**Crisp-DM Implementation for Elderly Social Protection**

|  |  |  |
| --- | --- | --- |
| **Keywords** |  | **ABSTRACT** |
| *Social Assistance, Crisp Dm, SVM, Classification.* |  | *In Jakarta, social assistance is available for specific categories of older adults but is given by Conditional Cash Transfer (CCT). Technical assistance is based on a database of beneficiaries and its mining due to the regulation registry. The sub-national social government maintains a Mining database. The author uses a database in one of the districts as a case study. The author uses Crisp DM to understand this social welfare problem because it has its logic of understanding, the logic and the details of pre-processing the data. Then, to decide, used a Support Vector Machine (SVM) as a model and tested it with cross-validation. To obtain the classification accuracy formed from the data processing results. This research aims to see the predictive results of the distribution of KLJ program assistance. Then, answer the form of modelling to create a decision system for the distribution of KLJ. For this reason, SVM is used to make decisions related to social assistance issues. In the SVM, which consists of two classes, it can be interpreted that one class contains “stop” while the other “receive”. However, there are offers bias in the data processing, with the "if" condition. The data processing reaches 100% accuracy and gets SVM kernel weight.* |
|  |
|  |  |

**INTRODUCTION**

Social welfare is addressed in both multidisciplinary and information systems. Social welfare uses computer technology to manage the delivery of social welfare by staff or others [1]. Welfare and social security are topics that have developed in the 20th century from a perspective that starts from political, economic, social, regional, and technical aspects [2]. Technically, the topic of welfare can be understood using the data mining method. Technical assistance is implemented based on the policies in force in an area. In the beginning, technical assistance was based on data. The preparation of providing aid based on administrative records was carried out by the State of Carolina but needed help to answer the existing questions [3]. Another case study concerns the technique of providing social assistance to children in the Philippines, highlighting the delay in updating the registry, thus hindering the process of passing the assistance [4]. After all, several comparative studies on data mining in social assistance exist. One of them is using algorithm C4.5 for social welfare recipients in Maligas Mountains [5]. The c4.5 algorithm can be used for profiling, while the area-based decision rule can also be used for the distributional process. However, this is difficult for local authorities, who still need computer technology. Therefore, the authors propose to use the SVM model of Crisp- DM [6] and test it by cross-validation.

Social support is always a challenge when it comes to data mining. This research would likely generate better ideas for dealing with social assistance in terms of information systems, such as updating beneficiaries, decision-making processes, and much more. The Crisp (DM) methodology is used in this research to evaluate the problems of older people because the steps in the industry can describe the challenges that occur in social assistance, particularly the community sector of services. In this article, the problems of older people are studied using the Crisp (DM) method because the steps in the industry can describe the issues that exist in social assistance, especially the community service industry.

Asda.

The basis of the aid distribution database is administrative records, and each region is responsible for providing social assistance. In the Jakarta region, the sub-national colonial government handles social problems in Jakarta, including the issue of older people. The assistance program for the elderly in Jakarta includes the following categories: over 60 years old; registered in the DTKS data (integrated social assistance data; data canter for poor and disadvantaged people, the basis for determining social assistance programs); has no regular income or is insufficient so that he is unable to meet his basic needs; has chronic pain and is bedridden (poorly ridden); is mentally and spiritually neglected. To provide for older people, the government provides them with social assistance. According to the data, the total population in 2020 is 10,562,088 people; older people are 907,738, 8.6% of the total population and 41,121 poor elderly. Homeless elderly are not counted in this figure because they are housed in nursing homes. The same year, 225,945 social assistance recipients for older people (KLJ) were served based on the poverty database [7]. Due to the large allocation of funds provided, it causes a small number of beneficiaries.

Sub-national social governments are interested in keeping databases of beneficiaries, actors providing social assistance, and policymakers regarding regional-level service. Conditional cash transfer has no evidence to suggest, but its design requires beneficiary verification, monitoring and dispute mechanisms heavily dependent on administrative resources [8]. On the other hand, a sub-national social government has limited access to data from different stakeholders, especially from the ministry. In providential, he had a local policy to keep a mining database.

Each country has a figure and information that is biased depending on the name of the policy program [9]. Brazil has Cadastro Unico, a national database integrated into one data which scoops national and separate databases from different targeting social protection to make Bolsa Familia (Brazil Cash Conditional Transfer). Unlike Brazil, Kenya developed IPRS (Integrated Population Registration System), a standalone MIS that comes from a unified registry that is desirable to provide recommendations and findings. The IPRS is used to compare to other databases used in program identification cards [10]. o Indonesia has a unified database called DTKS (henceforth Unified Database, UDB) held in the Social Ministry and managed by Sistem Informasi Kesejahteraan Sosial Next Generation (SIKS NG) containing social economic and demographic information [11]. When implementing social assistance programs, governments at the regional level are happy to maintain local databases at the regional level so that they do not accumulate Field [12].

Data mining is a process rather than something that is processed immediately. Data mining requires human involvement and analysis. In another discussion, Social Security has a data mining called Social Security Data Mining (SSDM). It is necessary to pay attention to the models, algorithms, and confidentiality of the data, assistance (profiling) policies, and domain-based decision rules [13]. Australia has a way of mining social security data. The government consists of information coordination (immigration data, customs, taxation, and banking systems), administrative profiles, business knowledge (designated for data warehouses, data reporting, and investigations), and information analysis (descriptive analysis and research of abnormal use of loans). Mining social assistance data can be done using two conceptual approaches, namely Comprehensive concept (political interpretation, economic interpretation, sociological interpretation, regional interpretation, regional interpretation) and technical performance (problem evaluation, method and policy modelling, enterprise-oriented analysis, correlation evaluation, infrastructure support, data-driven evaluation). The five essential data mining strategies on social security and welfare data are Modeling the effect of activity/activity sequence, mining impact target designs, Mining Positive and Negative consecutive patterns, Mining Attribute Combined examples and Identifying Impact Behavior for Intervention. [14]

There are at least five functions of data mining: a. Description, which is finding patterns and relationships in the data; b—estimation, which is determining the prediction of the variables; c. Prophecy is the same as estimation, except that data processing results are in time. Usually, there is a date attribute in predictions, d. Classification is a variable classified into a label (concept). e. Clustering > does not attempt to organise, estimate, or predict the value of the target variable. Grouping the data set into relatively homogeneous subsets or groups, the similarity of records within the cluster is maximised, and parallel to records outside that cluster is minimised [15]. Data mining is a large amount of data that correlates with the database. There are two data mining functions: predictive and descriptive [16].

Support Vector Machines (SVM) is a "state of the art" data mining ool for several reasons. SVM usually provide better results than other methods; they have no problem with local minima (the big problem with neural networks), and SVM does not require as many parameters to be specified as other methods, usually the capacity (explained later) and the kernel to use (and any parameters needed by the kernel). SVM can very quickly work with thousands of different features; they are usually swift, and finally, which makes a big difference, SVM uses a kernel function [17]. This technique, preprocessing data, can be a suitable form for learning. To generate two classes whose accuracy can be searched [18]. SVM also give an easy recommendation to stakeholders so they can Stop or Continue.

The original CRISP-DM model can still be seen in recent proposals, which focus on the traditional paradigm of a sequential list of stages from data to knowledge. CRISP-DM and the underlying data mining paradigm remain applicable to a much broader range of data science projects. Meanwhile, CRISP-DM is considered the de facto standard for analytics, data mining and data science projects [19]. A classic Crisp DM consists of six steps: business understanding, data understanding, data preparation, modelling, evaluation, and results. Each stage has a phase that describes the state of the business.

**METHODS**

In this subsection, the concept of crisp DM is explained. Also, the author uses the crisp-DM method to verify this social support. Using the crisp-DM, the industrial framework approach consists of 6 phases used as the basis for DSS: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

Diagram

Description automatically generated

**Figure 1. Crisp DM Method**

**Business Understanding**

Diagram

Description automatically generated

**Figure 2. Business Understanding Phases**

There are four processes in understanding business:

1. Explaining the business object: the business object in this paper is social assistance from the database, and older people are the target of this assistance. Social service is a cash transfer to the recipient, valid for one year. It can be terminated based on provisions, such as death, moving out of the region, assets, and care by a public nursing home.
2. Determine data mining achievements to determine success criteria from data mining: The target of the social recipient is the existing recipients so that he can hit the bull's eye.
3. Situation access: The database is managed by a regional stakeholder with the appropriate authority before the recipient must fill in all the required information. The recipient's announcement is more like an administrative file, except that it comes from social services and is confidential.
4. Plan activities: To find the existing data (recipients), all the information about the regulated is transferred to the data set. It should be noted that when creating tables that contain the basis of take-out regulations, they must have the same ID. Both data have the same values (id) and thus can be mined.

**Data Understanding**

Application

Description automatically generated with medium confidence

**Figure 3. Data Understanding Phases**

There are three processes to determine data: a) retrieve data and b) obtain and receive data. Since this data is confidential, a bureaucratic effort is required to release the data. c) Explain the contents of the database; 10 attributes in the database beneficiaries are text and number. This study uses data on beneficiaries for 2019-2020. The research was conducted within the Jatinegara sub-district for one month. For sampling, the researchers used an interview guide in the form of tables. This guide is used to review social welfare recipients to create a new database for comparison/analysis of data. The data is saved in Excel format. The researchers use RapidMiner with the SVM model to generate kernel weight in the study. The obstacle in this research is data licensing, which is difficult because it is confidential and bureaucratic. c) Data quality verification: The data taken is a dataset of residents with ID and only one ID for each data possession.

**Data Preparation**

Graphical user interface, application, table

Description automatically generated

**Figure 4. Data Preparation Phase**

The next step is data preparation. It consists of 5 processes: a) select data: select the data used; that is, database recipients of social assistance. b) clean data; cleaned data from noise (non-uniformity) and missing values; the databases were reduced from the double receiver. (c) create new data; create new records; create a single table describing a body of recipient information; (d). Data integration: Table merging is merging two or more tables containing the same object and creating new knowledge. The ID has a unique character and similarity, so it can be integrated to create a new attribute, the attribute statute, and existing knowledge. The game rule to get attribute stop of validation attributes made two classes. e). Data formats: unique identification of files (format/type of file storage). The file is stored in Microsoft Excel and local repositories, and folders are created for each district, so access takes a long time. Each database recipient saves the format as xls.

**Modelling**

Diagram

Description automatically generated

**Figure 5. Modelling Phase**

Consists of 3 processes: a) selection of a model; SVM was chosen based on data mining performance requirements, criteria that can generate two classes, more focused on dictating decisions; b) building a model; the author used SVM kernel weight-based on three attributes, c) preparing tests; separate training data and datasets.

**RESULTS**

The database of welfare recipients referred to by the author is from the Jatinegara district area. An overview of the database can be created with Rapid Miner.

**Table 1. Database Attribute**

|  |  |  |
| --- | --- | --- |
| Attributes | Type Data | Value |
| ID Administration | Real | 16 digit (1-0) |
| NAME | Text | A - Z |
| AGE | Polynominal | 60 - 100 |
| GENDER | Binominal | Male, Female |
| SUB-DISTRICT | Polynominal | Balimester - Kampung Melayu |
| YEAR | Binominal | 2019 & 2020 |
| PERCENTILE | Integer | 1-64 |
| STATUS | Polynominal | EXIST, NO FIND ADDRESSES, DIED |
| VALIDATION | Polynominal | EXIST, ADMINISTRATION RECORD, DISTRIC LETER |
| DISTRIBUTION | Binominal | RECEIVE & STOP |

The author uses a pivot table in Microsoft Excel. This gives a distribution table like the table below. The distribution attribute comes from "rule(if)". It is based on quality and validation. The if function makes the same thing out of binomial-type data as the SVM mentioned.

**Table 2. Social Assistance Recipients in Jatinegara by Subdistrict**



There are 5,443 KLJ recipients spread across the Jatinegara Sub-District. The representative age range has a median value of 12.50% in each village. In the 60-64 age group, the percentage is highest in Kampung Melayu sub-districts at 18.21%, the 65-69 age group is represented in Bidara Cina with 16.90% and the 70-74 age group at 17.63%. In Cipinang Besar Utara, the highest percentage was in the 75-79 years age group (29.36% of recipients), 80-84 years (19.70%), and 85-89 years (20%).

**The Altima used in data mining can be seen from the pre-processing flow image.**

A screenshot of a computer

Description automatically generated

**Figure 6. Flowchart Algorithm**

The author uses RapidMiner to run the SVM model—distribution of vector data for welfare recipients' SVM kernel weight.

Text

Description automatically generated

**Figure 7. Spread vector in Rapidminer Using SVM**

In Figure 7, A value of 1 means that the receiver should be stopped. However, there is a bias distribution in the vector between 0 and 1, namely, the status attribute died (0.716) and not found addresses (0.241), which should no longer receive assistance.

To test the model, the authors use a confusion matrix with a ratio of 0.8 for the training data and 0.2 for the test data. Then, using 1136 test data, an accuracy rate of 100% is obtained, shown in the following graph.

**Table 3. Confusion Matrix**



**CONCLUSIONS**

At the end of the article, you can use the 6-phase crisp DM as an approach to mining social welfare data. It is adding attributes using "rules" based on human logic. The "rule" created by humans is found in the guidelines of social welfare programs. Then, this knowledge is taught to the computer to form an understanding that is strung together so it can be classified. The basis for using the SVM model allows humans to make the correct and quick decisions. In addition, the model created by the SVM creates only two classes, so it seems to dictate a decision when implementing social welfare policy programs.

**REFERENCES**

[1] P. Henman and M. Adler, “Information technology and the governance of social security,” *Crit. Soc. Policy*, vol. 23, no. 2, pp. 139–164, 2003, doi: 10.1177/0261018303023002002.

[2] L. Cao, “Social security and social welfare data mining: An overview,” *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, no. 6, pp. 837–853, 2012, doi: 10.1109/TSMCC.2011.2177258.

[3] H.-C. (Monica) Kum, D. Duncan, K. Flair, and W. Wang, “Social welfare program administration and evaluation and policy analysis using knowledge discovery and data mining (KDD) on administrative data,” *dg.o 2003 Proc. 2003 Annu. Natl. Conf. Digit. Gov. Res.*, no. October 2015, pp. 1–6, 2003.

[4] E. L. Dadap-Cantal, A. M. Fischer, and C. G. Ramos, “Targeting versus social protection in cash transfers in the Philippines: Reassessing a celebrated case of social protection,” *Crit. Soc. Policy*, vol. 41, no. 3, pp. 364–384, 2021, doi: 10.1177/02610183211009891.

[5] R. Nur Adiha, S. R. Andani, and W. Saputra, “Application of Data Mining in Determining Social Assistance Recipients With C4.5 Algorithm in Maligas Mountain District,” *Int. J. Basic Appl. Sci.*, vol. 10, no. 3, pp. 88–99, 2021, doi: 10.35335/ijobas.v10i3.59.

[6] P. C. Ncr *et al.*, “Crisp-Dm,” *SPSS inc*, vol. 78, pp. 1–78, 2000.

[7] A. Surono, “Kartu Lansia Jakarta Edisi 2019,” *acurat.co*, 2019.

[8] R. Anderson and G. Mansingh, “Data Mining Approach to Decision Support in Social Welfare,” *Int. J. Bus. Intell. Res.*, vol. 5, no. 2, pp. 39–61, 2014, doi: 10.4018/ijbir.2014040103.

[9] V. Barca, *Integrating data and information management for social protection:*, no. October. 2017.

[10] S. Rao, “National databases of the poor for social protection,” no. April, 2016, doi: 10.13140/RG.2.1.1162.0241.

[11] OECD, *Social Protection System Review of Indonesia, OECD Development Pathways, OECD*. 2019.

[12] TNP2K, “Indonesia ’ s Unified Database for Social Protection Programmes Management Standards,” pp. 3–4, 2015.

[13] T. Balasubramanian, “Social security and social welfare data mining: An overview,” *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, no. 6, pp. 837–853, 2012, doi: 10.1109/TSMCC.2011.2177258.

[14] S. Yaliwal, P. Yaliwal, and N. Mirjankar, “Data Mining on Social Security and Social,” pp. 35–40, 2016.

[15] D. T. Larose and C. D. Larose, *DISCOVERING KNOWLEDGE IN DATA An Introduction to Data Mining Second Edition Wiley Series on Methods and Applications in Data Mining*. 2014.

[16] S. Agarwal, *Data mining: Data mining concepts and techniques*. 2014. doi: 10.1109/ICMIRA.2013.45.

[17] J. M. Gutiérrez, “and classifying documents,” no. 280485, 2010.

[18] M. H. Almaspoor, A. Safaei, A. Salajegheh, B. Minaei-Bidgoli, and A. A. Safaei, “Support Vector Machines in Big Data Classiication: A Systematic Literature Review Support Vector Machines in Big Data Classification: 1 A Systematic Literature Review 2 3,” 2021.

[19] F. Martinez-Plumed *et al.*, “CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories,” *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 8, pp. 3048–3061, 2021, doi: 10.1109/TKDE.2019.2962680.

|  |
| --- |
| **Copyright holder:**  (2023) |
| **First publication rights:**  International Journal of Social Service and Research (IJSSR) |
| **This article is licensed under:** |

**REVIEW**

|  |  |
| --- | --- |
| **Review Component** | **Review Comment** |
| Novelty (Research Novelty) | not enough |
| Objectivity ( English translation. Research Flow Suitability) | not enough |
| Method (Conformity Method) | complete it again |
| Scientific Impact (The existence of analysis points in the discussion chapter) | still not enough |
| Conclusion (Conclusions in accordance with the problems present in the abstract) | still not enough |
| References  (1) For bibliography at least the last 5 years for exact and 10 years for social sciences  (2) Primary references >80% of existing references | still not enough |