مجلة البحوث المالية والتجارية
المجلد (25) – العدد الثاني – إبريل 2024

# Classification of lung cancer Data using Support Vector Machine and Discriminant Analysis.

**Maha Farouk[1], and Abdel Rahim Awad Bassiouni[2]**

**(1) Department of Statistics, Mathematics, and Insurance-Faculty of Commerce – Tanta University, Tanta-Egypt.**

**(2) PhD Statistics, Faculty of Commerce, Tanta University.**

**Maha.ibrahim@commerce.tanta.edu.eg,**

**dr-abdelreheembassuny@outlook.co**



| | | |
|---|---|---|
| | تاريخ الإرسال | 30 _12_ 2023 |
| | تاريخ القبول | 06 -02-2024 |
| | رابط المجلة: | https://jsst.journals.ekb.eg/ |

## Abstract:

**This paper aims to compare the Support Vector Machine (SVM) and Discriminant Analysis (DA) in the classification process using a sample of 150 lung cancer patients due to accuracy, sensitivity, specificity, and misclassification rate. The patients were divided into two groups: the first group consists of 101 patients with early-stage lung cancer, and the second group contains 49 patients with advanced-stage lung cancer. This classification was based on a set of independent variables: age (x1), smoking (x2), occupation (x3), COVID-19 infection (x4), and treatment methods (x5). The results advocate the superiority of SVM over Discriminant Analysis. That is, SVM was found to achieve an accuracy rate of 90.76%, sensitivity of 97%, and specificity of 77.5%, resulting in a misclassification rate of 9.33%. On the other hand, Discriminant Analysis accuracy rate is 70%, sensitivity is 70.29%, specificity is 69.4%, and a misclassification rate is 30%.**

**Keywords: SVM, DA, classification, sensitivity, lung cancer**

**المستخلص:**

**يهدف هذاالبحث إلى مقارنة اسلوبي الة ناقل الدعم (SVM) والتحليل التمييزي (DA) في عملية التصنيف باستخدام عينة مكونة من 150 مريضًا بسرطان الرئة اعتمادا علي الدقة والحساسية ومعدل التصنيف الخاطئ. تم تقسيم المرضى إلى مجموعتين: المجموعة الأولى تتكون من 101 مريض بسرطان الرئة في مرحلة مبكرة، والمجموعة الثانية تحتوي على 49 مريضا بسرطان الرئة في مرحلة متقدمة. اعتمد هذا التصنيف على مجموعة من المتغيرات المستقلة: العمر (x1)، التدخين (x2)، المهنة (x3)، الإصابة بفيروس كورونا (x4)، وطرق العلاج (x5). تؤكد النتائج تفوق SVM على التحليل التمييزي. وهذا يعني أن SVM حقق معدل دقة 90.76%، وحساسية 97%، ونوعية 77.5%، كما ان معدل التصنيف الخاطئ قدره 9.33%. من ناحية أخرى، تبلغ نسبة دقة التحليل التمييزي 70%، والحساسية 70.29%، والنوعية 69.4%، ومعدل التصنيف الخاطئ 30%.**

**الكلمات المفتاحية: SVM، DA، التصنيف، الحساسية، سرطان الرئة**

## 1. introduction

**Artificial intelligence techniques are highly flexible in dealing with various fields, especially in the medical field, where they can identify and classify medical conditions. Its medical applications contain a vast amount of data for each medical case, making it challenging to use traditional methods for classification. Due to the difficulty of distinguishing and separating the stage of lung cancer for a patient, whether they are still in the early stages of the disease, which can facilitate limiting the spread of the disease and facilitate its elimination by determining the appropriate treatment method, or if the disease is in an advanced stage, requiring intensified doses to limit the spread of the disease and thus reduce expenses and decrease the mortality rate. Therefore, we need modern techniques and algorithms to perform such tasks. Several algorithms and techniques are used for classification, including both modern approaches like Support Vector Machine (SVM) and traditional ones like Discriminant Analysis. The selection of the appropriate technique for classification depends on the nature of the data and the objective of the classification process. Classification helps understanding complex data and making intelligent decisions based on the analysis of this data.**

**Using Discriminant Analysis SVM to classify data has been done for a long time till now. (Ovirianti et al., 2019) discovered an 88% classification rate after examining the SVM classification in the RBF kernel function. Using discriminant analysis, Baghrouch (2020) divided Algerian bank borrowers into good and poor groups according on how they repaid their debts. Support vector machines and linear discriminant analysis are employed at (Z. Rustam, et al., 2021) to categorize data on breast cancer. SVM was used by Rasheed (2022) to effectively classify linear and nonlinear brain MRI images. Furthermore, utilizing data from the Islamic Bank of Iraq for Investment and Development, Jaber and Ibrahim (2022) used SVM to categorize financial stock data into growing and declining patterns. They contend that SVM offers excellent accuracy and efficiency. Decision trees, random forests, support vector machines, and K-nearest neighbors are the four machine learning algorithms that (Phongying M., Hiriote S. (2023)) assess the effectiveness of diabetic classification models utilizing. The statistical discriminant approaches were reviewed by (Guedes, B. & Gomes, P., 2023), who emphasized their vital and useful function in classifying patients into various MDs groups. The linear discriminant analysis approac2023) used to categorize cardiac illness (Isnanto, R. R. et al., 2023)). Utilized datasets**

**come from the machine learning library at UCI. (R. Criswell, 2023) To find any distinctions between SVM and logistic regression, the classifications of the ACME insurance dataset were compared. Additionally, Alkhmiri (2023) used discriminant analysis, with a classification accuracy rate of 95.4%, to categorize data on newborns in Yemen into alive or died categories.**

**At this paper, we will study the traditional model, Discriminant Analysis, and SVM in the classification process. Despite the significant similarity between the two models in the idea of classification, where Discriminant Analysis relies on constructing a discriminant function that separates the data into two groups, SVM creates a hyperplane that linearly or non-linearly separates the data into two groups. However, there is no study that combines the two models in the classification process, which is the distinctive feature of the current study. The main objective of the research is to compare the Support Vector Machine (SVM) as a modern technique against Discriminant Analysis as a traditional model in classifying lung cancer patients into early-stage or advanced-stage disease. the utilization of new methods in the field of binary classification to keep up with the advancements in the fields of science and knowledge. Hence, the Support Vector Machine (SVM) technique was employed due to its significance and modernity in addressing nonlinear problems.**

**The remainder of this paper is organized as follows: Section 2 gives research Methodology, in section 3 real data application are present, conclusions are drawn in Section 4.**

## 2. Methodology:

**Machine learning is considered a branch of artificial intelligence that focuses on studying applications to achieve more accurate results (Langley & Simon, 1984). Machine learning can be divided into two main types as:**

$$\text{Machine learning}\begin{cases}\text{unsupervised learning}\\ \text{supervised learning}\begin{cases}\text{regression}\\ \text{classification}\end{cases}\end{cases}$$

**Here, we concern with classification using Support Vector Machine**

### 2.1. Support Vector Machine (SVM)

**Support Vector Machine (SVM) is a supervised learning algorithm developed in 1992 by Vapnik for data classification by finding a separating hyperplane.**

### a. Linear SVM

**Support Vector Machines (SVMs) are used to separate data that can be linearly separable by a hyperplane. A hyperplane is a straight line whose task is to separate a set of data points into two groups. The goal is to find the optimal values of the linear equation that best separates the data, with the closest points to the hyperplane representing support vectors. The distance between the closest point in each group and the hyperplane is called the margin. The objective is to maximize the distance between the closest point and the hyperplane, making it easier to classify new observations, as illustrated in figure (1).**

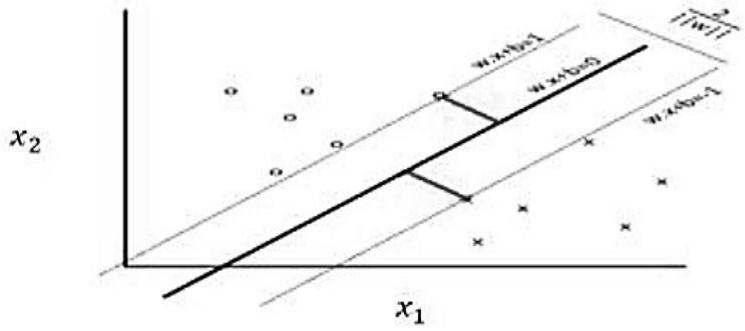



**Figure (1) Linear SVM support vector machine (Baraithiya;2019)**

### b. Nonlinear SVM

**Nonlinear SVM is Used to classify data that is characterized by being non-linearly separable and performs the process of converting it from linear to non-linear by adding a feature or a third dimension and retrying the linear separation. If it does not happen, a feature or a fourth dimension is added, and so on until the linear separation is done by a hyper plane. This method is called by the kernel function (Belloti, 2009) as shown in figure (2).**

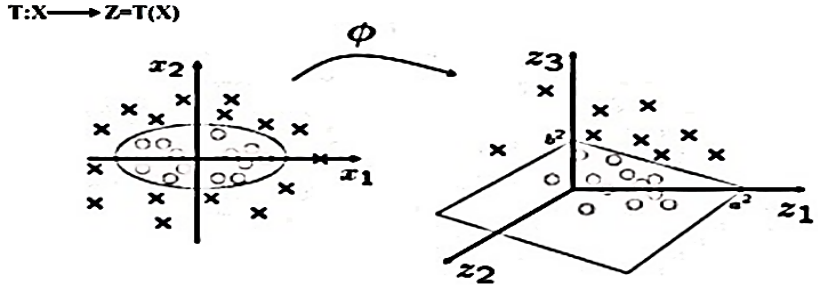



**Figure (2) Nonlinear SVM support vector function.**

**Where ∅ denotes the feature function, and there are multiple types of kernel functions used in the classification process, which are as shown in table (1) .(Huang et al, 2007)**

**Table (1) type of kernel function for classification**

| Function type | Mathematical formula |
|---|---|
| Gaussian (RBF) Kernel | $K(X_i, X_j)=exp^{(-\frac{\|X_i-X_j\|^2}{26^2})}$ |
| Polynomial Kernel | $K(X_i, X_j)= (1 + X_i^T . X_j)^d$ |
| Linear Kernel | $K(X_i, X_j)= X_i^t . X_j$ |
| Monomial | $K(X, y) = [X^T . y]^m$ |

## C. The mathematical formulation of the Support Vector Machine (SVM)

**SVM method performs the classification process by placing the data on either side of the hyperplane. The set of points that lie on one side and closest to the separating line is called the Support Vector. The equation for the first and second group vectors is shown below:**

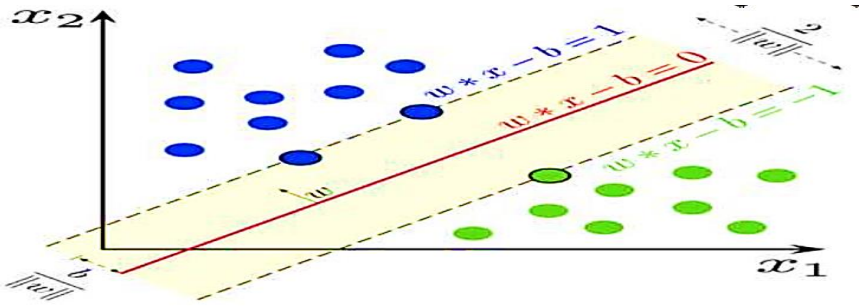
$$W^T . X_i + b = +1 \quad \text{for} \quad y = +1$$
$$i = 1, 2, \dots\dots . . N \quad (1)$$
$$W^T . X_i + b = -1 \quad \text{for} \quad y = -1$$
$$i = 1, 2, \dots\dots . . N \quad (2)$$

**The equation of the separating line or the hyperplane is:**

$$W^T . X_i += 0 \quad (3)$$

**The distance between the separating line and the closest point of the points of each group is called the margin. The greater the distance, the greater the probability of classifying new observations into one of the groups, as shown in figure (3).**

**Figure (3) SVM boundary plot (Noyum;2021).**




**Therefore, the initial formulation of SVM is formulated as**

$$L(w, b, \propto) = \frac{1}{2}\|w\|\frac{2}{2} - \sum_{i=1}^{N} \alpha_i\left(y_i\left(X_i^T w + b\right) - 1\right) \quad (4)$$

**When the equation is unpacked:**

$$L(w, b, \propto) = \frac{1}{2} W^T - \sum_{i=1}^{N} \propto_i y_i X_i^T W + b\left(\sum_{i=1}^{n} \alpha_i y_i + \sum_{i=1}^{n} \alpha_i\right)$$

$$\frac{\partial l\,(w, b, \propto)}{\partial W} = 0 \quad , W^* = \sum_{i=1}^{N} \alpha_i\, y_i\, X_i \qquad (5)$$

**Where $\alpha_i$ is the Lagrange multiplier:** $0 \le \alpha_i \le 1$

$$\frac{\partial L}{\partial b} = 0 \quad , \quad \sum_{i=1}^{N} y_i \alpha_i = 0$$

**The distance between the support vectors:** $\frac{2}{\|w^*\|}$

**To find the value of the bias term b:**

$$y(W^T X + b) = 1$$

**By substituting in Equation 5:** $y(\sum \alpha_i\, y_i\, x_i \;\; x + b) = 1$

**And multiply y,** $y^2(\sum \alpha_i\, y_i\, x_i \;\; x + b) = y$

**One of the support vector equations is y=∓1 and therefore**

$$y^2 = 1 = y - \sum_{i=1}^{n} \alpha_i\, y_i\, x_i \;\; x \qquad (6)$$

**By putting the equation in modified form**

$$b = \frac{1}{N} \sum_{i=1}^{N} \left(y - \sum_{i=1}^{n} \alpha_i\, y_i\, x_i\, x\right) \qquad (7)$$

**Where N is the total number of support vectors.**

**Therefore, the classification equation is:**

$$y = W\, x_i + b \qquad (8)$$

## 2.2 ) Discriminant Analysis:

**Discriminant analysis is a multivariate analysis method that relies on the discrimination between the sample data points relying on a linear combination of the independent variables called the discriminant function, which works to maximize the differences or disparities between the means of the groups resulted from the discrimination, so that the degree of homogeneity between the data points of each group increases. Then, the unknown data points are classified to one of the groups with the least possible classification error using discriminant functions (Pohar, Blas et al, 2004 & Johanson and Wichern, 2002). Discriminant analysis has two tasks, which are discrimination and then classification. To enable discriminant analysis to perform its first task, which is discrimination, we start by forming a few discriminant**

functions, which are the categories of the dependent variable minus one or the number of independent variables, whichever is less.

## A. Discriminant Function

For creating the Discriminant Function three steps must be done. First, we calculate the distance or difference between the means of each variable for the two groups as follows:

$$di = \bar{X}_{i(1)} - \bar{X}_{i\,(2)} = \begin{bmatrix} \bar{X}_{1(1)} & \bar{X}_{1\,(2)} \\ \bar{X}_{2(1)} & \bar{X}_{2(2)} \\ \bar{X}_{k(1)} - & \bar{X}_{k\,(2)} \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_k \end{bmatrix} \qquad (9)$$

$\bar{X}_{i\,(1)}$: means of the variables in the first group.

$\bar{X}_{i\,(2)}$ :means of the variables in the second group.

K: number of independent variables.

Second, creating the variance-covariance matrix between the two groups as follows:

$$V = \begin{bmatrix} V_{11} & V_{12} \ldots . & V_{1k} \\ V_{21} & V_{22} \ldots . & V_{2k} \\ V_{k\,1} & V_{k\,2} \ldots & V_{kk} \end{bmatrix}$$

were,

$$\text{Common variance} = V_{ii} = \frac{S_{i\,i\,(1)} + S_{i\,i\,(2)}}{n_1 + n_2 - 2} \qquad (10)$$

, and Covariance $= V_{ij} = \frac{S_{ij\,(1)} + S_{ij\,(2)}}{n_1 + n_2 - 2}$ $(11)$

Third, Building the discriminant function with standard coefficients takes the following form:

$$\hat{L} = \hat{\alpha}_1\, x_1 + \hat{\alpha}_2\, x_2 + \cdots \ldots \ldots . + \hat{\alpha}_k\, x_k \qquad (12)$$

whereas:

$\hat{L}$: is the standard discriminating value.

$x_k$: normalized discriminant variables for K variables.

$\hat{\alpha}$: The standardized discriminant coefficients are calculated as

$$\hat{\alpha} = V^{-1}\, d \qquad (13)$$

b. test the ability of the discriminant function to distinguish and separate groups.

To test the ability of the discriminant function to distinguish and separate groups, we use the following tests:

### 1- Wilk's Lambda test

The statistical assumptions for this test are:

The distinction on its ability is not a function, $H_0$: $M_1 = M_2$

, Differentiation is based on the ability of its function, $H_1$: $M_1 \neq M_2$

**Test statistic,** $\Lambda = \prod_{i=1}^{k} \frac{1}{1+\lambda_i}$ **(14)**

**Where $0 \leq \Lambda \leq 1$, if the value of Λ is close to one, the discriminant function does not have a strong ability to distinguish between the two groups. If the value of Λ is close to zero, the discriminant function has a strong ability to distinguish between the two groups. Wilk's Lambda statistic is used to select the variables that are included in the discriminant function. The variables with the lowest value of Wilk's Lambda statistic are the most important variables for distinguishing between the two groups. The variables with the highest value of the F statistic are also important for distinguishing between the two groups.**

## 2- Lawely – HoTelling (T):

**Test statistic:** $T^2 = \sum_{i=1}^{k} \lambda_i$ **(15)**

**We convert $T^2$ to F as follows:**

$$F = \frac{n_1 + n_2 - k - 1}{(n_1 + n_2 - 2)k} * T^2 \qquad (16)$$

**Its tabular value is:**

$F\alpha, (k-1, n_1 + n_2 - k - 1)$

**If the test statistic is greater than the tabular value, the function has a high ability to discriminate.**

**After the process of discrimination or separation comes the second task of discriminant analysis, which is classification, that is, classifying the new observations into one of the groups that have been distinguished, and this is done as follows:**

### 1. Cut of point:

**The discriminant value of the observations in the first group is determined as follows:**

$$\hat{L}_{1(1)} = \hat{\alpha}_1 x_{11} + \hat{\alpha}_2 x_{21} + \cdots \ldots \ldots \ldots + \hat{\alpha}_n n_1$$

$$\hat{L}_{2(1)} = \hat{\alpha}_1 x_{12} + \hat{\alpha}_2 x_{22} + \cdots \ldots \ldots \ldots + \hat{\alpha}_n n_2$$

**And so on:**

**The mean of the discriminant values for the first group and second group are calculated as:**

$$\bar{L}_{(1)} = \frac{\sum_{i=1}^{n1} L_{i\,(1)}}{n_1} \qquad , \bar{L}_{(2)} = \frac{\sum_{i=2}^{n2} L_{i\,(2)}}{n_2}$$

**The cut-off point is:** $\bar{\bar{L}} = \frac{\bar{L}_{(1)} + \bar{L}_{(2)}}{2}$ **(17)**

**To Classify or predict the belonging of a new observation to group (1) or group (2) with the least classification error as table (2) shows.**

**Table (2) Classification Role**

| Classification | $\bar{L}_{(1)} > \bar{L}_{(2)}$ | $\bar{L}_{(1)} < \bar{L}_{(2)}$ |
|---|---|---|
| If the discriminating value of the new item$> \bar{\bar{L}}$ | Group (1) | Group (2) |
| If the discriminating value of the new item $< \bar{\bar{L}}$ | Group (2) | Group (1) |
| If the discriminating value of the new item $= \bar{\bar{L}}$ | Randomly within either group | Randomly within either group |

**The classification table is as table (3) shows.**

**table (3) The classification table**

| the group | P (1) | N (2) |
|---|---|---|
| P (1) | TP | FN |
| N (2) | FP | TN |
| $N$ | $n_1$ | $n_2$ |

**Also, table (4) shows the classification indicators formulas of Percentage of correct classification in the first group ($S$), Specificity ($S_p$), Efficiency ($E_t$), and Misclassification error.**

**table (4), the formulas of ($S$), ($S_p$), and ($E_t$):**

| indicator | formulas |
|---|---|
| Percentage of correct classification in the first group | $S = \frac{TP}{TP + FP} \times 100$ |
| Specificity the ratio of correct classification in the second group | $S_P = \frac{TN}{FN + TN} \times 100$ |
| Efficiency Percentage of correct classification | $E_t = \frac{TP + TN}{n_1 + n_2} \times 100$ |
| Misclassification error the percentage of misclassification in the sample | $\mathrm{Me} = \frac{\mathrm{FN} + \mathrm{FP}}{N}$ |

## 2. Applied Study

**The paper's objective is to compare the support vector machine (SVM) as one of the modern techniques and the discriminant analysis as one of the traditional models in data classification, applied to a sample of 150 lung cancer patients, of which 101 patients are in the early stage of the disease and 49 patients are in the advanced stage of the disease.**

### 3.1. Model variables:

**The study model consists of the following variables as shown in table (5)**

**Table (5), The study variables**

| variable | |
|---|---|
| dependent (y) | If (y ≥+1), then the patient is in an early stage of the disease (Class 1) |
| | If (y ≤ -1), then the patient is in an advanced stage of the disease (Class 2) |
| $X_1$ | Age(numerical) |
| $X_2$ | Smoking (0 nonsmoker, 1 smoker) |
| $X_3$ | Occupation (1 employee, 2 worker, 3 retired) |
| $X_4$ | COVID-19 infection (1 infected, 2 not infected) |
| $X_5$ | Treatment methods (1 chemotherapy, 2 radiation, 3 both) |

### 3.2. Support machine (SVM)

**Using Stata 19 statistical software, based on Python and R language, to find the support vector function using the radial basic kernel function (RBF) with an error value of e=.1 and margin size of C=1. The results of using the support vector machine (SVM) to classify the patients into the first group or the second group were as shown in table (6).**

**Table (6) Classification of Support Vector Machine**

| Actually | Group size | Predict class 1 | Predict class 2 |
|---|---|---|---|
| class 1 | 101 | 98 | 3 |
| | | %97.03 | %2.97 |
| class 2 | 49 | 11 | 38 |
| | | %22.45 | %77.55 |

**From table (6), on a sample of 150 lung cancer patients, which were divided into two groups, with a size of 101 patients in the first group, 98 patients were correctly classified into the first group, and 3 patients**

**were incorrectly classified into the second group. The second group consists of 49 patients, of which 38 patients were correctly classified into the second group, and 11 patients were incorrectly classified into the first group. Therefore, the performance indicators of the classification process using the support vector machine are shown in table (7)**

**Table (7), The (SVM)classification indicators**

| indicator | value |
|---|---|
| Sensitivity | 97% |
| Specificity | 77.5% |
| Efficiency | 90.76% |
| Classification error | 9.33% |

**Where Sensitivity means that the patient is in early stage of the disease and is classified as being in the early stage of the disease, Specificity ($S_p$) is the patient is in the advanced stage of the disease and is actually classified as being in the advanced stage of the disease.**

**Support vectors:**

**The support vectors represent a set of observations, that is, the patients closest to the hyper plane in both groups. The support vectors in the first set that shown in table (8) are determined based on the outputs of the following function:**

$$y = W^{*}X_i + b$$

**Table (8): the support vectors in the first set, SVM**

| N0. support vector | observation | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|---|
| 1 | 3 | 0.71128 | 0.25637 | 0.25637 |
| 2 | 7 | -0.66632 | 0.46696 | -0.94537 |
| 3 | 8 | -0.28059 | -2.12725 | -0.94537 |
| 4 | 13 | -.72142 | .46696 | -.94537 |
| 5 | 18 | -.28059 | -2.12725 | -.94537 |
| 6 | 29 | .60107 | -2.12725 | .25637 |
| 7 | 30 | .43576 | -2.12725 | 1.45811 |
| 8 | 34 | -1.93370 | .46696 | .25637 |
| 9 | 36 | -.22548 | -.46696 | -.94537 |
| 10 | 46 | -.11528 | .46696 | .25637 |
| 11 | 52 | .38066 | -2.12725 | .25637 |
| 12 | 68 | -2.09901 | .46696 | -.94537 |
| 13 | 69 | -.61121 | .46696 | -.94537 |

| | | | | |
|---|---|---|---|---|
| 14 | 75 | -.94183 | .46696 | -.94537 |
| 15 | 77 | -.72142 | .46696 | -.94537 |
| 16 | 78 | -77652 | .46696 | -.94537 |
| 17 | 83 | -1.05204 | .46696 | -.94537 |
| 18 | 84 | -.44590 | -2.12725 | -.94537 |
| 19 | 85 | 1.37253 | .46696 | -.94537 |
| 20 | 93 | -1.76839 | .46696 | -.94537 |
| 21 | 99 | 1.20721 | .46696 | .25637 |
| 22 | 100 | 1.04190 | .46696 | .25637 |
| 23 | 101 | -.17038 | .46696 | .25637 |
| 24 | 102 | 2.36439 | .46696 | .25637 |
| 25 | 104 | .27045 | .46696 | .25637 |
| 26 | 107 | .93170 | .46696 | .25637 |
| 27 | 110 | -.28059 | .46696 | -.94537 |
| 28 | 116 | -.61121 | -2.12725 | 1.45811 |
| 29 | 119 | -.66632 | .46696 | -.94537 |
| 30 | 121 | -.44590 | .46696 | -.94537 |
| 31 | 125 | .10514 | .46696 | .25637 |
| 32 | 126 | -.06017 | .46696 | .25637 |
| 33 | 131 | .05003 | .46696 | .25637 |
| 34 | 134 | -2.0990 | -2.12725 | -.94537 |
| 35 | 137 | -.72142 | .46696 | -.94537 |
| 36 | 138 | -.44590 | -2.12725 | -.94537 |
| 37 | 141 | -1.10715 | .46696 | -.94537 |
| 38 | 142 | -.3908 | -2.12725 | 1.45811 |
| 39 | 144 | -2.2092 | .46696 | -.94537 |

| N0. support vectors | observation | $x_4$ | $x_5$ |
|---|---|---|---|
| 1 | 3 | -0.49833 | -1.77588 |
| 2 | 7 | -0.49833 | 0.73716 |
| 3 | 8 | -0.49833 | 0.73716 |
| 4 | 13 | -.49833 | .73716 |
| 5 | 18 | -.49833 | -.51936 |
| 6 | 29 | 1.99332 | .73716 |
| 7 | 30 | -.49833 | .73716 |
| 8 | 34 | 1.99332 | -1.77583 |
| 9 | 36 | .49833 | -.51936 |
| 10 | 46 | 1.99332 | .73716 |
| 11 | 52 | 1.99332 | .73716 |

| | | | |
|---|---|---|---|
| 12 | 68 | -.49833 | .73716 |
| 13 | 69 | -.49833 | .73716 |
| 14 | 75 | -.49833 | -1.77588 |
| 15 | 77 | 1.99332 | .73716 |
| 16 | 78 | -.49833 | .73716 |
| 17 | 83 | -.49833 | -.51936 |
| 18 | 84 | -.49833 | .73716 |
| 19 | 85 | 1.99332 | -.51936 |
| 20 | 93 | -.49833 | -.51936 |
| 21 | 99 | -.49833 | -1.77588 |
| 22 | 100 | -.49833 | -.51936 |
| 23 | 101 | -.49833 | .73716 |
| 24 | 102 | -.49833 | .73716 |
| 25 | 104 | 1.99332 | .73716 |
| 26 | 107 | -.49833 | -1.77588 |
| 27 | 110 | -.49833 | .73716 |
| 28 | 116 | -.49833 | .73716 |
| 29 | 119 | -.49833 | .73716 |
| 30 | 121 | 1.99332 | .73716 |
| 31 | 125 | -.49833 | .73716 |
| 32 | 126 | -.49833 | .73716 |
| 33 | 131 | -.49833 | .73716 |
| 34 | 134 | -.49833 | .73716 |
| 35 | 137 | 1.99332 | -.51936 |
| 36 | 138 | -.49833 | -.51936 |
| 37 | 141 | -.49833 | .73716 |
| 38 | 142 | -.49833 | .73716 |
| 39 | 143 | -.49833 | .73716 |

**Based on the previous table, from the first column, the number of support vectors in the first group was 39. The second column shows the observations that represent the support vectors when using the support vector machine (SVM). The remaining columns represent the values of the variables for those observations or patients. According to the statistical program, the standardized values of the variables (standardize) were used. The support vectors in the Second set are shown in table (9)**

**Table (9): The support vectors in the second group:**

| N0. support vectors | Observation | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|---|
| 1 | 31 | -1.32756 | .46696 | .25637 |
| 2 | 39 | 1.59294 | -2.12725 | -.94537 |
| 3 | 44 | .60107 | .46696 | 1.45811 |
| 4 | 45 | .49086 | .46696 | 1.45811 |
| 5 | 48 | -.06017 | .46696 | 1.45811 |
| 6 | 53 | .43576 | .46696 | 1.45811 |
| 7 | 54 | -.61121 | .46696 | 1.45811 |
| 8 | 64 | .60107 | .46696 | 1.45811 |
| 9 | 65 | .76638 | .46696 | 1.45811 |
| 10 | 98 | 1.20721 | -2.12725 | 1.45811 |
| 11 | 109 | .32555 | .46696 | -.94537 |
| 12 | 114 | .60107 | .46696 | 1.45811 |
| 13 | 115 | -.00507 | .46696 | 1.45811 |
| 14 | 122 | .81047 | .46696 | .25637 |
| 15 | 135 | -1.93370 | .46696 | -.94537 |
| 16 | 139 | 2.4195 | .46696 | 1.45811 |
| 17 | 143 | 1.42763 | -2.12725 | 1.45811 |
| 18 | 146 | 1.42763 | -2.12725 | 1.45811 |
| 19 | 149 | 2.5291 | .46696 | 1.45811 |

| N0. support vectors | observation | $x_4$ | $x_5$ |
|---|---|---|---|
| 1 | 31 | -.49833 | -.51936 |
| 2 | 39 | -.49833 | .73716 |
| 3 | 44 | -.49833 | -1.77588 |
| 4 | 45 | -.49833 | -1.77588 |
| 5 | 48 | -.49833 | .73716 |
| 6 | 53 | -.49833 | .73716 |
| 7 | 54 | -.49833 | .73716 |
| 8 | 64 | -.49833 | .73716 |
| 9 | 65 | -.49833 | .73716 |
| 10 | 98 | 1.99332 | -.51936 |
| 11 | 109 | -.49833 | -1.77588 |
| 12 | 114 | -.49833 | -1.77588 |
| 13 | 115 | -.49833 | -1.77588 |
| 14 | 122 | -.49833 | .73716 |

| | | | |
|---|---|---|---|
| 15 | 135 | -.49833 | .73716 |
| 16 | 139 | -.49833 | -1.77588 |
| 17 | 143 | -.49833 | .73716 |
| 18 | 146 | 1.99332 | .73716 |
| 19 | 149 | -.49833 | -.51936 |

**The number of support vectors in the second group was 19. The first column shows the observations that represent the support vectors when using the support vector machine (SVM), starting with observations (31, 39, 44, ..., 149). Therefore, the total number of support vectors in both groups is 58.**

**To know the observations (patients) in the first group that were misclassified to the second group, as are shown in table (10)**

**Table (10): misclassified observations in the first group**

| Observation | Value (y) |
|---|---|
| 44 | .0494066 |
| 54 | .0533826 |
| 65 | .0151714 |

**It is observed that there are three observations from the first group that were misclassified to the second group. This is called a classification error. As for the observations that belong to the second group and were misclassified as belonging to the first group, they had shown in table (11)**

**Table (11): misclassified observations in the second group**

| Observation | Value (y) | Observation | Value (y) |
|---|---|---|---|
| 36 | -.02647 | 110 | -.094729 |
| 77 | -.0434359 | 121 | -.065588 |
| 78 | -.043426 | 137 | -.0265206 |
| 81 | -.026149 | 144 | -.0562055 |
| 84 | -.02122 | 148 | -.0355268 |
| 85 | -.067996 | | |

**From the previous table, there are (11) patients from the second group, that is, in the late stage of the disease, who were misclassified as belonging to the first group, that is, they are still in the early stage of the disease. This is consistent with the outputs of the classification table. To classify new observations into one of the two groups, it is necessary to estimate the equation of the hyper plane, which is $y =$**

$W^*X + b$, and thus obtain the estimates of the weight vector (W) and the value of the bias (b), as shown in table (12).

Table (12), the weight vector (W) and the value of the bias (b)

| | $x_1$ age | $x_2$ smoking | $x_3$ occupation |
|---|---|---|---|
| $W_i$ | 1.944 | 4.1274 | 2.125 |
| | $x_4$ infection | $x_5$ treatment methods | b the bias |
| $W_i$ | .5711 | .34991 | 0.2694 |

Using the program (R), These values were randomly selected. Therefore, the following new observations were classified as shown in table (13).

Table (13), classification of new observations((SVM))

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | Predict | observation |
|---|---|---|---|---|---|---|
| 77 | 1 | 3 | 1 | 3 | Class 2 | 151 |
| 62 | 0 | 2 | 2 | 2 | Class 1 | 152 |
| 63 | 1 | 2 | 2 | 1 | Class 1 | 153 |
| 64 | 1 | 3 | 1 | 1 | Class 2 | 154 |

For example, if a patient is 77 years old, a smoker, retired, infected with COVID-19, and receiving both chemotherapy and radiation therapy, he would be classified as belonging to the second group, that is, in the late stage of the disease. And so on.

### 3.3 Discriminant Analysis:

Before beginning data classification, some tests must be done.

1. test the equality of the means of each variable from the two groups.

To test the equality of the means of each variable from the two groups and to ensure that there are statistically significant differences, the hypotheses are formulated as:

$$H_0: M_1 = M_2 \quad , \quad H_1: M_1 \neq M_2$$

From SPSS outputs in table (14), the significance of all independent variables is evident, where the Sig value is less than 0.05, thus rejecting the null hypothesis and accepting the alternative hypothesis that there

**are statistically significant differences between the means of each variable in both groups:**

**Table (14) Test of equality of means**

| variable | Groups | | Sig |
|---|---|---|---|
| | Class 1 | Class 2 | |
| $x_1$ | 47.222 | 61.5451 | .001 |
| $x_2$ | 0.8081 | 0.9431 | .059 |
| $x_3$ | 1.4848 | 2.378 | .001 |
| $x_4$ | 1.1818 | 1.9353 | .0441 |
| $x_5$ | 2.3535 | 3.5294 | .0201 |

## 2. Homogeneity of variance and covariance test

**The homogeneity of variance and covariance between the two groups is tested for statistically significant differences in the variance-covariance matrix using Box´s M, which states the following hypotheses.**

$$H_0: \Sigma_1 = \Sigma_2 \quad , \quad H_1: \Sigma_1 \neq \Sigma_2$$

**The test results were as shown in table (15)**

**table (15), Box´s M test results**

| Box´s M | Approx | $df_1$ | $df_2$ | Sig |
|---|---|---|---|---|
| 30.641 | 1.957 | 15 | 4230 | 0.15 |

**Given that the Sig value is greater than 0.05, the null hypothesis of equality or homogeneity of the variance-covariance matrix for both groups is accepted, and the conditions required for using discriminant analysis are met.**

**Create the discriminant function:**

**One of the most important characteristics of discriminant analysis is its high ability to exclude or retain independent variables based on their significance. This is done according to the largest value of F and the lowest value of Wilk's Lambda. Therefore, from the program outputs, the standardized discriminant function is formulated as follows:**

$$\hat{L} = \hat{\alpha}_1 x_1 + \hat{\alpha}_2 x_2 + \hat{\alpha}_3 x_3 + \hat{\alpha}_4 x_4 + \hat{\alpha}_5 x_5$$

$$\hat{L} = 0.327 x_1 + .143 x_2 + .810 x_3 + .023 x_4 + .234 x_5$$

**To test the ability of the function to distinguish, this is done through the Wilk's Lambda test statistic as shown in table (16)**

**Table (16) Willk's Lambda test:**

| Test of function | Willk's Lambda | Chi – square | df | Sig |
|---|---|---|---|---|
| 1 | .309 | 50.057 | 5 | .001 |

**From table (4), we note that the Wilk's Lambda statistic (Λ=.309) is close to zero. This is an indication of the high ability of the function to distinguish. We can also confirm this result by finding the canonical correlation coefficient as shown in table (17)**

**Table (17), the canonical correlation coefficient**

| Function | Eigen value | % of variance | Cumulative | Canonical correlation | $R^2$ |
|---|---|---|---|---|---|
| 1 | 1.411 | 100 | 100 | .745 | 0.555 |

**The value of the canonical correlation coefficient was 0.745. This is a high value, which increases the ability of the function to distinguish. In addition, the coefficient of determination is the square of the canonical correlation coefficient and means that the independent variables contribute 55.5% to the classification of the dependent variable.**

**Using discriminant analysis in classification, the Classification cases of every group are shown as in table (18)**

**Table (18): Classification table**

| | correctly classified | incorrectly classified | Total |
|---|---|---|---|
| Class 1 | 71 | 30 | 101 |
| Class 2 | 34 | 15 | 49 |

**Also, The ( DA)classification indicators were as shown in table (19)**

**Table (19), The( DA)classification indicators**

| indicator | value |
|---|---|
| Sensitivity | 70.29% |
| Specificity | 69.4% |
| Efficiency | 70% |
| Classification error | 30% |

**the arithmetic mean of the discriminating value in the first group $\bar{L}_{(1)} = -.457$, and the arithmetic mean of the discriminating value in the second group, $\bar{L}_{(2)} = 0.887$ and thus the cut-off point:**

$$\bar{\bar{L}} = \frac{\bar{L}_{(1)} + \bar{L}_{(2)}}{2} = .215$$

**Based on the discriminant function:**

$$\hat{L} = 0.327\,x_1 + .143\,x_2 + .810\,x_3 + .023\,x_4 + .234\,x_5$$

**After converting the new observations into standard scores and substituting in the following discrimination function, and since $\bar{L}_{(1)} < \bar{L}_{(2)}$ , if the discriminatory values of the observation are greater than the cut-off point, the second group is classified, and if it is the lowest classification in the first, then new views are categorized as shown in table (20)**

***Table (20),* classification of new observations((DA)**

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | |
|---|---|---|---|---|---|
| 77 | 1 | 3 | 1 | 3 | |
| 62 | 0 | 2 | 2 | 2 | |
| 63 | 1 | 2 | 2 | 1 | |
| 64 | 1 | 3 | 1 | 1 | |

| $\hat{L}$ | observation | Predict |
|---|---|---|
| 1.858 | 151 | Class 2 |
| .006616 | 152 | Class 1 |
| .10159 | 153 | Class 1 |
| 1.036 | 154 | Class 2 |

**That is, if there is a patient who is 77 years old, a smoker, retired, infected with COVID-19, and is undergoing chemotherapy and radiation therapy, then his discriminant value is 1.858, which is greater than the cutoff point. Therefore, the patient is classified as a member of group 2, i.e., in an advanced stage of the disease. However, if the patient is 62 years old, does not smoke, is employed, has not been infected with COVID-19, and is undergoing radiation therapy only, then his discriminant value is 0.006616, which is less than the cutoff point. Therefore, the patient is classified as a member of group 1, i.e., he is still in an early stage of the disease. And so on.**

## 4. conclusion

**The paper compares the Support Vector Machine (SVM) as one of the modern techniques and Discriminant Analysis (DA) in the classification process of a sample of 150 lung cancer patients. The patients were divided into two groups: the first group consisted of 101 patients with early-stage lung cancer, and the second group consisted of 49 patients with advanced-stage lung cancer. This classification was based on a set of independent variables: age ($x_1$), smoking ($x_2$), occupation ($x_3$), COVID-19 infection ($x_4$), and treatment methods ($x_5$). The research demonstrated the superiority of SVM over Discriminant Analysis.**

**A Comparison between Support Vector Machine SVM and Discriminant Analysis (DA) classification indicators are shown in table (21)**

**Table (21), the (DA) and (SVM)classification indicators**

| Indicators | Support vector machine | Discriminant Analysis |
|---|---|---|
| Accuracy | %90.76 | %70 |
| Sensitivity | %97 | %70.29 |
| Specificity | %77.5 | %69.4 |
| Misclassification rate | %9.33 | %30 |

**It is noted that all indicators give preference to the support vector machine (SVM) as one of the modern techniques over discriminant analysis as one of the traditional models in the classification process .so, the researcher's recommendations are it necessity to use SVM in statistical studies in the field of classification.**

## References:


**Ali, I., Saad, S., & Ahmed, S. (2016). "Using one-class SVM with spam classification". Iraqi Journal of Science, pp (501-506).**

**Al-Khamri, A. M. (2023). "The Use and Application of Discriminant Analysis in Predicting Preterm Birth in Al-Sa'iyah General Hospital for Maternity and Childhood, Sana'a Capital." Electric Journal of University of Aden for Basic and Applied Sciences, Vol (4), Issue (1).**

**Anikesh. (2018). "Types of machine learning ". Analytics Jobs. http:// Analytics Jobs.in/types-of- machine -learning.**

**Badr, H. (2019). "Using Support Vector Machine (SVM) Technique in Classification with Practical Application." Al-Mustansiriya Journal of Science and Education, Vol (20), Issue (5), University of Umm Al-Qura, Iraq.**

**Baghrouch, S. (2020). "The Use of Discriminant Analysis in Estimating the Risk of Intentional Loan Default in the National Agency for Microcredit Management." Nama Journal of Economics and Trade, Vol (4), Issue (1), Algeria.**

**Baraithiya, H., &Pateriya,R.K.(2019)."Classifiers ensemble for fake review detection". International Journal of Innovative Technology and Exploring Engineering,Vol (8),Pp(730-736 ).**

**Bellotti, T., & Crook, J. (2009)." Support vector machines for credit scoring and discovery of significant features". Expert systems with applications, Vol 36(2), pp (3302-3308).**

**(Criswell R. (2023)," Classification using Logistic Regression and Support Vector Machines" https://dc.ewu.edu/**

**Guedes B., Gomes P. (2023)," discriminant analysis in classification of anxiety disorders, 2023". Biometrics & Biostatistics International Journal. Vol (12), .http://dx.doi.org/10.15406/bbij.2023.12.00404**

**Huang, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. Expert systems with applications, Vol *33*(4), pp (847-856).**

**Isnanto R. R., Rashad I., and Widodo C. E. (2023)" Classification of Heart Disease Using Linear Discriminant Analysis Algorithm" E3S Web of Conferences. https://doi.org/10.1051/e3sconf/202344802053**

**jaber, M., & Ibrahim, A. (2022). "Stock Data Classification Using Support Vector Machine in Statistical Learning." Al-Rafidain**

**Journal of Science, University of Baghdad, College of Administration and Economics, Iraq, Vol (5), pp (104-117).**

**Johanson, R.A., and Wichern, D.W. (2002). "Applied multivariate statistical analysis" prentice hall upper saddle River, Nj.**

**Langley, P., & Carbonell, J. G. (1984). "Approaches to machine learning". Journal of the American Society for Information Science, Vol 35(5), pp (306-316).**

**Ovirianti, N. H., Zarlis, M., & Mawengkang, H. (2022). "Support Vector Machine Using a Classification Algorithm". Sinkron jurnal dan penelitian teknik informatika, Vol 7(3), pp (2103-2107).**

**Phongying M., Hiriote S. (2023)," Diabetes Classification Using Machine Learning Techniques" Computation, Vol 11(5), p(96); https://doi.org/10.3390/computation11050096.**

**Pohar, M., Blas, M., & Turk, S. (2004). Comparison of logistic regression and linear discriminant analysis: a simulation study. Metodoloski zvezki, Vol 1(1), p (143).**

**Rasheed, I. K., & Abd, H. T. (2022). "Segmentation of Magnetic Resonance Images of Brain Tumors using Support Vector Machine Method (Support Vector Machine)". Journal of Administration and Economics, Vol (133).**

**Rustam Z., Amalia Y., Hartini S., Saragih G.S., (2021)." Linear discriminant analysis and support vector machines for classifying breast cancer ". IAES International Journal of Artificial Intelligence, Vol. (10), pp. (253-256).**