

T. R.
VAN YUZUNCU YIL UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF STATISTICS

**COMPARISON OF LOGISTIC REGRESSION AND DISCRIMINANT
ANALYSIS MODELS IN SURVIVAL ANALYSIS OF BLOOD CANCER**

M. Sc. THESIS

PREPARED BY: Hawar Khan Murad HASAN
SUPERVISOR: Prof. Dr. Fevzi ERDOĞAN

VAN-2023

T. R.
VAN YUZUNCU YIL UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF STATISTICS

**COMPARISON OF LOGISTIC REGRESSION AND DISCRIMINANT
ANALYSIS MODELS IN SURVIVAL ANALYSIS OF BLOOD CANCER**

M. Sc. THESIS

PREPARED BY: Hawar Khan Murad HASAN

VAN-2023

ACCEPTANCE and APPROVAL PAGE

This thesis entitled “**COMPARISON OF LOGISTIC REGRESSION AND DISCRIMINANT ANALYSIS MODELS IN SURVIVAL ANALYSIS OF BLOOD CANCER**” presented by Hawar Khan Murad HASAN under the supervision of Prof. Dr. Fevzi ERDOĞAN in the department of Statistics has been accepted as an M.Sc. thesis according to Legislations of Graduate Higher Education on 28/12/2022 with unanimity/majority of votes members of the jury.

Chair: Prof. Dr. Sinan ÇALIK

Signature:

Member: Prof. Dr. Mahmut KARA

Signature:

Member: Prof. Dr. Fevzi ERDOĞAN

Signature:

This thesis has been approved by the committee of The Institute of Natural and Applied Science on 19/1/2023 with decision number

Signature

.....

Director of Institute

THESIS STATEMENT

All information presented in the thesis was obtained in the frame of ethical behavior and academic rules. In addition, all kinds of information that does not belong to me have been cited appropriately in the thesis prepared by the thesis writing rules.

Hawar Khan Murad HASAN



ABSTRACT

COMPARISON OF LOGISTIC REGRESSION AND DISCRIMINANT ANALYSIS MODELS IN SURVIVAL ANALYSIS OF BLOOD CANCER

HASAN, Hawar Khan Murad
M. Sc. Thesis, Department of Statistics
Supervisor: Prof. Dr. Fevzi ERDOĞAN
January 2023, 73 pages

This study was examined to deal with the classification of causing blood cancer disease (BCD), In this study, the comparison between one of the classical statistical methods Logistic Regression Analysis (LRA) with the methods of discriminant techniques Discriminant Analysis (DA) in order to classify blood cancer disease (BCD) for a group of presence and absence (BCD). The dataset contains two different data were conducted to achieve the goal of the study. First, it used secondary data that obtained from Hospital Nanakali in Erbil Northern Iraq in the period from 1st January 2017 to 31st December 2018 with a follow up period until 1st March 2021. Second, the data have been collected by questionnaire that has included the association between the type of presence blood cancer disease (BCD), this data contains 350 patients, including 17 variables, 16 of which are independent variables and a dependent variable (presence 1 and absence 0) of blood cancer disease depending, and after applied the previous methods. The rate of accuracy percentage of correct classification results 96.8% for LRA and 76.2% for DA.

Keywords: Blood cancer disease, Classification, Discriminant analysis, Logistic regression analysis.



ÖZET

KAN KANSERİNİN SAĞKALIM ANALİZİNDE LOJİSTİK REGRESYON VE DİSKRİMİNANT ANALİZİ MODELLERİNİN KARŞILAŞTIRILMASI

HASAN, Hawar Khan Murad
Yüksek Lisans Tezi, İstatistik Anabilim Dalı
Tez Danışmanı: Prof. Dr. Fevzi Erdoğan
Ocak 2023, 73 sayfa

Bu çalışma, kan kanseri hastalığına (KKH) neden olan sebeplerin ve durumların sınıflandırılması konusunu incelemek amacıyla yapılmıştır. Bu çalışmada, var olanlar ile olmayanlardan oluşan bir grup temelinde, kan kanseri hastalığına neden olan sebeplerin ve durumların sınıflandırılması için klasik istatistiksel yöntemlerden biri olan Lojistik Regresyon Analizi (LRA) ile ayırt edici tekniklerden oluşan bir yöntem olan Diskriminant Analizi (DA) yöntemleri karşılaştırılmak suretiyle incelenmiştir. Söz konusu araştırmanın amacına ulaşabilmesi için veri kümesinde iki farklı veri yer almaktadır. İlk olarak, 1 Ocak 2017 tarihinden 31 Aralık 2018 tarihine kadar olan dönem için Kuzey Irak'taki Erbil şehrinde faaliyet gösteren Nanakali Hastanesi'nden elde edilen ikincil veriler, 1 Mart 2021 tarihine kadar takip edilecek şekilde kullanılmıştır. İkincisi, söz konusu veriler, kan kanseri hastalığının (KKH) tipinin de sorulduğu ve bu tipin söz konusu kanser ile arasındaki ilişkiyi de içeren bir anket formu ile toplanmış olup bu veriler, 16 'sı bağımsız 17 değişkenden ve bir adet de bağımlı değişkenden (varlık 1 ve yokluk 0) oluşan toplamda 350 hastadan edinilen verileri içermektedir. Doğru sınıflandırma sonuçlarının doğruluk yüzdesi oranı Lojistik Regresyon Analizi (LRA) için % 96.8 ve Diskriminant Analizi (DA) için ise % 76,2 'dir.

Anahtar kelimeler: Kan kanseri hastalığı, Lojistik regresyon diskriminant analizi, Sınıflandırma.



ACKNOWLEDGMENTS

With my respect, I would like to thank some persons who have helped me during this master's course. At the beginning, I would like to thank my supervisor Prof. Dr. Fevzi ERDOGAN, he supported me and guided me to prepare of this thesis. As well as my gratitude and respect to staff of Administration and Economic college, particularly department of Statistics and Informatics. I offer this thesis to my father's soul (Said Murad) as he always supported me and guided me. I wish he could see the result of this thesis which I worked on this at his request, as he passed away by cancer. At the end, I praise the efforts of my husband (Khoshnaw) to help me, I also like to offer this thesis to my daughter (Mahsa Khan), my brother and my sisters (Hamin Khan, Hedi Khan and Hana Khan), and sister in law (Narmin Khan)

2023

Hawar khan Murad HASAN



TABLE OF CONTENTS

	Page
ABSTRACT	i
ÖZET	iii
ACKNOWLEDGMENTS	v
TABLE OF CONTENTS	xi
LIST OF TABLES	xi
LIST OF FIGURES	xiii
SYMBOLS AND ABBREVIATIONS	xv
LIST OF APPENDIX	xvii
1. INTRODUCTION	1
2. LITERATURE REVIEW	3
3. MATERIALS AND METHODS.....	9
3.1. Blood Cancer	9
3.1.1. The general that factors to the tendency to develop cancer	9
3.1.2. Symptoms of blood cancer	9
3.1.3. Type of blood cancer	10
3.1.4. Diagnosis of blood cancer	11
3.1.5. Treatment of blood cancer	11
3.2. The Logistic and Discriminant Models	12
3.2.1. Population	13
3.2.2. Variables	13
3.2.3. Estimation	13
3.2.4. Discriminate model.....	14
3.2.5. Logistic model	14
3.2.6. Classification table.....	14
3.2.7. Goodness - of-Fit	16
3.2.8. Coefficient of determination R2	16
3.2.9. The odds.....	17
3.2.10. Odds ratio (OR).....	17
3.2.11. Concept of logistic regression analysis.....	17

	Page
3.2.12. A theoretical overview of logistic regression	19
3.3. Principles of Logistics Regression.....	19
3.3.1. Logistic regression analysis tests	21
3.3.2. Likelihood ratio test	22
3.3.3. Wald test	23
3.3.4. Score test.....	24
3.4. The Value of Exp(B).....	24
3.5. Goodness- of-Fit	25
3.5.1. Chi-square test χ^2	25
3.5.2. The hosmer and lemeshow Test R^2_L	26
3.6. Coefficient of Determination	27
3.6.1. Cox and snell coefficient R^2	27
3.6.2. Nagelkerke coefficient R^2_N	27
3.6.3. Concept of discriminator analysis.....	28
3.7. A Theoretical Overview of Discriminant Varieties.	29
3.8. The Discrimination Model for Two Groups	30
3.9. Discriminate Analysis Tests	32
3.9.1. Wilkes's lambda test (Λ).	32
3.9.2. F-Test F	33
3.10. Standardized and Unstandardized Discriminant Coefficients	34
3.11. Goodness-of-Fit	35
3.11.1. Chi-square test pearson X^2	35
3.11.2. Wilkes's lambda test (Λ)	36
4. APPLICATION AND MAIN RESULTS	37
4.1. Subject and Purpose of the Application.....	37
4.2. Data Used in the Application.....	37
4.3. Descriptive Statistics.....	38
4.3.1. Age of patients	38
4.3.2. Gender of (BCD).....	39
4.3.3. Blood cancer disease of years (2017-2021)	39
4.3.4. Type of (BCD)	40

	Page
4.4. Selection of the Study Samples	42
4.4.1. Application using LRA	44
4.4.2. Estimating the LRA coefficients.....	46
4.4.3. Testing the discriminatory power of the LRA model	47
4.4.4. Application using DA	48
4.4.5. Estimating the discriminant model	50
4.4.6. Testing the discriminatory power of the discriminant model	50
4.5. Comparison and Discussion.....	51
5. CONCLUSION.....	53
REFERENCES	55
EXTENDED TURKISH SUMMARY (GENİŞLETİLMİŞ TÜRKÇE ÖZET).....	61
APPENDIX	553
CURRICULUM VITAE	73

LIST OF TABLES

Table	Page
Table 3.1. Classification table	14
Table 4.1. Description of the study variables	38
Table 4.2. Age of (BCD)	38
Table 4.3. Gender of (BCD)	39
Table 4.4. Years of (BCD)	40
Table 4.5. Type of blood cancer disease	41
Table 4.6. Summary statistics.....	41
Table 4.7. Description of variables.....	43
Table 4.8. The frequency distribution of the blood cancer disease	44
Table 4.9. Summary results of blood cancer data.....	44
Table 4.10. Model fitting information.....	45
Table 4.11. Pseudo R^2	45
Table 4.12. Result of LRT.....	46
Table 4.13. Results of fitting the LRA model to (BCD)	47
Table 4.14. The final classification results using LRA model	47
Table 4.15. VIF values of predictor variables	48
Table 4.16. Selection of discriminating variables depending on stepwise method.....	49
Table 4.17. Wilks lambda table	50
Table 4.18. Unstandardized canonical discriminant model coefficients	50
Table 4.19. The final classification results	51
Table 4.20. Performance criteria for comparing methods	51



LIST OF FIGURES

Figure	Page
Figure 3.1. Shows symptoms cells of blood cancer	10
Figure 3.2. Type of blood cancer.....	11
Figure 4.1. Gender of patients	39
Figure 4.2. (BCD) of years (2017-2021) by gender	40





SYMBOLS AND ABBREVIATIONS

Some symbols and abbreviations used in this study are presented below, along with descriptions.

Symbols	Description
Λ	Wilkes's lambda test
β	Beta
β_0	Fixed limit
H_0	Hypothesis
$\ln$	Natural logarithm
μ	Mu
P	Probability
R^2	Coefficient of determination
R^2_L	The hosmer and lemeshow
R^2_N	Nagelkerke coefficient
SE_β	Standard error
W^2	Wald test
χ^2	Chi-square

Abbreviations	Description
ALL	Acute Lymphoblastic Leukemia
AML	Acute Myeloid Leukemia
APL	Acute Promyelocytic Leukemia
BCD	Blood Cancer Disease
CLL	Chronic Lymphocytic Leukemia
CML	Chronic Myeloid Leukemia
DA	Discriminant Analysis
FN	False Negative Normal
FP	False Positive

Abbreviations	Description
LRA	Logistic Regression Analysis
LRT	Likelihood Ratio Test
N	Negative
OR	Odds Ratio
<i>P</i>	Positive
SE	Sensitivity
SP	Specificity
SS_R	Sum Squared Row
SS_T	Sum Squared Total
TN	True Negative
TP	True Positive

LIST OF APPENDIX

	Page
App 1. Questionnaire form	70





1. INTRODUCTION

Blood as a liquid material is a type of connective tissue which passes through artery that called blood vessels, blood can get all points of human body during its movement and has some important functions also. For a mature person, volume of blood is about 5-6 liters (about 6-8% of the persons' weight approximately). (Dean, L., 2005) Blood made up of two main parts plasma and cells, water is a main component of plasma, also cells are consisting of three main types of leukocyte, erythrocytes and platelets. Blood has multifunction, for example, leukocytes is the first defense line of the body and most important part of immunity, which quickly eat and destroy germs and foreign bodies. Furthermore, antibiotics were excreted by lymphocytes that devour germs, therefore, leukocyte conserves the body from different diseases and its danger (Henel and Schmitz, 2007).

There are several methods and approaches of statistical analysis for investigating of those variables that have not been studied before, especially after development and growth of prediction methods, in which discriminant and regression are within these methods. Logistics' target is to find out the variables that has a great effect on graded or accepted variables for classifying new situations by depending on preceding data. Because there are many variables in this field, it is suitable to use one of (discriminant analysis and logistic regression) methods in a medical field. (Hastie et al., 2009).

The exploration of the relationship between a group of independent variables and the response variable (the dependent variable), where the dependent variable is a categorical variable, has given rise to the usage of logistic regression. In addition, LR is separated into two categories: first, a binary logistic regression model with only two classes; and second, a multinomial logistic regression model with a dependent variable that is either ordinal or nominal and has two classes as well. The Logistic Regression Model, the first to use the logistic function, was widely used by researchers (Verhulst, 1920).

Discriminant analysis is a statistical technique for classifying the sample items into two or more groups and is used to analyze multivariate data. The discriminant function performs the classification process and consists of a linear combination of

independent variables (discrimination). Following the creation of the discriminating function, the new vocabulary of one of the groups under study is classified with the least amount of classification error possible. When there are two or more groups being investigated, linear discrimination is one of the forms of discrimination that is used when the variances for the groups being examined must be equal. Non-linear discrimination is used when the variances are not equal. (Eisenbeis., 1971).

The current thesis aims to investigate and compare two methods (Logistic Regression (LR), Discriminant Analysis (DA)) based on evaluation performance standards (Model Accuracy, Model Sensitivity, Model Specificity, Classification Table) in order to select the most efficient method (best classifier) for patients with blood cancer disease.

Finding the most significant independent variables influencing the blood of cancer patients for the forecasting of future cases and lowering the prevalence of this illness in the community by receiving the required healthcare.

The current thesis is divided into four pieces. Introduction and a review of the relevant literature are included in the first section. The second component, which is separated into two theoretical parts, covers sections on discrimination analysis and logistic regression. The implementation of two models, the interpretation of findings, and real data applications are all included in the third portion. The results are described and discussed in terms of categorization tables. The primary findings, overall thesis conclusions, and suggested actions are presented in the fourth part.

2. LITERATURE REVIEW

Aida Hadi Saleh's, Logistic Regression Analysis for the Evaluation of Leukemia Patient Survival (Sheikh Zayed Emergency Hospital in Baghdad 2010-2011- Iraq). Using the logistic regression approach, a sample of patients with this disease was taken. Age, gender, disease classification (acute myeloid leukemia, acute lymphoblastic leukemia, chronic myeloid leukemia, and chronic lymphocytic leukemia), and type of leukemia are risk factors (variables) associated with the severity of this disease. The treatment was categorized as (biological treatment and chemotherapy), the location of residency as (inside the governorate and outside the governorate), and the anemic status as (within the governorate and outside the governorate) (they do not suffer from anemia and suffer from anemia). The response variable represents the condition of the injured, which corresponds to the dependent variable. In the event of death, the value was one (1), while the value in the control condition was zero (0). The purpose of this study is to determine whether a logistic regression model can be used to investigate the survival time of leukemia patients and to discover the variables impacting the patient's condition and candidates for survival in the model. Among the findings of this study, the majority of patients with leukemia are younger than 20 years old, with a percentage of 42.47 percent; the majority of patients with leukemia are female, with a percentage of 63 percent; acute myeloid leukemia is the most prevalent, with a rate of 43.8 percent; and chronic lymphocytic leukemia is the rarest, with a rate of 4.3%. With a mortality rate of 63.63 percent, acute myeloid leukemia is the leading cause of death. And that the type of treatment offered to the patient varies according to the type of disease and the stage of the disease, and when the patient reaches an advanced state he uses chemotherapy, those who are treated with chemotherapy are the most likely to die, with a mortality rate of 72.72 percent. This study recommended conducting similar studies for different types of diseases and knowing the factors affecting each of those types, and adding other factors either by using this model or using the Cox model. Expanding the use of the two-response logistic model in Economic and social studies their previous uses were confined to the field of medical and educational sciences.

According to a 2006 study by Alwar Shah Abdul Karim, various cardiac diseases can be diagnosed using statistical discriminant analysis (Yarmouk teaching hospital, Al-Shahid Adnan and Baquba- Iraq). Building a probabilistic model to distinguish between them based on some indicators and disease-related data and utilizing variables with discriminatory features, addressing advanced (natural, quadratic, and logistic discriminatory), utilizing variables with discriminatory features, and applying them to the characteristics of two types of heart disease. The study sample consisted of 206 patients drawn from each of the following institutions: Yamouk Teaching Hospital, Martyr Adnan Hospital, and Baquba Hospital in accordance with the (independent) important influence variables: age, weight, height, low blood and gender is a variable (male=1, female=2). The study's findings showed that the variable ratio of cholesterol in blood was significant in all discriminatory models, and the positive indicator of the parameter represented rise in displays was 0.12%, which is commensurate with the nature of data of a confidential and pertinent nature. Therefore, even though some variables are not far from the normal model, it is clear that the estimation ratio of the quadratic model was crucial. This is also helpful with the test of disparity of variance.

Aida (2008) Utilizing discriminant analysis to identify some eye illnesses (Ibn al-haytham teaching hospital for eye diseases, Iraq). In this research featured analysis has been used to classify some eye diseases that represented by cataract, glaucoma, and retinal detachment that dependent variable is determined by a variety of disease-related independent factors, including gender, age, right eye pressure, left eye pressure, sugar, anemia, and urea. Through the application of the linear discriminatory function, this study attempted to diagnose the factors influencing eye illnesses and determine the impact of each of these factors. Among the findings of this study is that age plays a significant role in discriminating between cataract, glaucoma, and retinal detachment, particularly for glaucoma patients. Retinal detachment can occur at any age, followed by right eye pressure and left eye pressure, with sex and anemia being the two determinants. It had no substantial impact on the instances examined in the catastrophic categorization research. The study recommended diabetics maintain a blood sugar level as close to normal as possible, and every individual over the age of forty should undergo a blood test annually for early disease detection.

In a recent paper, authors Dip Nall et al. (2016) used a hybrid machine learning boosted regression algorithm and logistic regression, to identify key biomarkers associated with depression in the National Health and Nutrition Examination Study (2009–2010). They considered factors including gender, age, race, smoking, food security, poverty income ratio, body mass index, physical activity, alcohol use, medical conditions, and medication use when designing the questionnaire. (Dip Nall et al., 2016).

Cox and Thiel independently introduced the multinomial logit model in 1966 and 1969, considerably expanding the logit model's range of use and popularizing it. The multinomial logit was first connected to the theory of discrete choice, specifically the Luces choice axiom, by Daniel McFadden in 1973. He demonstrated how the multinomial logit followed from the assumption of independent of unrelated alternatives and interpreted the odds of alternatives as relation preferences (McFadden, 1973). This gave the logistic regression a theoretical basis.

Amal (2008) prepared a study entitled the use of logistic regression analysis to identify the causes of female breast cancer. The national center for radiotherapy and nuclear medicines case study, this research was done by using the logistic regression technique, population's simple random sample was chosen, for clients receiving care at the Khartoum-based national center for nuclear medicine, a sample size of 200 patients with breast cancer outside of this range was determined.

In their study from 2020, Yousra and Eman describe a comparison of two distinct methods of machine learning used to detect breast cancer: logistic regression and artificial neural networks (ANN). Several execution criteria, such as accuracy, were used to estimate the performance and trained models. The accuracy of the ANN model was 95.4%, whereas that of the logistic regression technique was 74.7%. The findings demonstrated that ANN is a more effective tool for predictive analysis than logistic regression.

Lei (2018) used the Logistic Regression Algorithm to identify data collection from breast cancer patients, the obtained dataset from UCI Wisconsin repository, in his title paper, "Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning." 90% accuracy was eventually attained by using all 32 characteristics in the dataset to train the model. When the two characteristics of maximum texture and maximum perimeter are picked, the classification accuracy is 96.5%, which is an increase

from the result obtained from the 32 features. The author then used a feature selection technique to choose key features from the 32.

Farid Al-Jaakni as Adman Gan conducted a study (2007) Using discriminatory multivariate statistical methods, families are categorized and divided according to their socioeconomic status in society (Damascus - Iraq). Utilizing one of the techniques for multivariate statistical analysis, this study classified and divided families according to the socioeconomic structure of the community. What are the challenging economic and social conditions in a society, which was the study's main problem? What economic and societal factors can be used to justify discrimination, such as severing families? To people at various economic and social levels, what is a society? The utilization of the download determines how important the study is. Discrimination is a crucial tool for planners to use when creating national (income tax) laws with reference to in order to achieve the principle of social justice for the families in society, money is paid out, subsidies are given out, and income is distributed.) On the economic and social circumstances that surround the family as it goes through a stage of change in the community. Finding the variables and elements that influence distribution and redevelopment was the goal of this study. Families' roles in society's social economic structure and their influence on its social economic goals who this family is in relation to. To challenge the approaches and procedures of multivariate analysis for data, such as the interdependent information on the research phenomena, to become appropriate.

According to Unaizah Hanum Binti Obaidellah, Alaá Rateb Mahmoud Al-shamash, Ph.D., in 2017[3], "Cancer" is the collective term for a set of more than 100 disorders. Cancer is a collection of distinct diseases, but they all begin when aberrant cells proliferate unchecked. Cancer can result in major health problems and even death if it is not treated. The detection techniques for lung, breast, and brain malignancies are reviewed in this essay. X-ray, ultrasound, magnetic resonance imaging, and computed tomography scan pictures are combined with artificial intelligence techniques such support vector machines, linear and logistic regression, neural networks, artificial neural networks, fuzzy logic, and adaptive neuro-fuzzy inference systems. The most significant and essential developments for the diagnosis of human cancer are imaging tools. The accuracy of this procedure and the ideal medical images for usage in each type of cancer were both thoroughly researched. When it comes to lung cancer, CT scans had the highest

accuracy (96.04%), which was obtained using feedforward backpropagation, while PET pictures had a greater accuracy of 97%, which was reached using an SVM classifier. The most accurate result for brain cancer was a 100% MRI scan that was obtained using a PPN.

According to the Breast Cancer Institute (BCI), breast cancer is one of the most serious diseases in the world and is extremely effective for women, Naresh Khuriwal Nidhi Mishra writes in 2018[6]. According to clinical specialists, identifying cancer in its earliest stages can save lives. According to cancer.net, personalized resources are available for more than 120 different cancer types and associated genetic disorders. In order to analyze cancer, machine learning techniques are primarily used. In this study, they used the Wisconsin Breast Cancer Database to develop an adaptive ensemble voting approach for detected breast cancer. The aim of this study compares and explains how ANN and logistic algorithms perform better when used with ensemble machine learning methods to detect breast cancer, even when the number of variables is constrained. when contrasted with associated literature-based study. A different machine learning method is used to demonstrate how the ANN technique with a logistic algorithm achieves 98.50% accuracy.



3. MATERIALS AND METHODS

3.1. Blood Cancer

Blood is an uncontrolled proliferation of blood cells, usually this proliferation of white blood cells, as their number in the blood exceeds the normal limit, which is from 5000 to 10,000 cells per mm, in leukemia the number of cells is from 15,000 to 30,000 cells per mm, and this proliferation may reach more than 100,000 cells per mm or even be less than 5000 cells per mm (i.e., less than the normal number). The difference in this number is due to the stage of leukemia and the extent of its progression in the body. Abnormal white blood cells are produced in the bone marrow, which is the factory of all blood cells in the body. Reducing the number of manufactured blood cells and before the white blood cells are released from the blood, which leads to a temporary lack of white blood cells in the blood at the beginning of the disease, but the marrow of the body quickly releases these cancerous cells into the blood, thus increasing the number of white blood cells in the blood and this increase is usually a steady increase.

3.1.1. The general that factors to the tendency to develop cancer

- 1- Genetic abnormalities
- 2- Radiation exposure is another
- 3- Chemotherapy and radiation
- 4- Chemicals
- 5- Blood conditions that can cause leukemia
- 6- Viruses

3.1.2. Symptoms of blood cancer

One of the effects of this increase is that the patient suffers from a lack of hemoglobin and thus becomes anemic in the blood, so he begins to suffer from fatigue

from any effort he makes, nausea, and a tendency to garlic and laziness. With the continuation of the disease, the patient suffers from a lack of platelets, so his blood becomes delayed in clotting and there are several bruises in the skin as a result of the body's inability to stop clots. Frequent bleeding from the nose or gums and the increase in white blood cells to a large degree in the blood prevents the red blood cells carrying oxygen from reaching the body's cells, causing the patient with various small clots with pains in the abdomen and chest. This increase also weakens the body's resistance to various diseases, so you find that the patient often comes with several bacterial or viral infections before he is diagnosed with such a disease.

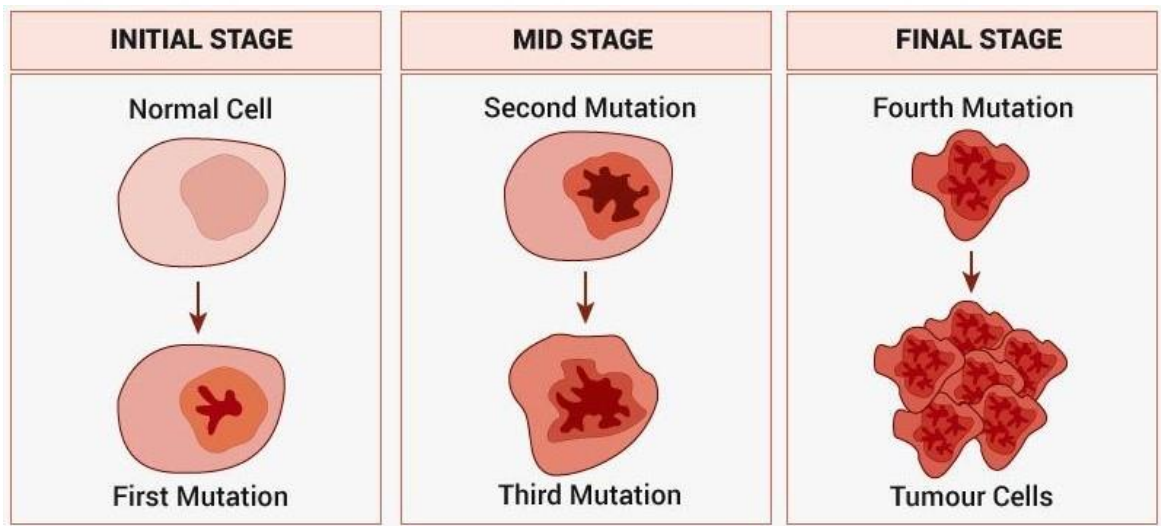


Figure 3.1. Shows symptoms cells of blood cancer.

3.1.3. Type of blood cancer

- 1- Acute leukemia
- 2- Chronic leukemia
- 3- Lymphocytic leukemia
- 4- Myeloid leukemia

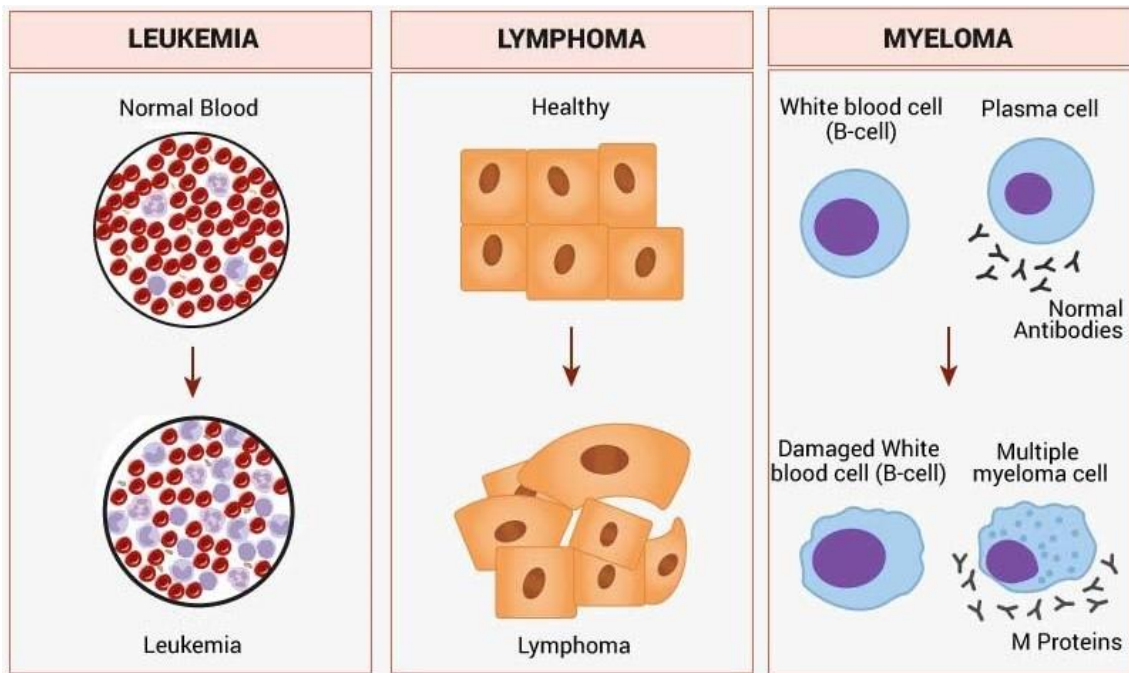


Figure 3.2. Type of blood cancer.

3.1.4. Diagnosis of blood cancer

Blood cancer is diagnosed by doing routine blood tests in a highly qualified laboratory in order to read the blood slice and see if it contains abnormal cells. If abnormal cells are found in them, we resort to taking a sample of bone marrow, and the best place to take the sample is the pelvic bone, as it is a safe place from any problems. From the study of bone marrow cells, it is estimated whether there are cancerous cells or not. If found, the cancer cells are stained with several special dyes, and the result of these special dyes determines what type of leukemia is diagnosed specifically. A sample is taken to study the chromosomes of leukemia. Many treatments depend heavily on the type of chromosomal changes that cause leukemia, and then the patient is given the appropriate treatment for this type.

3.1.5. Treatment of blood cancer

Blood cancer treatment is different from chronic leukemia, each of them has its own treatment. With the progress of medicine and the discovery of scientists, the genetic changes in every cancer have become possible to fix these genetic changes, for some

types now have a curative treatment with medicines that may not be a cure. A biological chemical that affects only to a large degree on cancer cells without any effect on the rest of the cells of the body. We mention, for example, that a type of acute leukemia called (M3) we can cure it by giving high amounts and doses of vitamin A. And another type of chronic leukemia (CML) is called Leukemia Oid Myel Chronic), it can be treated in an effective way by giving the patient a special enzyme inhibitor and it is commercially known as the drug "Gleevec", these two types in particular have achieved great success in the last ten years, and development required most of the rest of the species in the work. Bone marrow transplantation to replace it with the patient's bad blood. Many of these operations are successful now. In the near future, most types of leukemia will be able to be treated very effectively.

3.2. The Logistic and Discriminant Models

Binary logistic and discriminatory models are essential models in data reconciliation. Still, they are not widely applied in scientific fields. We find that many variables are related, but their correlation coefficients are weak, indicating either no relationship between them or it is not linear. Scientifically, there are several correlations between variables, but they are not linear and have zigzag forms, such as the parabola or the exponential function. There are various approaches to studying non-linear data, including translating it into two linear forms and analyzing them using linear regression. However, this method is not one of them. It is held responsible that the linear model's assumptions are not met. In linear regression, the values surrounding the regression line must conform to a normal distribution with a mean of zero and a variance of one; however, these assumptions are rarely valid after data processing. Therefore, it is necessary to understand certain concepts, sample size, and variable modeling by defining the form of the model equation and techniques of analysis depicted in logistic regression analysis and discriminant analysis. We believe it is vital to introduce the following concepts and definitions for clarity in the following paragraphs:

3.2.1. Population

This phrase is used in statistics to describe a group of units (group) (such as individuals, plants, cities, hospitals, and others) for which the results or statistical model must be concealed. In statistics, the community comprises all the units of the phenomenon under study. One of the purposes and objectives of social science research is to comprehend and explain social phenomena and make predictions. And to do so, we must translate our conceptualizations of these phenomena into actual data by analogy.

3.2.2. Variables

Variables are the specifications or characteristics measured on people in a given sample. Thus, the variable is that specification, characteristic, or quantity whose value changes from one element to another or from observation to another individual. If one wanted to measure the heights of kids in a classroom, one should obtain several measurements, each representing the length of one student, indicating that the distance is varied. If you record the temperature in a specific city every day for a month, you will obtain a set of values representing the temperature; consequently, the temperature in that city is a variable. Here are some other variables: Blood pressure, blood sugar level, sex, the number of children each family has in a village, an individual's economic and social position, smoking, and age were all considered.

3.2.3. Estimation

The concept of prediction is related to science in that it is one of its objectives. In addition to description and interpretation, science tries to establish the foundations and principles for existential phenomena so that they can be predicted before they occur, based on criteria and indicators that the notion of the thing to be expected is subjected to. The concept of causation is another concept related to forecasting. As is the case when forecasting meteorological conditions and temperatures based on real-time indicators(data), scientific prediction of phenomena was not a recent development. The covenant has an extension in the history of science, beginning with Laplace's concepts of

probability and his attempt to develop a model for forecasting weather conditions, and extending to modern ideas and rules in forecasting, which include logistic regression and discriminatory analysis, where the emergence of logistic and prejudiced regression equations and models is associated with the problem of applying linear regression models in predicting taxonomic distributions.

3.2.4. Discriminate model

It is the model that can differentiate between groups (separation of observations) and place each perspective in the appropriate group. And that the discriminatory model is a mathematical model that can be built using the indicators of a sample whose observations were randomly selected from two distinct groups, and that this model permits us to choose any individual (statement) and determine its group.

3.2.5. Logistic model

It is a statistical technique for analyzing the correlation between the dependent variable of the nominal level and one or more independent variables, which are also known as accompanying variables or explanatory variables and can have any type of measurement level.

3.2.6. Classification table

It is a table that displays the proportion of observed cases with a trait and the proportion of observed cases without that trait, in contrast to the proportion of expected cases with the trait and the proportion of expected cases without that trait. This allows the table to display the proportion of cases that have been correctly classified and the proportion of cases that have been classified incorrectly (Soderstorm and Leitner, 1997).

A valid assumption based on the use of this table is that if the model predicts the categorization of cases in the form of a criterion, this provides evidence that the model fits the observed data. The classification table's typical format is as follows:

Table 3.1. Classification table

Classification		Expect		Total
		Positive (P)	Negative (N)	
Observe	Positive (P)	True Positive (TP)	False Negative (FN)	P
	Negative (N)	False Positive (FP)	True Negative (TN)	P'
Total		Q	Q'	1

3.2.6.1. Sensitivity (SE)

It is one of the statistics used in analyzing classification tables, and it is defined as the value of the probability that the expected classification will be positive for the case that is actually positive, and it is calculated according to the following equation:

$$SE = \frac{TP}{TP + FN} \quad (3.1)$$

3.2.6.2. Specificity (SP)

It is one of the statistics applied in this study for classification tables and is defined as the value of the possibility that a case with an actual negative classification would result in the predicted classification being negative, and it is given according to the following equation:

$$SP = \frac{TN}{(FP + TN)} = \frac{TN}{P'} \quad (3.2)$$

3.2.6.3. Hit ratio

The value of the likelihood of proper classification is one of the statistics used in the examination of classification tables. It is also referred to as the efficiency ratio. Accordingly, if the efficiency, denoted by the symbol EF , is defined as: $EF = TP + TN$, then the appropriate classification ratio, also referred to as the efficiency ratio, is equal to: (Cizek and Fitzgerald, 1999, Fraas and Newman, 2003).

$$\text{Hit Ratio} = \frac{EF}{Total} = \frac{TP + TN}{P + P'} = \frac{TP + TN}{Q + Q'} \quad (3.3)$$

3.2.7. Goodness - of-Fit

It serves as a gauge of how well a model fits the data used to build it, and the quality of the fit typically depends on the consistency between the model's predicted and actual data quality.

3.2.8. Coefficient of determination R^2

It is the proportion of the contribution of the independent (explained) variables in the events of changes in the dependent variable (the outcome variable) in the model. The symbol R^2 denotes it, and when its value equals the correct one, it means that when. Furthermore, there is a perfect relationship between the independent variables and the dependent variable, and this indicates that all errors are equal to zero, meaning that: $0 = e = e = \dots = e$, and this means that the true values coincide with the estimated values, a very bad and unacceptable condition, but it is not Logical and rare, its value is zero, which indicates that the values of the independent variables are equal to zero, meaning that: $0 = X = X = \dots = X$, and this means that the independent explanatory variables have no contribution to the model. The coefficient of determination is one of the most widely used measures to measure the strength of the quality of the model. This coefficient measures the percentage of variance that the model explains to the total variance in the dependent variable (dependent) values. Among the characteristics of the coefficient of determination are the following:

- 1- The coefficient of determination only accepts positive values.
- 2- The coefficient of determination has values between 0 and 1 ($1 > R^2 > 0$)
- 3- The coefficient of determination does not always indicate whether a straight-line model is appropriate.

3.2.9. The odds

It is a manner of expressing the likelihood that something will occur as opposed to not occurring, i.e.

$$Odds = \frac{p}{(1 - p)} \quad (3.4)$$

Where:

Odds: is the possibility that something will occur.

P: is the possibility that something will occur, and

$(1 - p)$: is the likelihood that it won't.

3.2.10. Odds ratio (OR)

The probability of an occurrence is the chance that it will occur divided by the possibility that it won't. The odds of one variable (Q) divided by the odds of another variable (2) is the weighting ratio, which is expressed as follows:

$$OR = \frac{Odds1}{Odds2} = \frac{Q_1}{Q_2} \quad (3.5)$$

Where:

OR: The weighting rate

Q1: The initial weighting factor is this.

Q2: The second weighting factor is this.

3.2.11. Concept of logistic regression analysis

Lea (1997) observes that despite the fact that numerous statistical approaches have been created to evaluate data with taxonomic dependent variables, logistic regression analysis offers a number of benefits that make it suited for application in such situations.

The significance of logistic regression analysis lies in the fact that it is a more potent instrument, as it provides a test of the significance of the transactions and also provides the researcher with an estimate of the extent of the influence of the independent variable on the binary response variable. It is superior to the other variable in explaining the emergence of the desired outcome, and logistic regression analysis can incorporate qualitative factors and interaction limitations. The use of logistic regression analysis necessitates a number of assumptions, one of which is that the dependent variable has only one possible value for each individual, i.e., the individual cannot belong to two categories simultaneously. Furthermore, there should be no issue with multiple linear correlation in the data. And the lack of multicollinearity, which indicates that there are no significant correlations between the collection of independent variables. The logistic regression model is governed by the response variable (dependent), which can take only one of two possible values (0 or 1), and there must be no outliers in the alive and deceased independent variables (which is interpreted as either yes or no, or sick and Not sick, or injured and uninfected, and so on). On how to predict a categorical outcome variable, and among the factors that might lead a researcher to use the logistic regression model, we can list the following:

1. Predicting the likelihood that a given event will occur based on a list including one or more interpretative (predictive) factors.
2. Determine the odds ratio that quantifies the significance of a predictive variable in the answer.
3. Predicting the probability of an occurrence in a dangerous scenario, like logistic regression is extensively used in the health and medical industries to predict the potential of infection and identify risk and susceptibility variables associated with illnesses.
4. The order of relative significance of the predictive factors in interpreting the response variable (dependent).
5. Predicting and explaining the values of qualitative variables, since it is regarded the most appropriate statistical approach for research that intend to create predictive models using qualitative dependent variables. Therefore, it is utilized in several spheres of life, including the medical and social sciences. It is also used in marketing to estimate the likelihood that a consumer would purchase a product or refrain from doing so.

3.2.12. A theoretical overview of logistic regression

In a few words, a logistic regression is a multiple regression in which the outcome variable is categorical in pairs and the predictive variables are continuous or categorical, which, in layman's terms, means that we can predict that an individual belongs to either of the two categories based on the information provided. And as a basic illustration, we may investigate the characteristics that indicate a person's gender (whether he is male or female). We could assess your level of sloth, vanity, alcohol intake, and daily flatulence. Using the logistic model, you may discover that all of these characteristics indicate a person's gender, but this approach also enables us to predict whether a certain individual is male or female. Therefore, if we randomly select a person with high scores for laziness, vanity, alcohol use, and burping times per day, the logistic model may indicate that this individual is likely to be male. The logistic regression model has vital uses in medical research, including the formulation of models describing the sorts of variables that might indicate whether a tumor is malignant or benign. The patient database may be utilized to determine the parameters influencing the prediction of the tumor's malignant development. Then, these parameters may be assessed for a new patient and included into a logistic model to estimate the chance of tumor development. If the likelihood of tumor development is sufficiently low, the physician may opt not to undertake a costly and unpleasant operation since it may be unnecessary. And as a result of these uses, does an individual get blood cancer? In this study, the independent variables of sex, age, type of blood cancer, type of treatment blood cancer, history family and smoking, and activity were utilized to predict whether or not these individuals will acquire blood cancer disease. That continues. In the social sciences, we seldom face such life-threatening decisions, and because the logistic model is a highly valuable tool, it is regrettable that the logistic regression model is neglected in many texts. This study aims to reestablish equilibrium by explicating the concepts of the binary logistic model and its implementation processes.

3.3. Principles of Logistics Regression

In the logistic model, instead of predicting the value of the variable Z from the predictive variable G_1 or a group of several predictive variables (G_n), we predict the

probability of Z knowing the value of G_1 or values of (G_n) . And in the case of one predictor variable G_1 (the independent variable), then:

$$P = \frac{e^{\beta_0 + \beta_1 g_1}}{1 + e^{\beta_0 + \beta_1 g_1}} \quad (3.6)$$

Where:

P is a Probability, Z=1 and represents the logistic curve of the relationship G_1 with probability P, and at 1-P the probability Z=0

$$1 - P = 1 - \frac{e^{\beta_0 + \beta_1 g_1}}{1 + e^{\beta_0 + \beta_1 g_1}} \quad (3.7)$$

$$1 - P = \frac{1}{1 + e^{\beta_0 + \beta_1 g_1}} \quad (3.8)$$

The probability distribution of the number of deaths in the sample also assumes a Bernoulli distribution when the dependent variable is binary (death or life), with the probability of success (P) when (Z=1) and the chance of failure (1-P) when (Z=0). According to the Bernoulli distribution, the dependent variable has a curved form resembling the letter S when there is a direct relationship between it and the explanatory variables.

$$P(Z) = \frac{P}{1 - P} = \frac{\frac{e^{\beta_0 + \beta_1 g_1}}{1 + e^{\beta_0 + \beta_1 g_1}}}{\frac{1}{1 + e^{\beta_0 + \beta_1 g_1}}} = e^{\beta_0 + \beta_1 g_1} \quad (3.9)$$

Therefore, the simplest form of the equation of the logistic model can predict the Z probability of occurrence

$$P(Z) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 g_1)}} \quad (3.10)$$

whereas:

$P(Z)$: represents the probability of occurrence of Z

e : base natural logarithm

β_0 : fixed limit

G_1 : represents the first predictive variable

β_1 : coefficient of the predictive variable G_1

Several (independent) predictive variables can be added to Equation No. (3.10) in the following form:

$$P(Z) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 g_1 + \beta_2 g_2 + \dots + \beta_n g_n)}} \quad (3.11)$$

Equation No. (3.10) is referred to as the basic binary logistic model, while equation No. (3.11) is referred to as the multiple binary logistic models, which describes the probability of occurrence Z in terms of Z . (that is, the probability that 0.1 of a case belongs to a certain category). Therefore, the equation yields a probability value between 0 and 1 as its solution. A score near to 0 indicates that Z is unlikely to occur, whereas a value close to 1 indicates that Z will occur with high probability. Each variable is assigned its own coefficient. And while conducting the analysis and needing to estimate these coefficients in order to solve the equation, these parameters are calculated by fitting models to the observed data.

3.3.1. Logistic regression analysis tests

After estimating the logistic model, it is required to assess the significance of the variables included in the model to determine which should remain and which should be removed.

These are the tests:

1. Likelihood Ratio
2. Wald Test
3. Score Test

3.3.2. Likelihood ratio test

This test is known as the logarithm of likelihood statistic (Log Likelihood) that the logistic model predicts the probability of an event occurring for a specific individual (we call that $P(Z_i)$ is the probability of Z occurring for an individual I based on whether or not the event will occur for that individual) (we can call this Z_i is actual score for Person No. 1).

The value of Z for a certain individual is either 0 (the outcome did not occur) or 1 (the outcome did occur), and the expected value $P(Z)$ is a number between 0 (there is no chance of the event occurring) and 1 (the outcome did occur) (the outcome will happen). We can utilize observed and anticipated values to evaluate the model's fit. This measurement is referred to as the Log Likelihood test:

$$\text{Log likelihood} = \sum_{i=1}^n \{Z_i \ln [P(Z_i)] + (1 - Z_i) \ln [1 - P(Z_i)]\} \quad (3.12)$$

The Likelihood logarithm is derived from the sum of the probabilities associated with expected and actual outcomes. Consequently, the probability log statistic is comparable to the sum of squares for residuals in multiple regression in that it indicates the amount of unjustified information remaining after the model has been fitted to the data. Consequently, the big values of the probability logarithm statistic suggest a poor fit of the statistical models, since the larger the value of the probability logarithm, the greater the number of invalid observations. It is possible to determine the logarithm of each model's probability and compare them by examining the difference between their log probability values. One use is comparing the state of the logistic model to a particular reference state. The most common reference case is the model including simply the constant term. In the logistic model, what would be the best prediction for predicting the outcome? The result is composed of 0 and 1 numbers, rendering the mean value meaningless. In logistic regression, the reference model is the one that predicts the outcome the majority of the time. And this is the only logistics model with set limits. Adding one or more predictive variables to the model is possible if the model improvement calculation includes the following:

$$g^2 = -2LL(baseline) - 2LL(new) \quad (3.13)$$

That is, simply subtract the LL value of the new model from the LL value of the reference model (the model containing only the fixed term) (the model that includes the predictive variable that was added). This test is implemented by estimating two data models, a first model that includes all the variables, and a second model that includes a subset of those variables, then comparing which of the two models has greater significance than the other, using the logarithm of the function of the greatest possibility of the two models, and concluding that the model with the smallest value is superior, but that this difference is statistically significant.

3.3.3. Wald test

This test is known as the Wald statistic, which is a test statistic with a known probability distribution (chi-square distribution) and is used to determine the discrete contribution of each of the predictive variables to the logistic regression model and to test whether the β coefficients for the predictive variables in the model have different values. Countless times the value 0. If the value of the parameter is different from 0, we may claim that the predictive variable significantly contributes to predicting the value of the result Z, and the greater the Wald value, the greater the proportional contribution of the variable. In other words, the Wald test examines the null hypothesis that the logistic regression coefficient associated with the explanatory variable (G) is equal to zero (Salih, 2010).

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

$$W^2 = \left(\frac{\beta}{SE_{\beta}} \right)^2 \quad (3.14)$$

whereas:

W: Wald statistic

β : The parameter value of the predictive variable

SE_{β} : the associated standard error.

The Wald statistic W^2 division of β^2 (Poston, 2004), but if the Wald. statistic value is calculated W Instead of W^2 ; the equation would be:

$$W^2 = \left(\frac{\beta}{SE_{\beta}} \right) \quad (3.15)$$

3.3.4. Score test

Rao's efficient statistic is used to quantify and test Wald's statistic, although its calculation is simpler. This test is also used to determine whether or not the parameter of the logistic regression model for the explanatory variable is equal to 0. It evaluates the relevance of each variable individually and serves as a criterion for incorporating the variable in the logistic model (Saleh, 2014).

3.4. The Value of Exp(B)

The value of (Exp(B)) is one of the crucial values in the interpretation of the logistic model; it indicates the change in probability caused by a change in the unit amount of the predictive variable. Consequently, it is comparable to the coefficient β in the logistic regression, but simpler to comprehend (because it does not require a logarithmic transformation). When the predictive variable is categorical, Explanation of Exp is simpler (B). Imagine that we are attempting to forecast whether a man will acquire heart disease as a result of smoking or not smoking. The chances (Odds) of an occurrence are the probability of the occurrence divided by the probability of the event. The calculation for non-occurrence is as follows:

$$Odds = \frac{P(event)}{p(noevent)} \quad (3.16)$$

$$P(event|Z) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 g_1)}}$$

$$P(noevent|Z) = 1 - P(event|Z)$$

This proportionate change in odds is denoted by $\text{Exp}(B)$, which can be interpreted as follows in terms of the change in odds:

If the value is larger than 1, it suggests that the probability of the result increases as the predictive variable increases. If the number is less than 1, it implies that the probability of the occurrence reduces as the predictive variable increases.

3.5. Goodness- of-Fit

The measure of the model's fit quality is the proximity of the observed values to the estimate line. The fit indicates whether the statistical model is appropriate for the research sample's data, and the quality of fit assesses the convergence of observed values and expected values. Among the quality of fit tests utilized are:

3.5.1. Chi-square test χ^2

It was proposed by Pearson, with the symbol χ^2 , and is defined by the following formula.

$$\chi^2 = \sum_{i=1}^n \frac{(G_i - Z_i)^2}{Z_i} \quad (3.17)$$

whereas:

Z_i : Viewed values

G_i : Expected values

n: number of views

And the χ^2 statistic distribution resembles the chi-square distribution χ^2 with a degree of freedom equal to the number of observations minus the number of model parameters (p). If the value of the test statistic is greater than the given threshold of significance, i.e., if it is not significant, then the observed values will equal the predicted values, indicating that the model fits the data well, i.e., that the model is consistent with the data.

3.5.2. The hosmer and lemeshow test R^2_L

The hosmer and lemeshow quality of fit statistic examines the hypothesis that the observed data differ significantly from the model's predicted values. Therefore, we desire an insignificant result for this test, as this would imply that the model does not deviate significantly from the observed data. It represents the proportion of the model's 2LL value to the initial 2LL value. In other words, it is the proportion of the model's justification to the amount to which it was initially necessary. R^2_L is determined by dividing the value of chi-square χ^2 of the model after entering the variables into it (based on the logarithm of probability) by the original 2LL (the logarithm of the probability of the model before to inserting the predictive variables into it).

$$R^2_L = \frac{\chi^2}{-2LL(\text{Original})} \quad (3.18)$$

Whereas:

χ^2 : reflects the value of the difference between the logarithm of probability when the model contains only the fixed term and the logarithm of probability when variables have been added to the model.

2LL (Original) -: indicates the logarithmic value of the Likelihood (weighting function)) when the model contains only the fixed term R2L: Represents the relative decrease in the absolute value of the log- Likelihood, a measure of how much the mismatch has improved due to the inclusion of predictive variables. Its value might range between 0 (indicating that the predictive variables are ineffective in predicting the outcome variable) and 1 (indicating that the predictive variables are useful in predicting the outcome variable) (which indicates that the model ideally predicts the outcome). According to Saleh (2014), this test is used to measure the degree of fit and determine whether the model adequately represents the data. The Chi-square test of good fit χ^2 is used to analyze the differences between the actual and anticipated values, using a degree of freedom equal to the number of observations minus the number of parameters in the logistic regression model. This test requires a high sample size because it is ineffective with small sample sizes. If the test statistic is greater than the significance threshold given

by the researchers, it implies that the observed values match the expected values, hence indicating the model's accuracy.

3.6. Coefficient of Determination

R^2 is the coefficient of determination?

The (R) statistic can be used to examine the robustness of the logistic model, i.e., the proportional contribution of the influencing factors included in the estimated model to the dependent outcome variable (injury) as measured by the following coefficients:

3.6.1. Cox and snell coefficient R^2

A version of the coefficient of determination R^2 for the logistic model that is dependent on the logarithm of the probability of a model [LL(new)], the logarithm of the probability of the original reference model [LL(baseline)], sample size n, and the standard deviation of the sample.

It is computed according to the following formula

$$R_{CS}^2 = 1 - \left(e^{\frac{1}{2}(\text{new}) - (\text{baseline})} \right)^{\frac{2}{n}} \quad (3.19)$$

This parameter is infamous for failing to achieve its maximum value of 1. Therefore, its modified form is referred to as Nagelkerke's R squared.

3.6.2. Nagelkerke coefficient R^2_N

A replica of the logistic model's coefficient of determination R^2 , which is a modified version of the Cox & Snell coefficient that solves the problem of this statistic's failure to achieve its maximum value. So, I suggest Nagelkerke edited the following in 1991:

$$\begin{aligned}
 R_N^2 &= \frac{R_{CS}^2}{1 - e^{\left[\frac{2LL(\text{baseline})}{n}\right]}} \\
 R_N^2 &= \frac{R_{CS}^2}{1 - \left(e^{\frac{-\text{baseline}}{2}}\right)^{\frac{2}{n}}}
 \end{aligned} \tag{3.20}$$

3.6.3. Concept of discriminator analysis

Discriminator analysis is a statistical approach for classifying sample objects into two or more groups using multivariate data. The discriminant model is a linear structure of the independent variables, and it is the model that executes the discrimination process. And the classification procedure follows the formation of the discriminatory model since it relies on it to classify the new vocabulary for one of the studied groups with the lowest possible classification error. Discriminant analysis is an essential multivariate analytic technique. The input variables are evaluated in light of the application of these approaches. In the model in a coherent manner, taking into account the interrelationships between these variables, it also seeks to form a statistical model that depicts the interrelationship between the various variables, and its significance is primarily due to its effectiveness in distinguishing between observations by using many variables, by finding linear combinations. Variables differentiation is the combination of a set of variables; the use of discriminatory analysis requires the availability of a number of assumptions, such as that the data of the explanatory variables have a multivariate normal distribution, that the averages are different in each group of aggregates, and that the variances are the same for all aggregates.

Among the factors that may lead a researcher to select a discriminating analysis are the following:

1. Distinguishing between two or more societies that overlap based on a number of independent variables that express distinctive qualities in these societies.
2. Classification of any single item or observation of a society based on the values of its independent variables, regardless of which society it belongs to.

3. Determining the relative significance of the independent variables in the discrimination model based on their capacity to differentiate between the communities under study

3.7. A Theoretical Overview of Discriminant Varieties

By transforming the dependent variables into factors or base dimensions, the problem of having many values for evaluating statistical significance can be reduced. Multiple regression is based on fitting a linear model to a set of data to predict an outcome variable (or dependent variable). Each predictive variable (or independent variable) contributes uniquely to the linear model. The only difference is that we are interested in the inverse problem (i.e., predicting the independent variable from a set of dependent variables). Variables are the linear structures of dependent variables, which are sometimes known as factors or latent variables. These unique functions will be investigated using a mathematical approach that increases the maximum, so that the first discriminant function (S_1) is a linear function of the dependent variables that maximizes the differences across groups. As a result, the ratio of systematic variance to unsystematic variance (justified to unsystematic variance) SSM/SSR will be greatest for the first variable, while the ratio for subsequent variables SSR will be less. Remember that this ratio corresponds to the F-ratio in a univariate analysis of variance, and when we examine the characteristic function, we obtain the maximum possible value of F to the first distinct subordinate. This variable can be written according to the linear regression equation because it is a linear combination of dependent variables:

First. This variable can be written according to the linear regression equation (because it is a linear combination of the dependent variables):

$$\begin{aligned}
 Z &= c_0 + c_1 G_1 + c_2 G_2 + \dots + c_p G_p \\
 S_1 &= c_0 + c_1 G_1 + c_2 G_2 + \dots + c_p G_p \\
 S_1 &= c_0 + c_1 G_1 + c_2 G_2
 \end{aligned}
 \tag{3.21}$$

Equation (3.21) shows the multiple regression equation for the sake of two predictive variables and extends this equation to show that a similar form of this equation can describe the distinct functions.

3.8. The Discrimination Model for Two Groups

The discrimination model is a model that can be formulated based on the indicators of the sample whose items were randomly selected and placed in two different groups. By this model, we can test the item and determine its belonging to any group. Assuming that the two groups, that should be compared, have the same covariance and covariance matrix Σ and their average vectors μ_1 and μ_2 and the observations for the two groups $(G_{11}, G_{12}, \dots, G_{1n_1})$ and $(G_{21}, G_{22}, \dots, G_{2n_2})$ respectively and the vector Z_{ij} consists of the observations P of the dependent variables.

The discrimination model is a mixture of any combinations of these variables, and the formulas of the discrimination model for the two groups are as follows:

$$Z_1 = c_1 G_{1i1} + c_2 G_{1i2} + \dots + c_p G_{1ip} \quad i = 1, 2, \dots, n_1 \quad (3.22)$$

$$Z_1 = c_1 G_{2i1} + c_2 G_{2i2} + \dots + c_p G_{2ip} \quad i = 1, 2, \dots, n_2 \quad (3.23)$$

n_1 and n_2 observations of the two groups

Therefore, the values of a must be found that maximize the standard difference between the averages of the two groups $(\bar{Z}_1 - \bar{Z}_2)$ that can be negative, so it must be squared to make the difference ratio as large as possible. If we symbolize the ratio of the differences between the two groups to the differences within the two groups with the symbol Q , since:

$$Q = \frac{[\bar{Z}_1 - \bar{Z}_2]^2}{S_Z^2} = \frac{[\bar{Z}_1 - \bar{Z}_2]^2}{\sum_{i=1}^2 \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_1)^2} \quad (3.24)$$

If we have the variables $G_1, G_2, \dots, G_p$, the discrimination model takes the following form:

$$Z = c_0 + c_1 G_1 + c_2 G_2 + \dots + c_p G_p \quad (3.25)$$

Where Z represents a combination of explanatory variables called the discrimination model, where the coefficients ($c_0, c_1, c_2, \dots, c_p$) are estimated to make the discrimination model give the best discrimination between the two groups by making the square of the difference between the averages of the two groups to the variance (the sum of squares within the groups) Q as large as possible.

$$Q = \frac{\text{Between groups}}{\text{Within groups}} = \frac{[\bar{Z}_1 - \bar{Z}_2]^2}{\sum_{i=1}^2 \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_i)^2} \quad (3.26)$$

Where:

$$\bar{Z}_1 = \hat{c}_1 G_{11} + \hat{c}_2 G_{12} + \dots + \hat{c}_{1p} G_{1p} \quad (3.27)$$

$$\bar{Z}_2 = \hat{c}_1 G_{21} + \hat{c}_2 G_{22} + \dots + \hat{c}_{2p} G_{2p} \quad (3.28)$$

In the case of two groups, we have one discrimination model, and in the case of three groups, we have two discrimination models.

After extracting the coefficients ($\hat{c}$), the observation is classified into one of the two groups depending on the midpoint of the two groups (L), which makes the possibility of classification as low as possible.

$$L = \frac{\hat{Z}_1 + \hat{Z}_2}{2} \quad (3.29)$$

The observation is classified into the first group if it is $\hat{Z} > L$

The observation is classified into the second group if it is $\hat{Z} < L$

The observation is randomly classified into the first or second group if it is $\hat{Z} = L$

As:

$$\hat{Z} = \hat{c}_1 G_1 + \hat{c}_2 G_2 + \dots + \hat{c}_p G_p \quad (3.30)$$

3.9. Discriminate Analysis Tests

When it is desired to distinguish between the two groups, we can test the hypothesis that states equalization of averages the two groups.

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

And to test the hypothesis of the equality of the covariance matrix and the covariance of the two groups:

$$H_0: \Sigma_1 = \Sigma_2$$

$$H_1: \Sigma_1 \neq \Sigma_2$$

After estimating the discrimination model, its significance must be tested, and the variables included in the model, which must remain in the model, should be tested and which of them are not significant, must be deleted. These tests are:

1. Wilkes's Lambda Test (Λ).
2. Test F.
3. Hotelling-Lawley Test (T^2)
4. Pillai-Bartlett (V).

3.9.1. Wilkes's lambda test (Λ)

It is the product of the unjustified variance on each of the distinct dependent

variables, so it represents the ratio of the error variance to the total variance $\frac{SS_R}{SS_T}$

for each of the variables, and it takes the following formula:

$$\Lambda = \prod_{i=1}^s \frac{1}{1 + \lambda_i} \quad (3.31)$$

Where λ represents the special values (Eigenvectors) for each of the distinct variables, and S is the number of variables. The large special values, which represent a large effect in themselves, give a small value to Wilkes's Lambda, and therefore we obtain statistical significance. When the value of Wilkes's Lambda is small, the value of (Λ) ranges between zero and one. If it is close to or equal to one, this indicates that the averages of the groups are equal and there is no distinction between groups, and this means that the computed discrimination model is failing. But if its value is close to zero, this indicates the strength of discrimination. This test is used to select the constituent variables included in the discrimination equation model by selecting the variables that have the lowest value of Wilkes's Lambda and the highest value (F), and to measure the degree of divergence between the two groups, and the extent to which the discrimination model is able to distinguish.

3.9.2. F-test F

In view of the difficulty of obtaining the tabular value because there are no tables specified to the Wilkes criterion, another alternative criterion was developed by Roy, and this criterion is a measure of F and its mathematical formula is

$$F = \frac{1-\Lambda^{\frac{1}{S}}}{\Lambda^{\frac{1}{S}}} * \frac{ms-2\lambda}{p(k-1)} \quad (3.32)$$

With the degrees of freedom as:

$$df_1 = p(k-1), df_2 = ms - 2\lambda$$

$$m = N - \frac{1}{2}(p+k)$$

$$s = \left(\frac{p^2(1-k)^2 - 4}{(1-k)^2 + p^2 - 5} \right)^{\frac{1}{2}}$$

$$\lambda = \frac{p(k-1) - 2}{4}$$

and

K: number of the groups.

P: number of the variables.

Λ : test value of Wilkes's Lambda and is calculated in the (3.31) formula.

The test F is used to examine the statistical significance of the ability of the discrimination model to separate between groups. If the calculated value of F is greater than the tabulated value of F under a certain level of significance and the degree of freedom (df2, df1); this means that the discrimination model has the ability to distinguish between the groups.

3.10. Standardized and Unstandardized Discriminant Coefficients

Standardized coefficients are represented by the ($\hat{c}_n$) which is shown in the following equation:

$$\hat{Z} = \hat{c}_1 G_1 + \hat{c}_2 G_2 + \dots + \hat{c}_n G_n \quad (3.33)$$

$\hat{Z}$: is the standardized discriminated value.

G_n : is the standardized discriminated variable.

$\hat{c}_n$: is the standardized discriminated factor.

n: is number of the standardized discriminated variables in the discriminated equation and is equal to (number of the influential variables-1).

The non-standard discriminant coefficients are used in the formation of the discriminant model instead of the standard discriminant coefficients. This is because the discriminant variables of the groups are shown in real values and ratios and not in standard values. It should be noted that the non-standard discriminant coefficients do not give the relative importance of the discriminant variables because they are derived from the raw data, namely, the real values of the discriminant variables. The non-standard coefficients are represented by the values of (d) shown in the following equation:

$$Z = d_0 + d_1 G_1 + d_2 G_2 + \dots + d_n G_n \quad (3.34)$$

Where:

z : is sign of the non-standard discriminant equation.

d_0 : is fixed.

G_n : are the non-standard discriminant variables.

d_n : are the non-standard discriminant coefficients

3.11. Goodness-of-Fit

The measure of the fit quality of the model means how close the observed values are to the estimate line. Goodness-of-fit means whether the statistical model is suitable for the data of the study sample, and the quality of goodness-of-fit measures the convergence between the observed values and the expected values of the model. Among the tests used in the quality of fit are:

3.11.1. Chi-square test pearson X^2

It is proposed by Pearson and is signalized by the symbol X^2 and defined by the following formula:

$$X^2 = \sum_{i=1}^n \frac{(G_i - Z_i)^2}{Z_i} \quad (3.35)$$

Where:

G_i represents the viewed repetition.

Z_i represents the expected repetition.

n is number of the views in the table.

And the distribution of the X^2 statistics approximates the Chi-square X^2 distribution with degree of freedom = number of cells - number of features in the model. The Chi-Square test detects whether there is a significant relationship between two categorical variables (descriptive).

3.11.2. Wilkes's lambda test (Λ)

It is a ratio that justified by the model to the amount whose justification was initially required, i.e., the ratio of the error variance to the total variance $\frac{SS_R}{SS_T}$ for each of the variables, and Λ is calculated from the following equation:

$$\Lambda = \prod_{i=1}^s \frac{1}{1 + \lambda_i} \quad (3.36)$$

Λ is a measure of the extent to which mismatch has improved as a result of including predictive variables. The value of (Λ) ranges from zero to one. Its value can vary between 0 (meaning that the predictive variables are useless in predicting the outcome variable) and 1 (which indicates that the model perfectly predicts the outcome).

4. APPLICATION AND MAIN RESULTS

4.1. Subject and Purpose of the Application

In this section is deals with the collected data and its analysis, we shall apply two methods of classification logistic regression analysis and discriminant analysis. We evaluated the classification accuracy of the models using the confusion matrix, then we fitted the final model in order to determine the most important independent variables affecting blood cancer diseases patients. All the implementations of the study on real data application are carried out using SPSS package version 25.

4.2. Data Used in the Application

Study two different data were conducted to achieve the goal of the study. First, it used secondary data that achieved from Nanakali in Erbil Northern Iraq in the period from 1st January 2017 to 31st December 2018 with a follow up period until 1st April 2021. Data included the number of Blood Cancer disease (BCD), age of patient, sex of patient and type blood cancer of patient. Second, the application concerning the comparison LRA and DA was performed using the (BCD). The data have been collected by questionnaire that has included the association between the type of presence blood cancer disease (BCD) were labeled, such that 1 for presence blood cancer and 0 for non- presence (BCD). In this section, table (4.1) shows a description of the study variables, all the variables and coding methods were defined Of variables.

Table 4.1. Description of the study variables

Num	Variable Name	Definition of Variable	Coding of Variable	Types of Variables
1	Sex	Sex of patients	1 male 2 females	Nominal
2	Age	Age of patients.	Class 1-9 10-19 20-29 90-99	Numerical
3	Years	Years of patients between (2017-2021)	NA	Numerical
4	Type of Blood Cancer	Consists of (7) types of blood cancer	X ₁ (AML) X ₂ (Acute leukemia) X ₃ (CML) X ₄ (CLL) X ₅ (CPL) X ₆ (ALL) X ₇ (AML2)	Numerical

4.3. Descriptive Statistics

4.3.1. Age of patients

The patient's age was grouped into ten categories (Table 4.2). Among these categories, patients within the age group of 1-9 are responsible for the large number of accidents (24%). Patients with the age group of 90-99 are the smallest in all the four measurements (1.38%).

Table 4.2. Age of (BCD)

Age Group	Frequency	Percent (%)
1-9	154	24
10-19	110	16.54
20-29	62	9.5
30-39	43	6.59
40-49	62	9.5
50-59	70	10.73
60-69	77	11.8
70-79	46	7.05
80-89	19	2.91
90-99	9	1.38
Total	652	100.0

4.3.2. Gender of (BCD)

As shown in Table 4.3, out of the 652 number of patients, males represented 55.2% of all the patients in the data, while females constituted 44.8%. It shows that males more involved than females to (BCD).

Table 4.3. Gender of (BCD)

Gender	Frequency	Percent (%)
Male	360	55.2
Female	292	44.8
Total	652	100.0

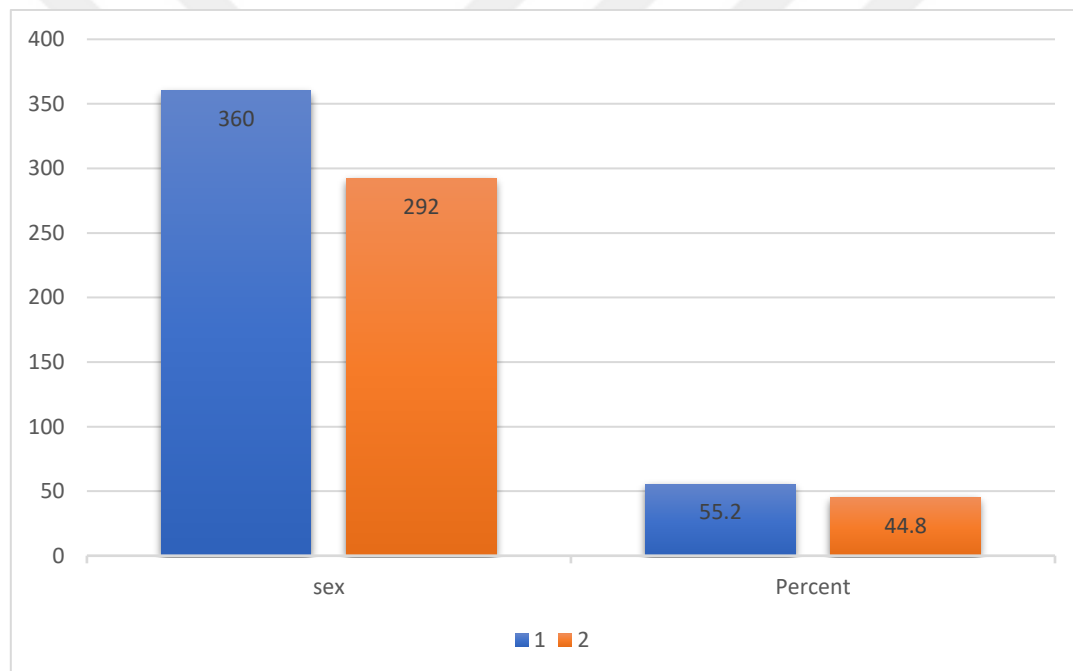


Figure 4.1. Gender of patients.

4.3.3. Blood cancer disease of years (2017-2021)

As it is shown in Table 4.4, the data of the study records that most of the (BCD) between years (2017-2021), the study concluded that 32 patients (4.9 %) have 2017 years, 157 patients (24.1 %) have 2018 years, 148 patients (22.7 %) have 2019 years, 149 patients (22.9 %) have 2020 years and 166 patients (25.5 %) have 2021 years.

Table 4.4. Years of (BCD)

Years	Frequency	Percent (%)
2017	32	4.9
2018	157	24.1
2019	148	22.7
2020	149	22.9
2021	166	25.5
Total	652	100

Furthermore, out of 32 (4.9 %) of participants who have 2017 years comprises of 13 male and 19 female, in addition, out 157 (24.1 %) of them who have 2018 years consist of 99 males and 58 females, out 148 (24.1 %) of patients who have 2019 years consist of 72 male and 76 female, out 149 (22.7 %) of patients who have years 2020 consist 83 male and 66 female and out 166 (25.5 %) of patients who have 2021 years consist 93 male and 73 female (Figure 4.2).

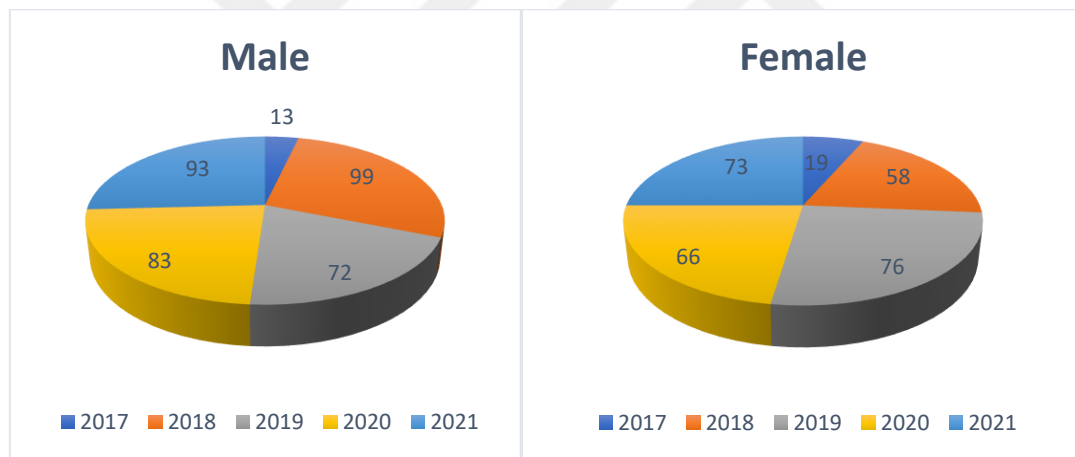


Figure 4.2. (BCD) of years (2017-2021) by gender.

4.3.4. Type of (BCD)

Table 4.5 provides the type (BCD) consists of (7) types of blood cancer, the study concluded that 147 patients (22.5 %) have X1 (AML), 129 patients (19.8 %) have X2 (Acute leukemia), 61 patients (9.4 %) have X3 (CML), 88 patients (13.5 %) have X4 (CLL), 20 patients (3.1 %) have X5 (CPL), 202 patients (31 %) have X6 (ALL) and 5 patients (0.8 %) have X7 (AML2).

Table 4.5. Type of blood cancer disease

Type of blood cancer	Frequency	Percent (%)
X ₁	147	22.5
X ₂	129	19.8
X ₃	61	9.4
X ₄	88	13.5
X ₅	20	3.1
X ₆	202	31
X ₇	5	0.8
Total	652	100

The following Table 4.6, shows the results for summary statistics of variable and Pearson's chi-square test for variables:

H_0 : There is no difference in classification between the disease who made (gender, age, type of cancer) and does not make (gender, age, type of cancer).

H_1 : There is difference in classification between the disease who made (gender, age, type of cancer) and does not make (gender, age, type of cancer).

Table 4.6. Summary statistics

Variable	Number	Mean	Std. Deviation	x2	P-value
Gender	652	1.45	0.498	9.298	0.054***
Age	652	4.40	2.488		
1-9	652	14.45	2.515		
10-19	652	24.08	2.853		
20-29	652	34.56	2.848		
30-39	652	44.19	2.757	98.001	0.024***
40-49	652	53.49	2.691		
50-59	652	64.79	2.769		
60-69	652	73.02	2.856		
70-79	652	82.32	2.428		
80-89	652	97	3.969		
90-99	652	29.4	2.49		
Type of Cancer	652	25.8	2.234		
X1	652	12.2	2.113		
X2	652	17.6	2.33	72.944	0.000***
X3	652	4	2.204		
X4	652	40.4	2.698		
X5	652	1	0.474		
X6	652				
X7	652				

Table 4.6 shows that the p-value $0.054 \leq 0.05$ which indicates that there is a significant difference between the two groups (male and female), and the p-value $0.024 \leq 0.05$ which indicates that there is a significant difference between the classes of age, and the p-value $0.000 \leq 0.05$ which indicates that there is a significant difference between type of blood cancer disease (BCD).

4.4. Selection of the Study Samples

This data contains 350 patients, including 17 variables, 16 of which are independent variables and a dependent variable (presence 1 and absence 0) of blood cancer disease depending. The studied samples consist of 59 absence blood cancer disease and 291 presence blood cancer disease. The age of patients ranged from 2 to 92 years; with a mean $\pm$ SD of 38.79 ± 24.16 years respectively, also study consist of 165 (56%) males and 135 (44%) females .

Table (4.7) shows a description of the study variables which is composed of a dependent variable and 16 independent variables used in this study, all the variables and coding methods of each variable were defined by the medical accreditation of (BCD) used in this thesis.

Table 4.7. Description of variables

No.	Variable Name	Definition of Variable	Coding of Variable	Types of Variables
1.	Class	Classification of patients.	1 presence of (BCD). 0 absence of (BCD).	nominal
2.	Sex (X ₁)	Sex of patients	1 male. 2 females.	nominal
3.	Age (X ₂)	Age of patients.	NA	numerical
4.	Smoking (X ₃)	Smoking for patients	1 smoker 2 non smoker	nominal
5.	Family History (X ₄)	Cancer is common-many people have someone is their family who	1 family history 0 non	numerical
6.	AML (X ₅)	Acute myeloid leukemia is a type of blood cancer, it starts in your bone marrow.	1 AML 0 non-AML	numerical
7.	CML (X ₆)	Chronic myeloid leukemia is a type of blood cancer that affects the blood and bone marrow.	1 CML 0 non-CML	numerical
8.	Acute leukemia (X ₇)	is a family of serious medical conditions relating to an original diagnosis of leukemia	1 Acute leukemia 0 non-Acute leukemia	numerical
9.	CLL (X ₈)	Chronic lymphocytic leukemia is a type of blood cancer that affects white blood cells	1 CLL 0 non-CLL	numerical
10.	APL (X ₉)	Acute promyelocytic leukemia is a subtype of acute myeloid leukemia	1 APL 2 non-APL	numerical
11.	ALL (X ₁₀)	Acute lymphoblastic leukemia is a type of cancer that affects white blood cells	1 ALL 0 non-ALL	numerical
12.	AML (X ₁₁)	Acute myeloblastic leukemia is a type of cancer of the myeloid line of blood cells	1 AML 0 non-AML	numerical
13.	Surgery (X ₁₂)	Is a type of cancer treatment	1 Surgery 0 non-Surgery	numerical
14.	Radio (X ₁₃)	Is a type of cancer treatment	1 Radio 0 non-Radio	numerical
15.	Chemo (X ₁₄)	Is a type of cancer treatment	1 Chemo 0 non-Chemo	numerical
16.	Hormone (X ₁₅)	Is a type of cancer treatment	1 Hormone 0 non-Hormone	numerical
17.	Immune (X ₁₆)	Is a type of cancer treatment	1 Immune 0 non-Immune	numerical

Notes: - Most of the definitions of the independent variables used in the study were taken from the website of the american international laboratory of medicine. <https://pubchem.ncbi.nlm.nih.gov>.

Table 4.8. The frequency distribution of the blood cancer disease

Class Name	Class Size	Class Distribution
presence of (BCD)	291	83.1%
absence of (BCD)	59	16.9%

The above table shows that the data set has 350 cases, actually comprised of divided in two groups, such that the first group with $n_1=291$ which displays the 83.1% of observation, and the second group with $n_2=59$ which displays the 16.9% of observation. The next part of the summary result shows table 4.9 gives the results Summary of statistical analysis which displays the mean and the standard deviation of the predictors

Table 4.9. Summary results of blood cancer data

Ser.	Predictor Variable	Mean	Standard Deviation
1	X_1	1.50	0.501
2	X_2	45.67	18.818
3	X_3	0.34	0.473
4	X_4	0.28	0.448
5	X_5	0.26	0.779
6	X_6	0.12	0.720
7	X_7	0.33	1.294
8	X_8	0.39	1.220
9	X_9	0.13	0.968
10	X_{10}	0.24	1.101
11	X_{11}	0.19	1.263
12	X_{12}	0.35	0.867
13	X_{13}	1.32	2.313
14	X_{14}	1.55	2.602
15	X_{15}	1.03	2.439
16	X_{16}	1.17	2.802

4.4.1. Application using LRA

Logistic regression analysis was done on the blood cancer disease, after dividing the data into two groups (presence and absence) we begun bulding the model based on the training dataset which includes 350 cases. We must now examine the model to understand how well it fits the data and generalizes to other observations. Next, we see the deviance residuals, which are a measure of model fit. This part of the output shows

the distribution of the deviance residuals for individual cases used in the model, by model fitting information shows table 4.10.

Table 4.10. Model fitting information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2Log Likelihood	Chi-Square	D.F.	Sig
Intercept Only Final	316.143	238.070	16	0.00
Final	78.073			

Table 4.10. summarizes, we can see that coefficients are significant ($p\text{-value} \leq 0.05$). The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the independents variable urea and creatinine are statistically significant. that -2Log Likelihood value for only the intercept final was 316.143 and final models was 78.073 for those subjects that could be analyzed using OLR without problem . The value of the Chi-square was 238.070.

The correlation measurement provided by OLR analysis that is shown in table 4.11, we found that values Cox & Snell R^2 (0.493) also Nagelkerke R^2 (0.827) tell us the proportion of variance in the dependent variable that is explained by the predictors. They are acceptable values because they approach number one. Therefore, our model fits depending on the difference between the model and the observed data.

Table 4.11. Pseudo R^2

Cox & Snell R Square	Nagelkerke R Square
0.493	0.827

The Likelihood Ratio test of model regression coefficients is used to investigating that explanatory variables included for evaluting and identifying the relationship between an independent variable and the dependent variable. Practically the objective of table 4.12 shows the results of the test, nine variable most significant predictors were selected properly (X_{14} , X_{11} , X_{10} , X_9 , X_8 , X_7 , X_6 , X_5 and X_1), if the step was to add variable, the inclusion sense if the significance of the change is less than 0.05, if the step was to remove a variable, the exclusion justifies sense if the significance of cgange is greter than 0.10.

Table 4.12. Result of LRT

Effect	Model Fitting Criteria Variable -2Lo Likelihood of Reduced Model	Mean Chi-Square	LRT D.F.	Sig.
Intercept	81.791	81.299	1	0.00
X ₁₄	84.538	6.465	1	0.01
X ₁₁	88.213	10.140	1	0.01
X ₁₀	144.881	66.808	1	0.00
X ₉	99.973	21.900	1	0.00
X ₈	173.148	95.075	1	0.00
X ₇	174.420	96.347	1	0.00
X ₆	121.228	43.155	1	0.00
X ₅	182.041	103.968	1	0.00
X ₁	82.608	4.535	1	0.03

4.4.2. Estimating the LRA coefficients

LRA model is a well-known statistical technique for determining the relationship between the classification of a patient and many explanatory variables. LRA model provides an estimate of the predictors' effect for other clarification variables. In the table 4.13 summarizes the calculated results of Wald test of the independent variables in the model compares them with value $\alpha = 0.05$. The nine most important predictors were selected respectively (X₁₄, X₁₁, X₁₀, X₉, X₈, X₇, X₆, X₅ and X₁) variables, where the significant, therefore the variables which will reject the null hypothesis and accept the alternative hypothesis in logistic regression method. The idea is to test the hypothesis that the coefficient of an independent variable in the model is not significantly different from zero. If the test fails to reject the null hypothesis, this suggests that removing the variable from the model.

Table 4.13. Results of fitting the LRA model to (BCD)

Variable	β	Std. Error	Wald	Sig.
Intercept	0.053	0.012	4.326	0.000
X ₁₄	-1.889	0.762	6.143	0.013
X ₁₁	0.033	1.003	0.009	0.000
X ₁₀	-6.582	1.349	23.811	0.006
X ₉	5.059	1.377	13.491	0.000
X ₈	6.670	1.121	35.425	0.000
X ₇	-7.391	1.395	28.060	0.000
X ₆	-5.119	1.052	23.665	0.000
X ₅	-7.425	1.363	29.699	0.000
X ₁	1.552	0.788	3.878	0.04

That is, the estimated logistic regression model is:

$$\ln \left(\frac{\pi}{1-\pi} \right) = 0.053 + 1.552(X_1) - 7.425(X_5) - 5.119(X_6) - 7.391(X_7) + 6.670(X_8) + 5.059(X_9) - 6.582(X_{10}) + 0.33(X_{11}) - 1.889(X_{14})$$

4.4.3. Testing the discriminatory power of the LRA model

To analyze the effectively of the method in the classification, we can calculate the classification results are compared to real data. Table 4.14 shows the classification table and evaluation criteria for the logistic regression model of the training dataset.

Table 4.14. The final classification results using LRA model

Cases	Predicted Group Membership		Total
	absence of (BCD)	presence of (BCD)	
absence of (BCD)	51	8	59
presence of (BCD)	3	288	291
Model Accuracy	96.8%		Total= 350
Model Sensitivity	86.4%		
Model Specificity	98.9%		

From table 4.14. which analysis of the classification table, shows that the logistic regression method successfully classified, as can be seen, the model accuracy of (96.8%), and sensitivity and specificity are both above (86.4% and 98.9%). That is, the model can

accurately predict based on the independent predictors entered by (86.4%). The model specificity was (98.9%), that is to say, it can accurately predict patients with out BCD based on the independent variable entered by (98.9%).

4.4.4. Application using DA

Before defining and creating a DA model for two groups, it is necessary to perform some tests to make sure that multicollinearity (linear interference) is not a problem in the data. The presence of multicollinearity demonstrates that the DA model has a problem, and the absence of multicollinearity indicates that the set of independent variables do not have any strong correlations. Due of the DA model's dependence on several linear correlations. The outcome of solving the multiple linear correlation issue for the initial data is shown in Table 4.15 above (linear interference). The Variance Inflation Factor (VIF) test is utilized in this case, and according to Berenson et al., a result of $VIF > 10$ indicates multicollinearity is present and the assumption is broken (2002) The VIF values for each predictor are presented in Table 4.15.

Table 4.15. VIF values of predictor variables

Ser.	Predictor variables	VIF Values
1	X_1	1.217
2	X_2	1.074
3	X_3	1.218
4	X_4	1.045
5	X_5	1.028
6	X_6	1.241
7	X_7	1.071
8	X_8	1.070
9	X_9	1.031
10	X_{10}	1.056
11	X_{11}	1.071
12	X_{12}	1.241
13	X_{13}	1.256
14	X_{14}	1.177
15	X_{15}	1.124
16	X_{16}	1.194

Note that all the VIF values in table 4.15 are relatively small, which are varying from a high $X_{13}=1.256$ for the number of Radio, that is a kind of blood cancer treatment to a low $X_5=1.028$ for Acute Myeloid Leukemia, which is a kind of blood cancer disease (BCD). Thus, on the basis of the VIF criteria, which are values of the variance inflation

factor and they are less than 10, there is no directory of multicollinearity within the set of the variable predictor or predictors, i.e, anyone is able to continue with the analysis. Before creating the discriminant model, Wilks Lambda statistics with F-test are carried out here for examining the equality of the group and that is to specify the alterable which significantly contribute to differentiate the groups and to measure each predictor's potential, also. It is equal to the proportion of the total variance in the discriminant scores not clarified by differences among the groups. Smaller values of Wilks Lambda mark greater discriminatory ability of the function, Rencher (2002), As we talked about in section (3.4), that we possess 16 predictors, the stepwise method can be useful to pick up the best alterables automatically to be used in the model. Ten most important predictors were selected in row $X_2, X_4, X_5, X_7, X_8, X_{10}, X_{13}, X_{13}, X_{14}$, and X_{15} and others were removed from analysis as they were not significant. Table 4.16 presents the selection of discriminating variables depending on stepwise method and performing Wilks Lambda statistic.

Table 4.16. Selection of discriminating variables depending on stepwise method

Variable Entered	Wilks' Lambda	D.F.1	D.F.2	D.F3	Statistic	Exact F D.F.1	D.F.2	Sig.
X_{14}	0.926	1	1	343	27.379	1	343	0.00
X_{13}	0.884	2	1	343	22.350	2	342	0.00
X_4	0.855	3	1	343	19.207	3	341	0.00
X_5	0.832	4	1	343	17.131	4	340	0.00
X_{15}	0.820	5	1	343	14.927	5	339	0.00
X_{12}	0.807	6	1	343	13.473	6	338	0.00
X_7	0.795	7	1	343	12.392	7	337	0.00
X_8	0.785	8	1	343	11.508	8	336	0.00
X_2	0.774	9	1	343	10.877	9	335	0.00
X_{10}	0.763	10	1	343	10.362	10	334	0.00

Additionally, Wilks Lambda can be used to gauge how well each function distinguishes between blood cancer cases and presence. We can infer from table 4.17 that the discriminant function is significant and that the corresponding function adequately explains group membership.

Table 4.17. Wilks lambda table

Wilks Lambda	Chi-square	D.F.	Sig.
0.763	91.334	10	0.000

4.4.5. Estimating the discriminant model

Table 4.18 provides the unstandardized canonical discriminant function coefficients that are required by the classification formula in DA. The discriminant scores are obtained by adding the constant to the sum of products of the unstandardized coefficients and the observations.

Table 4.18. Unstandardized canonical discriminant model coefficients

Variables	Coefficients of Function
X ₂	.014
X ₄	.767
X ₅	.601
X ₇	.217
X ₈	.230
X ₁₀	.244
X ₁₂	.279
X ₁₃	.159
X ₁₄	.142
X ₁₅	.107
(Constant)	-1.838

The is, the estimated discriminant model is:

$$G = -1.838 + 0.014 + 0.767 + 0.601 + 0.217 + 0.230 + 0.244 + 0.279 + 0.159 + 0.142 + 0.107$$

4.4.6. Testing the discriminatory power of the discriminant model

Performed here to evaluate the efficiency of the discriminatory model of the estimated function. Table 4.19 shows the final results of classification.

Table 4.19. The final classification results

Cases	Predicted Group Membership		Total
	Absence of Blood Cancer	Presence of Blood cancer	
Absences of (BCD)	56	3	59
Presence of (BCD)	80	211	291
Total			350
Model Accuracy	76.2%		
Model Sensitivity	94.9%		
Model Specificity	72.5%		

From table 4.19. which analysis of the classification table, shows that the discriminant analysis model successfully classified, as can be seen, the model accuracy of (76.2%), and sensitivity and specificity are both above (94.9% and 72.5%). That is, the model can accurately predict based on the independent predictors entered by (94.9%). The model specificity was (72.5%), that is to say, it can accurately predict patients with out BCD based on the independent variable entered by (72.5%).

4.5. Comparison and Discussion

In this section, we explored the classification of utilizing the logistic regression and discriminant analysis. Evaluation measures are available to compare power functions between two or more models. Using the Accuracy information criteria, Sensitivity information criteria and Specificity information criteria to compare modeling might also be beneficial. The results show that the LRA is the best model as it demonstrated that Accuracy model of 82.2% of the original cases were accurately classified to their related groups. Table 4.20 shows model accuracy, model sensitivity, model specificity of classification for testing data set on blood cancer.

Table 4.20. Performance criteria for comparing methods

Methods	Logistic Regression	Discriminant Analysis
Model Accuracy	96.8%	76.2%
Model Sensitivity	86.4%	94.9%
Model Specificity	98.9%	72.5%

Table 4.20 shows that whereas support discriminant analysis methods' accuracy was 76.2%, logistic regression analysis's accuracy was 96.8%. According to the precision of the method—more exact is best—this indicates the preference for LRA. The sensitivity standard for LRA was 86.4%. The model's performance was identical to that of utilizing DA, giving DA performance in accordance with the sensitivity criterion. Additionally, the specificity requirement for LRA was 98.9%, compared to 72.5% for DA. implies that the first approach has a definite preference. In the conclusion, we discovered that the LRA approach significantly outperformed the DA results in the classification of BCD patients at the Erbil Nanakali Hospital of the sample that was examined for this thesis.



5. CONCLUSION

- 1- The statistical analysis had been used secondary data that achieved from Nanakali in Erbil Northern Iraq in the period from 1st January 2017 to 31st December 2018 with a follow up period until 1st April 2021. Data included the number of Blood Cancer disease (BCD), used for categorical variables to indicate frequencies and percentages. Pearson's chi-square used to find the association between the categorical variables. Based on the empirical results, the variables age of patient, sex of patient and type blood cancer of patient, are the significant.
- 2- The most of the patients with blood cancer are from the age of 10 and less than 19 years, as thier percentage is 16.54%
- 3- The most of the patients with blood cancer are from cases were recorded in 2021, as thier perecentage is 25.5%
- 4- The Acute lymphoblastic leukemia (ALL) disease is more common than all other type of blood cancer disease, with perecentage is 31%
- 5- The results obtained for data form, we explored the classification utilizing the logistic regression analysis (LRA) and discriminant analysis (DA). There are evalution measure to compare power between two or more models, using the accuracy, sensitivity and specificity informations criteria for compare modeling might be benficial. The results shows that the LRA is the best model becuase accurasy model was of 96.8%, sensitivity model was of 97.2% and specificity model was of 94.4%, that the value of models are more than DA .
- 6- The results indicated that the factors had greatest effect on the data of patients with BCD using the two methods (LRA, DA) that the variables (Acute myeloid leukemia and chemo) are the most effective and significant variables.



REFERENCES

- Abdel-Karim, N. D., 2006. **Using Statistical Discriminatory Method for Diagnosing Some Heart Diseases**. Master of science, to college management and economics, Mustansiriyah University.
- Abdulhafedh, A., 2022. Comparison between common statistical modeling techniques used in research, including: discriminant analysis vs logistic regression, ridge regression vs LASSO, and decision tree vs random forest. **Open Access Library Journal**, **9** (2): 1-19.
- Abdul-Kareem, A. D., 2006. Using of statistical discriminate methods for the diagnosis of some heart diseases. **Kirkuk University Journal - Scientific Studies**, **1** (2): 16-28.
- Abuosi, A., Atinga, R. A., 2013. Service quality in healthcare institutions: establishing the gaps for policy action. **International Journal of Health Care Quality Assurance**, **26** (5): 481-492.
- Agresti, A., 2007. **Categorical Data Analysis. Second edition**. New York: Johnson Wiley and Sons, Inc.
- Anderson, T. W., 1984. **An Introduction to Multivariate Statistical Analysis**. 2nd Edition, John and Sons, New York, USA.
- Andrew, S. L., Kai, U. E., Yusuke, T., Adeera, L., Josef, C., Jerome, R., 2005. Definition and classification of chronic kidney disease: a position statement from kidney disease: **Improving Global Outcomes (KDIGO) Kidney Inter**, **67** (6): 2089-2100.
- Bhatla, N., Jyoti, K., 2012. An analysis of heart disease prediction using different data mining techniques. **International Journal of Engineering**, **1** (8): 1-4.
- Chen, S-T., Hsiao, Y-H., Huang, Y-L., Kuo, S-J., Tseng, H-S., Wu, H-K., Chen, D-R., 2009. Comparative analysis of logistic regression, support vector machine and artificial neural network for the differential diagnosis of benign and malignant solid breast tumors by the use of three-dimensional power doppler imaging. **Korean Journal of Radiology**, **10** (5): 464-471.
- Cizek, G. J., Fitzgerald, S. M., 1999. An Introduction to logistic regression. **Measurement and Evaluation in Counseling and Development**, **31** (4): 223-45.
- Cizek, G. J., Fitzgerald, S. M., 1999. Methods, plainly speaking: an introduction to logistic regression. **Measurement and Evaluation in Counseling and Development**, **31** (4): 223.
- Dayton, C. M., 1992. **Logistic Regression Analysis**. Department of Measurement Statistics and Evaluation, University of Maryland: 474-574.
- Dean, L., 2005. **Blood Groups and Red Cell Antigens [Internet]**. Bethesda (MD): National center for biotechnology information (US).
- Denis, D. J., 2018. **SPSS Data Analysis for Univariate, Bivariate, and Multivariate Statistics**. John Wiley and Sons.
- Eisenbeis, R. A., 1971. Discriminant analysis and classification procedures. Zeitschrift für die gesamte staatswissenschaft/ **Journal of Institutional and Theoretical Economics**: 500-521.
- Fagerland, M. W., Hosmer, D. W., Bofin, A. M., 2008. Multinomial goodness-of-fit tests for logistic regression models. **Statistics in Medicine**, **27** (21): 4238-4253.
- Ferrer, A., Arce, J., Wang, L., 1999. **Comparing the Classification Accuracy among Nonparametric, Parametric Discriminant Analysis and Logistic Regression**

- Methods.** Paper presented at the annual meeting of the American educational research association, Montreal, Quebec, Canada.
- Ferrer, J., Wang, L., 1999. **Comparing the Classification Accuracy among Nonparametric, Parametric Discriminant Analysis and Logistic Regression Methods.** Paper presented at the annual meeting of the American educational research association, Montreal, Quebec, Canada.
- Fraas, J. W., Newman, I., 2003. **Ordinary Least Squares Regression, Discriminant Analysis, and Logistic Regression:**
- Garson, G. D., 2014. **Logistic Regression: Binary and Multinomial.** Asheboro, NC. Statistical publishing associates.
- Goodman, L. A., Kruskal, W. H., 1979. Measures of association for cross classifications. Measures of association for cross classifications. **Springer:** 131-146.
- Greenwood, C., Farewell, V., 1988. A comparison of regression models for ordinal data in an analysis of transplanted-kidney function. **The Canadian Journal of Statistics/La Revue Canadienne de Statistique**, **16** (4): 325-335.
- Haggstrom, G. W., 1983. Logistic regression and discriminant analysis by ordinary least squares. **Journal of Business and Economic Statistics**, **1**: 229-238.
- Harrell, F. E., 2001. Regression modeling with applications to linear models, logistic regression, and survival analysis. **Springer Science-Verlag, New York.**
- Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning. Cited on, 33.
- Heena, N., Shruti, G., Shubham, S., Subashka, R., 2019. Detection analysis of various types of cancer by logistic regression using machine learning. **International Journal of Engineering and Advanced Technology (IJEAT)**, **9** (1): 2249-8958
- Henel, G., Schmitz, J. L., 2007. Basic theory and clinical applications of flow cytometry. **Laboratory Medicine.**, **38**: 428-436.
- Hosmer, D. W., Lemeshow, S., Klar, J., 1988. Goodness-of-fit testing for the logistic regression model when the estimated probabilities are small. **Biometrical Journal**, **30**: 911-924.
- Johnson, R. A., Wichern, D. W., 2007. **Applied Multivariate statistical analysis.** Sixth edition, Pearson, prentice Hall.
- Khuriwal, N., Mishra, N., 2018. **Breast Cancer Diagnosis Using Adaptive Voting Ensemble Machine Learning Algorithm.** IEEMA Engineer infinite conference: 1-5.
- Krzanowski, W. J., 1977. The performance of Fisher's linear discriminant function under non-optimal condition. **Technometrics**, **19** (2): 191-200.
- Lei, L., 2018. Research on logistic regression algorithm of breast cancer diagnosis data by machine learning. **International Conference on Robots and Intelligent System (ICRIS)**, **15**: 157-160.
- Pampl, F. C., 2000. **Logistic Regression Primer.** Sage university paper series on quantitative applications in the social sciences. Beverly Hills, CA: Sage: 07-132
- Rencher, A. C., Schimek, M. G., 1997. Methods of multivariate analysis. **Computational Statistics**, **12** (4): 422-422.
- Saleh, A. H., 2008. The use of characteristic analysis to diagnose some eye diseases. (Ibn Al-Haytham teaching hospital for eye diseases - (Iraq). **Journal of Management and Economics, University of Karbala**, **0** (67): 264-286.

- Saleh, A. H., 2014. Logistic regression analysis to study the survival time of leukemia patients. *Iraq Journal of Administration and Economics, University of Karbala*, **3** (9): 261-280.
- Yousra, A., Mohammed, E., Saleh, G., 2020. Investigating the applicability of logistic regression and artificial neural networks in predicting breast cancer. *Journal of Al-Qadisiyah for Computer Science and Mathematics*, **12** (2): 63-73.



EXTENDED TURKISH SUMMARY (GENİŞLETİLMİŞ TÜRKÇE ÖZET)

KAN KANSERİNİN SAĞKALIM ANALİZİNDE LOJİSTİK REGRESYON VE DİSKRİMİNANT ANALİZİ MODELLERİNİN KARŞILAŞTIRILMASI

HASAN, Hawar Khan Murad
Yüksek Lisans Tezi, İstatistik Anabilim Dalı
Tez Danışmanı: Prof. Dr. Fevzi Erdoğan
Ocak 2023, 73 sayfa

Bu çalışma, kan kanseri hastalığına (KKH) neden olan sebeplerin ve durumların sınıflandırılması konusunu incelemek amacıyla yapılmıştır. Bu çalışmada, var olanlar ile olmayanlardan oluşan bir grup temelinde, kan kanseri hastalığına neden olan sebeplerin ve durumların sınıflandırılması için klasik istatistiksel yöntemlerden biri olan Lojistik Regresyon Analizi (LRA) ile ayırt edici tekniklerden oluşan bir yöntem olan Diskriminant Analizi (DA) yöntemleri karşılaştırılmak suretiyle incelenmiştir. Söz konusu araştırmanın amacına ulaşabilmesi için veri kümesinde iki farklı veri yer almaktadır. İlk olarak, 1 Ocak 2017 tarihinden 31 Aralık 2018 tarihine kadar olan dönem için Kuzey Irak'taki Erbil şehrinde faaliyet gösteren Nanakali Hastanesi'nden elde edilen ikincil veriler, 1 Mart 2021 tarihine kadar takip edilecek şekilde kullanılmıştır. İkincisi, söz konusu veriler, kan kanseri hastalığının (KKH) tipinin de sorulduğu ve bu tipin söz konusu kanser ile arasındaki ilişkiyi de içeren bir anket formu ile toplanmış olup bu veriler, 16 'sı bağımsız 17 değişkenden ve bir adet de bağımlı değişkenden (varlık 1 ve yokluk 0) oluşan toplamda 350 hastadan edinilen verileri içermektedir. Doğru sınıflandırma sonuçlarının doğruluk yüzdesi oranı Lojistik Regresyon Analizi (LRA) için % 96.8 ve Diskriminant Analizi (DA) için ise % 76,2 'dir.

Anahtar kelimeler: Diskriminant analizi, Kan kanseri hastalığı, Lojistik regresyon analizi, Sınıflandırma.

1. GİRİŞ

Yıllık istatistikler göz önüne alındığında kan kanserinin (Lösemi) bütün Dünya çapında görülen yaygın bir hastalık olduğu ortaya çıkmaktadır. Bu hastalığa yakalanan insanların bir kısmının hayatını kaybettiği de yine bu istatistiklerden görülebilmektedir. Bu çalışmanın amacı yaş, cinsiyet, aile geçmişi, kilo, boy, sigara kullanımı ve spor değişkenleri esas alınarak bu hastalığı kategorize etmek, kategorize edilmiş bağımlı değişkenlerin (sonuç), lösemi hastasının hastalık tipi, tedavi tipi, hastalığa yakalanma ve yakalanmama oranları ve bu verilerin uygun istatistiksel yöntemler kullanılarak modellenmesidir. Etkin faktörler (değişkenler), lösemi tipinin ve tedavisinin sınıflandırılmasındaki olasılıkların belirlenmesinde bilinebilir. Bu, verileri belirli varsayımlar altında tanımlayan istatistiksel bir model oluşturarak yapılır. Çalışmanın ana fikri gelişmiş istatistiksel modeller ve ayrımcılık kullanılarak lösemi türlerini ve tedavi yöntemlerini sınıflandırmasıdır. Kansere yakalanmak yaşamı tehdit eden bir hastalık olarak kabul edilmektedir. Hastalığın türünün teşhis edilmesi aynı zamanda kişinin hayatını kurtarmaya ve hastalığın ciddi vakalarını önlemeye yardımcı olur. Lösemnin erken evre teşhisi. Bu nedenle bu çalışmada iki model (Lojistik regresyon analizi) ve (Ayrım Analizi) karşılaştırılmıştır. Bu iki model aracılığıyla, verileri lösemi hastaları ve lösemi olmayan hastalar olarak sınıflandıran bir anket formuna dayalı olarak lösemi verilerini analiz ediyoruz.

Özellikle tahmin yöntemlerinin gelişmesinden sonra, önceden incelenmemiş, içinde diskriminant ve regresyonun da bulunduğu bu yöntemler içerisinde değişkenlerin araştırılması için kullanılan çeşitli istatistiksel analiz yöntemleri ve yaklaşımları mevcuttur. Lojistiğin hedefi, önceki verilere bağlı olarak yeni durumları sınıflandırmak için derecelendirilmiş veya kabul edilmiş değişkenler üzerinde büyük etkisi olan değişkenleri bulmaktır. Bu alanda birçok değişken olduğu için bir tıp alanında (Diskriminant Analizi ve Lojistik Regresyon Analizi) yöntemlerinden birinin kullanılması uygundur.

Bağımlı değişkenin kategorik bir değişken olduğu bir grup bağımsız değişken ile yanıt değişkeni (bağımlı değişken) arasındaki ilişkinin araştırılması lojistik regresyon kullanımına yol açmıştır. Buna ilaveten, Lojistik Regresyon (LR) kendi içerisinde iki kategoriye ayrılmıştır: Bunlardan birincisi, sadece iki sınıfa sahip ikili lojistik regresyon

modeli ve ikincisi ise, sıralı veya nominal olan bağımlı bir değişkene sahip ve yine iki sınıfa sahip olan çok terimli lojistik regresyon modelidir. Lojistik fonksiyonun ilk ilke defa kullanıldığı Lojistik Regresyon Modeli, sonraları araştırmacılar tarafından yaygın olarak kullanılmıştır (Verhulst, 1920).

Yapılan anket sonucu aşağıdaki bulgulara ulaşılmıştır. Lösemi vakalarının büyük çoğunluğu 2017 – 2021 yılları arasında görülmüştür. En fazla vakanın kayıt altına alındığı yıl ise 2021’dir. Ayrıca, Akut Limfoblastik Lösemi (ALL) en yaygın olan lösemi türüdür. Radyasyon tedavisi de bu iki modeli tedavi etmek ve karşılaştırmak için yaygın olarak kullanılmaktadır. Lojistik Regresyon Analizi modeli etkin faktörlerin belirlenmesi ve sınıflandırılmasında %8.096’lık bir oranla Ayırım Analizinden daha etkilidir.

2. KAYNAK BİLDİRİŞLERİ

Aida Hadi Saleh'in, Lösemili Hasta Sağkalımının Değerlendirilmesi için Lojistik Regresyon Analizi (Şeyh Zayed Acil Hastanesi, Bağdat, 2010 -2011 - Irak). Lojistik regresyon yaklaşımı kullanılarak, bu hastalığa sahip hastalardan örnekler alındı. Yaş, cinsiyet, hastalık sınıflaması (akut miyeloid lösemi, akut lenfoblastik lösemi, kronik miyeloid lösemi ve kronik lenfositik lösemi) ve löseminin tipi, bu hastalığın şiddeti ile ilişkili risk faktörleridir (değişkenlerdir). Söz konusu tedavi (biyolojik tedavi ve kemoterapi), ikamet yeri (valilik sınırları içinde ve dışında) ve anemik durum (valilik sınırları içinde ve dışında) olarak kategorize edildi (anemiden muzdarip olmayanlar ve anemiden muzdarip olanlar). Yanıt değişkeni, bağımlı değişkene karşılık gelen yaralı durumunu temsil eder. Ölüm halinde bu değer bir (1), kontrol durumunda ise söz konusu değer sıfır (0) olur. Bu çalışmanın amacı, lösemi hastalarının sağkalım sürelerini araştırmak ve modelde hastanın durumunu ve sağkalım adaylarını etkileyen değişkenleri keşfetmek için lojistik regresyon modelinin kullanılıp kullanılamayacağını belirlemektir. Bu çalışmanın bulguları arasında, lösemili hastaların çoğunluğunu yüzde 42,47 oran ile 20 yaşından küçüklerin oluşturduğu; lösemili hastaların çoğunluğunun yüzde 63 oran ile kadınlar olduğu, akut miyeloid löseminin yüzde 43,8 oran ile en sık görülen lösemi tipi olduğu, bunun yanı sıra ise kronik lenfositik löseminin ise sadece yüzde 4,3 oran ile en nadir görülen lösemi tipi olduğu yer almaktadır. Akut miyeloid lösemi ise yüzde 63,63 ölüm oranı ile en önde gelen ölüm nedenidir. Ve hastaya sunulan tedavi türünün hastalığın

tipine ve evresine göre deđiřtiđi ve hasta ileri bir evreye ulařtıđında hastaya kemoterapi aldırıldıđı, kemoterapi ile tedavi edilenlerin ölüm oranının yüzde 72,72 ile en yüksek seviyede olduđu da bu çalışmada yer alan bilgiler arasındadır. Bu çalışmada, farklı hastalık türleri için benzer çalışmaların yapılması ve bu tiplerin her birini etkileyen faktörlerin neler olduđunun öğrenilmesi ve bu model ya da Cox modeli kullanılarak başka faktörlerin de çalışmalara ilave edilmesi önerilmiştir. Ekonomik ve sosyal çalışmalarda iki yanıtlı lojistik modelin kullanım alanının genişletilmesi önerilmiştir çünkü bu modelin eskiden bu yana kullanımı tıp ve eğitim bilimleri alanları ile sınırlı kalmıştır.

3. MATERYAL VE YÖNTEM

Bu çalışmada, söz konusu çalışmanın amacına ulaşabilmesi için iki farklı veri kullanılmıştır. İlk olarak, 1 Ocak 2017 tarihinden 31 Aralık 2018 tarihine kadar olan dönem için Kuzey Irak'taki Erbil şehrinde faaliyet gösteren Nanakali Hastanesi'nden elde edilen ikincil veriler, 1 Nisan 2021 tarihine kadar takip edilecek şekilde kullanılmıştır. Söz konusu veriler, Kan Kanseri Hastalığı (KKH) olanların sayısını, söz konusu hastaların yaşlarını, cinsiyetlerini ve söz konusu hastaların sahip oldukları kan kanseri tipini içermektedir. İkincisi, Lojistik Regresyon Analizi (LRA) ile Diskriminant Analizi (DA) karşılaştırmasına ilişkin uygulama Kan Kanseri Hastalığı (KKH) kullanılarak gerçekleştirilmiştir. Kan Kanseri Hastalığı (KKH) tipini ve arasındaki ilişkiyi içeren anket verileri, kan kanseri hastalığı olanları (1) ile ve kan kanseri hastalığı olmayanları (2) ile etiketlemek suretiyle toplanmıştır.

3.1. Kan Kanseri

Kan kanseri, genellikle beyaz kan hücrelerinin çođalması şeklinde kendini gösteren, kan hücrelerinin kontrolsüz bir şekilde çođalmasıdır. Söz konusu çođalma durumu gerçekleştiğinde yani lösemide, beyaz kan hücrelerinin kandaki sayısı mm başına 5000 ila 10000 hücre aralığındaki normal sınırı aşarak yani artarak, beyaz kan hücrelerinin kandaki sayısı mm başına 15000 ila 30000 hücre aralığına ulaşmaktadır ve bu çođalma hali mm başına 100000 hücreden de fazla bir miktara ulaşabilir veya tam aksi yönde gerçekleşerek mm başına 5000'den de daha düşük bir miktara (yani, normal sınır

alt limit değerinden daha da az bir miktara) düşebilir. Söz konusu sayıdaki farklılıklar, lösemnin evresinden ve vücuttaki ilerlemesinin derecesinden kaynaklanmaktadır. Vücuttaki tüm kan hücrelerinin fabrikası olan kemik iliğinde anormal beyaz kan hücreleri üretilir. Üretilen kan hücrelerinin sayısının azaltılması ve söz konusu hastalığın başlangıcında kandaki beyaz kan hücrelerinin geçici olarak eksikliğine yol açan beyaz kan hücrelerinin kandan salınmasından önce, kemik iliği bu kanserli hücreleri kana hızlı bir şekilde salar ve bu şekilde kandaki beyaz kan hücrelerinin sayısını artırır ve bu artış genellikle sabit bir artıştır.

3.2. Kanser Gelişme Eğilimini Etkileyen Eeael Faktörler

- 1- Genetik (kalıtsal) anormallikler (anomaliler)
- 2- Radyasyona maruz kalmak gibi ekstrem bir durum
- 3- Kemoterapi ve radyasyon
- 4- Kimyasallar
- 5- Lösemiye neden olabilecek kan durumları
- 6- Virüsler

3.3. Kan Kanseri Tipi

- 1- Akut lösemi
- 2- Kronik lösemi
- 3- Lenfositik lösemi
- 4- Miyeloid lösemi

4. BULGULAR

4.1. Lojistik Regresyon Analizi (LRA) Kullanarak Uygulama

Verilerin iki ayrı gruba (varlık ve yokluk) ayrılmasından sonra kan kanseri hastalığı (KKH) üzerinde lojistik regresyon analizi (LRA) yapıldı. 350 vakayı içeren eğitim veri kümesini temel alarak modeli inşa etmeye başladık. Şimdi, verilerle ne kadar

uyumlu olduğunu ve diğer gözlemlere ne derece genelleme yapma yetisine sahip olduğunu anlamak için söz konusu modeli incelemeliyiz. Daha sonra, model uyumunun bir ölçüsü olan sapma artıklarını görürüz. Çıktının bu kısmı, modelde kullanılan bireyler bazında sapma artıklarının dağılımını göstermektedir. Söz konusu modelin uygunluk bilgileri 4.10 sayılı Çizelge gösterilmektedir.

Çizelge 4.10 Model uygunluk bilgileri

Model	Model Uygunluk Kriterleri	Olasılık Oranı Testleri		
	-2Log Olasılığı	Ki - Kare	D.F	Anlamlı
Sadece Son Kesişim Sonucu	316.143	238.070	16	0.00
Sonuç	78.073			

Çizelge 4.10 da katsayıların anlamlı olduğu görülmektedir ($p \leq 0.05$). Lojistik regresyon katsayıları, bağımsızlar değişkeni üre ve kreatinindeki bir birimlik artış için sonucun log olasılıklarındaki değişimi istatistiksel olarak anlamlıdır. Sadece kesişim sonucu için -2Log olasılık değeri 316.143 ve OLR kullanılarak sorunsuz bir şekilde analiz edilebilen denekler için nihai modeller 78.073 'tür. Ki-kare değeri 238.070 'dir.

Çizelge 4.11 'de gösterilen OLR analizi ile sağlanan korelasyon ölçümü sonucunda ön göstergeler ile açıklanan bağımlı değişkendeki varyans oranını bize gösteren Cox & Snell R^2 (0.493) ve Nagelkerke R^2 (0.827) değerlerini bulduk. Bunlar bir sayısına yaklaştıkları için iyi değerlerdir. Modelimizin ne kadar iyi uyduğu, model ile gözlemlenen veriler arasındaki farka bağlıdır.

Çizelge.4.11. Pseudo R^2

Cox ve Snell R Kare	Nagelkerke R Kare
0.493	0.827

Model regresyon katsayılarının Olasılık Oranı testi, bağımsız değişken ile bağımlı değişken arasındaki ilişkiyi değerlendirmek ve tanımlamak için açıklayıcı değişkenlerin dahil edilip edilmediğini araştırmak için kullanılır. Uygulamada, Çizelge 4.12 'nin amacı testin sonuçlarını göstermektir. Dokuz değişken en önemli ön göstergeler, düzgün bir

şekilde seçildi (X_{14} , X_{11} , X_{10} , X_9 , X_8 , X_7 , X_6 , X_5 ve X_1) söz konusu adım değişken eklemekse, değişimin anlamlılığı 0,05'ten az ise dahil etme eğilimini, ancak söz konusu adım bir değişkeni kaldırmaksa, değişimin anlamlılığı 0.10 'dan büyük ise hariç tutma eğilimini ortaya çıkarır.

Çizelge 4.12. LRT sonucu

Etkisi	Model Uygunluk Kriterleri Değişkeni	Ortalama Ki-kare	LRT	
			D.F.	Anlamlı
	-2Lo Düşük Model Olasılığı			
Kesişme	81.791	81.299	1	0.00
X_{14}	84.538	6.465	1	0.01
X_{11}	88.213	10.140	1	0.01
X_{10}	144.881	66.808	1	0.00
X_9	99.973	21.900	1	0.00
X_8	173.148	95.075	1	0.00
X_7	174.420	96.347	1	0.00
X_6	121.228	43.155	1	0.00
X_5	182.041	103.968	1	0.00
X_1	82.608	4.535	1	0.03

4.1.1. Diskriminant analizi (DA) kullanarak uygulama

İki grup için bir Diskriminant Analizi (DA) modeli tanımlamadan ve oluşturmadan önce, çoklu doğrusallığın (doğrusal çıkarsama) verilerde bir sorun yaratmadığından emin olmak için bazı testler yapmak gerekir. Çoklu doğrusallığın varlığı Diskriminant Analizi (DA) modelinin bir sorunu olduğunu, çoklu doğrusallığın olmaması ise bağımsız değişken kümesinin herhangi bir güçlü korelasyona sahip olmadığını göstermektedir. Bunun nedeni, Diskriminant Analizi (DA) modelinin birkaç doğrusal korelasyona bağımlılığıdır. İlk veriler için çoklu doğrusal korelasyon sorununun çözülmesinin sonucu yukarıdaki Çizelge 4.15 'te gösterilmiştir (doğrusal çıkarsama). Bu durumda Varyans Enflasyon Faktörü (VIF) testinden yararlanılır ve Berenson ve arkadaşlarına göre Varyans Enflasyon Faktörü (VIF) > 10 durumunun sonuçlarından bir tanesi çoklu doğrusallıktır ve varsayımın bozulduğunu gösterir. (2002) Her ön gösterge için Varyans Enflasyon Faktörü (VIF) değerleri Çizelge 4.15 'te sunulmuştur.

Çizelge 4.15. Ön gösterge değişkenlerinin varyans enflasyon faktörü (VIF) değerleri

Sıra No	Ön Gösterge Değişkenleri	Varyans Enflasyon Faktörü (VIF) Değerleri
1	X_1	1.217
2	X_2	1.074
3	X_3	1.218
4	X_4	1.045
5	X_5	1.028
6	X_6	1.241
7	X_7	1.071
8	X_8	1.070
9	X_9	1.031
10	X_{10}	1.056
11	X_{11}	1.071
12	X_{12}	1.241
13	X_{13}	1.256
14	X_{14}	1.177
15	X_{15}	1.124
16	X_{16}	1.194

Çizelge 4.15 'teki tüm Varyans Enflasyon Faktörü (VIF) değerlerinin Radyo sayısı için yüksek $X_{13}=1.256$ ' dan nispeten küçük olduğunu, bu değerlerin, bir tür kan kanseri hastalığı (KKH) olan Akut Miyeloid Lösemi için olan $X_5=0.028$ 'e göre uygulanan bir tür kan kanseri tedavisi için ise düşük olduğu görülmektedir. Bu nedenle, varyans enflasyon faktörünün değerleri olan ve 10 'dan küçük olan Varyans Enflasyon Faktörü (VIF) kriterlerine dayanarak, değişken ön gösterge veya ön göstergelerin kümesi içinde çoklu doğrusallık dizini yoktur, yani herkes bu noktadan sonra analize devam edebilir. Diskriminant modelini oluşturmadan önce, grubun eşitliğini incelemek yani grupların farklılaşmasına önemli ölçüde katkıda bulunan değiştirilebilirliği belirtmek ve her bir ön göstergenin potansiyelini ölçmek için bu aşamada F-testli Wilks' Lambda istatistikleri elde edilecek söz konusu testler gerçekleştirilir. Bu, gruplar arasındaki farklılıklarla açıklanması mümkün olmayan diskriminant skorlarındaki toplam değişim oranına eşittir. Wilks' Lambda'nın daha küçük değerleri, fonksiyonun daha büyük ayırt edici becerisini gösterir. Rencher (2002). Bölüm (3.4)' te bahsettiğimiz gibi, 16 ön göstergeye sahip olduğumuz, kademeli yöntem, söz konusu modelde otomatik olarak kullanılacak en iyi değiştirilebilirleri almak için yararlı olabilir. En önemli on ön gösterge sırayla seçildi $X_2, X_4, X_5, X_7, X_8, X_{10}, X_{13}, X_{13}, X_{14}, ve X_{15}$ ve diğerleri anlamlı olmadığı için analizden

çıkarıldı. Çizelge 4.16 'da kademeli yöntem ve Wilks' Lambda istatistiğine bağlı olarak ayırt edici değişkenlerin seçimi gösterilmiştir.

Çizelge 4.16 Kademeli yöntemle ilgili olarak ayırt edici değişkenlerin seçimi

Değişken Girilen	Wilks' Lambda	D.F.1	D.F.2	D.F.3	İstatistik	Tam F D.F.1	D.F.2	Anlam lı
X ₁₄	0.926	1	1	343	27.379	1	343	0.00
X ₁₃	0.884	2	1	343	22.350	2	342	0.00
X ₄	0.855	3	1	343	19.207	3	341	0.00
X ₅	0.832	4	1	343	17.131	4	340	0.00
X ₁₅	0.820	5	1	343	14.927	5	339	0.00
X ₁₂	0.807	6	1	343	13.473	6	338	0.00
X ₇	0.795	7	1	343	12.392	7	337	0.00
X ₈	0.785	8	1	343	11.508	8	336	0.00
X ₂	0.774	9	1	343	10.877	9	335	0.00
X ₁₀	0.763	10	1	343	10.362	10	334	0.00

4.2. Karşılaştırma ve Tartışma

Bu bölümde, lojistik regresyon ile diskriminant analizini kullanma sınıflandırmasını inceledik. İki veya daha fazla model arasındaki güç fonksiyonlarını karşılaştırmak için değerlendirme önlemleri mevcuttur. Modellemeyi karşılaştırmak için doğruluk bilgi kriterlerini ve hassasiyet bilgi kriterlerini ve özgüllük bilgi kriterlerini kullanmak da faydalı olabilir. Sonuçlar, orijinal vakaların %82.2 'sinin doğruluk modelinin ilgili gruplarına doğru bir şekilde sınıflandırıldığını gösterdiğinden, Lojistik Regresyon Analizi (LRA) 'nin en iyi model olduğunu göstermektedir.

Çizelge 4.20 'de kan kanseri ile ilgili test veri kümesi için model doğruluğu, model hassasiyeti, model özgüllüğü sınıflandırması gösterilmektedir.

Çizelge 4.20 Yöntemleri karşılaştırmak için performans kriterleri

Yöntemler	Lojistik Regresyon	Diskriminant Analizi
Model Doğruluğu	%96,8	%76,2
Model Hassasiyeti	%86,4	%94,9
Model Özgüllüğü	%98,9	%72,5

Çizelge 4.20 'de destek diskriminant analiz yöntemlerinin doğruluğunun %76,2 iken lojistik regresyon analizinin doğruluğunun %96.8 olduğu görülmektedir. Yöntemin hassasiyetine göre (ne kadar hassas ise o kadar iyi) sonuçlar, Lojistik Regresyon Analizi (LRA)'nin daha iyi bir tercih olduğunu göstermektedir. Lojistik Regresyon Analizi (LRA) için hassasiyet standardı %86.4 'ydı. Söz konusu modelin performansı, Diskriminant Analizi (DA) kullanımıyla aynıydı ve hassasiyet kriterine uygun olarak Diskriminant Analizi (DA) performansı veriyordu. Buna ilaveten, Lojistik Regresyon Analizi (LRA) için özgüllük gereksinimi %72.5 olan Diskriminant Analizi (DA) ile karşılaştırıldığında % 98,9 'dur. Bu sonuç da ilk yaklaşımın kesinlikle tercih edilmesi gerektiği anlamına gelmektedir. Sonuç olarak, Lojistik Regresyon Analizi (LRA) yaklaşımının, bu tez için incelenen örneklemin Erbil Nanakali Hastanesi'ndeki Kan Kanseri (KK) hastalarının sınıflandırılmasında Diskriminant Analizi (DA) sonuçlarından önemli ölçüde daha iyi performans gösterdiğini tespit ettik.

5. TARTIŞMA VE SONUÇ

İstatistiksel analiz, 1 Ocak 2017 tarihinden 31 Aralık 2018 tarihine kadar olan dönem için Kuzey Irak'ın Erbil kentindeki Nanakali Hastanesi'nden elde edilen ikincil veriler kullanılarak 1 Nisan 2021 'e kadar takip edilmik suretiyle gerçekleştirilmiştir. Kan Kanseri Hastalığından (KKH) muzdarip kişilerin sayısını içeren veriler, frekansları ve yüzdeleri belirtmek için kategorik değişkenler için kullanılıyordu. Kategorik değişkenler arasındaki ilişkiyi bulmak için Pearson Ki-Kare Testi kullanılmıştır. Ampirik sonuçlara göre, hastanın yaşı, cinsiyeti ve hastanın kan kanseri tipi değişkenleri anlamlıdır.

Kan kanseri hastalarının çoğu 10 yaşında veya 10 yaşından büyük ve 19 yaşından küçüktür, ve bu hastalar genelin %16.54 'ünü oluşturmaktadır.

Kan kanserli hastaların çoğu 2021 yılında kaydedilen vakalardandır ve bunların genele oranı %25.5 'tir.

Akut lenfoblastik lösemi (ALL) hastalığı diğer tüm kan kanseri hastalıklarından daha yaygındır ve genele oranı %31 'dir.

Veri formu için elde edilen sonuçları, Lojistik Regresyon Analizi (LRA) ile Diskriminant Analizi (DA) kullanarak sınıflandırmayı araştırmak suretiyle elde ettik. İki veya daha fazla model arasındaki gücü karşılaştırmak için belli bir değerlendirme ölçütü

vardır ve karşılaştırma modellemesi için doğruluk, hassasiyet ve özgüllük bilgileri kriterlerini kullanmak faydalı olabilir. Sonuçlar, Lojistik Regresyon Analizi (LRA) 'nın en iyi model olduğunu, çünkü doğruluk modelinin %96.8, hassasiyet modelinin %97.2 ve özgüllük modelinin %94.4 olduğunu, ve elde edilmiş olan bu modellerin değerlerinin Diskriminant Analizi (DA) için elde edilen değerlerden daha yüksek olduğunu göstermektedir.

Sonuçlar, iki yöntem (Lojistik Regresyon Analizi (LRA), Diskriminant Analizi (DA) kullanılarak Kan Kanseri (KK) hastalarının verileri üzerinde söz konusu faktörlerin en büyük etkiye sahip olduğunu ve değişkenlerin (akut miyeloid lösemi ve kemo) etkili ve anlamlı değişkenler olduğunu göstermiştir.





APPENDIX

App 1. Questionnaire form

Hawar khan Murad HASAN
VAN YUZUNCU YIL UNIVERSITY
DEPARTMENT OF STATISTIC



Questionnaire Form

The following form is consist of a few parts quations about a scientific MA research entitled (Comparison Logistic Regression And Discriminant Analysis In Analyzing The Classification Of Blood Cancer) your answers to the whole quatiions correctly is very important to the reserach.

1. Age of patient _____
2. Sex of patient: Male (1) _____, Female (2) _____
3. Marital status: Single _____, Married _____
4. Level of Education: Illiterate (1) _____, secondary (2), _____, High school (3) _____ institute (4) _____, University (5) _____, MA (6) _____, PHD (7) _____
5. The area where he lives (adders code) is? Erbil (1) ____ Suleimani (2) ____ Duhok (3) ____ Other (4) ____
6. Are you smoking?
Yes (1) _____, No (0) _____
7. What is the type of blood cancer?
Acute lymphocytic leukemia (ALL). (1) _____
Acute myelogenous leukemia (AML). (2) _____
Chronic lymphocytic leukemia (CLL). (3) _____
Chronic myelogenous leukemia (CML). (4) _____
8. Did you have Surgery? Yes (1) _____ No (0) _____
9. Did you have Chemotherapy? Yes (1) _____ No (0) _____
10. Did you have Radiotherapy? Yes (1) _____ No (0) _____

11. Did you have Bone marrow transplantation? Yes (1) _____ No (0) _____
12. Did you have Stem cells Transplantation? Yes (1) _____ No (0) _____
13. The spread of the tumor (Proliferation,), includes: Diffuse (1) _____ (Not Spread), (0) _____



CURRICULUM VITAE

Hawarkhan Murad HASAN hold his Bachelor's Degree in Statistics June 2016 from Salahaddin University, Iraq. He has started master of science education at Department of Statistics of Van Yuzuncu Yil University in 2019. He has been working as University of Salahaddin in 2017.





VAN YUZUNCU YIL UNIVERSITY
THE INSTITUTE OF NATURAL AND APPLIED SCIENCES
THESIS ORIGINALITY REPORT

Date: 19/1/2023

Thesis Title:

**COMPARISON LOGISTIC REGRESSION AND DISCRIMINANT ANALYSIS
IN ANALYZING THE SURVIVAL OF BLOOD CANCER**

The title of the mentioned thesis,
above having total 73 pages with cover page, introduction, main parts and conclusion, has been checked
for originality by Turnitin computer program on the date of 19/1/2023 and its detected similar rate was
15 % according to the following specified filtering

Originality report rules:

- Excluding the Materials and Methods
- Excluding the Bibliography,
- Excluding the publications obtained from the thesis,
- Excluding the text parts less than 7 words (Limit match size to 7 words)

I read the Thesis Originality Report Guidelines of Van Yuzuncu Yil University for Obtaining and Using
Similarity Rate for the thesis, and I declare the accuracy of the information I have given above and my
thesis does not contain any plagiarism; otherwise I accept legal responsibility for any dispute arising in
situations which are likely to be detected.

Sincerely yours,

19/1/2023

Name and Surname: Hawar khan Murad HASAN

Student ID# 19910001261

Department: Statistics

Program: M. Sc. ☒

Ph.D. ☐

APPROVAL OF SUPERVISOR
SUITABLE

APPROVAL OF THE INSTITUTE
SUITABLE

Prof. Dr. Fevzi ERDOĞAN