International Journal of Financial, Administrative, and Economic Sciences (IJFAES)
Vol. (4), No. (4)
IJFAES
April 2025
المجلة الدولية للعلوم المالية والإدارية والاقتصادية
الإصدار (4)، العدد (4)

Application of Machine Learning Techniques for Classifying Household Standards of Living in Sinnar State, Sudan: A Comparative Study of Discriminant Analysis and Decision Trees

Abdalrahim Ahmed Gissmalla Mustafa
Dept. of Information Management, Riyadh, Maximus KSA Company, Saudia Arabia
Ph.D. in Applied Statistics, Statistics, Sinnar University, Sudan
Abdalrhim192@gmail.com

Abstract

This study aims to classify households in Sinnar State, Sudan, into high, medium, and low standard-of-living groups based on selected economic, demographic, and social variables. Using a two-stage cluster sampling method, data were collected from 800 households across 23 administrative units. A structured questionnaire was used to gather primary data, and SPSS software was employed for statistical analysis. Discriminant analysis and decision tree models were applied to identify key factors affecting household classification and assess model accuracy. Results showed significant differences between groups, confirming the suitability of discriminant analysis. The most influential variables in the first discriminant function were income sufficiency, car ownership, and health expenditure sources. The decision tree model slightly outperformed discriminant analysis, achieving 72% classification accuracy compared to 71.6% from the discriminant model. Despite the close performance, both methods effectively distinguished household living standards.

The study concludes that advanced statistical techniques such as discriminant analysis and decision trees are useful for socio-economic classification and can support better-targeted development policies. It recommends applying these models

to guide government interventions and focusing on income-generating initiatives to enhance household welfare.

Keywords: Standard of Living Classification, Discriminant Analysis, Decision Tree Model, Socio-Economic Factors, SPSS Analysis, Income Sufficiency.

Introduction

The standard of living reflects the general well-being and socio-economic status of individuals or households, commonly influenced by factors such as income, education, health, and demographic structure. Analyzing these determinants helps governments and organizations craft effective policies to promote social development and reduce inequality. In the field of multivariate statistics, discriminant analysis is frequently used to classify observations into predefined groups based on several predictor variables (Rencher, 2002). This method is particularly useful when the aim is to understand which factors contribute most to group differentiation.

Discriminant analysis has been widely applied in social science and economic research to study phenomena where group membership is categorical, and the influencing factors are continuous or categorical. One of its advantages is that it does not require prior assumptions about the group structure, unlike other modeling techniques. Instead, it works by maximizing the separation between groups based on the linear combinations of predictor variables (Hair et al., 2010). In this study, it serves to classify households in Sinnar State, Sudan, into different standard-of-living levels and to identify the most influential factors that distinguish between them.

Understanding these factors is critical in the context of Sudan, where regional disparities in development and service provision often affect household well-being. The findings derived from discriminant analysis not only highlight the underlying socio-economic structure but also assist planners and decision-makers in designing more efficient strategies for improving living standards across different areas of the state.

Study Problem

The problem of the study is to classify households into standard of living groups with high, medium and low standard based on some economic, demographic, and social factors, so the problem of the study can be identified in the following questions:

1. How to discriminate between three groups of households according to their standard of living high, medium and low standard of living?
2. To use decision trees to classify Sinner's administrative unit based on standard of living.
3. Is the discriminate analysis batter than decion Trees?

Objectives of Study

- To classify households into three categories based on some of their economic, demographic, and social traits.
- To use decision trees to categorize Sinnar's administrative unit based on income.

Methodology of Study

Sampling Methods:

A two-stage cluster sample, known as the "double stage sample," was used to select samples from households in which the paterfamilias of Sinnar state. Firstly, the locality was considered a cluster, and all 23 administrative units of the state were included in the study. In the second stage of sampling, from each cluster (administrative unit), households were selected using simple random sampling.

Sources of Data:

The sources of data collection are dependent on preliminary data about the questionnaire for the most important factors affecting the standard of living in Sinner State.

International Journal of Financial, Administrative, and Economic Sciences (IJFAES)
Vol. (4), No. (4)
IJFAES
April 2025
المجلة الدولية للعلوم المالية والإدارية والاقتصادية
الإصدار (4)، العدد (4)

Sample Size:

The samples size for this study was determined using the statistical formula of:

The Discriminant Function for Several Groups:

For k groups (samples) with $\boldsymbol{n_i}$ observations in the ith groups, e transform each observation vector $\boldsymbol{y_{ij}}$ to obtain $z_{ij} = a'\bar{y}_{ij} = \ i = 1,2,\dots,k;\ j = \ 1,2,\dots n_i$ and find the means $\bar{z}_1 = a'\bar{y}_i$, where $\bar{y}_i = \sum_{j=1}^{n_i} \frac{y_{ij}}{n_{ij}}$. We seek the vector a that maximally separates$\bar{z}_1, \bar{z}_2 \dots, \bar{z}_k$. To express separation among$\bar{z}_1, \bar{z}_2 \dots, \bar{z}_k$, in the form

$$\frac{(\bar{z}_1 - \bar{z}_2)^2}{{s_s}^2} = \frac{[a'(\bar{y}_1 - \bar{y}_2)]^2}{a'S_{P1}a} = \frac{[a'(\bar{y}_1 - \bar{y}_2)((\bar{y}_1 - \bar{y}_2)]^2}{a'S_{P1}a}$$

To extend to k groups, we use the H matrix from MANOVA in place of $(\bar{y}_1 - \bar{y}_2)((\bar{y}_1 - \bar{y}_2)'$ and E in place of $'S_{P1}a$ to obtain

$$\lambda = \frac{a'Ha}{a'Ea}$$

Which can also be expressed as

$$\lambda = \frac{SSH(z)}{SSE(z)}$$

Where SSH (z) and SSH (z) are the between and within sum of squares for z. We can write in this form

$$a'Ha => \lambda a'\ Ea$$

$$a'(Ha - \lambda Ea) = 0$$

We examine value of $\boldsymbol{\lambda}$ and a that are solutions of (3-46) in a search for the value of a that result in maximum $\boldsymbol{\lambda}$**.** The solution a' =0 is not permissible because it gives $\boldsymbol{\lambda = 0/0}$ in Eq other solutions are found from

International Journal of Financial, Administrative, and Economic Sciences (IJFAES)
Vol. (4), No. (4)
IJFAES
April 2025
المجلة الدولية للعلوم المالية والإدارية والاقتصادية
الإصدار (4)، العدد (4)

Ha-$\boldsymbol{\lambda}$ **Ea = 0**

Which can be written in the form

$(E^{-1}H - \lambda \mathrm{I})a = 0$

The solutions of Eq are the eigenvalues$\lambda_1, \lambda_2 \ldots\ldots\ldots \lambda_s$

Associated eigenvectors $a_1, a_2 \ldots\ldots\ldots a_s$ of $E^{-1}H$. We consider them to ranked $\lambda_1, > \lambda_2 > \cdots \ldots\ldots > \lambda_s$. The number of (nonzero) eigenvalues s is the rank of **H,** which can be found as the smaller of $k-1$ or p. Thus, the largest eigenvalues λ_1 is the maximum value of $\lambda \frac{a'Ha}{a'Ea}$. And the coefficient vector that produce the maximum is the corresponding eigenvector a_1. Hence the discriminant function that maximumly separates the mean is $z_1 = a'y$; that is, z_1 represents that dimension or direction that maximally separates that means.

From that s eigenvectors $a_1, a_1 \ldots\ldots, a_s$ of $E^{-1}H$ corresponding to $\lambda_1, \lambda_2 \ldots\ldots\ldots \lambda_s$ we obtain **s** discriminant functions $z_1 = a'_1 y, z_2 = a'_2 y, z_s = a'_s y$, which show the dimension or directions of difference among $\bar{y}_1, \bar{y}_2, \ldots. \bar{y}_k$.Theses discriminant functions are uncorrelated, but they are orthogonal ($a'_i . a'_j = 0 \; for \; i \neq j$ because $E^{-1}H$id not symmetric.

The relative importance of each discriminant function z_1 can be assessed by it eigenvalue as a proportion of the total:

$$\frac{\lambda}{\sum_{j=1}^{s} \lambda}$$

By this criterion, two or three discriminant functions will often suffice to describe the group difference. The discriminant functions associated with small eigenvalue can be neglected.

International Journal of Financial, Administrative, and Economic Sciences (IJFAES)
Vol. (4), No. (4)
IJFAES
April 2025
المجلة الدولية للعلوم المالية والإدارية والاقتصادية
الإصدار (4)، العدد (4)

Test of Significance for the Several – Group Case:

In order to hypotheses, we need the assumption of multivariate normality. This was not explicitly required for development of discriminant functions.

A test of significance for each discriminant function is also available we use the Wilks Λ -test for significant difference among mean vectors.

$$\Lambda_I = \prod_{i=1}^{s} \frac{1}{1+\lambda_1}$$

Which is distributed as $\Lambda_{p,k-1,N-k}$, where $N = \sum_i n_i$ for an unbalanced design or $N = kn$ in the balanced case. Since Λ_I is small if one or more λ_i 's are large, Wilks' Λ tests for significance of the eigenvalues and thereby for discriminant functions. The s eigenvalues represented s dimensions of separation of thee mean vectors $\bar{y}_1, \bar{y}_2, \ldots\ldots\ldots \bar{y}_k$ .we are interested in which, if any, of these dimensions are significant. In the context of discriminant functions, Wilks' Λ is more useful than then other three MONOVA test statistics, because it can be used on subset eigenvalues.
In addition to the exact test provided by critical value for Λ we can use the χ^2 –approximation for Λ_I given by:

$$\chi^2 = -\left[N - 1 - \frac{1}{2}(p+k)\right]\sum_{i=1}^{s} \ln(1+\lambda_i)$$

Which is approximately $\boldsymbol{\chi^2}$ with $p(k-1)$ degrees of freedom. The test statistic Λ_I and its approximation (3-55).

And test the significance of $\lambda_{2,}\lambda_3, \ldots\ldots\lambda_s$, we delete λ_1 from Wilks' Λ and the associated $\boldsymbol{\chi^2}$ -approximation to obtain

International Journal of Financial, Administrative and Economic Sciences, London Vol (4), No (4), 2025
https://doi.org/10.59992/IJFAES.2025.v4n4p12 E-ISSN 2977-1498

International Journal of Financial, Administrative, and Economic Sciences (IJFAES)
Vol. (4), No. (4)
IJFAES
April 2025
المجلة الدولية للعلوم المالية والإدارية والاقتصادية
الإصدار (4)، العدد (4)

$$\Lambda_2 = \prod_{i=2}^{s} \frac{1}{1+\lambda_1}$$

$$v_2 = -\left[N - 1 - \frac{1}{2}(p+k)\right] \ln \Lambda_2 = [N - 1 - \frac{1}{2}(p+k)] \sum_{i=2}^{s} \ln(1 + \lambda_i)$$

Which is approximately χ^2 with $(p-1)(k-2)$ degree of freedom. If this test lead to rejection of H_0, we conclude that at leastλ_2 is significant along with the associated discriminant function $z_2 = a'_2 y$. We can continue in this fashion, testing each λ_i in turn a test fails to reject H_0

Stepwise Selection of Variables:

In many applications, a larger number of dependent variables is available and the experimenter would like to discard those that are redundant (in the presence of the other variables) for separating the groups. Stepwise is limited to procedures that delete or add variables one at a time. We emphasize that we are selecting dependent variables (**y's**), and there for the basic model (one-way MANOVA) does not change.in subset selection in regression, on the other hand, we select independent variables with a consequent alteration of the model.

If there are no variables for which we have a priori interest in testing for significance, we can do a data-directed search for variables that best separate the groups. Such a strategy is often called stepwise discriminant analysis, although it could more aptly be called stepwise MANOVA.

Classification Rules:

To develop a classification rule for an observation **y** for multiple group case involves $\pi_1, \pi_2 \dots \dots, \pi_k$ populations and $f_i(\boldsymbol{y})$ for $(i = 1,2, \dots k)$ probability density functions.

Furthermore, we have $p_1, p_2 \dots p_k$ prior probabilities that an observation is form

International Journal of Financial, Administrative and Economic Sciences, London Vol (4), No (4), 2025
https://doi.org/10.59992/IJFAES.2025.v4n4p12 E-ISSN 2977-1498

International Journal of Financial, Administrative, and Economic Sciences (IJFAES)
Vol. (4), No. (4)
IJFAES
April 2025
المجلة الدولية للعلوم المالية والإدارية والاقتصادية
الإصدار (4)، العدد (4)

population π_1 where the $\sum p_i = 1$. The cost of assigning an observation to population π_i where it belongs to π_i is represented as $C(\boldsymbol{j}|\ \boldsymbol{i})$ for $j = 1,2, \dots, k$. Finally, we let $P(\boldsymbol{j}|\ \boldsymbol{i})$ denote the probability of classifying an observation into π_i given that it should be in π_i.

Then the $P(j|\ i)$ =1-$\sum_{j=1}^{k} P(\boldsymbol{j}|\ \boldsymbol{i})$ $for\ i \neq j$. With notation for the k group case, the total probability of misclassification (TPM) and the expected cost of misclassification become

$$TPM = \sum_{i=1}^{k} p_i \left[\sum_{\substack{j=1 \\ j \neq 1}}^{k} P(j|\ i) \right]$$

$$ECM = \sum_{i=1}^{k} p_i \left[\sum_{\substack{j=1 \\ j \neq 1}}^{k} C(j|\ i) P((j|\ i) \right]$$

Assuming known probability density functions, the optimal rule for classifying an observation **y** into one k populations that minimizes the TPM is to allocate **y** to π_i if

$$p_i\ f_i(\boldsymbol{y}) > p_i f_i\ (y) for\ all\ j = 1,2 \dots k$$

So that $\boldsymbol{p_i}\ \boldsymbol{f_i}$ is maximum. If all $\boldsymbol{p_i}$ are equal, the rule is called the maximum likelihood rule since $f_i(\boldsymbol{y})$ is likelihood for an observation y using Bayes' theorem, it is also seen to be equivalent to assigning observations based upon a maximum posterior probability assuming $\pi_i = N_p(\mu_i, \Sigma)$ and letting

$$D_i^2(y) = (y - \bar{y}_i)' S^{-1} (y - \bar{y}_i)$$

The rule in (3-69) becomes: Assig y to π_i for which

International Journal of Financial, Administrative and Economic Sciences, London Vol (4), No (4), 2025
https://doi.org/10.59992/IJFAES.2025.v4n4p12 E-ISSN 2977-1498

$$L_i\ (y) = \ -D_i^2(y)/2 + LogP_i)$$

Is a maximum. This is easily established by evaluating the log $P_i f_i(\boldsymbol{y})$, ignoring constant terms, and estimating $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\iota}}$ with $\bar{\boldsymbol{y}}_{\boldsymbol{i}}$ and $\boldsymbol{\Sigma}$ by the unbiased common estimate **S**. equivalently, one may assign an observation to π_i to the population with the maximal posterior probability

$$\boldsymbol{P}(\boldsymbol{\pi_i}|\mathbf{y}) = \frac{p_i e^{-D_i^2(y)/2}}{\sum_{j=1}^{k} p_j e^{-D_i^2(y)/2}}$$

Given the all, the populations are multivariate normal. With common covariance matrix $\boldsymbol{\Sigma}$ and equal prior probabilities $\boldsymbol{p_i}$, classification based upon a maximum value of $-\boldsymbol{D_i^2}(\boldsymbol{y})$ or that **y** is closest to $\bar{y}_i$ is equivalent to classifying an observation based upon Fisher's **S** discriminant functions using the eigenvectors of $|\boldsymbol{H} - \boldsymbol{\lambda E}| = \boldsymbol{0}$. **(**Richard A Johnson, Dean W. Wichern2007**)**

Classification and Regression Trees (CART):

Classification and Regression Trees (CART) is a relatively new method of data analysis developed by a group of American statisticians (Breiman et al., 1984). The aim of CART is to classify observations into a subset of Known classes or to predict levels of regression functions CART is a non- parametric tool which is designed to represent decision rules in a form of so called binary trees binary trees spit a learning sample parallel to the coordinate axis and represent the resulting data clusters hierarchically starting from a root node for the whole learning sample itself and ending with relatively homogenous buckets of observations.

Regression trees are constructed in a similar way but the final buckets do not represent classes but rather approximations to an unknown regression function at a particular point of the independent variable. In this sense, regression trees are estimates via a non-parametric regression model.

International Journal of Financial, Administrative, and Economic Sciences (IJFAES)
Vol. (4), No. (4)
IJFAES
April 2025
المجلة الدولية للعلوم المالية والإدارية والاقتصادية
الإصدار (4)، العدد (4)

CART Measures:

A more formal framework on how spilt and where to split needs to be developed. Suppose there are n observations in the learning sample and n_j is the overall number of observations belonging to class j, j = 1,… , J. The class probabilities are:

$$\pi(j) = \frac{n_j}{n}, j = 1, \dots, J$$

$\pi(j)$ is the proportion of observations belonging to a particular class. Let n (t) be the number of observations at node t and n_j(t) the number of observations belonging to the j-th class at t. The frequency of the event that an observation of the j-th class falls into node t is:

$$p(j,t) = \pi(j)\frac{n_{j(t)}}{n_j}$$

The proportion of observations at t are $p(t) = \sum_{j=1}^{j} p(j,t)$ he conditional probability of an observation to belong to class j given that it is at node t is:

$$p(j|t) = \frac{p(j,t)}{p(t)} = \frac{n_j(t)}{n(t)}$$

Define now a degree of class homogeneity in a given node. This characteristic a CART measure I (t) – will represent a class homogeneity indicator for a given tree node and hence will help to find optimal splits. Define an impurity function ι (t) which is determined on $(p1, \dots \dots pj) \in [0,1]^j \; with \sum_{j=1}^{J} p_{j=1}$ so that:

1. ι has a unique maximum at point $\left(\frac{1}{J}, \frac{1}{J}, \dots \dots \frac{1}{J}\right)$
2. ι has a unique minimum at points (1,0,0,…..0),(0,1,0,…0),..(0,0,0,…1)

International Journal of Financial, Administrative, and Economic Sciences (IJFAES)
Vol. (4), No. (4)
IJFAES
April 2025
المجلة الدولية للعلوم المالية والإدارية والاقتصادية
الإصدار (4)، العدد (4)

3. ι is a symmetric function of $p1, \ldots . pj$

Each function satisfying these conditions is called an impurity function. Given ι, define the impurity measure I (t) for a node t as:

$$i(t) = t\{p(1|t), p\ (2|t), \ldots . . p(J|t)\} \ldots$$

Denote an arbitrary data split by s, then for a given node t which we will call a parent node two child nodes arise t_l and t_R representing observations meeting and not meeting the split criterion s. a fraction p_L of data from t falls to the left child node and p_R =1-p_L is the share of data in t_R

A quality measure of how well split s works is:

$$\Delta_i(s,t) = i(t) - P_l i(t_l) - p_{R^i}(t_R) \ldots \ldots$$

The higher the value of $\Delta_i(s,t)$ he better split we have since data impurity is reduced. In order to find an optimal split s it is natural to maximize$\Delta_i(s,t)$for different splits s, the value of i(t) remains constant, hence it is equivalent to find. (W. HArdle.Simar, 2007).

Data Analysis and Discussion

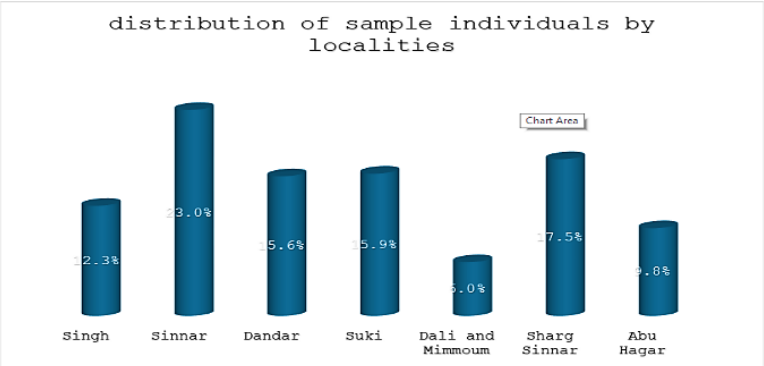

Figure (1): distribution of sample individuals by localities

International Journal of Financial, Administrative and Economic Sciences, London Vol (4), No (4), 2025
https://doi.org/10.59992/IJFAES.2025.v4n4p12 E-ISSN 2977-1498

International Journal of Financial, Administrative, and Economic Sciences (IJFAES)
Vol. (4), No. (4)
IJFAES
April 2025
المجلة الدولية للعلوم المالية والإدارية والاقتصادية
الإصدار (4)، العدد (4)

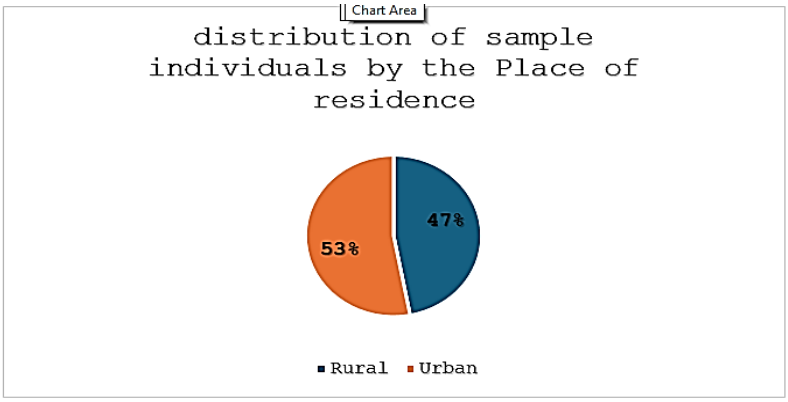

Figure (2): distribution of sample individuals by the Place of residence

The table above shows that 53% of the respondents are from urban areas and 47% are from rural areas.

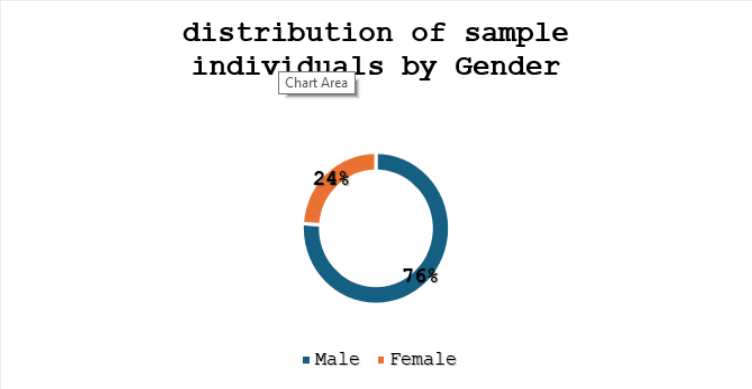

Figure (3): distribution of sample individuals by Gender

The table above shows that 76% of the respondents were male, and 24% were female.

International Journal of Financial, Administrative and Economic Sciences, London Vol (4), No (4), 2025
https://doi.org/10.59992/IJFAES.2025.v4n4p12 E-ISSN 2977-1498

International Journal of Financial, Administrative, and Economic Sciences (IJFAES)
Vol. (4), No. (4)
IJFAES
April 2025
المجلة الدولية للعلوم المالية والإدارية والاقتصادية
الإصدار (4)، العدد (4)

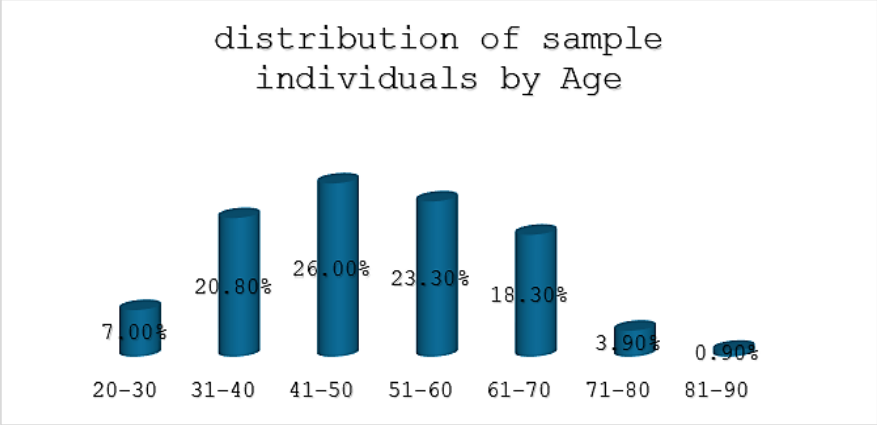

Figure (4): distribution of sample individuals by Age

The table above shows that (26%) of the sample between (41-50) years, (23.3%) of the respondents between (51-60) years, (20.8%) of the respondents between (31-40) years, (18.3%) of the respondents between (61-70) years, (7%) of the respondents between (20-30) years, (0.9%) of the respondents between (81-90) years, The results show that the majority of the samples in the age groups (41–50), (51–60) The survey's goal was to question paterfamilias, and we can see in our society that the majority of paterfamilias their age are in these groups.

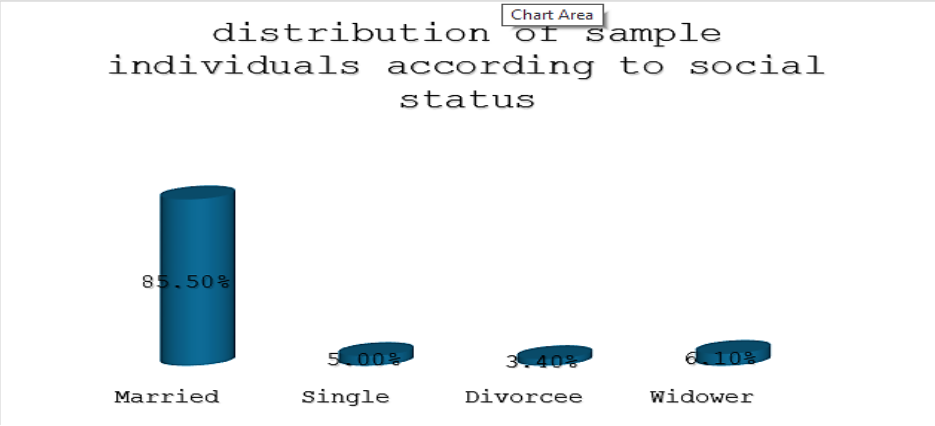

Figure (5): distribution of sample individuals according to social status

International Journal of Financial, Administrative and Economic Sciences, London Vol (4), No (4), 2025
https://doi.org/10.59992/IJFAES.2025.v4n4p12 E-ISSN 2977-1498

International Journal of Financial, Administrative, and Economic Sciences (IJFAES)
Vol. (4), No. (4)
IJFAES
April 2025
المجلة الدولية للعلوم المالية والإدارية والاقتصادية
الإصدار (4)، العدد (4)

The table above shows that the majority of the sample is married at 85.5%, while the proportion of unmarried was 5%, and the proportion of divorced and widowed was 9.5%.

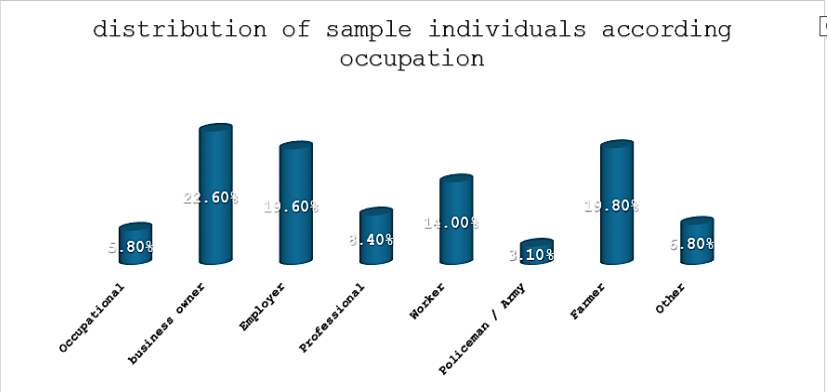

Figure (6): distribution of sample individuals according to occupation

The table above shows that the distribution of sample according to the occupation, 22.6% of the

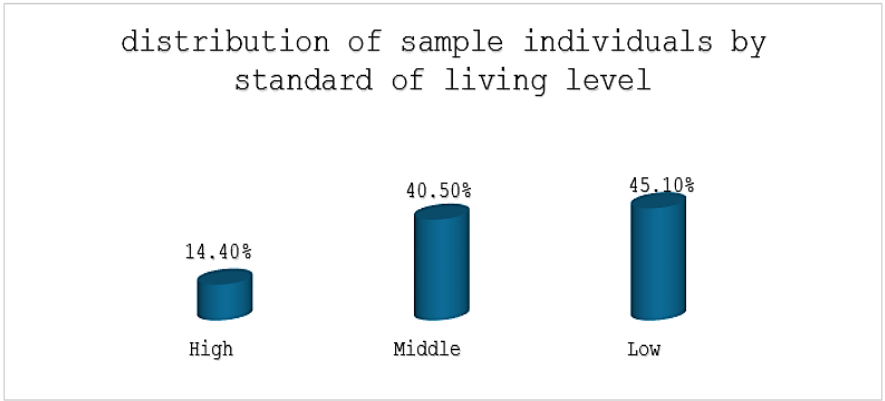

Figure (7): distribution of sample individuals by standard of living level

The table above shows that 45.1% of households are at a low standard of living, 40.5% of households are at a medium level of living, and only 14.4% of households are at a high standard of living.

https://doi.org/10.59992/IJFAES.2025.v4n4p12 E-ISSN 2977-1498

International Journal of Financial, Administrative, and Economic Sciences (IJFAES)
Vol. (4), No. (4)
IJFAES
April 2025
المجلة الدولية للعلوم المالية والإدارية والاقتصادية
الإصدار (4)، العدد (4)

Table (1): Tests of Equality of Group Means

Variables	Wilks' Lambda	F	df1	df2	Sig.
Place of residence	.949	21.210	2	797	.000
Gender	.993	2.781	2	797	.063
Educational level	.995	1.945	2	797	.144
Occupation	.960	16.407	2	797	.000
The number of less family members from 15 years	.999	.540	2	797	.583
Family type	.990	4.169	2	797	.016
Type of housing ownership?	1.000	.026	2	797	.974
What kind of house does the family live in?	.921	33.974	2	797	.000
How many rooms are in the house?	.947	22.187	2	797	.000
What type of house roof?	.881	53.968	2	797	.000
What type of house floor?	.890	49.179	2	797	.000
What is the main source of cooking fuel?	.969	12.787	2	797	.000
What type of toilet?	.926	31.851	2	797	.000
What is the main source of drinking water for the family?	.996	1.780	2	797	.169
What main source of water does the family use to cook food?	.998	.707	2	797	.494
How is drinking water saved inside the house?	.969	12.679	2	797	.000
Car	.845	72.854	2	797	.000
Refrigerator	.938	26.291	2	797	.000
air conditioner	.898	45.293	2	797	.000
freon air conditioner	.977	9.194	2	797	.000
computer/ laptop	.943	24.197	2	797	.000
smart screen	.969	12.628	2	797	.000
I-pad	.978	9.057	2	797	.000
monthly household income	.830	81.898	2	797	.000
Is the income of the head of household sufficient for living expenses?	.544	333.616	2	797	.000
Food	.992	3.098	2	797	.046
Education	.976	9.762	2	797	.000
personal needs	.963	15.209	2	797	.000
Do you resort to permanent borrowing to provide the living expenses for the family?	.829	82.154	2	797	.000
House	.979	8.490	2	797	.000
Land	.924	32.701	2	797	.000
Store	.913	38.166	2	797	.000
How many friends do you communicate with at least once a month?	.993	2.655	2	797	.071
Do the family use the family planning methods?	.978	9.140	2	797	.000
Which kind of methods are used in family planning?	.997	1.002	2	797	.368
If you feel sick, break or injured, where do you go?	.998	.633	2	797	.531
How far is the nearest hospital or health centre to your home?	.977	9.244	2	797	.000
What are the sources of health spending?	.960	16.718	2	797	.000
Has health care effected family spending in the last 12 months?	.909	39.962	2	797	.000
How many times a mother has received health care during pregnancy in the last 12 months?	.982	7.143	2	797	.001

Wilk's lambda (also called the U statistic) and the F ratio are used to assess whether group means differ significantly for a given variable, based on one-way ANOVA. Wilk's lambda ranges from 0 to 1, with lower values indicating greater differences between group means. A lambda value below 0.95 suggests significant group differences. Variables with the lowest Wilk's lambda is most important in the discriminant function. In this case, the most critical variable is whether the household head's income is sufficient for living expenses.

Table (2): discriminant functions

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.413	700.031	24	.000
2	.869	111.042	11	.000

The table above shows that the Wilks' Lambda for the tests of functions 1 through 2 using Chi-square, the chi-square test = 700.031, DF = 24, and significant = 0.000, which is less than the level of significance of 0.05. The Wilks' Lambda for the test of function 2 using Chi-square is 111.042, DF = 111.042, and significant = 0.000, which is less than the level of significance of 0.05. This means the two discriminant functions are statistically significant.

Table (3): discriminant eigenvalue and the canonical correlation

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1.107a	87.4	87.4	.725
2	.159a	12.6	100.0	.371

The table presents eigenvalues for two discriminant functions, which measure how well the independent variables explain variance in the dependent variable. Function 1 has an eigenvalue of 1.0107 and accounts for 87.4% of the variance, making it more significant than Function 2, which has an eigenvalue of 0.159 and explains 12.6% of the variance. The canonical correlation, indicating the strength of the relationship between predictors and groups, is higher for Function 1 (0.725) than for Function 2 (0.371), further highlighting the greater importance of Function 1.

International Journal of Financial, Administrative, and Economic Sciences (IJFAES)
Vol. (4), No. (4)
IJFAES
April 2025
المجلة الدولية للعلوم المالية والإدارية والاقتصادية
الإصدار (4)، العدد (4)

Table (4): Canonical Discriminant Function Coefficients

Variables	Function	
	1	2
Place of residence (X1)	-.265	.606
occupation(X2)	.004	.147
Is there any family member who works other than the head of the family? (X3)	.126	.827
What type of house floor? (X4)	.069	-.312
car(X5)	.541	-.906
air conditioner(X6)	.297	.160
monthly household income(X7)	-.003	.006
Is the income of the head of household sufficient for living expenses? (X8)	2.076	.976
personal needs(X9)	.396	-.224
house(X10)	.112	-.744
What are the sources spending on health? (X11)	-.002	-.401
Has health care effected the spending of family in the last 12 months? (X12)	-.320	-.263
(Constant)	-4.759	.162

The table shows that the unstandardized discriminant coefficients for the variables entered in analysis these functions are:

$$Z_1 = -4.759 - .265X_1 + .004X_2 + .126X_3 + .069X_4 + .541X_5 - .003X_6 + .297X_7 + 2.076X_8 + .396X_9 + .112X_{10} - .002X_{11} - .320X_{12}$$

$$Z_2 = .162 + .606X_1 + .147X_2 + .827X_3 - .312X_4 - .906X_5 + .160X_6 + .006X_7 + .976X\_8 - .224X\_9 - .744X_{10} - .401X_{11} - .263X_{12}$$

Table (5): classification Table

What is evaluation of the standard of living for your family?			Predicted Group Membership			Total
			high	middle	low	
Original	Count	High	88	25	2	115
		Middle	67	182	75	324
		Low	11	47	303	361
	%	High	76.5	21.7	1.7	100.0
		Middle	20.7	56.2	23.1	100.0
		Low	3.0	13.0	83.119	100.0
a. 71.6% of original grouped cases correctly classified.						

International Journal of Financial, Administrative and Economic Sciences, London Vol (4), No (4), 2025
https://doi.org/10.59992/IJFAES.2025.v4n4p12 E-ISSN 2977-1498

International Journal of Financial, Administrative, and Economic Sciences (IJFAES)
Vol. (4), No. (4)
IJFAES
April 2025
المجلة الدولية للعلوم المالية والإدارية والاقتصادية
الإصدار (4)، العدد (4)

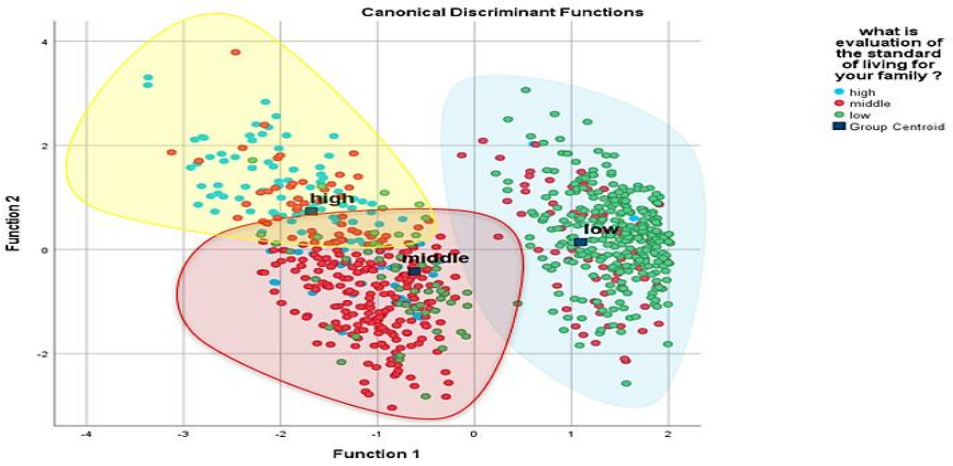

Figure (8): canonical discriminant function

The discriminant analysis correctly classified 71.6% of the 800 households into high, middle, or low standards of living—accurately identifying 573 households, while 227 were misclassified. It performed best in identifying low standard of living households, correctly classifying 83.9% (303 out of 361). For high standard of living, 76.5% (88 out of 115) were correctly classified. The accuracy was lowest for the middle standard of living, with only 56.2% (182 out of 324) correctly identified.

Decision Trees (CHAID):

The decision tree procedure offers several different methods for creating tree models, the CHAID has been used in the research.

CHAID is an abbreviation for "Chi-squared Automatic Interaction Detection." At each step, CHAID chooses the independent (predictor) variable that has the strongest interaction with the dependent variable. The categories of each predictor are merged if they are not significantly different with respect to the dependent variable.

International Journal of Financial, Administrative, and Economic Sciences (IJFAES)
Vol. (4), No. (4)
IJFAES
April 2025
المجلة الدولية للعلوم المالية والإدارية والاقتصادية
الإصدار (4)، العدد (4)

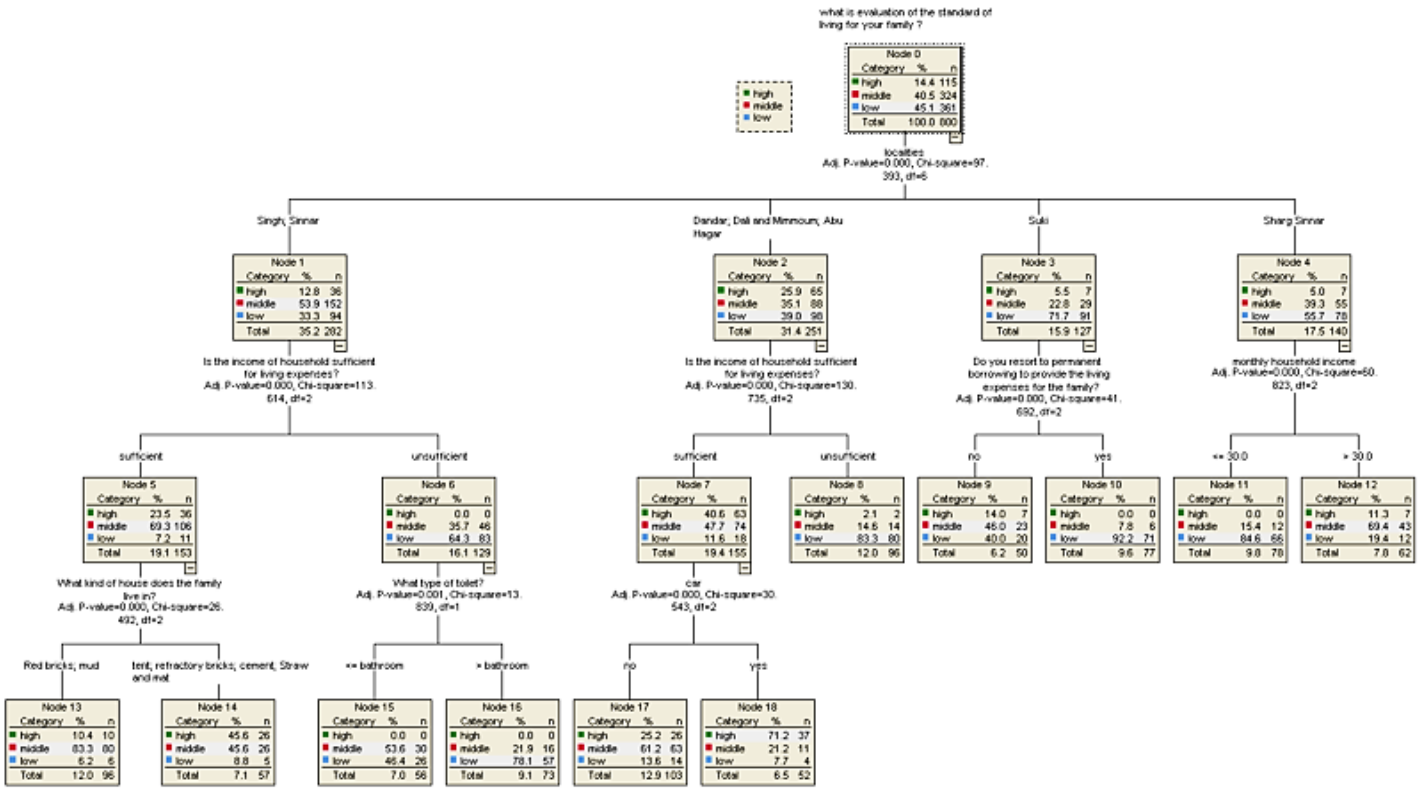

Figure (9): Tree diagram for standard of living model

The tree diagram represents how Sinnar localities are divided into four main nodes:

- Node 1: Singh and Sinnar.
- Node 2: Dender, Dali, Mithmom, and Abu Hgar.
- Node 3: Suki.
- Node 4: Sharq Sinnar.

At each node, the most important predictors of standard of living are identified:

- Node 1 & 2: The most significant predictor is whether household income is sufficient for expenses (Chi-square = 113.614 and 130.735, p-value = 0.000). This

splits the nodes into "sufficient" and "insufficient" income groups.

- Node 3: The key predictor is whether households borrow to cover expenses (Chi-square = 41.692, p-value = 0.000), dividing into "yes" and "no."
- Node 4: Monthly income is the strongest predictor (Chi-square = 60.823, p-value = 0.000), splitting into "income < 30" and "income > 30."

Subsequent splits include:

- Node 5: Split by house type (red bricks vs. tents/other materials).
- Node 6: Split by toilet type.
- Node 7: Split by car ownership.

Nodes 8 to 12 were not further divided due to lack of statistically significant relationships.
The risk and classification tables assess how effectively the model classifies the data.

Table (6): Risk estimation

Estimate	Std. Error
.280	.016

Risk estimate of 0.28 indicates that the category predicted by the model (low standard of living, medium standard of living, and high standard of living) is wrong for 28% of the cases. As a result, the chance of misclassifying a standard of living is 28%.

Table (7): Classification

Classification				
Observed	Predicted			
	High	middle	low	Percent Correct
High	37	76	2	32.2%
Middle	11	265	48	81.8%
Low	4	83	274	75.9%
Overall Percentage	6.5%	53.0%	40.5%	72.0%

The table shows that the model correctly classifies 72% of the standard of living of a household.

The model could be able to classify 81.8% of households that belong to the medium standard of living. That means it succeeded in classification in 265 households and filed in 59 households.

The model could be able to classify 75.9% of households that belong to the low standard of living. That means it succeeded in the classification of 274 households and filed in 57 households.

The model could be able to classify 32.2% of households that belong to the low standard of living. That means it succeeded in classification in 37 households and filed in 78 households.

Results

1. There are statistically significant differences between independent variables in the three groups using the F test. This means the use of discriminating functions in classification.
2. The discriminate functions are statistically significant where they have been tested using the Chi-square test, meaning the two functions are able to distinguish households in Sinnar state into groups with a standard of living (high, middle, and low).
3. The result showed that there are three variables in the first function that have the greatest impact on the discrimination of living levels (is the income sufficient for living expenses, having a car, and what are the sources of health spending).
4. The results revealed that the Decision Tree model achieved a slightly higher classification accuracy for new observations, with an accuracy of 72%, compared to the Discriminant Analysis model, which achieved an accuracy of 71.6%. The performance of both models was relatively close.

Recommendation

1. Apply the discriminant function developed in this study to classify households according to their standard of living, enabling the government to design and implement targeted income-enhancing projects for each category.
2. Taking advantage of advanced statistical methods from discriminate analysis, logistics models, and classification trees to classify between more than two groups in all areas of knowledge.
3. Creating new jobs to achieve an abundance in income rather than depleting enterprises, so as to be able to cover the monthly spending on food, housing, and health.

References

1. Afifi, A., & May, S. (2012). Practical multivariate analysis (5th ed.). Taylor & Francis Group, LLC.
2. Afifi, A., May, S., Donatello, R. A., & Clark, V. A. (2020). Practical multivariate analysis (6th ed.). Taylor & Francis Group, LLC.
3. Ahmed. (2014). Classification of Syrian provinces by household consumption using cluster analysis. Tishreen University Journal - Economic and Legal Sciences Series, 37(2).
4. Hamad, A. K. S. (2018). Classification of the Iraqi provinces of some transitional diseases using measures (CCC, Delta) in the cluster analysis. University of Salah Eddin, Department of Scientific Publishing, 22(5), 187–206.
5. Izenman, A. J. (2008). Modern multivariate statistical techniques: Regression, classification, and manifold learning. Springer-Verlag.
6. Everitt, B., & Hothorn, T. (2011). An introduction to applied multivariate analysis with R. Springer-Verlag New York.
7. Huberty, C. J., & Olejnik, S. (2006). Applied MANOVA and discriminant analysis (2nd ed.). Wiley-Interscience.

8. Czepiel, S. A. (2002). Maximum likelihood estimation of logistic regression models: Theory and implementation. Retrieved from http://czep.net
9. Al Mekhlafi, F. A. I. (2019). Classification and discrimination of Yemeni provinces by sources of income using cluster analysis and discriminant analysis. Taiz University Research Journal, 19.
10. Johnson, R. A., & Wichern, D. W. (2007). Applied multivariate statistical analysis (6th ed.). Prentice Hall.
11. Ho, R. (2006). Handbook of univariate and multivariate data analysis and interpretation with SPSS. Taylor & Francis Group, LLC.
12. Ahmad, Z., & Ejaz, Z. (2011). Classification of households with respect to poverty by using cluster analysis. In ICCS-11, Lahore, Pakistan.