Global Journal of Economics and Business (GJEB)

Volume.13, Issue.2 pp: 176-185
*April (2023)*

Article 6

# Using the Discriminant Analysis to Classify the Income of Households in Sinnar State, Sudan (2021)

**Abdalrahim Ahmed Gissmalla, Adel Ali Ahmed**

| Accepted<br>قبول البحث | Revised<br>مراجعة البحث | Received<br>استلام البحث |
|---|---|---|
| 2023/4/4 | 2023 /3/12 | 2023 /2/27 |

DOI: https://doi.org/10.31559/GJEB2023.13.2.6

رفـــاد REFAAD
www.refaad.com

*Global Journal of Economics and Business (GJEB)*

المجلة العالمية للاقتصاد والأعمال

Journal Homepage: https://www.refaad.com/Journal/Index/2

E-ISSN 2519-9293 | P-ISSN 2519-9285


# Using the Discriminant Analysis to Classify the Income of Households in Sinnar State, Sudan (2021)

# استخدام التحليل التمييزي لتصنيف دخل الأسر بولاية سنار- السودان (2021)

**Abdalrahim Ahmed Gissmalla[1], Adel Ali Ahmed[2]**

**عبد الرحيم أحمد قسم الله [1]، عادل علي أحمد[2]**

[1] Lecturer of Statistics, Department of Statistics and Econometrics, University of Sinnar, Sudan

[2] Professor of Statistics, Department of Applied Statistics and Demography, University of Gezira, Sudan

[1] أستاذ محاضر- تخصص الإحصاء التطبيقي- قسم الإحصاء والاقتصاد القياسي- جامعة سنار- السودان

[2] بروفسيور- تخصص الإحصاء التطبيقي- قسم الإحصاء والتطبيقي والديمغرافيا- جامعة الجزيرة- السودان

[1] Abdalrhim192@gmail.com

**Abstract:**

The purpose of this study was to distinguish between sufficient and insufficient income and to identify the most discriminating factors that influence income. The data was obtained from households in Sinnar through a structured questionnaire addressed to the heads of families, a sample of (800) households (417) had sufficient incomes, and (383) had insufficient incomes. Discriminate analysis and decision trees were applied with the help of the (SPSS) program. The results suggested that the discrimination model applied had a good fit with the data obtained from the sample and that 7 of the 24 variables used in discrimination were statistically significant. The most important discriminating variables were the evaluation of the standard of living and borrowing to cover the family's living expenses. The research showed that the possible error in discriminate function model specificity does not exceed 14.2% compared to decision trees where the possible error does not exceed 14.5%. The research study recommended the use of a statistical discrimination model to discriminate between a sufficient income and insufficient income and the use of decision trees to classify the administrative unit of Sinnar according to income.

***Keywords:*** *discriminant; decision trees; Income; classification; Sinnar state.*

**الملخص:**

الغرض من هذه الدراسة هو التمييز بين الدخل الكافي وغير الكافي وتحديد العوامل الأكثر تمييزًا التي تؤثر على دخل الأسر. تم جمع البيانات من الأسر في ولاية سنار من خلال استمارة استبيان موجه إلى أرباب الأسر، حيث تم اختيار عينة من (800) أسرة (417) لديها دخل كاف و (383) لديها دخل غير كاف. تم استخدام التحليل التمييزي وشجرة القرار أو ما تسمى بشجرة الانحدار باستخدام برنامج (SPSS). أوضحت النتائج أن نموذج التمييزي المطبق كان ملائمًا بشكل جيد للبيانات التي تم الحصول عليها من العينة، وأن 7 متغيرات من 24 متغير من المتغيرات المستخدمة في الدالة التمييزية كانت ذات دلالة إحصائية حيث كانت أهم المتغيرات التمييزية هي تقييم مستوى المعيشة، الاقتراض لتغطية نفقات معيشة الأسرة كما أظهرت نتائج الدراسة أن الخطأ المحتمل في الدالة التمييزية لا يتجاوز 14.2% مقارنة بأشجار القرار (شجرة الانحدار) حيث لا يتجاوز الخطأ المحتمل 14.5%. أوصت الدراسة باستخدام نموذج التمييز الإحصائي للتمييز بين فئات الدخل الكافي والدخل غير الكافي واستخدام أشجار القرار (شجرة الانحدار) لتصنيف الوحدة الإدارية لسنار حسب الدخل.

***الكلمات المفتاحية:*** *التمييز؛ شجرة القرار؛ الدخل؛ التصنيف؛ ولاية سنار.*

## Introduction:

Discrimination and classification are multivariate techniques concerned with separating distinct sets of objects (or observations) and allocating new objects (observations) to previously defined groups. Discriminant analysis is rather exploratory. As a separative procedure, it is often employed on a one-time basis in order to investigate observed differences when causal relationships are not well understood. Classification procedures are less exploratory in the sense that they lead to well-defined rules that can be used for assigning new objects. Classification ordinarily requires more problem structure than discrimination does. The first goal of discrimination and classification is to describe, either graphically (in three or fewer dimensions) or algebraically, the differential features of objects (observations) from several known collections (populations). We try to find "discriminants" whose numerical values are such that the collections are separated as much as possible. And the second goal is to sort objects (observations) into two or more labeled classes. The emphasis is on deriving a rule that can be used to optimally assign new objects to the labeled classes (Johnson & Wichrn,2007).

We shall follow convention and use the term "discrimination" to refer to Goal One. This terminology was introduced by RA Fisher in the first modern treatment of separating problems. A more descriptive term for this goal, however, is separation. We shall refer to the second goal as classification or allocation. A function that separates objects may sometimes serve as an allocator, and, conversely, a rule that allocates objects may suggest a discriminatory procedure. In practice, goals one and two frequently overlap, and the distinction between separation and allocation becomes blurred.

### Statement of study:

The problem of the study is to classify households into income groups with sufficient income and insufficient income based on some economic, demographic, and social factors, so the problem of the study can be identified in the following questions:

- How to discriminate between two groups of households: those with sufficient income for living expenses and those with insufficient income?
- To use decision trees to classify Sinner's administrative unit based on income.

### Objectives of study:

- To categorize households into two income groups 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.

### Sample size:

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

$$n_0 = \frac{Z^2 p \times (q)}{d^2}$$

Where:
$n_0 \equiv$ The required sample size.
$p \equiv$ The proportion of households (assumed the income is sufficient in the household is 50%).
$Z \equiv$ The standard score corresponds to a 95% confidence level (and is thus equal to 1.96).
d ≡ The margin of error (estimated at 5%).
With a design effect of (2) for the multistage nature of cluster sampling, accordingly, the sample size for the study was (800) households.

## Discriminant analysis:

Discriminant analysis techniques are used to classify individuals into one of two or more alternative groups (or populations) based on a set of measurements. The populations are known to be distinct, and each individual

belongs to one of them. These techniques can also be used to identify which variables contribute to making the classification. Thus, as in regression analysis, we have two uses: prediction and description (Donatello, 2020).

The first is description of group separation, in which linear functions of the variables (discriminant functions) are used to describe or elucidate the difference between two or more groups. The goals of descriptive discriminant analysis include identifying the relative contribution of $P$ variables to the separation of the groups and finding the optimal plane one which the points can be projected to best illustrate the configuration of groups.

The second is Prediction or allocation of observations to groups, in which linear or quadratic functions of the variables (classification functions) are employed to assign an individual sampling into one of the groups. The measured values in the observation vector for an individual or object are evaluated by classification functions to find the groups to which the individual most likely belongs (Rencher, 2002).

**Classification with two multivariate normal populations:**

The sample mean vectors and covariance matrices are determined by:

$$\underset{(P\times 1)}{\bar{X}_1} = \frac{1}{n_1}\sum_{j=1}^{n_i} X_{ij}\,, \qquad S_{(P\times p)} = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(X_{1j}-\bar{X}_1)\,(X_{1j}-\bar{X}_1)'$$

$$\underset{(P\times 1)}{\bar{X}_1} = \frac{1}{n_2}\sum_{j=1}^{n_i} X_{ij}\,, \qquad S_{(P\times p)} = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(X_{2j}-\bar{X}_2)\,(X_{2j}-\bar{X}_2)'$$

Since it is assumed that the percent populations have the same covariance matrix $\Sigma$, the sample covariance matrices $S_1$ and $S_2$ are combined (pooled) to derive a single, unbiased estimate of $\Sigma$.

$$S_{pooled} = \left[\frac{n_1-1}{(n_1-1)+(n_2-1)}\right]S_1 + \left[\frac{n_2-1}{(n_1-1)+(n_2-1)}\right]S_2$$

Is an unbiased estimate of $\Sigma$ if the data matrices $X_1$ and $X_2$ contain random samples from the populations $\pi_1$ and $\pi_2$ respectively.

Fisher's linear discriminant function for two groups is:

$$\hat{y} = \hat{a}'x = (\bar{x}_1 - \bar{x}_2)'S^{-1}{}_{pooled}\,x$$

This linear discriminant function is fisher's linear function, which maximally separates the two populations, and the maximum separation in the sample is:

$$D^2 = (\bar{x}_1 - \bar{x}_2)'S^{-1}{}_{pooled}\,x(\bar{x}_1 - \bar{x}_2)$$

Fisher's solution to the separation problem can also be used to classify new observations.

**Test of significance:**

For two populations, the maximum relative separation that can be obtained by considering linear combinations of the multivariate observations is equal to the distance $D^2$ . This is convenient because $D^2$can be used, in certain situations, to test whether the population means $\mu_1$and $\mu_2$ differ significantly. Consequently, a test for differences in mean vectors can be viewed as a test for the "significance" of the separation that can be achieved.

Suppose the populations $\pi_1$ and $\pi_2$ are multivariate normal with a common covariance Matrix$\Sigma$, we can test $H_0$: $\mu_1 = \mu_2$ versus $H_1$: $\mu_1 \neq \mu_2$ is accomplished by referring: (Wiley,2004)

$$D^2 = \left(\frac{n_1+n_2-P-1}{(n,+n_2-2)P}\right)\left(\frac{n_1 n_2}{n_1+n_2}\right)$$

**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 therefore, 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.

We first decriable an approach that is usually called forward selection. In the first step, calculate $\Lambda(y_i)$ for each variable and choose the one with minimum $\Lambda(y_i)$ (or maximum associated F). In the second step, calculate $\Lambda(y_i/y_1)$ for each of the $p-1$ variables not entered at first step, where $y_1$ indicates the first variable entered. For the second variable, we choose the one with minimum $\Lambda(y_i/y_1)$ (or maximum associated partial F), that is, the

variable that adds the maximum separation to the one entered at step. Denote the variable entered at step 2 by $y_2$. In the third step, calculate $\Lambda(y_i/y_1, y_2)$ for each of the $\boldsymbol{p-2}$ remaining variables and choose the one that minimize $\Lambda(y_i/y_1, y_2)$ (or maximizes the associate partial F).
Continue this process until some predetermined threshold value falls below $\boldsymbol{F_{in}}$.

A stepwise procedure follows a similar sequence, except that after a variable has entered, the variables previously selected are re-examined to see if each still contributes a significant amount. The variable with the smallest partial F will be removed if the partial F is less than second threshold value, $\boldsymbol{F_{out}}$ .If $\boldsymbol{F_{out}}$ is the same as $\boldsymbol{F_{in}}$, there is a very small possibility that the procedure will cycle continuously without stopping. This possibility can be eliminated by using a value of $\boldsymbol{F_{out}}$ slightly less than $\boldsymbol{F_{in}}$(Wiley, 2002).

**Evaluating Classification Functions:**

One important way of judging the performance of any classification procedure is to calculate its "error rates " or misclassification probabilities. When the forms of the parent populations are known completely, misclassification probabilities can be calculated with relative ease because parent populations are rarely known.

$$TMP = P_1 \int_{R_2} f_1(x)\,\mathrm{d}x + P_2 \int_{R_1} f_1(x)\,\mathrm{d}x$$

The smallest value of this quantity obtained by a judicious choice of $R_1$ and $R_2$ is called the optimum error rate (ORE).

$$ORE = P_1 \int_{R_2} f_1(x)\,\mathrm{d}x + P_2 \int_{R_1} f_1(x)\,\mathrm{d}x$$

- **Data analysis and discussion of results**

A sample of 800 households in Sinnar State was withdrawn using the two-stage cluster sample, which is one of the most appropriate samples for this study and achieves the main objectives of this study because the size of the population is large, widespread, and heterogeneous in characteristics. So, we use this type of sample to ensure that the sample representing the population is represented the best.

**Table (1): Distribution of sample individuals according to Place of residence**

| Place of residence | Frequency | Percent |
|---|---|---|
| Rural | 376 | 47.0% |
| Urban | 424 | 53.0% |
| **Total** | **800** | **100.0%** |

**Source:** researcher's calculation

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

**Table (2): Distribution of sample individuals according to Gender**

| Gender | Frequency | Percent |
|---|---|---|
| Male | 609 | 76 % |
| Female | 191 | 24% |
| **Total** | **800** | **100.0%** |

**Source:** researcher's calculation

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

**Table (3) Distribution of sample individuals according to Age**

| Age | Percent | Frequency |
|---|---|---|
| 20-30 | 56 | 7.0% |
| 31-40 | 166 | 20.8% |
| 41-50 | 208 | 26.0% |
| 51-60 | 186 | 23.3% |
| 61-70 | 146 | 18.3% |
| 71-80 | 31 | 3.9% |
| 81-90 | 7 | .9% |
| **Total** | **800** | **100.0%** |

**Source:** researcher's calculation

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

**Table (4): Distribution of sample individuals according to Educational level**

| Educational level | Frequency | Percent |
|---|---|---|
| Illiterate | 83 | 10.4% |
| Reads and writes | 184 | 23.0% |
| Basis / Primary | 93 | 11.6% |
| Intermediate level | 110 | 13.8% |
| Secondary | 175 | 21.9% |
| Diploma | 88 | 11.0% |
| Bachelor | 29 | 3.6% |
| High Diploma | 14 | 1.8% |
| Master | 15 | 1.9% |
| PHD | 9 | 1.1% |
| **Total** | **800** | **100.0%** |

**Source:** researcher's calculation

The high level of education increases the standard of living of the individual as well as the cultural level in various economic, social, and health aspects, and the respondents have been surveyed on the school grade that they have completed, as shown in table (4-5). The first thing to note is the drop in the percentage of university and postgraduate education to 19.4% which hurts the living situation.

**Table (5): Distribution of sample individuals according to social status**

| social status | Frequency | Percent |
|---|---|---|
| Married | 684 | 85.5% |
| Single | 40 | 5.0% |
| Divorcee | 27 | 3.4% |
| Widower | 49 | 6.1% |
| **Total** | **800** | **100.0%** |

**Source:** researcher's calculation

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

**Table (6): Distribution of sample individuals according to occupation**

| Occupation | Frequency | Percent |
|---|---|---|
| Occupational | 46 | 5.8% |
| business owner | 181 | 22.6% |
| Employer | 157 | 19.6% |
| Professional | 67 | 8.4% |
| Worker | 112 | 14.0% |
| Policeman / Army | 25 | 3.1% |
| Farmer | 158 | 19.8% |
| Other | 54 | 6.8% |
| **Total** | **800** | **100.0%** |

**Source:** Calculations based on the data file

The distribution of the sample by occupation is shown in the table above. 22.6% are "business owners," 19.8% are "farmers," 19.6% are "employers," 14% are "workers," 8.4% are "professionals," 6.8% are "other," 5.8% are "occupational," and 3.1% are "Policemen/Army."

**Table (7): Distribution of sample individuals according to Family type**

| Family type | Frequency | Percent |
|---|---|---|
| Extended family | 322 | 40.3% |
| Small family | 478 | 59.8% |
| **Total** | **800** | **100.0%** |

**Source:** researcher's calculation

The table above shows the distribution of the sample according to family type; approximately 60% of the sample consists of small families with parents and children, and the other 40% are extended families.

**Table (8): Distribution of sample individuals according to Is family income enough for household expenses?**

| Income | Frequency | Percent |
|---|---|---|
| Sufficient | 417 | 52.1% |
| Insufficient | 383 | 47.9% |
| **Total** | **800** | **100.0%** |

**Source:** researcher's calculation

The table above shows that 52% of families' income is sufficient for household expenses while 48% of families' income is insufficient for household expenses.

**Table (9): Tests of Equality of Group Means**

| Variables | Wilks' Lambda | F | df1 | df2 | Sig. |
|---|---|---|---|---|---|
| Localities | .998 | 1.615 | 1 | 798 | .204 |
| Place of residence | .956 | 37.006 | 1 | 798 | .000 |
| Gender | .992 | 6.711 | 1 | 798 | .010 |
| Educational level | .998 | 1.798 | 1 | 798 | .180 |
| social status | .995 | 3.981 | 1 | 798 | .046 |
| Occupation | .959 | 33.935 | 1 | 798 | .000 |
| Family size | .995 | 4.375 | 1 | 798 | .037 |
| The number of fewer family members from 15 years | .999 | .452 | 1 | 798 | .501 |
| a family member who works other than the paterfamilias | .971 | 23.741 | 1 | 798 | .000 |
| The number of unemployed more than 15 years old | 1.000 | .206 | 1 | 798 | .650 |
| Family type | .986 | 10.990 | 1 | 798 | .001 |
| monthly household income | .877 | 112.127 | 1 | 798 | .000 |
| Food | .995 | 4.046 | 1 | 798 | .045 |
| Education | .989 | 8.931 | 1 | 798 | .003 |
| Health | .996 | 3.139 | 1 | 798 | .077 |
| personal needs | .982 | 14.459 | 1 | 798 | .000 |
| borrowing to provide the living expenses for the family | .710 | 325.702 | 1 | 798 | .000 |
| House | .995 | 3.657 | 1 | 798 | .056 |
| Land | .951 | 41.099 | 1 | 798 | .000 |
| Store | .916 | 73.394 | 1 | 798 | .000 |
| Other | .996 | 3.229 | 1 | 798 | .073 |
| the main breadwinner of the family | 1.000 | .235 | 1 | 798 | .628 |
| evaluation of the standard of living | .575 | 588.953 | 1 | 798 | .000 |
| what is the main shopping place? | .982 | 14.309 | 1 | 798 | .000 |

**Source:** researcher's calculation

Wilk's lambda and the F ratio are used to test the equality of the means of the groups for the same variable. Wilk's lambda for each predictor is equal to the ratio of the within-group sum of squares to the total sum of squares. It is estimated from a one-way analysis of variance by considering the status variable as an independent variable and the predictor variable as a dependent variable. Wilk's lambda is also known as the U statistic. The range of Wilk's lambda value is 0 to 1. If a variable's Wilk coefficient is less than 0.95, it is revealed that the group means are significantly different. The larger the value, the smaller the significance, and the smaller the value, the greater the significance is ensured.

The table shows that all of the significant differences are between the groups. The lowest Wilk's lambda presents the highest importance in the discriminant function. Hence, the most important variable is the discriminant function, which is whether the evaluation of a standard of living is.

**Table (10): Wilks' Lambda**

| Test of Function(s) | Wilks' Lambda | Chi-square | Df | Sig. |
|---|---|---|---|---|
| 1 | .464 | 609.891 | 7 | .000 |

**Source:** Calculations based on the data file

The table above shows Wilks' Lambda for the tests of function using Chi-square, the chi-square test = 609.891, DF = 7, and significant = 0.000, which is less than the level of significance of 0.05. This means the discriminant function is statistically significant.

**Table (11): The Eigenvalue and Canonical Correlation**

| Function | Eigenvalue | % of Variance | Cumulative % | Canonical Correlation |
|---|---|---|---|---|
| 1 | 1.155[a] | 100.0 | 100.0 | .732 |

a. First 1 canonical discriminant function was used in the analysis.

**Source:** researcher's calculation

The table above shows the eigenvalue of the discriminant function, which verifies the relationship between the independent and dependent variables. The largest eigenvalue tells us the most variance in the dependent variable is explained by the function as seen in the table. The eigenvalue of the function is 1.155, and 100% of the variance among the two groups can be explained by this function. The relationship between the predictors and groups is called the canonical correlation. The canonical correlation of the discriminant function is 0.732.

**Table (12): Standardized Canonical Discriminant Function Coefficients**

| Variables | Function(1) |
|---|---|
| Occupation | .139 |
| monthly household income | -.121 |
| borrowing to provide the living expenses for the family | -.514 |
| House | -.128 |
| Store | .133 |
| evaluation of the standard of living | .710 |
| the main shopping place | .097 |

**Source:** researcher's calculation

The table shows the standardized canonical discriminant function coefficient variables and provides one function, which measures the relative importance of the selected variables, and the sign indicates the direction of the relationship. The larger absolute value corresponds to greater discriminating ability.

In this function, the strongest predictor is (the evaluation of the standard of living) with a value of (0.710). The second one is (borrowing to provide the living expenses for the family) with a value of (0.514).

**Table (13): Canonical Discriminant Function Coefficients**

| Variables | Function(1) |
|---|---|
| Occupation (X1) | .066 |
| monthly household income(X2) | -.002 |
| resort to permanent borrowing to provide the living expenses for the family (X3) | -1.246 |
| house(X4) | -.257 |
| store(X5) | .353 |
| evaluation of the standard of living (X6) | 1.321 |
| what is the main shopping place (X7) | .169 |
| (Constant) | -.484 |

The table shows the unstandardized discriminant coefficients for the variables entered into the analysis for the discriminant function.

$$Z = -.484 + .066X_1 - .002X_2 - 1.246X_3 - .257X_4 + .353X_5 + 1.321X_6 + .169X_7$$

**Table (14): Classification Table**

| Is the income of the household sufficient for living expenses? | | | Predicted Group Membership sufficient | insufficient | Total |
|---|---|---|---|---|---|
| Original | Count | Sufficient | 336 | 81 | 417 |
| | | insufficient | 33 | 350 | 383 |
| | % | Sufficient | 80.6 | 19.4 | 100.0 |
| | | insufficient | 8.6 | 91.4 | 100.0 |

**Source:** researcher's calculation

The table shows that 85.8% of the observations were classified correctly into sufficient incomes and insufficient incomes, which means approximately 686 households were classified correctly out of 800 households, and 114 households were misclassified.

The discriminant functions could be able to classify 80.6% of households that have sufficient income. That means it succeeded in the classification of 336 households and filed 81 households.

The discriminant functions could be able to classify 91.4% of households that have insufficient income. That means it succeeded in the classification of 350 households and filed 33 households.

## 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 concerning the dependent variable (IBM SPSS Decision Tree21, 1989,2012).

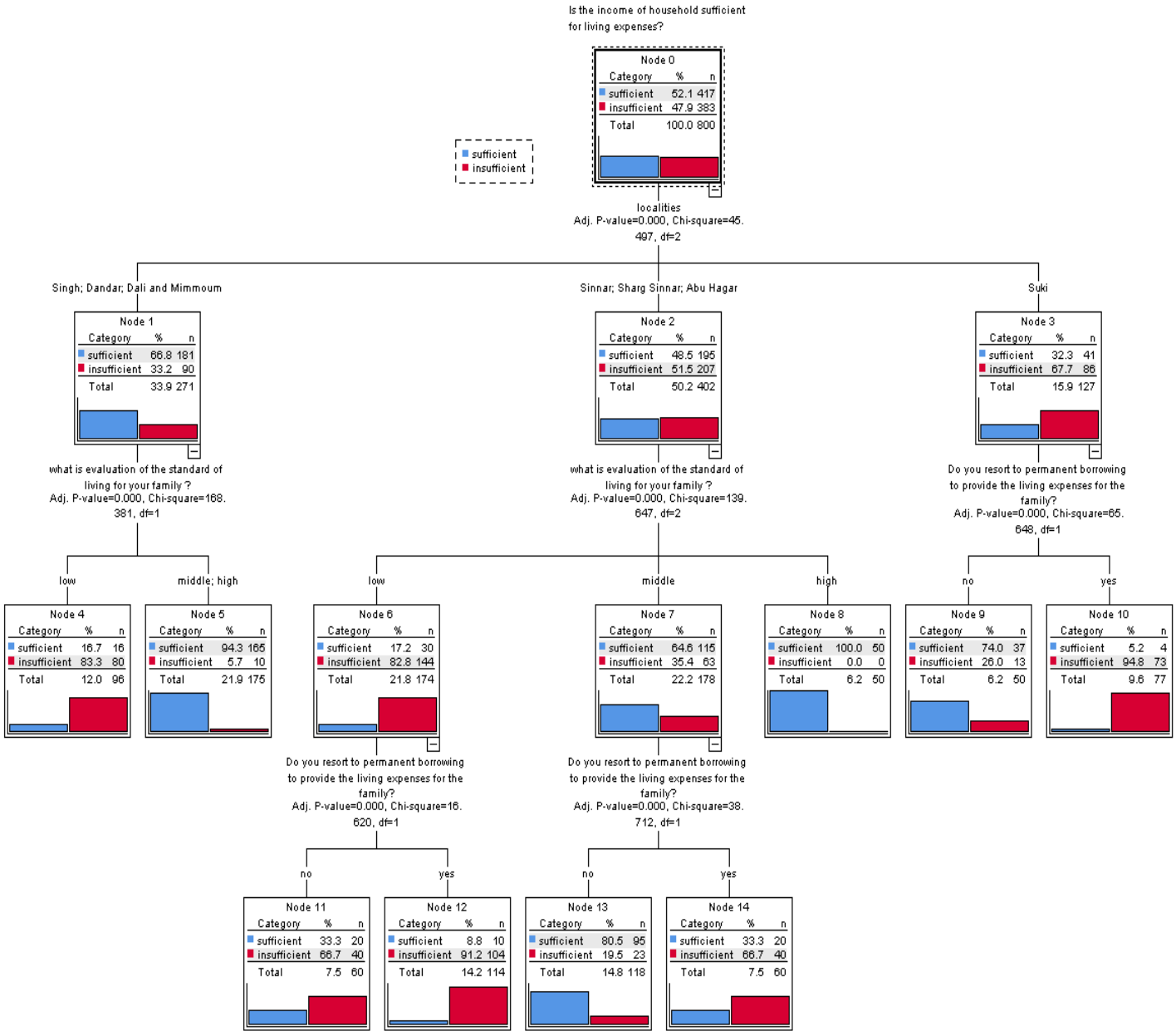


**Figure (1): Tree diagram for income of the household**
**Source:** researcher's calculation

This tree diagram is a graph representation of the tree model. The tree diagram shows that the Sinnar localities have been divided into 3 nodes according to income (node 0). Node 1 contains Singh, Dsnder, Dali, and Msimuom, node 2 contains Sinnar, Sharg Sinnar, and Abu Hagar, and node 3 node contains only one locality (Suki).

Node (1) is divided into two nodes. Node 4 contains the low level of standard of living according to income, and Node 5 contains the middle and high levels of standard of living according to income. It is the best predictor of income. The Chi-square =168.381 and the P-value is 0.000, which is less than 0.05. That means there is a relationship between nodes 4 and 5 and node 1.

Node (2) is divided into three nodes. Node 6 contains the low level of standard of living according to income, Node 7 contains the middle level of standard of living according to income, and Node 8 contains the high level of standard of living according to income. It is the best predictor of income. The Chi-square =139.647 and the P-value is 0.000, which is less than 0.05. That means there is a relationship between node 6, node 7, node 8, and node 1.

Node (2) is divided into three nodes. Node 6 has a low standard of living in income, Node 7 has a middle level of standard of living about income, and Node 8 has a high level of standard of living about income. It is the best

predictor of income. The Chi-square =139.647 and the P-value is 0.000, which is less than 0.05. That means there is a relationship between node6, node7, node8, and node1.

Node (3) is divided into two nodes. Node 9 contains households that do not borrow to provide the living expenses for the family according to their income, and Node 10 contains households that borrow to provide the living expenses for the family according to their income. It is the best predictor of income in node 3. The Chi-square = 65.648 and the P-value is 0.000, which is less than 0.05. That means there is a relationship between nodes 9 and 10 and node 3.

The sixth node (6) is divided into two nodes. Node 11 contains households that do not borrow to provide the living expenses for the family at a low level of standard of living according to their income, and Node 12 contains households that borrow to provide the living expenses for the family at a low level of standard of living according to their income. It is the best predictor of income in node 3. The Chi-square = 16.620 and the P-value is 0.000, which is less than 0.05. That means there is a relationship between nodes 11 and 12 and node 6.

The seventh node (7) is divided into two nodes. Node 13 contains households that do not borrow to provide the living expenses for the family at a middle level of standard of living according to their income, and Node 14 contains households that borrow to provide the living expenses for the family at a middle level of standard of living according to their income. It is the best predictor of income in node 3. The Chi-square = 38.712 and the P-value is 0.000, which is less than 0.05. That means there is a relationship between nodes 13 and 14 and node 7. And we can see from the diagram that there is no statistical significance in the variables that entered into the model in node 8. That is why it has not been divided, as well as nodes 4,5, 9, and 10.

**Table (15): Risk estimation**

| Risk | |
|---|---|
| **Estimate** | **Std. Error** |
| .145 | .012 |

**Source:** researcher's calculation

The risk estimate of 0.14 indicates that the category predicted by the model sufficient incomes to insufficient incomes is approximately wrong for 14.5% of the cases. As a result, the chance of misclassifying an income is 14%.

**Table (16): Classification**

| Observed | Predicted | | |
|---|---|---|---|
| | **sufficient** | **insufficient** | **Percent Correct** |
| Sufficient | 347 | 70 | 83.2% |
| Insufficient | 46 | 337 | 88.0% |
| Overall Percentage | 49.1% | 50.9% | 85.5% |

**Source:** researcher's calculation

The table shows that the model correctly classifies 85.5% of the household income.

## Result:

- There are significant differences between the independent variables in the two groups using the F Test.
- The discriminant function was tested using the chi-square test, which means it can classify households with sufficient and insufficient income in Sinnar state.
- The most important variables that contributed to the classification between the two groups are the evaluation of the standard of living, borrowing to provide the living expenses for the family, occupation, having a store, having a house, monthly household income, and the main shopping place.
- The error rate in the classification was small, and that indicates the quality of the discriminant function in the classification, where the misclassification of the sample was 15% and the accuracy of the classification was 85%.
- The accuracy of the classification of new observations by the discriminant function and decision tree is similar.

## Recommendations:

- Apply the discriminant function that has been reached in this study to classify the households' income, so that government can provide projects to increase the income.
- Use the decision trees to classify the administrative unit of Sinnar according to income.
- Use the discriminant function to broaden the study to cover all of Sudan.

## References:

Afifi, Abdelmonem, Susanne, May, Robin, A. Donatello & Virginia, A.Clark (2020). *practical multivariate analysis*. sixth edition, Taylor & Francis Group, LLC, International Standard Book Number-13: 978-1-138-70222-6 (Hardback) ,London, New York.

Alan, Julian Izenman, (2008). *Modern multivariate Statistical Techniques, regression, classification, and manifold learning*. Springer-Verlag New York.

Alvin, C. Rencher. (2002). *Methods Applied multivariate analysis*. Neil H. Timm, Springer-Verlag New York, Inc.

CARL, J. HUBERTY& Stephen Olejnik, (2002). *Applied MANOVA and Discriminant Analysis*. second edition, Wiley, Interscience Published Simultaneously in Canada.

Geoffrey, J. McLachlan. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley series in probability and statistics, Hoboken, New Jersey.

IBM SPSS, (Decision Tree21, 1989,2012), IBM Corporation, U.S. Government Users Restricted Rights - Use, duplication, or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Richard, A. Johnson & Dena W. Wichrn. (2007). *Applied multivariate statistical analysis*. sixth edition, New Jersey, SUA.

Robert, Ho. (2006). Handbook of univariate and multivariate data analysis and interpretation with SPSS Central Queensland University, Rockhampton, Australia, Taylor & Francis Group, LLC United States of America.

Rodrigo, C. Barros. (2015). *Automatic Design of Decision-Tree Induction Algorithms*. Alex A. Freitas, Springer Cham Heidelberg New York Dordrecht London.

Wolfgang Karl, Leopold Simar,(2007), Applied multivariate statistical analysis, Springer, Berlin Heidelberg New York.