**Abdul Malek bin Abdul Rahman - Manb 171144**

**MANB 2153 - Assignment 2 Part 1**

**Part 1 - Classification Using Naive Bayes, Decision Tree & KNN on Iris Dataset**

import pandas as pd

import numpy as np

from sklearn.model\_selection import train\_test\_split

df = pd.read\_csv('iris.csv',delimiter = ',')

df.head(4)

iris\_input = df.iloc[:,0:3]

iris\_target = df.iloc[:,3]

iris\_YY = []

for cat in iris\_target:

if cat == 'Iris-setosa':

cls = 1

elif cat == 'Iris-versicolor':

cls = 2

else:

cls = 3

iris\_YY.append(cls)

iris\_X = np.asarray(iris\_input)

iris\_y = np.asarray(iris\_YY)

np.random.seed(0)

indices = np.random.permutation(len(iris\_X))

iris\_X\_train = iris\_X[indices[:-10]]

iris\_y\_train = iris\_y[indices[:-10]]

iris\_X\_test = iris\_X[indices[-10:]]

iris\_y\_test = iris\_y[indices[-10:]]

**KNN Classifier**

from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier()

knn.fit(iris\_X\_train, iris\_y\_train)

KNeighborsClassifier(algorithm='auto', leaf\_size=30, metric='minkowski',

metric\_params=None, n\_jobs=1, n\_neighbors=5, p=2,

weights='uniform')

predict\_result\_iris = knn.predict(iris\_X\_test)

iris\_y\_test

print('The accuracy of the Knn classifier on training data is: {:.2f}'.format(knn.score(iris\_X\_train, iris\_y\_train)))

print('The accuracy of the Knn classifier on test data is: {:.2f}'.format(knn.score(iris\_X\_test, iris\_y\_test)))

**Naive Bayes Classifier**

from sklearn.naive\_bayes import GaussianNB

iris\_X\_train, iris\_X\_test, iris\_y\_train, iris\_y\_test = train\_test\_split(iris\_X, iris\_y)

NB = GaussianNB()

NB.fit(iris\_X\_train, iris\_y\_train)

print("The accuracy of the Naive Bayes classifier on training data is:", NB.score(iris\_X\_train, iris\_y\_train) , sep = "")

print("The accuracy of the Naive Bayes classifier on testing data is: ", NB.score(iris\_X\_test, iris\_y\_test) , sep = "")

**Decision Tree Classifier**

from sklearn import tree

decision\_tree = tree.DecisionTreeClassifier(criterion='gini')

decision\_tree.fit(iris\_X\_train, iris\_y\_train)

print('The accuracy of the Decision Tree classifier on training data is: {:.2f}'.format(decision\_tree.score(iris\_X\_train, iris\_y\_train)))

print('The accuracy of the Decision Tree classifier on test data is: {:.2f}'.format(decision\_tree.score(iris\_X\_test, iris\_y\_test)))

**Part 2 - 10 Journal Reviews on Classification Application - KNN, Naive Bayes and Decision Tree**

**Journal 1 - Enhancing WO3 gas sensor selectivity using a set of pollutant detection classifiers**

In this study, the researcher is using a novel hybrid approach with the main objective of solving the problem of cross-selectivity of gases in electronic nose (E-Nose) using the combination of classifiers for support vector machine (SVM) and K-Nearest Neighbors (KNN) methods. The fusion approach of using two classifiers is taken in order to improve the accuracy of the classification method conducted for gas discrimination. An array of three gas metal oxide sensors was applied to identify the presence of Oxone, ethanol and acetone. PCN, KNN and SVM are trained as the pattern recognition method to classify the data cluster related to different gases. It is found from the study that the system using combination of machine learning technique is more effective in identifying the pollutant gases compared to individual classification model. A combination of supervised and unsupervised approaches is used in relative to other standard approaches.

The method used in this study started with three WO3 sensors of E-Nose system applied for data acquisition to identify the three pollutant gases. Next, two transient parameters, derivative and integral were extracted for each gas response. After that, the principal component analysis (PCA) was applied to extract the most relevant sensor data and dimensionality reduction. The new coordinates calculated by PCA were applied as inputs for classification by SVM method. Lastly, KNN Classification method is used to calculate only the support vector (SV)s.

It is proven that the fusion method used led to highest classification rate of 100 percent compared to accuracy achieved by the individual classifiers namely, KNN, SVM-Linear, SVM-RBF, SVM-Polynomial with ratings of 89, 75.2, 80 and 79.9 percent respectively. Data collected in this study was the measurement sensitivity of resistance layer of the sensors using a source-meter (Keithley). It allows choosing the number of sensors tested in the matrix and the measurement parameters. Meanwhile, the sensor responses are captured in real time.

**Journal 2 - Shallow learning model for diagnosing neuro muscular disorder from splicing variants**

Muscular dystrophies are a cluster of ailments that cause imperfections and progressive deterioration of skeletal muscles that help for locomotion. It comprises of 30 types of disorders that vary in age of onset severity, inheritance pattern and in the form of exaggerated. In this research, the cloned gene sequences are synthesized based on mutation position and its location on the chromosome by using the positional cloning approach.

The central theme in this study is to spot out discriminative descriptors from diseased gene sequences based on splicing variants and to furnish with effective machine learning solution for predicting the type of muscular dystrophy disease with the splicing mutations. Multi-class classification is worked out through data modelling of gene sequences. The synthetic mutational gene sequences are designed as the diseased gene sequences are not readily obtainable for this intricate diseased. An eminent muscular dystrophy disease prediction model is built using supervise learning techniques in scikit-learn environment. The data set is constructed with the extracted features as numpy array. The data are normalized by transforming the features values into the range between 1 and 0 aid in scaling the input attributes for a model. Naive Bayes, decision tree, K-nearest neighbour and SVM learned models are developed using phyton library framework in scikit-learn.

Data set for this research are collected from various type of genes associated with the five types of neuromuscular disorder studied. A review is made from 55 genes that are associated with five types of muscular dystrophy like DMD, BMD, EMD, LGMD and CMT. Several types of mutated sequences based on mutations like missense, non-sense, synonymous, insertion duplication, deletion mutations and splicing mutations are collected. For each category of muscular dystrophy disease, 120 synthetic mutated gene sequences are generated and a corpus consisting of 600 sequences for all five categories of muscular dystrophy is developed.

**Journal 3 - Support vector machine and K-nearest neighbour for unbalanced fault detection**

In manufacturing industry, efficient fault detection of the machinery or its parts will save lots of money and give competitive egde to the manufacturers especially those running continuous process. Maintenance costs contribute signficantly to the total operating cost of all manufacturing. Hence, it can make or break the corporation especially when there is no proper maintenance plan and actions taken. The purpose of this study is to design an appropriate approach for detecting unbalanced fault in rotating machines using KNN and SVM Classifiers.

The detection system used is a fault diagnostic approach based on signal processing, feature extraction and fault classification. Vibration signals were acquired from a designed experimental system with three conditions namely, no load, balanced load and unbalanced load. FFT techniques was applied to transform the vibration signals from time domain into frequency domain. A total of 29 feature parameters were extracted from FFT amplitude of the signals. KNN and SVM classifiers were used to identify and categorized these three conditions.The performance of two classifiers used were obtained under different values of their parameter. The result obtained exhibited that the proposed approached can be used effectively for detecting unbalanced condition inn rotating machine.

The data set consisting of 29 features and 1,044 samples for each condition, no load, balanced load and unbalanced load were divided into two subsets, training data set and testing data set. Frequency domain used in the study refers to analysis of the vibration signal as a function of frequency. Processed signals consisting of large set of data for each sample, henceforth some statistical functions should be used to reduce feature vector.

**Journal 4 - New colour SIFT descriptors for image classification with applications to biometrics**

Biometric application for image classification, indexing and retrieval will depend on the powerful information provided by colour features. The information retained is also significant in identification of object and natural scene categories as well as geographical features from images. The choice of a colour space is important for many computer vision algorithms. Different colours spaces display different colour properties. With the large variety of available colour spaces, the inevitable question that arises is how to select the colour space that produces the best result for a particular computer vision task. Colour features such as the colour histogram, colour texture and local invariant features provide varying degrees of success against image variations such as viewpoint and lighting changes, clutter and occlusions. The central theme of this study is a new oRGB-SIFT descriptor and its integration with other colour SIFT features to produce the novel colour SIFT Fusion (CSF) and the colour grey scale SIFT Fusion (CGSF) descriptors for image classification with special purpose for biometric application.

Classification process is implemented using a novel EFM-KNN classifier, which combines the Enhanced Fisher Model (EFM) and the K Nearest Neighbour (KNN) decision rule. The effectiveness of the proposed descriptors and classification method are evaluated using 20 image categories from two large scale, grand challenge data set, the Caltech 256 data base and the UPOL Iris data base. Caltech database comprised of 30,607 images divided into 256 categories and a clutter class. Each category contains at least 80 images, a maximum of 827 images while the mean number of images per category is 119. The first set of experiments assesses the overall classification performance of the eight descriptors on the Biometric Dataset with 20 categories. The second set of experiments evaluate the classification performance using the PCA and the EFM-KNN methods.

The fusion of both Colour SIFT descriptors (CSF) and Colour Greyscale SIFT descriptor (CGSF) show significant improvement in the classification performance, which indicates that various colour-SIFT descriptors and greyscale-SIFT descriptors are functioning separately for image classification process.

**Journal 5 - Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques**

Near Infrared (NIR) spectroscopy is commonly used in the industry like oil and gas as a non-destructive measurement technique for many chemical compounds, including products of petroleum refining and petrochemicals, food items, pharmaceuticals, combustion products and chemicals. NI has proved its efficiency for both applications, laboratory and industrial applications. One of the major applications of NI spectroscopy is for categorizing of petroleum products like gasoline and such since petroleum refining and petrochemicals consists of different type of hydrocarbons, whose components could be estimated and identified by NI method. The study conducted in this paper compared the abilities of nine different multivariate classification methods, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), regularized discriminant analysis (RDA), soft independent modelling of class analogy (SIMCA), partial least squares (PLS) classification, K-Nearest Neighbor (KNN), support vector machine (SVM), probabilistic neural networks (PNN) and multilayer perception (MLP) for gasoline classification. In all cases, NIR spectroscopy was found to be effective for gasoline classification purposes while KNN, SVM and PNN techniques for classification were found to be effective ones.

Three set of NIR spectra (450, 415 and 345) were used for classification of three different samples of gasoline into 3, 6 and 3 classes respectively according to their sources (refinery or process) and type. PNN method is the most effective ways of gasoline classification using NIR spectra analysis. Nevertheless, KNN and SVM classifiers have also shown adequate results while the former is much easier to understand, compute and program.

Data collected in this study are spectroscopy data in the form of measurement of the wavelength of each spectrum for three gasoline samples. The parameters used for experiment are spectral range, sample volume, optical path, spectral resolution, numbers of scan, cell material, time of one measurement and number of measurements per sample.

**Journal 6 - A personalized recommender system based on web usage mining and decision tree induction**

As online purchase is becoming a trend for many people, a personalize product recommendation is an enabling mechanism designed to overcome information overload. Collaborative filtering has been known to be one of the most successful recommendation methods, but its application in e-commerce has exposed well-known limitations such as sparsity and scalability, which would lead to poor recommendation. The objective of this study is to propose a personalized recommendation methodology by which one could experience an effective and quality recommendation when doing personal purchasing via internet market place.

The suggested methodology is based on a variety of data mining techniques such as web usage or application mining, decision tree induction, association rule mining and the product taxonomy. For the evaluation of the methodology, a recommender system using intelligent agent and data warehousing technologies were implemented. Customer preferences and the product association are automatically learned from clickstream or web usage data while in order to avoid the poor recommendations that will lead to disappointing outcome, customers who are likely to purchase the recommended products are selected using the decision tree induction.

The web usage data of particular customer as well as product purchase data and customer related data are the central components of the methodology used in the study. There has been several customer behaviour model for e-commerce, which have different analysis purpose. The overall process of web usage mining is generally divided into two main tasks, data preparation and pattern discovery. The data preparation tasks build a server session file where each session is a sequence of request of different types made by a single user for every visit made to the relevant site. Pattern discovery task involve the discovery of association rules, sequential pattern, usage cluster and page cluster.

**Journal 7 - An assessment of the effectiveness of decision tree methods for land cover classification**

Over the period of few years, the usage of Decision Tree (DT) to classify remotely sensed data has increased. It is computationally fast, make no statistical assumptions and can handle data that are represented over different measurement scales. In this study, separate test and training data sets from two different geographical areas and two different sensors, multispectral Landsat ETM+ and hyperspectral DAIS, are used to evaluate the performance of univariate and multivariate DTs for land cover classification. The level of classification accuracy achieved by the DT is compared to results from back-propagating Artificial Neural Network (ANN) and the maximum like hood (ML) classifiers. The outcome of this study showing that the performance of univariate DT is acceptably good in comparison with that of other classifiers, except with high dimensional data. Classification accuracy increases linearly with training dataset size to a limit of 300 pixels per class in this case. While boosting produces an increase in classification accuracy of between 3% and 6%, the use of attribute selection method does not appear to be justified in terms of accuracy increases. It is also found from the study that neither the univariate DT nor the multivariate DT performed as well as the ANN and ML classifiers with high dimensional data.

The maximum like hood, multi-layer back propagation neural network and decision tree procedures are used in this research. Unlike conventional statistical and neural or connectionist classifiers, which use all available features simultaneously and make a single membership decision for each pixel, the DT uses a multi stage or sequential approach to the problem of label assignment. The labelling process is considered to be a chain of simple decisions based on the results of sequential tests rather than a single complex decision. Sets of decision sequences form the branches of the DT, with tests being applied at the nodes.

Two contrasting data sets are being used in this study. The first is a medium-resolution (Landsat ETM+) image of part of Eastern England near the town of Littleport. Official field data printouts, which record the crop or crop grown in each field for the year 2000 were collected from farmers and their representatives’ agencies and other parts of the area were surveyed on the ground to assembled the ground reference data. Random sampling methods being used to collect separate training and test data sets in both study areas using ground reference data generated from field observations and in the case of the Littleport study area, from official farm records.

**Journal 8 - Spam filtering with Naive Bayes - Which Naive Bayes?**

Naive Bayes is recognised as a popular choice in commercial and open source anti-spam email filters. The purpose of this study is to evaluate five different versions of Naive Bayes and compared their classifying output on six new, non-encoded datasets that contained ham messages of particular Enron users and fresh spam messages. The experimental procedure emulates incremental training of personalized spam filters and plotting of ROC curves that allow comparison made on different versions of Naive Bayes classifier over the entire trade-off between true positives and true negative. The result of this study shown that the two non-common Naive Bayes Classifier used in spam filtering literature, Flexible Bayes (FB) and multi-nominal Naive Bayes with Boolean attributes give credible performance from the experiments compared to the rests, while the latter is slightly better among the twos. The effectiveness was achieved over large attributes set or 3,000 attributes.

The six ham messages collection from Enron users were each paired with one of the three spam collected. These six datasets emulate different situations faced by real users. Users start with a small amount of training messages and retains the filter as new messages arrive. This incremental retraining and evaluation differ significantly from the cross-validation experiments that commonly used to measure the performance of learning algorithms and which have been adopted on many various spam filtering experiments.

Following the Enron investigation, personal files of approximately 150 Enron employees were made publicly available. These files include a large number of personal e-mail messages, which has been used to create e-mail classification benchmarks. The collection was further augmented with spam messages collected in 2005, leading to a benchmark with 43,000 ham and approximately 50,000 spam messages. In addition, spam messages obtained from four different sources, the SpamAssassin corpus, the Honeypot project, the spam collection of Bruce Guenter (BG) and spam collected by one of the authors, Georgios Paliouras.

**Journal 9 - Intrusion detection using Naive Bayes Classifier with feature reduction**

The growing numbers of people using computers and surfing the web all around the world in the networked machines has led to an increase in unauthorized activity, not only from external but also from internal attackers, including among others like disgruntled employees and people abusing their privileges for personal gain. The main purpose of this study is to identify important reduced input features in building intrusion detection system (IDS) that is computationally efficient and effective. In order to achieve this, the researcher investigate the performance of three standard feature selection methods using Correlation-based Feature Selection, Information Gain and Gain Ratio.

Feature Vitality Base Reduction Method was chosen as the procedure to identify important reduced input features. Meanwhile, Naive Bayes was selected as a classifier on reduced datasets for intrusion detection. Results obtained from the Empirical study concluded that selected reduced attributes give better performance to design IDS that is efficient and serve the purpose of network intrusion detection.

A dynamic model called "Intelligent Intrusion Detection System" was designed in the research using AI approach for intrusion detection. This include neural networks and fuzzy logic with networks profiling, that uses simple data mining techniques to process the network data. The system combines anomaly, misuse and host-based detection. Simple Fuzzy roles allow constructing if-then rules that reflect the common ways of describing security attacks.

The dataset used in this study is NSL-KDD labelled dataset. This dataset suggested to solve some inherent problems the earlier data set KDD99. NSL-KDD dataset contains one type of normal data and 22 different types of attacks which falls into one of four categories comprising of DoS, probe, R2L and U2R. Total of 62,986 records extracted out of 1,250,973 NSL-KDD datasets connections available for training and testing.

**Journal 10 - Naive Bayes Classifier for rapid assignment of rRNA sequences into new bacterial taxonomy**

The objective of this paper is to classify the bacterial 16S rRNA sequences into the new higher order taxonomy as proposed in Bergey's Taxonomic Outline of the Prokaryotes. With this taxonomical advance from domain to genus with confidence estimates for each assignment, the RDP Classifier or Naive Bayes Classifier is tested with a corpus of 23,095 rRNA sequences as assigned by the NCBI into their alternative higher order taxonomy. The RDP Classifier is identified to be suitable both for the analysis of single rRNA sequence and for the analysis of libraries of thousand of sequences. rRNA based analysis is accepted as a cornerstone method in microbial diversity exploration and bacterial identification.

The Ribosomal Database Project II (RDP) provides data, tools and services related to rRNA sequences to the research community. The Classifier being tested using exhaustive leave-one-out testing on the entire Bergey corpus. The corpus contained 453 genera represented by single sequence. In addition, they repeated the testing method on the small contiguous regions of 400, 200, 100 and 50 bases at random.

The data used were collection of 300,000 bacterial rRNA sequence over 4 years period. The list is still evolving as species are re-evaluated and discrepancies are resolved. As bacterial taxonomy continues to evolve, so does the Classifier's performance in tandem.

**References**

R.M Balabin, R.Z Safieza and E.I Lomakina (2010), *Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques*, Analytica Chimica Acta (2010), doi:10.1016/j.aca.2010.05.013

R Faleh, S Gomri, M Othman, K Aguir and A Kachouri (2018), *Enhancing WO3 gas sensor selectivity using a set of pollutant detection classifiers*, Sensor Review, Vol. 38 Issue: 1, pp. 65-73.

Y. H Cho, J. K Kim and S. H Kim (2002), *A personalized recommender system based on web usage mining and decision tree induction,* Expert System with Applications, 23 (2002), pp. 329-342.

S Kalimuthu and V Vijayakumar (2017), *Shallow learning model for diagnosing neuro muscular disorder from splicing variants,* World Journal of Engineering, Vol. 14, Issue: 4, pp. 329-336.

S Mukherjee and N Sharma (2012), *Intrusion Detection using Naïve Bayes Classifier with Feature Reduction,* Procedia Technology, 4 (2012), pp. 119-128.

V Metsis, I Androutsopoulos and G Paliouras (2006), *Spam Filtering with Naïve Bayes – Which Naïve Bayes?* CEAS 2006 – Third Conference on Email and Anti-Spam, July 27-28.

A Moosavian, H Aahmadi, B Sakhaei and R Labaffi (2014), *Support vector machine and K-nearest neighbour for unbalanced fault detection,* Journal of Quality in Maintenance Engineering, Vol. 20, Issue: 1, pp. 65-75.

M Pal and P. M Mather (2003), *An assessment of the effectiveness of decision three methods for land cover classification,* Remote Sensing of Environment, 86 (2003), pp. 554-565.

A Verma, C Liu and J Jia (2011), *New colour SIFT descriptors for image classification with applications to biometrics,* Int. Journal Biometrics, Vol. 3, No. 1, 2011.

Q Wang, G. M Garrity, J. M Tiedje and J. R Cole (2007), *Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy,* Applied and Environmental Microbiology, Aug. 2007, pp. 5261-5267.