Table of Contents

1.Introduction

Diabetes is becoming a global metabolic disorder and the number of cases soaring dynamically. When the cure is impossible then prevention is the best measure that we can take. But to prevent, first we need to identify the factor and disease properly. Health care sectors have large data bases with potential information about the disease and patient health history. Combining all the available information and data analysis may help to recognize the hidden pattern or early prediction of the disease.

Classification is a supervised method to build a model which predict the correct label for input data. Using machine learning algorithm, a predictive model is fully trained by training data and then it is evaluated based on test data. A dataset that has been originated from "National Institute of Diabetes and Digestive and Kidney Diseases" which in total has 768 instances with 9 attributes. Among which there are eight independent features (Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age) and one dependent feature (outcome). The dependent feature has been used to classify the data set or giving labels. Here **0** denote the person does not have diabetes and **1** denote the person have diabetes. Our goal is to use machine learning technique to build a proper model which can be used to predict whether a patient is diabetic or not based on different factors.

2. State of art (Classifiers)

A diabetes data set has been used to build a predictive model by applying classification algorithm and then the classifier´s performances has been analysed. We have compared the performance of two classifier K-Nearest Neighbors Algorithm (KNN) and Support vector machine (SVM) algorithm based on accuracy and confusion matrix.

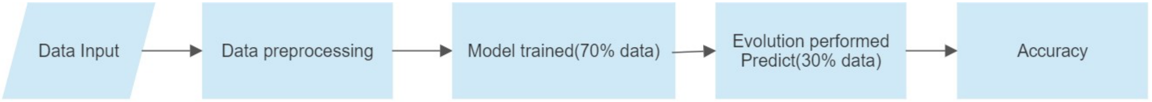
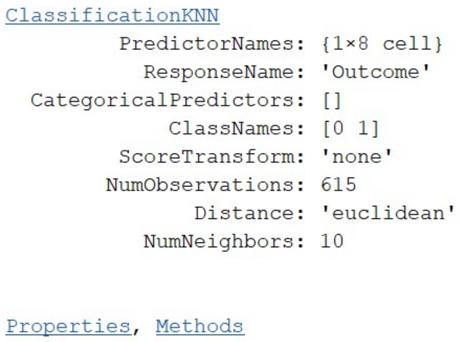


Fig- Classification of diabetes dataset

2.1. kNN algorithm- K-nearest neighbors algorithm is known as an effective lazy classifier as it does not use a learning model to predict data rather it calculates the similarity between incoming data and available data. It classifies any data by searching for the nearest neighbor and combine them together to form one cluster. The parameter k decides how many neighbors should be considered for observation. It predicts the class of new data based on similarity and calculating the distance between closest data. The algorithm is highly sensitive to noisy data and missing data. For that first the unnecessary information or value has been removed. Here KNN algorithm has been applied in the original data set for classification of classes (0,1). First the data has been normalization and the data has been splitted into 80 percent training set and 20 percent test set using hold-out validation. We have calculated the Euclidean distance between the data and K=10 has been taken to build the model and the accuracy rate was 69% for original features.



```
ClassificationKNN
           PredictorNames: {1×8 cell}
             ResponseName: 'Outcome'
    CategoricalPredictors: []
               ClassNames: [0 1]
           ScoreTransform: 'none'
          NumObservations: 615
                 Distance: 'euclidean'
             NumNeighbors: 10


  Properties, Methods
```

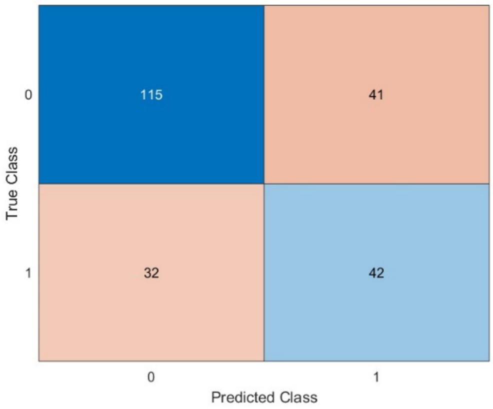
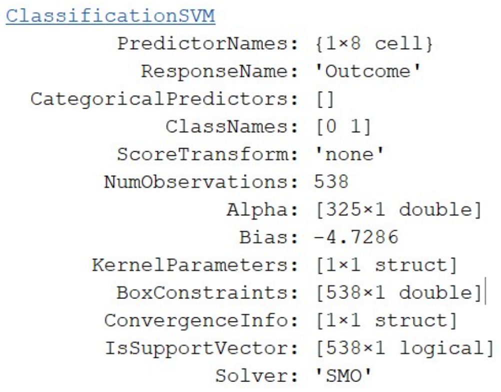


Fig- Basic kNN classifier

Fig- Confusion Matrix

2.2. SVM algorithm- Support vector machine (SVM) algorithm demonstrates very high performance for binary classification by generating hyperplane that separate classes after the transformation of input data into high dimensional space. One special property of SVM is it significantly reduce the empirical classification error by maximizing the geometric margin. For applying SVM algorithm to our predictive model, the data has been normalized first and then the data has been splitted into 70 percent train set and 30 percent test set using hold-out validation. Linear hyper-plane has been generated using SVM algorithm. Usually, the dimension of hyperplane depends on the number of features. After training the prediction model, evaluation has been performed based on the prediction of test data. Accuracy rate was 77.8261 in original data set.



```
ClassificationSVM
         PredictorNames: {1×8 cell}
           ResponseName: 'Outcome'
  CategoricalPredictors: []
             ClassNames: [0 1]
         ScoreTransform: 'none'
        NumObservations: 538
                  Alpha: [325×1 double]
                   Bias: -4.7286
       KernelParameters: [1×1 struct]
         BoxConstraints: [538×1 double]
        ConvergenceInfo: [1×1 struct]
        IsSupportVector: [538×1 logical]
                 Solver: 'SMO'
```

Fig -Basic classification model (SVM)

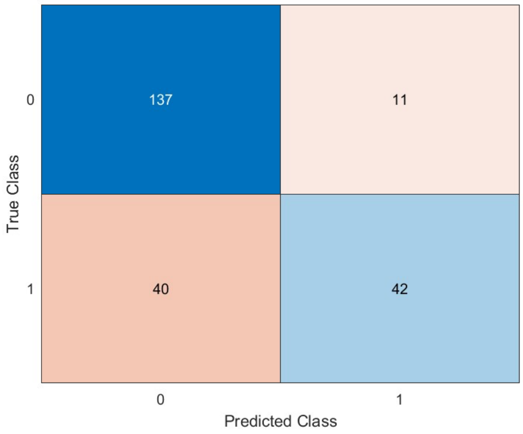


Fig- Confusion Matrix

3. Feature selection (Data Preprocessing)

Feature selection is a technique of reducing input variable to improve the performance of predictive model. It helps to visualize the classification model in two dimension and reduce the time and space complexity. Data visualization is the first step for effective feature selection. Here we have used filter method for preprocessing the data. The correlation of variables with the outcome helps us to choose the best features which manipulates the data. Here it is observed that Glucose level have high correlation with outcome and BMI also have better correlation compared to other features. Also, BMI have certain influence on SkinThickness which is highly correlated with insulin level. When both discriminants have same impact on output then we can filter out one of the feature.

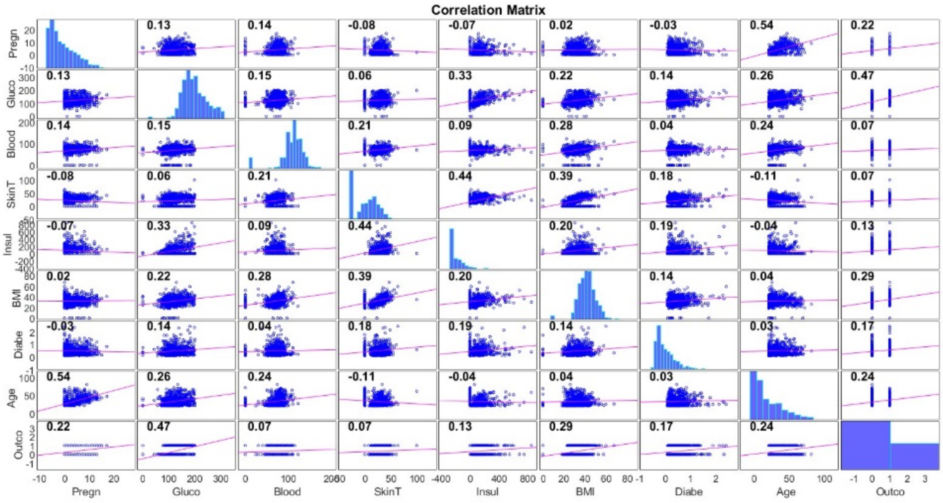


Fig-Correlation of features with outcome and other input features

3.1. KNN algorithm (Filtering)- k-Nearest Neighbors algorithm is a very simple and effective algorithm that predict model with less time but in high dimensional data it does not perform well. Also, it is costly to calculate the distance of each data on large dataset. After applying the feature selection method, the accuracy rate has significantly increased to 72.60%.

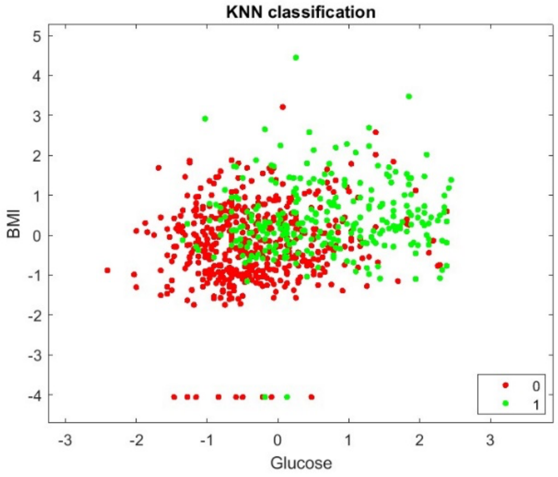


Fig- kNN visualization

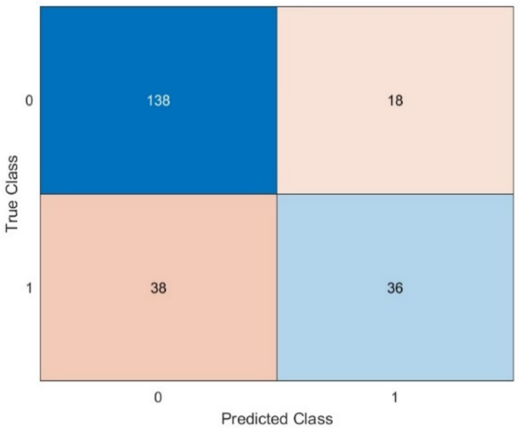
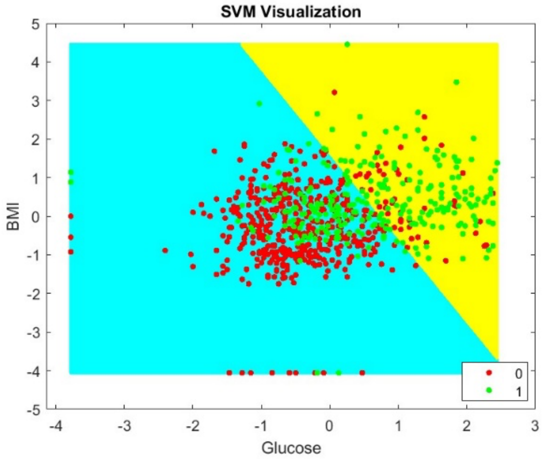


Fig- Confusion matrix

3.2. SVM algorithm (Filtering)- The main task of SVM algorithm is to find the optimal hyperplane the separates two classes. Though SVM consistently deliver higher performance in the field of supervised learning, but it is very sensitive to how the cost parameter and kernel parameter are set. At the same time, it is always difficult to deal with many features but sometimes more information trains a model properly and decrease the misclassification rate. Below the decision surface showing the linear hyperplane that has separated two classes. Applying feature selection, the accuracy rate of the model was 73%.



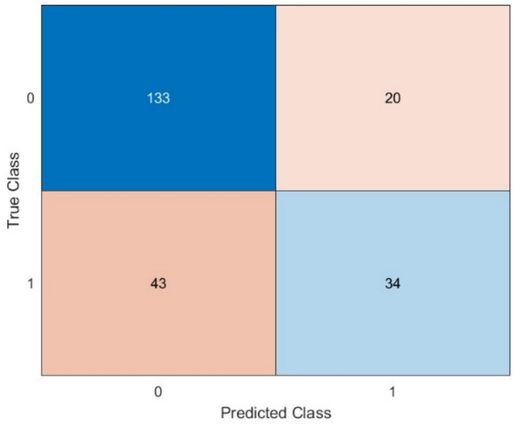
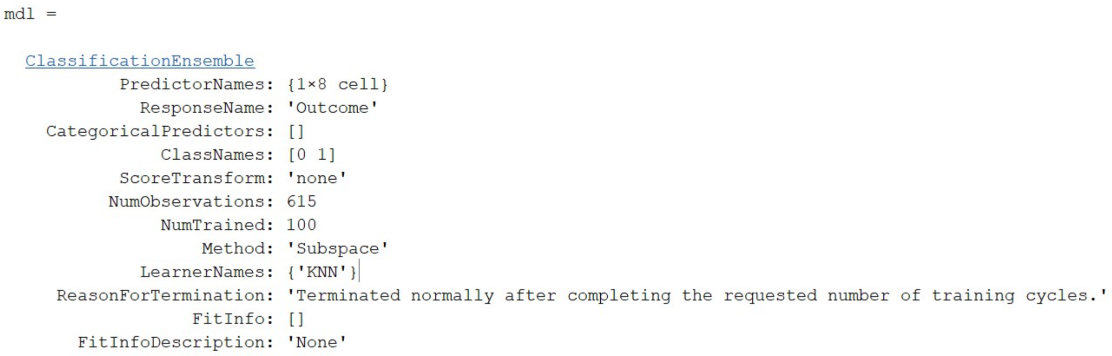


Fig- Decision surface of support vector machine

Fig- Confusion matrix

4. Advanced classification method

Ensemble method combines the prediction of several base estimator that is made of given learning algorithm to improve accuracy. For evaluation and increasing the performance of KNN we have used an ensemble of subset of kNN with subspace method. kNN has been used as a weak learner and different value of parameter k has been used to create weak classifier of different sample from 80% data set. Further the accuracy has been tested on remaining 20% data. The accuracy rate was 78%.



```
mdl =

  ClassificationEnsemble
           PredictorNames: {1×8 cell}
             ResponseName: 'Outcome'
    CategoricalPredictors: []
               ClassNames: [0 1]
           ScoreTransform: 'none'
          NumObservations: 615
               NumTrained: 100
                   Method: 'Subspace'
             LearnerNames: {'KNN'}
     ReasonForTermination: 'Terminated normally after completing the requested number of training cycles.'
                  FitInfo: []
       FitInfoDescription: 'None'
```

Fig-Basic ensemble classifier model of kNN

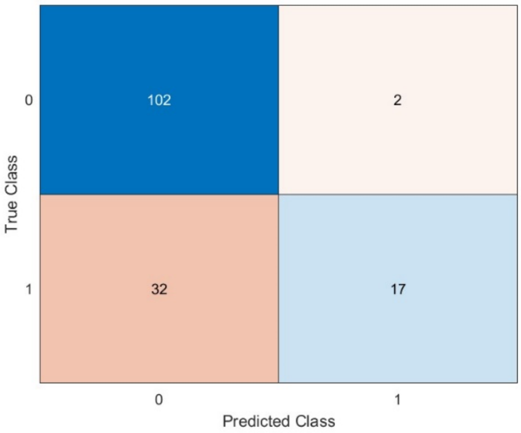


Fig- Confusion Matrix

5. Reduction complexity and execution time.

The time complexity of KNN algorithm is O(nd) and time complexity of SVM is O(n3).
For the diabetes data set, KNN took 4.72sec whereas SVM took 3.76 sec for training purpose. It is observed that SVM provides better speed compared to kNN in term classification. The performance of KNN depends on one hyperparameter whereas the performance of SVM depends on two hyperparameter.

6. Evaluation

Performance Classifier Model	Precision	Recall	Accuracy	F1 score
kNN	0.74	0.78	0.69	0.76
SVM	0.93	0.77	0.778	0.84
kNN (filtering)	0.88	0.78	0.75	0.83
SVM (filtering)	0.87	0.75	0.73	0.80
Ensemble	0.98	0.76	0.78	0.85

Fig- Performance evaluation of different classifier

In the above table, we have compared the performance of five classifier using three algorithms. For evaluation purpose we have considered 4 metrics. Precision identifies the rate of correctly positive predictions. On the other hand, accuracy rate gives us the idea about how many times a model has identifies the correct value with respect to total value. Recall reflects how many positive cases the classifier correctly predicted. F1 score combines both precision and recall.

It is observed KNN algorithm works better with less features. Its accuracy and precision both have increased with less features. With feature selection the performance of kNN have increased by 6%. On the other hand, SVM showed higher performance without feature selection. Its prediction rate has decreased significantly with less features. It can be said that KNN algorithm work better with large data and less features, but SVM need more information to predict more accurately.

After comparing kNN algorithm with SVM it is found that for the specific data set(diabetes), in general SVM has performed better. kNN algorithm performed average with outliers but SVM can handle outliers very well. So, to increase the performance of kNN algorithm we have tried the multiple learner's strategy to combine the result of different kNN classifier. It is observed that applying ensemble learning algorithm we got the highest precision and accuracy rate.

Reference

1. https://www.kaggle.com/datasets/mathchi/diabetes-data-set

2.https://www.researchgate.net/publication/285663733_Data_classification_using_support_vector_machine

3.https://www.researchgate.net/publication/291556969_Ensemble_of_a_subset_of_kNN_classifiers

4. https://www.geeksforgeeks.org/support-vector-machine-algorithm/