**MACHINE LEARNING PROJECT**

**Milestone #1**

**A. Dataset Building and Normalization**

1. **Source of Dataset**

UCI Machine Learning Repository – Center for Machine Learning and Intelligent Systems

http://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set

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

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

1. **Description of Dataset**

Number of original features: 20

Sample size: 1151

This dataset contains features extracted from the Messidor image set to predict whether an image contains signs of diabetic retinopathy or not. All features represent either a detected lesion, a descriptive feature of a anatomical part or an image-level descriptor. The underlying method image analysis and feature extraction as well as our classification technique is described in Balint Antal, Andras Hajdu: An ensemble-based system for automatic screening of diabetic retinopathy, Knowledge-Based Systems 60 (April 2014), 20-27.

1. **Name, Description and type of each feature**
2. Result of Quality Assessment--------------------------------------------------------------------------------------------Binary
3. Result of Pre-screening---------------------------------------------------------------------------------------------------Binary

3-8. Results of MA detection-------------------------------------------------------------------------------------------------Integer

9-16. Results of MA detection for exudates----------------------------------------------------------------Continuous [0 to 1]

17. Euclidean distance of the center of the macula and the center of the optic disc---------Continuous [0 to 1]

18. The diameter of the optic disc-------------------------------------------------------------------------Continuous [0 to 1]

19. The binary result of the AM/FM-based classification-------------------------------------------------------------Binary

20. Class label--------------------------------------------------------------------------------------------------------------------Binary

The Continuous attributes were normalized to the range [0 to 1] to attain a linear and more robust relationship.

1. **Processed/Normalized dataset**

asdsada

1. **Number of features that were removed and the reason for their removal**

None

**B. Performance of kNN, Decision Trees, Bayesian Network on Different Feature Sets**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Complete Feature Set** | **Feature Selection using Forward Search** | **Feature Selection using Backward Search** | **Top 9 features (with higher coefficients) based on 3 Components of Principal Components Analysis (PCA)** |
| **kNN (k=1)** | **61.25 %** |  |  |  |
| **kNN (k=3)** | **62.29 %** |  |  |  |
| **kNN (k=5)** | **63.16 %** |  |  |  |
| **kNN (k=7)** | **64.73 %** |  |  |  |
| **kNN (k=9)** | **63.95 %** |  |  |  |
| **Decision Trees** | **53.69 %** |  |  |  |
| **Bayesian Network (e.g. Naïve Bayes)** | **60.30 %** |  |  |  |

**Table 1 – Accuracy**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Complete Feature Set** | **Feature Selection using Forward Search** | **Feature Selection using Backward Search** | **Top 9 features (with higher coefficients) based on 3 Components of Principal Components Analysis (PCA)** |
| **kNN (k=1)** | **38.75 %** |  |  |  |
| **kNN (k=3)** | **37.71 %** |  |  |  |
| **kNN (k=5)** | **36.84 %** |  |  |  |
| **kNN (k=7)** | **35.27 %** |  |  |  |
| **kNN (k=9)** | **36.05 %** |  |  |  |
| **Decision Trees** | **46.31 %** |  |  |  |
| **Bayesian Network (e.g. Naïve Bayes)** | **39.70 %** |  |  |  |

**Table 2 – Classification**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Complete Feature Set** | **Feature Selection using Forward Search** | **Feature Selection using Backward Search** | **Top 9 features (with higher coefficients) based on 3 Components of Principal Components Analysis (PCA)** |
| **kNN (k=1)** | **0.225** |  |  |  |
| **kNN (k=3)** | **0.247** |  |  |  |
| **kNN (k=5)** | **0.265** |  |  |  |
| **kNN (k=7)** | **0.296** |  |  |  |
| **kNN (k=9)** | **0.281** |  |  |  |
| **Decision Trees** | **0.069** |  |  |  |
| **Bayesian Network (e.g. Naïve Bayes)** | **0.233** |  |  |  |

**Table 3 – Kappa**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Complete Feature Set** | **Feature Selection using Forward Search** | **Feature Selection using Backward Search** | **Top 9 features (with higher coefficients) based on 3 Components of Principal Components Analysis (PCA)** |
| **kNN (k=1)** | **61.31 %** |  |  |  |
| **kNN (k=3)** | **62.43 %** |  |  |  |
| **kNN (k=5)** | **63.33 %** |  |  |  |
| **kNN (k=7)** | **64.91 %** |  |  |  |
| **kNN (k=9)** | **64.15 %** |  |  |  |
| **Decision Trees** | **53.46 %** |  |  |  |
| **Bayesian Network (e.g. Naïve Bayes)** | **62.12 %** |  |  |  |

**Table 4 – Weighted Mean Recall**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Complete Feature Set** | **Feature Selection using Forward Search** | **Feature Selection using Backward Search** | **Top 9 features (with higher coefficients) based on 3 Components of Principal Components Analysis (PCA)** |
| **kNN (k=1)** | **61.26 %** |  |  |  |
| **kNN (k=3)** | **62.39 %** |  |  |  |
| **kNN (k=5)** | **63.30 %** |  |  |  |
| **kNN (k=7)** | **64.88 %** |  |  |  |
| **kNN (k=9)** | **64.12 %** |  |  |  |
| **Decision Trees** | **53.47 %** |  |  |  |
| **Bayesian Network (e.g. Naïve Bayes)** | **68.07 %** |  |  |  |

**Table 5 – Weighted Mean Precision**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Complete Feature Set** | **Feature Selection using Forward Search** | **Feature Selection using Backward Search** | **Top 9 features (with higher coefficients) based on 3 Components of Principal Components Analysis (PCA)** |
| **kNN (k=1)** |  |  |  |  |
| **kNN (k=3)** |  |  |  |  |
| **kNN (k=5)** |  |  |  |  |
| **kNN (k=7)** |  |  |  |  |
| **kNN (k=9)** |  |  |  |  |
| **Decision Trees** |  |  |  |  |
| **Bayesian Network (e.g. Naïve Bayes)** |  |  |  |  |

**Table 6 – AUC**

**C. Discussion of Results**

Discuss the results, determine which classifier is better and explain the reasons for choosing such classifier.

**D. Problems Encountered**

Discuss the problems that you had encountered starting from Dataset Building until Model Building and how did you address such problems.

***Note: name your doc and excel file with the surname of group members. Ex. Lim\_Cruz.docx***