The breast cancer report

Project members names:
1- سكشن ٣ علوم الحاسب محمد صلاح صدقة علي ابو النجا

Problem definition:
The analysis can be done with breast cancer dataset for the disease diagnosis,treatment prediction, and prognosis of diseases, the risk awareness prediction and survival of breast cancer. The advantage of using data mining in the disease diagnosis is that it can mine the large medical dataset within a short time. The prediction accuracy is high and reliable. The result will be based on the various attributes in multidimensional view. Neural networks is a promising approach in many classification systems and business forecasting. It is a bio inspiration approach, which works like a brain in dealing with information processing. This method works based on three stages that are1.architecture or model 2. Learning algorithm 3. Activation functions.

data description:

Attributes 2 through 10 have been used to represent instances.
Each instance has one of 2 possible classes: benign or malignant.

1. Wolberg,~W.~H., \& Mangasarian,~O.~L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In {\it Proceedings of the National Academy of Sciences}, {\it 87}, 9193--9196.

-- Size of data set: only 369 instances (at that point in time)
-- Collected classification results: 1 trial only
-- Two pairs of parallel hyperplanes were found to be consistent with 50% of the data
-- Accuracy on remaining 50% of dataset: 93.5%
-- Three pairs of parallel hyperplanes were found to be consistent with 67% of data
-- Accuracy on remaining 33% of dataset: 95.9%

2. Zhang,~J. (1992). Selecting typical instances in instance-based learning. In {\it Proceedings of the Ninth International Machine

Learning Conference} (pp. 470--479). Aberdeen, Scotland: Morgan Kaufmann.
-- Size of data set: only 369 instances (at that point in time)
-- Applied 4 instance-based learning algorithms
-- Collected classification results averaged over 10 trials
-- Best accuracy result:
-- 1-nearest neighbor: 93.7%
-- trained on 200 instances, tested on the other 169
-- Also of interest:
-- Using only typical instances: 92.2% (storing only 23.1 instances)
-- trained on 200 instances, tested on the other 169
4. Relevant Information:
Samples arrive periodically as Dr. Wolberg reports his clinical cases.
The database therefore reflects this chronological grouping of the data.
This grouping information appears immediately below, having been removed from the data itself:
Group 1: 367 instances (January 1989)
Group 2: 70 instances (October 1989)
Group 3: 31 instances (February 1990)
Group 4: 17 instances (April 1990)
Group 5: 48 instances (August 1990)
Group 6: 49 instances (Updated January 1991)
Group 7: 31 instances (June 1991)
Group 8: 86 instances (November 1991)

Total: 699 points (as of the donated datbase on 15 July 1992)
Note that the results summarized above in Past Usage refer to a dataset of size 369, while Group 1 has only 367 instances. This is because it originally contained 369 instances; 2 were removed. The following statements summarizes changes to the original Group 1's set of data:
Group 1 : 367 points: 200B 167M (January 1989)
Revised Jan 10, 1991: Replaced zero bare nuclei in 1080185 & 1187805
Revised Nov 22,1991: Removed 765878,4,5,9,7,10,10,10,3,8,1 no record
: Removed 484201,2,7,8,8,4,3,10,3,4,1 zero epithelial
: Changed 0 to 1 in field 6 of sample 1219406
: Changed 0 to 1 in field 8 of following sample:
: 1182404,2,3,1,1,1,2,0,1,1,1
5. Number of Instances: 699 (as of 15 July 1992)
6. Number of Attributes: 10 plus the class attribute
7. Attribute Information: (class attribute has been moved to last column)

Attribute Domain
-- --
1. Sample code number id number
2. Clump Thickness 1 - 10
3. Uniformity of Cell Size 1 - 10
4. Uniformity of Cell Shape 1 - 10
5. Marginal Adhesion 1 - 10
6. Single Epithelial Cell Size 1 - 10
7. Bare Nuclei 1 - 10
8. Bland Chromatin 1 - 10
9. Normal Nucleoli 1 - 10
10. Mitoses 1 - 10
11. Class: (2 for benign, 4 for malignant)
8. Missing attribute values: 16
There are 16 instances in Groups 1 to 6 that contain a single missing
(i.e., unavailable) attribute value, now denoted by "?".
9. Class distribution:
Benign: 458 (65.5%)
Malignant: 241 (34.5%)

Data preprocessing:

Data mining is the process of extracting patterns from data; these patterns may be discovered depending on the data mining tasks that are applied on the dataset. The two basic data mining tasks are: descriptive data mining tasks which help to understand the characteristic properties of dataset and predictive data mining tasks which are used to perform predictions based on available dataset. Predictive data mining is the chosen data mining task for this study. data mining applications can use different parameters
to examine data which includes; association (patterns that define the relationship between data), sequence/pattern analysis (patterns where one event leads to another), classification (identification of new patterns with predefined targets) and clustering (grouping of identical of smaller objects). The basic steps include:

- Problem definition is the definition of the goals and objectives and the identification of tools to be used to build the defined model.
- Data exploration is the recommendation for useful dataset if the existing dataset does not meet the required need for analysis.
- Data preparation is the process of cleaning and transforming data to remove missing and invalid data and validation of data for robust analysis.
- Modeling is based on the desired outcomes and data. This involves the use of data mining algorithms (for this study; naïve bayes, decision trees and multi-layer

perceptron) in meeting the necessary objectives-which for the purpose of this study is classification.

- Evaluation and deployment is the analysis and interpretation of the results of analysis to create recommendations for consideration.

Data mining algorithms used and why? Principal Component Analysis (PCA) is an unsupervised, non-parametric statistical technique primarily used for dimensionality reduction in machine learning. High dimensionality means that the dataset has a large number of features. ... PCA can also be used to filter noisy datasets, such as image compression . Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process. So to sum up, the idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.

STEP 1: STANDARDIZATION The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.

Once the standardization is done, all the variables will be transformed to the same scale. **STEP 2: COVARIANCE MATRIX COMPUTATION** The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to each other, The covariance matrix is a $p \times p$ symmetric matrix (where p is the number of dimensions) that has as entries the covariances associated with all possible pairs of the initial variables. For example, for a 3-dimensional data set with 3 variables x, y, and z, the covariance matrix is a 3×3 matrix of this from:

Covariance Matrix for 3-Dimensional Data Since the covariance of a variable with itself is its variance (Cov(a,a)=Var(a)), in the main diagonal (Top left to bottom right) we actually have the variances of each initial variable. And since the covariance is commutative (Cov(a,b)=Cov(b,a)), the entries of the covariance matrix are symmetric with respect to the main diagonal, which means that the upper and the lower triangular portions are equal.

STEP 3: COMPUTE THE EIGENVECTORS AND EIGENVALUES OF THE COVARIANCE MATRIX TO IDENTIFY THE PRINCIPAL COMPONENTS Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the ***principal components*** of the data. Before getting to the explanation of these concepts, let's

first understand what do we mean by principal components. Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. These combinations are done in such a way that the new variables (i.e., principal components) are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components. So, the idea is 10-dimensional data gives you 10 principal components, but PCA tries to put maximum possible information in the first component, then maximum remaining information in the second and so on, until having something like shown in the scree plot below.

The **main goal** of a **PCA** analysis is to identify patterns in data; **PCA aims** to detect the correlation between variables. If a strong correlation between variables exists, the attempt to reduce the dimensionality only makes sense. PCA's goal is to reduce the

curse of dimensionality. It will reduce the features in such a way that it retains most of the information of the features in its principal components. An important thing to note that these principal components are orthogonal to each other in the vector space. **visualization techniques for exploring data** While doing data analyses, two important observations were made. Firstly, the mean values of cell radius, perimeter, area, compactness, concavity and concave points can be used in the classification of cancer. Larger values of these parameters tend to show a correlation with malignant tumors. Secondly, the mean values of texture, smoothness, symmetry or factual dimension do not show a particular preference of one diagnosis over the other. In any of the histograms, there are no noticeable large outliers that explain further cleanup. one can easily infer that the class distribution that is 357 benign tumors and 212 malignant tumors. **Data Visualization**

Violin Plot The blue region on the left part of the vertical line indicates malignant tumor and the right part indicates a benign tumor. Let's interpret the plot as illustrated , in texture mean feature, the median of the Malignant and Benign looks like separated so it can be good for classification. However, in thefractal dimension mean feature, median of the Malignant and Benign does not look like separated so it does not give good information for classification. Let's interpret one more thing about the plot as shown in Fig. 8, variable of concavity worst and concave point worst looks like similar but how to decide whether they are correlated with each other or not. (Not always true but, basically if the features are correlated with either of it can bedropped). **Join Plot** From one can easily conclude that concavity worst and point worst predicted from violin plot turn out to be correlated. Hence, one feature is dropped instead of taking both the features. In order to compare two features deeper, let's use joint plot as shown in Fig. 9. Look

at this in joint plot below, it is really correlated. Pearson value is correlation value and 1 is the highest. Therefore, 0.86 looks enough to say that they are correlated.

Swarm Plot

Up to this point, some analyses and discoveries on the present data has been made already. a similar analysis to violin plot is carried illustrating the first 10 and last ten features to analyses the features with respect to data. In this plot .the variance can be seen more clearly. In these two plots which feature looks like more clear in terms of classification. Here from the area worst feature in the last swarm plot looks malignant and benign are separated not totally but mostly.

Evaluation :The performance evaluation criteria allow the measurement of the accuracy of the models developed using the training dataset. The results of the classification are recorded on a confusion matrix. A confusion matrix is a square which shows the actual classification along the vertical and the predicted along the vertical. All correct classifications lie along the diagonal from the north-west corner to the south-east corner also called True Positives (TP) and True Negatives (TN) while other cells are called the False Positives (FP) and False Negatives (FN). If the unlikely case is considered positive then likely and benign are called negatives, if likely is considered as positive then unlikely and benign are considered negatives and the same also applies if benign is called the positive. These values are used to determine the following evaluation criteria. The error rates of the developed models using both classifiers were also determined alongside with the performance evaluation criteria mentioned above.

Graphs of experiments:

(Join Plot)
(Violin Plot)
(Swarm Plot)

Result tabe: