**PROJECT SUMMARY**

**DATA SET DESCRIPTION:**

* **Abstract**: The data set entitled “Diabetic Retinopathy Debrecen Data Set” comprises of features that are pulled out from a Messiodor image set to prognosticate whether there are traces of diabetic retinopathy in the image or not.
* **Source**:

1. Dr. Balint Antal, Department of Computer Graphics and Image Processing  
   Faculty of Informatics, University of Debrecen, 4010, Debrecen, POB 12, Hungary  
   antal.balint '@' inf.unideb.hu
2. Dr. Andras Hajdu, Department of Computer Graphics and Image Processing  
   Faculty of Informatics, University of Debrecen, 4010, Debrecen, POB 12, Hungary  
   hajdu.andras '@' inf.unideb.hu

* **Data Set Information:** The data set comprises of features that are pulled out from a Messiodor image set to prognosticate whether there are traces of diabetic retinopathy in the image or not. The dimensions or features of data corresponds to either a discovered lesion, a descriptive feature of an anatomical part or an image-level descriptor.
* **Feature Information:**

|  |  |
| --- | --- |
| Feature No. | Information given by feature |
| 0 | The binary result of quality assessment. 0 = bad quality 1 = sufficient quality |
| 1 | The binary result of pre-screening, where 1 indicates severe retinal abnormality and 0 its lack. |
| 2-7 | The results of MA detection. Each feature value stand for the number of MAs found at the confidence levels alpha = 0.5, . . ., 1, respectively |
| 8-15 | contain the same information as 2-7 for exudates. However, as exudates are represented by a set of points rather than the number of pixels constructing the lesions, these features are normalized by dividing the number of lesions with the diameter of the ROI to compensate different image sizes. |
| 16 | The Euclidean distance of the centre of the macula and the centre of the optic disc to provide important information regarding the patient’s condition. This feature is also normalized with the diameter of the ROI. |
| 17 | The diameter of the optic disc. |
| 18 | The binary result of the AM/FM-based classification |
| 19 | Class label. 1 = contains signs of DR  (Accumulative label for the Messidor classes 1, 2, 3), 0 = no signs of DR. |

**Abbreviations used**: MA: Microaneurysm, ROI: Region of Interest, DR: Diabetic Retinopathy, AM: Amplitude Modulation, FM: Frequency Modulation

**CITATION:** Balint Antal, Andras Hajdu: An ensemble-based system for automatic screening of diabetic retinopathy, Knowledge-Based Systems 60 (April 2014), 20-27.  
The dataset is based on features extracted from the Messidor image dataset.

**EXPERIMENTS CONDUCTED:**

Here are conducted 3 experiments in total for this project:

* We experimented with K-Nearest Neighbour as a classifier.
* We experimented with Naïve Bayes as a classifier.
* We experimented with Support Vector Machine as a classifier.

**RESULT DISCUSSION**

**SVM (SUPPORT VECTOR MACHINES):**

* The Data Set was first Divided into Train and Test set with a 75:25 split.
* Linear and RBF Kernel both have been compared along with their cross validated models with hyper parameter tuning
* SVM implementation in sklearn for linear kernel without cross validation gives a model that yields an error percentage of 29.5138888889%
* SVM implementation in sklearn for rbf kernel without cross validation gives a model that yields an error percentage of 30.90277777777778%
* 5-fold cross validation yields lower error for linear kernel, but 10-fold yields even lower error for rbf kernel (can be observed by changing the values for cv in both instances to 5 and 10)
* C interval for linear kernel, C and gamma for RBF kernel have deliberately been restricted to a few values in the gridsearchcv since it is a brute force method and lower values reduce computational times and are convenient in observing the results quickly
* SVM implementation in sklearn for Linear kernel with cross validation and C value 1000 gives a model that yields an accuracy percentage of: 72.5694445
* SVM implementation in sklearn for linear kernel with cross validation and C value 10, gamma/spread value of, 0.1 gives a model that yields an error percentage of: 74.3055556
* ROC curve implementation is provided for best of these two kernels and not both

**KNN (K-Nearest Neighbours)**:

* The Data Set was first Divided into Train and Test set with a 75:25 split.
* Before estimating optimal parameters, the accuracy of KNN on Test Set is 64.23611111%
* The Training Data then went under 5-fold Cross Validation while setting the parameter to 20.
* The accuracy rates for the different folds were: [63.005, 58.959, 59.537, 65.697, 68.023]
* While the Average accuracy was: 63.04%
  + On comparison of the confusion matrix before and after setting optimal parameters it was noticed that the number of True Negatives increased by 20 and number of True Positives decreased by 8. But the over accuracy was increased.
  + We then performed K-Fold Cross Validation with Parameter Selection using Grid Search for the odd values of the parameter from [1-99].
  + The most optimal Parameter estimated through Grid-Search was 33 Nearest Neighbours, which got us the accuracy of 66.39% on the training set
* Using the most optimal parameters on the test set, we got an accuracy of 68.40%

**Naive Bayes:**

* + The Data Set was first Divided into train and Test set with 75:25 split
* Naïve Bayes implementation in sklearn without cross validation gives a model that yields an error percentage of 36.45833333333333
  + Naive Bayes does not really have any hyperparameters to tune, also that is the reason it can perform decently across various classification problems.
  + Cross Validation on the results brings about only a fraction of improvement in the performance of the classifier model, this could be since we consider each feature/dimension is independent of others (which might lead to loss of important information anyways)
  + GaussianNB is used because the data is continuous - numerical, also we assume it to be normally distributed.
  + Final error that the cross validated model of Naive Bayes classifier yields is : 36.11111111111111
  + The only tuning we can probably do is to perform feature selection PCA to better the performance of the classifier.

**Conclusive Comments:**

After conducting all the 3 Experiments on the data set used from UCI repository, these observations were collected by analysing the results:

* Training times for Naïve Bayes and KNN are much shorter than that of SVM and quadratic optimization problem plays a big part in that
* Accuracy for Non-Linear kernel in SVM is significantly higher compared to the other two for this problem
* Training times for KNN would usually be very low since the distances for the test data are calculated only during the testing time which forms the core of classification process
* Training time for linear Kernel is significantly higher than RBF kernel after tuning, probably due to the data set and might be because the underlying quadratic optimization problem for linear kernel was much more difficult to compute
* AUC score for SVM is also higher than the other two classifiers signifying that it generally performs better across various decision boundaries and not just the best one for which the confusion matrix has been calculated.

**Project Members**

|  |  |  |
| --- | --- | --- |
| Akashdeep Singh | singa190@mcmaster.ca | 400353314 |
| Sulaiman Ghouse Mohiuddin | mohius3@mcmaster.ca | 400418055 |
| Vikram Chawla | chawlv3@mcmaster.ca | 400354764 |