# Prediction of Household Income and Expenditure Using Data Mining

Ashmy Thomas
*P G Scholar, Amal Jyothi College of Engineering*
Kanjirappally, Kerala.
ashmythomas@mca.ajce.in

Ms. Nimmy Francis
*Asst. Professor, Amal Jyothi College of Engineering*
Kanjirappally, Kerala.
nimmyfrancis@amaljyothi.ac.in

**ABSTRACT- Data mining is a field of computer technology which is useful for extracting and coming across styles in big information set. Its miles very usually carried out to any shape of large facts set or processing of information which incorporates collection, extraction and analysis in addition to any utility of pc decision guide device. The statistics mining discover the more than one agencies the information, which could then be used to reap greater prediction consequences by means of way of a selection help machine. Records mining can be implemented to expect the household profits and charges so that you can be beneficial in dealing with the expenses in the future. This paper shows special class algorithms of facts mining used to expand a information mining version for predictions of charges and income taking instructional qualification and variety of people incomes within the family because the base.**

**Keywords—income, expenditure, total earning people, WEKA tool.**

## I. Introduction

One of the main issues that the families are facing now a days are the effective management of the expenses. One who manages the expenses are considered as an ideal person in the society. The principle goal of the proposed gadget is to find out whether or not any sample inside the to be had facts and mine them which might also assist to expect the fees primarily based on their instructional qualifications. Facts mining is a multidisciplinary field masking facts era, device getting to know, information, pattern recognition, information retrieval, neural networks, records structures, artificial intelligence and records visualization. The software of statistics mining is not unusual inside the prediction of economic region. Financial data mining is an emerging field that can be effectively applied in the economic field. It uses numerous thoughts and concept, like category and clustering. When used effectively, a massive amount of financial data can be driven for analysis and modelling to help in effective planning, decision-making for the future.

## II. Relationship between Household income and expenditure

### A. Economic Division

The economic and social diversions and charge variance typically purpose modifications in monetary and social measurement which might be related to the overall boom of the citizen . The variant in expenditure, spending among families or even their academic qualification will finally affect variations in human way of life. There could be a huge difference among the rich and the poor. center elegance and poor humans tends to be affected extra in this example. despite the fact that their profits is low, there can be extra charges which may affect their monetary balance.

### B. Relationship

Income can be from any sources varying from agricultural sector to high scale industrial sector. Even though income depends on many factors, most probably educational standards can be a major factor affecting the whole more the educational level less the poverty or expenses.

But there are some common cause for the expenses. this includes the education, stationary requirements, day to day needs and emergency situations like hospital cases, which is unpredictable. This irregular expenses can result in a very complicated situations for the poor families.

## II. ANALYZING DATA

In order to research facts we need to comply with some pre requirements. There is a want to gather the right dataset at the primary. After amassing the dataset we need to offer the information to a statistics mining algorithm. Therefore deciding on a records mining device follows the dataset series. After imparting the enough facts to the information mining device, use numerous algorithms to evaluate the outcomes.

## III. Data Collection

Detailed income and expenses are collected. Main parameters include the Monthly household income, Monthly household expenses, Number of family members, EMI or rent amount. Annual household income, highest qualified member in the family and class. All the data excluding the highest qualified member and class is numeric. Data set is collected in an excel sheet.

| | MthlyHHI | MthlyHHE | NoofFlyM | EmiorRen | AnnualHH | HighestQu | NoofEarni | class |
|---|---|---|---|---|---|---|---|---|
| 2 | 5000 | 8000 | 3 | 2000 | 64200 | UnderGra | 1 | no |
| 3 | 6000 | 7000 | 2 | 3000 | 79920 | Illiterate | 1 | no |
| 4 | 10000 | 4500 | 2 | 0 | 112800 | UnderGra | 1 | yes |
| 5 | 10000 | 2000 | 1 | 0 | 97200 | Illiterate | 1 | yes |
| 6 | 12500 | 12000 | 2 | 3000 | 147000 | Graduate | 1 | yes |
| 7 | 14000 | 8000 | 2 | 0 | 196560 | Graduate | 1 | yes |
| 8 | 15000 | 16000 | 3 | 35000 | 167400 | PostGradt | 1 | no |
| 9 | 18000 | 20000 | 5 | 8000 | 216000 | Graduate | 1 | no |
| 10 | 19000 | 9000 | 2 | 0 | 218880 | UnderGra | 1 | yes |
| 11 | 20000 | 9000 | 4 | 0 | 220800 | UnderGra | 2 | yes |
| 12 | 20000 | 18000 | 4 | 8000 | 278400 | UnderGra | 2 | yes |
| 13 | 22000 | 25000 | 6 | 12000 | 279840 | Illiterate | 1 | no |
| 14 | 23400 | 5000 | 3 | 0 | 292032 | Illiterate | 1 | yes |
| 15 | 24000 | 10500 | 6 | 0 | 316800 | Graduate | 2 | yes |
| 16 | 24000 | 10000 | 4 | 0 | 244800 | Graduate | 2 | yes |
| 17 | 25000 | 12300 | 3 | 0 | 246000 | Graduate | 1 | yes |
| 18 | 25000 | 20000 | 3 | 3500 | 261000 | Graduate | 1 | yes |
| 19 | 25000 | 10000 | 6 | 0 | 258000 | UnderGra | 3 | yes |
| 20 | 29000 | 6600 | 2 | 2000 | 348000 | Graduate | 1 | yes |
| 21 | 30000 | 13000 | 4 | 0 | 385200 | Graduate | 1 | yes |
| 22 | 30500 | 25000 | 5 | 5000 | 351360 | UnderGra | 1 | yes |
| 23 | 32000 | 15000 | 4 | 0 | 445440 | Professio | 1 | yes |
| 24 | 34000 | 19000 | 6 | 0 | 330480 | Professio | 1 | yes |
| 25 | 34000 | 25000 | 3 | 4000 | 469200 | Professio | 1 | yes |
| 26 | 35000 | 12000 | 3 | 0 | 466200 | Graduate | 1 | yes |

Fig.1. Dataset that is used for the implementations.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 27 | 35000 | 25000 | 4 | 0 | 449400 | Professio | 2 | yes |
| 28 | 39000 | 8000 | 4 | 0 | 556920 | UnderGra | 1 | yes |
| 29 | 40000 | 10000 | 4 | 0 | 412800 | UnderGra | 1 | yes |
| 30 | 42000 | 15000 | 4 | 0 | 488880 | Graduate | 1 | yes |
| 31 | 43000 | 12000 | 4 | 0 | 619200 | Graduate | 1 | yes |
| 32 | 45000 | 25000 | 6 | 0 | 523800 | Graduate | 3 | yes |
| 33 | 45000 | 40000 | 6 | 3500 | 507600 | Professio | 2 | yes |
| 34 | 45000 | 10000 | 2 | 1000 | 437400 | PostGradu | 1 | yes |
| 35 | 45000 | 22000 | 4 | 2500 | 610200 | PostGradu | 1 | yes |
| 36 | 46000 | 25000 | 5 | 3500 | 596160 | Graduate | 1 | yes |
| 37 | 47000 | 15000 | 7 | 0 | 456840 | Professio | 4 | yes |
| 38 | 50000 | 20000 | 4 | 0 | 570000 | Professio | 1 | yes |
| 39 | 50500 | 20000 | 3 | 0 | 581760 | Professio | 2 | yes |
| 40 | 55000 | 45000 | 6 | 12000 | 600600 | Graduate | 2 | yes |
| 41 | 60000 | 10000 | 3 | 0 | 590400 | PostGradu | 1 | yes |
| 42 | 60000 | 50000 | 6 | 10000 | 590400 | Graduate | 1 | yes |
| 43 | 65000 | 20000 | 4 | 5000 | 647400 | Illiterate | 2 | yes |
| 44 | 70000 | 9000 | 2 | 0 | 756000 | Graduate | 1 | yes |
| 45 | 80000 | 20000 | 4 | 0 | 1075200 | Graduate | 1 | yes |
| 46 | 85000 | 25000 | 5 | 0 | 1142400 | UnderGra | 2 | yes |
| 47 | 90000 | 48000 | 7 | 0 | 885600 | PostGradu | 3 | yes |
| 48 | 98000 | 25000 | 5 | 0 | 1152480 | Professio | 2 | yes |
| 49 | 100000 | 30000 | 6 | 0 | 1404000 | Graduate | 3 | yes |
| 50 | 100000 | 50000 | 4 | 20000 | 1032000 | Professio | 2 | yes |
| 51 | 100000 | 40000 | 6 | 10000 | 1320000 | PostGradu | 1 | yes |
| 52 | | | | | | | | |

Fig.2. Dataset

The next step is to create the corresponding CSV file of the Dataset. CSV(Comma Separated Value) format will lay the data in a table of rows and columns and a comma is used to separate values on a row.

After creating the CSV file , we will convert the corresponding file into arff file(Attribute Relation File Format) , where a header is used that provides metadata about the data types in the columns. The arff file will be processed in the Weka tool for data processing and classification.

### C. WEKA Tool

WEKA- which is an open source software which provides tools for data preprocessing, implementation of several Machine Learning algorithms which we use in our day to day life , and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems to extract the required patterns and apply them.
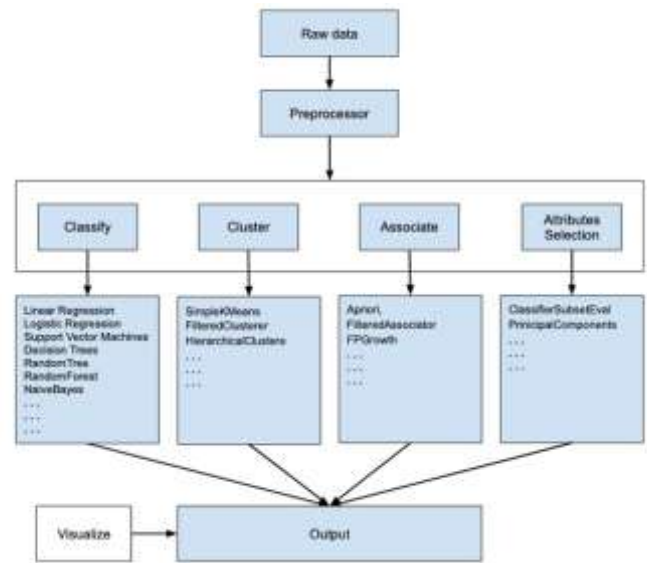


Fig.3. Basic work flow of WEKA Tool

Relying on the form of machine studying model that we use to expand, we should select any of the options which include Classify, Cluster, or partner. We are able to automatically choose any of the identical. WEKA presents the implementation of several algorithms and we can analyze every one in every of them separately. As stated, pick out an algorithm of your personal choice and set the favored enter and run it at the dataset. You may observe diverse models at the same dataset. You can then examine the outputs of different fashions and choose the excellent that meets your prediction.

- To install WEKA visit the official site, download the exe file and install it in the system.

- Open the WEKA tool and you will see the subsequent screen.



Fig.4. Explorer window

### D. Implementation

Fig.5. Dataset after converting into CSV format

- The above data shows the CSV file. This file is then inserted to the Weka tool extract the best output.

- Open the Weka tool and after that we need to click on at the Explorer. Pick the pre technique and from there select the dataset in CSV format which you need.

- Click the 'open file' and select the appropriate dataset you need.
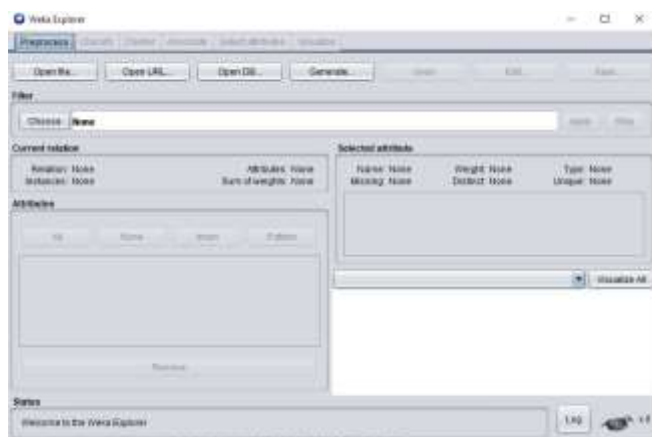
## IV. EXPERIMENT RESULT AND DISCUSSION



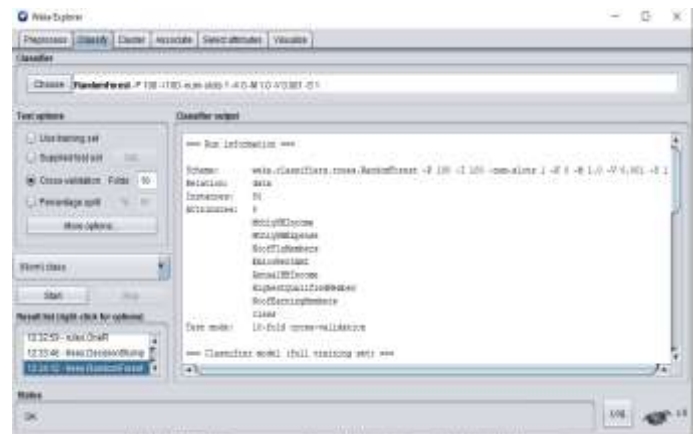Fig. 6. Preprocessing window



Fig.7. After preprocessing data.



Fig. 8. After applying the Random Forest Algorithm.

Random Forest Algorithm is a type of supervised mastering algorithm, which may be used for classification and regression. It is higher than single decision tree.

- Correctly Classified Instances   45     90 %

- Incorrectly Classified Instances 5     10    %

- Kappa statistic     0

- Mean absolute error     0.1296

- Root mean squared error     0.2693

- Relative absolute error     66.2162 %

- Root relative squared error     88.5377 %

  Total Number of Instances     50

Table 1. Detailed Accuracy by class

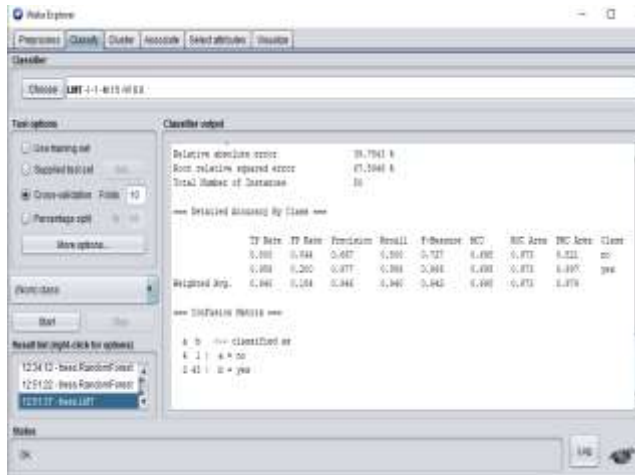|  | TP Rate | FP Rate | Precision | Recall |
|---|---|---|---|---|
|  | 0.000 | 0.000 | ? | 0.000 |
|  | 1.000 | 1.000 | 0.900 | 1.000 |
| Weighted Avg | 0.900 | 0.900 | ? | 0.900 |

Fig. 9. After applying Logistic model tree (LMT)

LMT is a classification model that is associated with supervised gaining knowledge of set of rules which mixes the properties of both decision tree and logistic regression.

- Incorrectly Classified Instances    5    10%
- Correctly Classified Instances    45    90%
- Kappa statistic    0
- Mean absolute error    0.1296
- Root mean squared error    0.2693
- Relative absolute error    66.2162 %
- Root relative squared error    88.5377 %

    Total Number of Instances    50

Table 2. Detailed Accuracy by class

|  | TP Rate | FP Rate | Precision | Recall |
|---|---|---|---|---|
|  | 0.800 | 0.044 | 0.667 | 0.800 |
|  | 0.956 | 0.200 | 0.977 | 0.956 |
| Weighted Avg | 0.940 | 0.184 | 0.946 | 0.940 |

## V. RESULT ANALYSIS

We have collected dataset and converted it to the format which is understandable by WEKA tool. After that we have applied various classification algorithm on collected dataset to predict the income and expense ratio using the WEKA tool. As per the result shown in table 1, figure 5 and 6 we found that LMT algorithm gives the highest accuracy among all other algorithms.

## VI. CONCLUSION AND FUTURE SCOPE

Information mining carry out the very vital function for prediction of earnings and expenditure of a own family. statistics mining techniques and techniques helps own family participants for taking important choice-making method. in this research paper, we have implemented numerous class algorithms of facts mining for predicting. To perform the information mining analysis we've got amassed facts set which in particular encompass monthly earnings fees, month-to-month income, every year income, every year fees, highest academic qualifications and total variety of human beings incomes within the family.

Via applying this predictions a human beings can plan numerous techniques to control their costs. We also got here to recognize that people with more academic qualification have more profits.

## VII. REFRECES

[1] Household income and expenditure relationships: A simultanous Equation approach, Chinese Business Review, ISSN 1537-1506 June 2011, Vol. 10, No. 6, 395-405.
[2] Maximizing Student Efficiency Using Data Mining Technique, Volume2, issue 6.
[3] Household income and expenditure statistics, S eventeenth International Conference of Labour Statisticians Geneva, 24 November-3 December 2003.
[4] An analysis of income and expenditure pattern of working women in the contex of emerging consumer culture.
[5] WEKA tutorial by tutorials point.