OORJA
ISSN - 0974-7869 (Print)
ISSN - 2395-6771 (Online)

ANALYZING INCOME INEQUALITY USING CLASSIFICATION TECHNIQUES AND VISUALIZATION

Ms. Ritu Khandelwal, Assistant Professor, International School of Informatics & Management, Jaipur

Abstract

Extreme wealth and income inequality are a big concern, especially in the United States. The potential to end poverty is a strong justification for lowering the world's rising economic disparity. The idea of universal economic disparity promotes a country's economic stability and guarantees sustainable development. The governments of many nations have been working hard to address this issue and offer an ideal answer. This paper focus on the issue of income inequality is addressed using machine learning. Machine learning is a technique that uses artificial intelligence, and other learning techniques to find the pattern of data and related knowledge from several large datasets. Orange is both a free and open-source application for data analysis and data visualization. We can now forecast future earnings with the quick advancements in storage capacity and computer performance. This study's issue will be the classification of adult datasets utilizing the orange tool. the various classification techniques applied to an adult sensory dataset utilizing the orange tool.

For this, the UCI Adult Dataset has been used. To tie up current learners and add some pre-processing to create new versions, this paper addresses the highlights. In this study, different categorization techniques are contrasted with an analysis of the outcome using a confusion matrix. It comprises a few performance indicators, such as recall, an area under the curve (AUC), an F1 score, and precision. Using main features, classification is applied to determine a person's yearly income that falls higher than $50,000 or less than $50,000. The main purpose of this study is to give comprehensive analysis of relevant methods on dataset. The decision tree model, Naive Bayes, k-Nearest Neighbour(KNN), and support vector machine(SVM) were used for the comparative study of this paper. One of those, the decision tree recorded the maximum accuracy of 98.4%, surpassing the standard accuracy of earlier works. Attained levels of precision and recall are 98.6% and 99.3%, respectively.

Keywords: Machine learning, Decision tree, Naive Bayes, KNN, SVM, Data mining

Introduction

The ability of machine learning to transform vast volumes of data into informative knowledge and information using statistical, mathematical, artificial intelligence, and learning approaches has drawn attention in society in recent years. A set of procedures called machine learning examines the additional value of a data set in the form of knowledge that has not previously been learned by hand. In recent years, the issue of income disparity has received a lot of attention. The goal of eliminating this problem does not appear to be limited to improving the lot of the poor. Many people struggle with the rise of economic disparity and demand for a fair distribution of wealth. This paper emphasizes thorough study and the essential factors for raising a person's income. Such an analysis aids in focusing attention on the key areas that can considerably raise one's income level. The aim is to estimate, using data from the census, if a person's income is larger than $50,000 based on factors including age, education, and marital status. The decision tree model, naive Bayes, KNN, and SVM

come under the classification technique(Khandelwal et al., 2020). Open-source data mining software is required to create classification algorithms on datasets. The software utilized is a Pythonbased Orange utility. Orange is a data mining tool that is helpful for exploratory data analysis and visual programming. The widgets that make up an orange are its various parts. MacOS, Windows, and Linux are all supported by this data mining program.

The classification of a dataset used to analyze annual income that was collected from the UCI Machine Learning Repository is the issue that this study looks at (Chakrabarty & Biswas ,2018). It is anticipated that the knowledge gained from doing this research will benefit those who do or do not earn more than $50,000 annually. Several things influence a person's yearly income. It makes intuitive sense that factors like the person's age, gender, occupation, and amount of education have an impact. The prediction task is to ascertain whether a person earns more than $50,000 annually. This paper is organized as follows: an introduction, literature study, Research Methodology, Results, and discussion and conclusion.

Literature Review

Researchers have attempted to estimate income levels in the past using machine learning models. Many machine learning models, including Logistic Regression, Naive Bayes, Decision TreesKNN, SVM, Gradient Boosting, and 6 configurations of Activated Neural Networks, were employed by researchers to investigate and analyze the Adult Dataset. They also performed a comparative examination of how well they predicted outcomes. A Random Forest Classifier method was used to forecast people's income levels(Khandelwal et al., 2020), (Khandelwal & Virwani, 2019). To scale up the accuracy, this work has used complicated algorithms like XGBOOST, Random Forest, and stacking of models for prediction tasks, including Logistic Stack on XGBOOST and SVM Stack on Logistic. With the Adult Dataset, researchers attempted to mimic Bayesian Networks, Decision Tree Induction, Lazy Classifier, and Rule Based Learning Approaches and gave a comparative analysis of the predicted performances. Some have made an effort to pinpoint the key data elements that could help reduce the complexity of the various machine-learning models used for categorization tasks (Khandelwal & Virwani, 2019). To anticipate income levels, some researchers have done a comparative study of four different machine learning techniques: naïve bayes, classification and regression trees, random forests, and support vector machines (Deepajothi & Selvarajan, 2012). In paper, income prediction data based on the Current Population Survey provided by the U.S. Census Bureau are generated and evaluated using Principal Component Analysis (PCA) and the Support Vector Machine approach.

Research Methodology

Using Orange software and the classification approach, the investigation was carried out. Using the UCI Machine Learning Repository as the data source .

Normal procedures are utilized in machine learning algorithms, for example, contingent likelihood assessment, scoring of traits, data separating and selection, irregular sampling, and others. Orange digs into these techniques in parts and constructs its strategies by collecting parts utilizing calculations.

Before applying data mining strategies, datasets is cleaned and ready from its crude state. While this issue is normally present with any information, data miners regularly work with more chaotic information than analysts and psychometricians; rather than seriously recorded test or review information, information excavators frequently work with log information or learning the board

OORJA
ISSN - 0974-7869 (Print)
ISSN - 2395-6771 (Online)

framework information recorded in structures that are not quickly amiable to analysis. The methodology targets data preparation, data pre-processing, model development, model evaluation, then data visualization.

Knowing the key features is essential since they play a crucial role in the process' accuracy [7]. The class attribute is created from the "type" attribute. There are both category and numerical data types in the dataset. Information that is unavailable for an item is referred to as missing value (case). The accuracy and quality of the data will decline as it is processed as a result of the missing value resulting from the object's information not being provided, being hard to find, or not being there.

Result and Discussion

Data preparation for the model evaluation of each classifier is elaborated using Orange:

A. Data preparation

Firstly the data preparation phase will start. The adult_sample dataset is taken to understand the workflow of data analysis which includes data processing, modeling, evaluation, and visualization. Initially, the dataset is loaded using a file widget that shows the name and type of attribute with all values shown in Fig. 1. It defines no. Number of instances(recorded). There are two methods to load them.

Initially, the dataset will be loaded using the *File* widget that shows the name and type of attribute with all values that have shown in Fig. 1. It defines no. of instances(records). It has two ways to load it.

The Data Table shows the complete data. In this example no. of instances 977. Data Sampler incorporates all sampling methods whatever the user wants to apply to the input dataset.

- A fixed proportion of data finds a selected range of data (e.g. 80% of all the data)
- The fixed sample size divides the dataset into fixed numbers. of instances, and it returns only one set at a time.
- Cross-validation partitions data instances into no. of instances, the user can select the number of folds (subsets) and the fold they wish to use as a sample.
- Bootstrap infers the sample from the population statistic, and then a connection is established with preprocessing. Using the available options, the data can be cleaned and filtered. Fig.2 shows the establishment of a connection between these widgets.

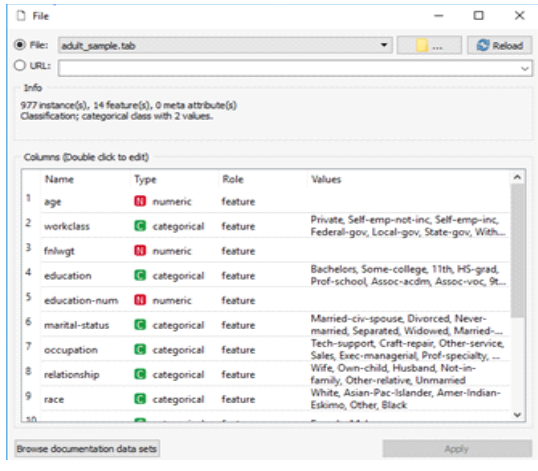

Figure I : Loading Dataset

B. Data pre-processing

After applying sampling 977 instances are divided into two sets of training dataset and test dataset.792 instances move to sample data. Then, both have gone through the preprocess widget to split no. instances in the training and testing datasets. In total, 730 instances were used for the training dataset.

C. Data Modeling Using Classification

Fig. 3 shows the complete workflow analysis. Fig. 4 shows the construction of a classifier or model, using which the prediction results can be generated. All models were based on classification techniques. The same connections can be established using clustering or association-rule mining techniques.

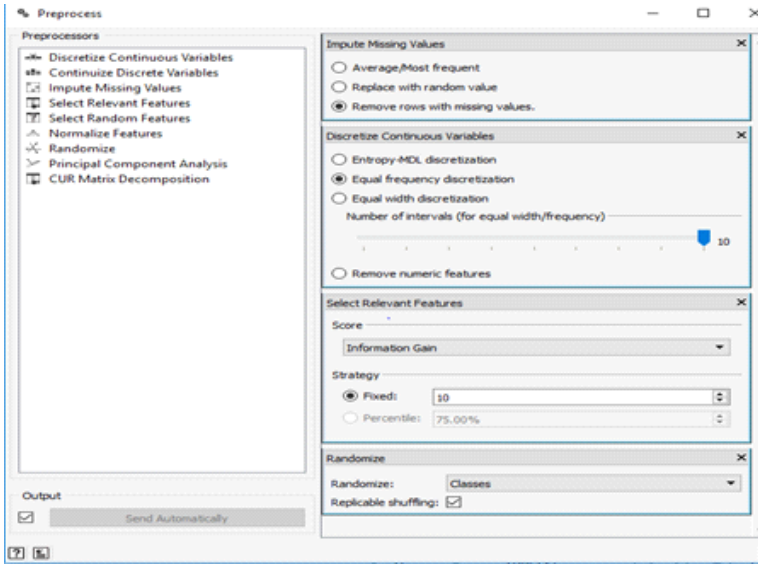

Figure II :Data pre-processing using Pre-process Widget

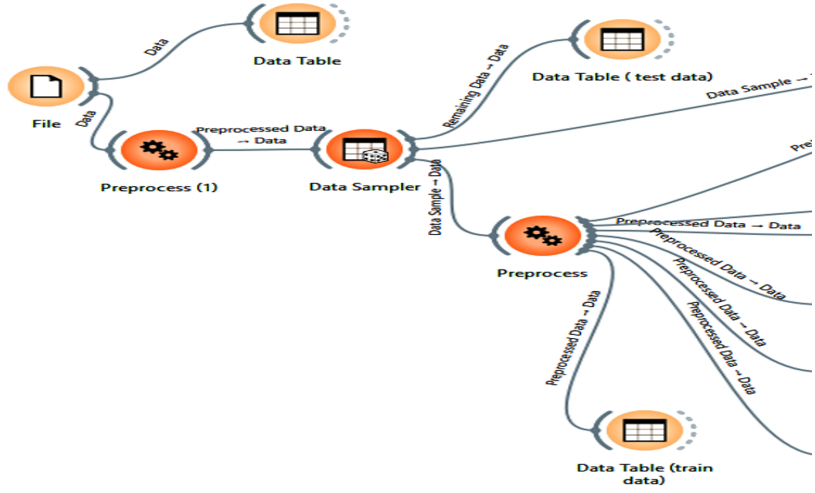

Figure III : Workflow of the Data Analysis Process

OORJA
ISSN - 0974-7869 (Print)
ISSN - 2395-6771 (Online)

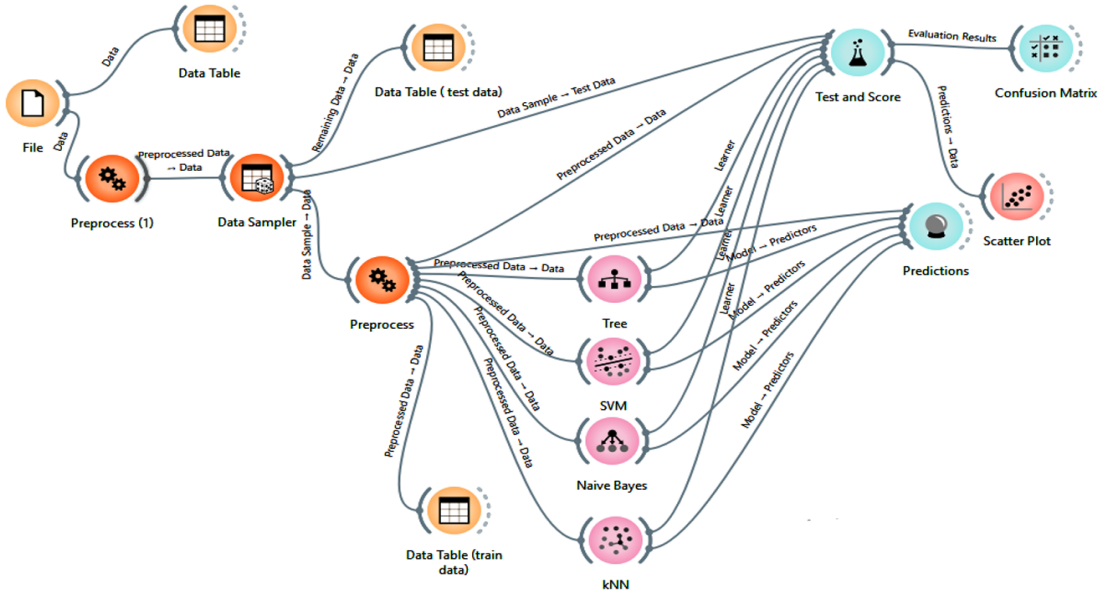

Figure IV :Workflow of Classification Model

D. Model Evaluation Using the prediction widget, the user can predict the value of the target class label for the forthcoming dataset so that decisions can be made effectively. The test and score widget show the values of different parameters, based on which the accuracy of the model can be measured

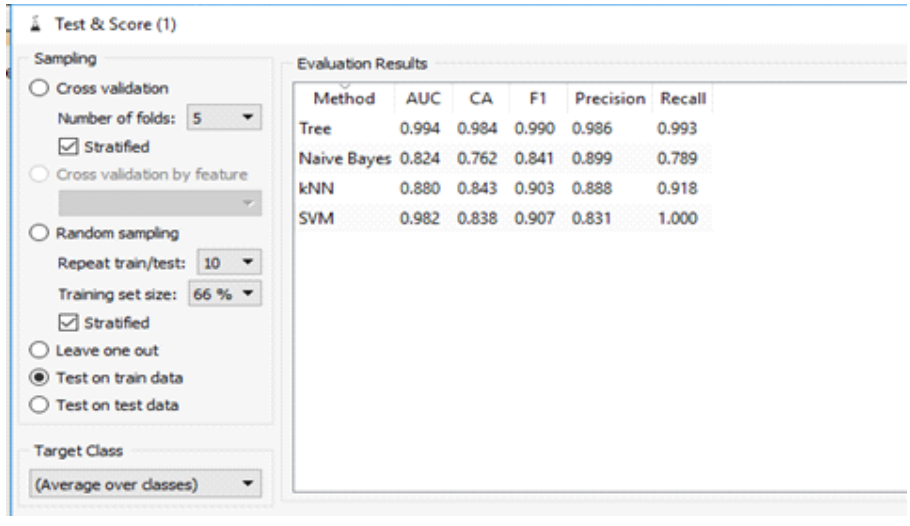

Figure V : Result analysis using Test & score widget

Fig. 5 shows the classification accuracy of the decision tree obtained by choosing any sampling method. These results can be measured by changing the options provided. Fig. 6 shows the ratio of the predicted and actual class instances.

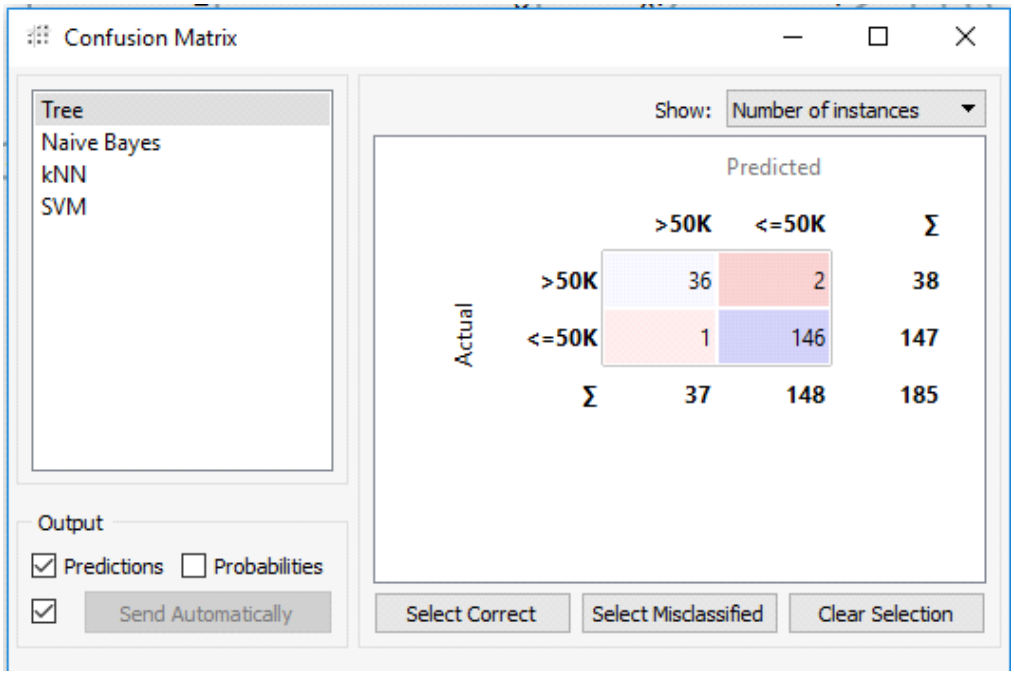

Figure VI :correctly classified and misclassified no. of instances

E. Visualization in Orange

Data visualization is a procedure used to address information-utilizing pictures. Currently, it is becoming advantageous to obtain business. Notwithstanding their true capacity, the advantages of data visualization are sabotaged today by an overall absence of comprehension. Nothing in the field of business insight today can bring us closer to satisfying the guarantee of knowledge in the work environment than data perception. The objective of data visualization is to convey data more intelligently using graphical pictures. representation does not imply that it needs to be drawn out to be useful or colossally tasteful to appear appealing. Several widgets show the characteristics and visualizations of the attributes of a given dataset. Fig. 7 shows the distribution of all data.

OORJA
ISSN - 0974-7869 (Print)
ISSN - 2395-6771 (Online)

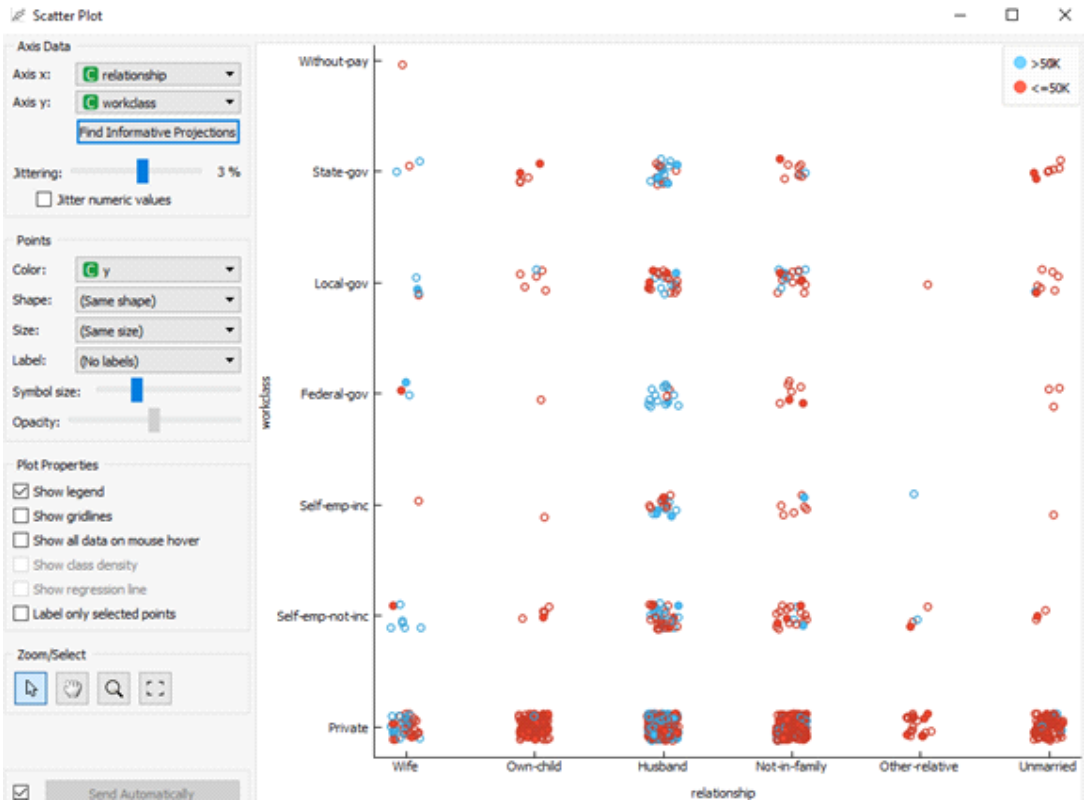

Figure I :Relationship between Attributes Relationship and Work Class

Discussion and Conclusion

The classification model is developed using various classifiers that predicts the value of the dependent variable, income, for various conditional states of the independent variables. It is possible to estimate if future income worth will be little or great. accuracy of predictions is evaluated by contrasting them with the real (Reference) value after extrapolating the income values from the test dataset. Models have been assessed to get the best possible match between predicted and actual data. After using the test dataset to forecast the income values, each model's accuracy is tested by comparing the predicted values to the real (Reference) value. The goal of the model is to have the largest percentage of anticipated and actual values match. The confusion matrix shows the accuracies of the decision tree, Naive Bayes classifier, KNN, and SVM at 0.98, 0.76, 0.84, and 0.83, respectively. Prediction from Decision Tree, some rules are derived that help to predict income greater than 50k or not. Finally, we can conclude that the best model for predicting income is the decision tree model, because it shows the highest accuracy (98.4%) among the three models. The most amazing tool for almost any type of analysis is Orange, and using Orange to visualize datasets is enjoyable. The development of roots marks the beginning of tree growth (located at the top). The information is then split depending on characteristics that can be used as leaves. Creating decision rules from constructed decision trees is known as decision rule formation. The decision tree can be

used to derive the rule by following the branches from root to leaf. The accuracy of the Decision Tree classification process is determined based on the dataset processing utilizing the tree approach employing Orange software. The value of precision is 0.98, indicating that the Decision Tree method is good.

References

Chakrabarty, N., & Biswas, S. (2018, October). A statistical approach to adult census income level prediction. In 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) (pp. 207-212). IEEE.

Deepajothi, S., & Selvarajan, S. (2012). A Comparative Study of Classification Techniques On Adult Data Set 1. International Journal of Engineering Research & Technology (IJERT), 1(8), 1–8.

Income Classification using Adult Census Data (CSE 258 Assignment 2) | Semantic Scholar. (n.d.). Retrieved March 26, 2023, from https://www.semanticscholar.org/paper/Income-Classification-using-Adult-Census-Data-(-CSE-Chockalingam-Shah/3dd5e9f335511efbb81 d65f1d6d 4995019f8b5fd

Khandelwal, R., Goyal, H., & Shekhawat, R. S. (2020). Comparative Analysis of Machine Learning Techniques Using Predictive Modeling. Recent Advances in Computer Science and Communications, 15(3), 1136–1147. https://doi.org/10.2174/2666255813999200904164539

Khandelwal, R., & Virwani, H. (2019). Comparative Analysis for Prediction of Success of Bollywood Movie. SSRN Electronic Journal, 104–111. https://doi.org/10.2139/ssrn.3350907