      168  43.1
..           ...      ...            ...            ...      ...   ...
763           10      101             76             48      180  32.9
764            2      122             70             27        0  36.8
765            5      121             72             23      112  26.2
766            1      126             60              0        0  30.1
767            1       93             70             31        0  30.4

     DiabetesPedigreeFunction  Age
0                       0.627   50
1                       0.351   31
2                       0.672   32
3                       0.167   21
4                       2.288   33
..                        ...  ...
763                     0.171   63
764                     0.340   27
765                     0.245   30
766                     0.349   47
767                     0.315   23

[768 rows x 8 columns]
```

Figure 3: Output after Regression

GIET UNIVERSITY JOURNAL, 2024, VOL.3, NO.3, PP. -112-120

Manuscript received 28th April, 2024; revised 3rd May, 2024; Accepted 6th May, 2024

```
[ ] print(Y)

    0      1
    1      0
    2      1
    3      0
    4      1
          ..
    763    0
    764    0
    765    0
    766    1
    767    0
    Name: Outcome, Length: 768, dtype: int64
```

Accuracy

```
[ ] # accuracy score on the training data
    X_train_prediction = classifier.predict(X_train)
    training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

[ ] print('Accuracy score of the training data :', training_data_accuracy)

    Accuracy score of the training data : 0.7866449511400652

[ ] # accuracy score on the test data
    X_test_prediction = classifier.predict(X_test)
    test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

[ ] print('Accuracy score of the test data :', test_data_accuracy)

    Accuracy score of the test data : 0.7727272727272727
```

Figure 4: Accuracy after the logistic regression

Manuscript received 28th April, 2024; revised 3rd May, 2024; Accepted 6th May, 2024

```
input_data = (10,115,0,0,0,35.3,0.134,29)

#changing the input data to numpy array
input_data_as_numpy_array = np.asarray(input_data)

# reshape the array as we are predicting for one instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

# standardize the input data
std_data = scaler.transform(input_data_reshaped)
print(std_data)

prediction = classifier.predict(std_data)
print(prediction)

if (prediction[0] == 0):
  print('The person is not diabetic')
else:
  print('The person is diabetic')
```

```
[[ 1.82781311 -0.184482   -3.57259724 -1.28821221 -0.69289057  0.41977549
  -1.02042653 -0.36084741]]
[1]
The person is diabetic
/usr/local/lib/python3.7/dist-packages/sklearn/base.py:451: UserWarning: X does not have valid feature names, but StandardScaler was fitted with feature names
  "X does not have valid feature names, but"
```

Figure 5: Accuracy comparisons

4. Conclusion

After using all these patient records of female, we are able to build a machine learning model to accurately predict whether or not the patients in the dataset have diabetes or not along with that we were able to draw some insights from the data via data analysis and visualization. By early prediction, we can:

- Protect the patients from different diseases.
- Protect from eye blindness.
- Prevent from kidney and other body organs affects.
- Manage the levels of diabetes

5. References

1. K.A. Manjula, P.Karthikeyan, "Heart disease Prediction using Ensemble based Machine Learning Techniques", International Conference on Trends in Electronics and Informatics 2019.
2. R. Hafezi, A. N. Akhavan, "Forecasting Heart disease Changes: AUT Journal of Modeling and Simulation, 2018.
3. A. K. Agarwal, Swati Kumari, " Heart disease Prediction using Machine Learning", International Journal of Trend in Scientific Research and Development,2020.
4. X. Yang, "The Prediction of Heart disease Using ARIMA Model", 2nd International Conference on Social Science,2019.
5. R. Ghule, Abhijeet Gadhave, " Heart disease Prediction using Ensemble based Machine Learning Techniques",2022.
6. A.C. Shravani, Divya, "Predicting Future Heart disease using Machine Learning Approach", International Journal of Advanced Computer Science and Applications, 2017.