### **Assignment 1**

### **Group Members:**

|  |  |
| --- | --- |
| **Name** | **Matric Number** |
| SAGAL MOHAMED YUSUF | 22093010 |
| TARSVINI A/P RAVINTHER | 17193844 |
| RABITA BHUIYA TANAYA | 22111473 |
| ZHANGXUAN | 22091244 |

### **TASK 1: Theoretical Understanding**

#### **a) Overview of Data Classification:**

**Data Classification:** One supervised learning approach in data mining and machine learning is data classification, the purpose of which is to organize data sets into predetermined categories. In this step, we use characteristics derived from a training dataset to build a model that can guess the class labels of fresh, unknown data. During training, the model discovers how features relate to class labels; afterwards, it uses this information to classify new instances [1].

**Importance of Data Classification in Data Mining based on Decision Tree Models:**

A key component of data mining and machine learning is data classification, and one of the most often used and successful techniques for this kind of work is the use of Decision Tree models. This section explores the role that data classification plays in data mining, with a focus on the advantages and uses of Decision Tree models.

**Understanding Data Classification:**

Data classification is the process of organizing information into sets with established boundaries. The approach is trained on a labeled dataset, which contains both the features and the matching class labels; it is a sort of supervised learning. The end goal is to construct a model that, given some features, can guess what class new, unseen data pieces belong to.

**Decision Tree Models:**

In a Decision Tree, which resembles a flowchart, each node stands for a feature-based decision, each branch for the decision's conclusion, and each leaf node for a class label. For numerical and categorical data, Decision Trees are the way to go because they are easy to understand and use.

**Importance in Data Mining:**

**1. Decision Making:**

**Predictive Insights:** Predictions provided by Decision Tree models are essential for decision-making processes since they are clear and easy to understand. Decision Trees, for instance, can assess a borrower's propensity to fail on a loan, allowing financial institutions to make educated lending decisions [2]. Financial institutions can learn what causes defaults by looking at the decision routes and changing their policies appropriately.

**Operational Efficiency:** Operational efficiency is enhanced through the automation of decision processes using Decision Tree models. In the medical field, for example, a Decision Tree can categorize patients according to their symptoms and test results, allowing for the early detection and treatment of conditions like diabetes and thyroid disorders.

**2. Pattern Recognition:**

**Fraud Detection:** Identifying fraudulent actions is of utmost importance in the realms of banking and e-commerce. In terms of spotting trends that might be a sign of fraud, decision tree models really shine. By looking at things like transaction amount, location, and frequency, they may sort actions as normal or suspicious in transaction data. This is useful for preventing and detecting fraud in a timely manner.

**Customer Segmentation:** In marketing, successful customer segmentation relies on the ability to recognize trends in consumer behavior. Improved consumer involvement and happiness can result from segmentation that allows firms to target certain demographics with targeted marketing campaigns.

**3. Data Organization:**

**Structured Data:** Decision trees are useful for data organization and structure, which makes data more manageable and retrievable. To better understand and meet the needs of customers, decision trees can be used in customer relationship management (CRM) to group client interactions and feedback into categories.

**Efficient Data Retrieval:** Decision trees make data retrieval more efficient by categorizing data. Documents, emails, and other content can be more easily found and retrieved through classification in information systems.

**4. Risk Management:**

**Credit Scoring:** When it comes to money and banking, risk management is king. Credit scoring makes extensive use of Decision Trees to categorize loan applicants according to their creditworthiness. Financial organizations can enhance their risk assessment and portfolio management using Decision Trees, which analyze features such as income level, job status, and credit history [3].

**Insurance Underwriting:** By categorizing policyholders according to risk criteria, Decision Trees aid underwriting in the insurance industry. Insurers can ensure their business sustainability by setting suitable premiums and reducing losses.

**5. Targeted Marketing:**

**Personalized Campaigns:** Businesses can use Decision Trees to find out who to advertise to. Businesses can target certain populations with relevant, individualized marketing messages by segmenting their clientele according to demographics, shopping habits, and other personal preferences. Because of this, marketing campaigns are more successful and result in better conversion rates.

**Market Segmentation:** By categorizing buyers according to shared traits, Decision Trees facilitate market segmentation. Businesses can improve consumer happiness and loyalty by catering to certain market segments with tailored products and services.

**6. Resource Optimization:**

**Operational Efficiency:** Decision trees can foretell when machines will break down, allowing for preventative maintenance and less downtime in manufacturing. By guaranteeing that machinery runs efficiently, this optimizes resource utilization and boosts productivity.

**Supply Chain Management:** Supply chain management makes use of decision trees to classify items and providers according to performance indicators. By locating trustworthy vendors and top-notch goods, this aids companies in streamlining their supply chain processes.

**7. Enhancing Security:**

**Intrusion Detection:** By distinguishing between legal and fraudulent network activity, Decision Trees aid in the detection of intrusions in the field of cybersecurity. This ensures that sensitive information and systems are protected by allowing for rapid reactions to suspected security breaches.

**Access Control:** By sorting users into categories and then providing access according to those categories, Decision Trees can implement access control. By doing so, we can improve data security and make sure that only authorized workers may access essential information.

**8. Improving Customer Experience:**

**Personalized Recommendations:** Decision Trees can study online shoppers' actions to generate tailored suggestions for products. By making it easier for shoppers to discover items that are a good fit for their tastes, this improves the shopping experience and boosts customer happiness.

**Customer Support:** Customer support tickets can be more efficiently resolved by using Decision Trees to classify them and assign them to the right support team. Customer support becomes more efficient and satisfaction levels rise as a result.

**9. Compliance and Reporting:**

**Regulatory Compliance:** By sorting data according to regulatory criteria, Decision Trees helps firms comply with regulatory requirements. To ensure they comply with anti-money laundering legislation, financial institutions can categorize transactions to identify and report any questionable behavior.

**Reporting:** Organizations cannot accomplish regulatory compliance, performance monitoring, and strategic planning without proper data classification, which in turn allows for the generation of comprehensive and precise reports.

**10. Scientific Research:**

**Data Analysis:** Decision trees are useful in scientific research for sifting through large datasets in search of correlations and trends. Genomic research, astronomy, and environmental science are just a few areas that can benefit from this.

**Hypothesis Testing:** By classifying experimental data, Decision Trees help in hypothesis testing and enable researchers to validate hypotheses and derive relevant conclusions [4].

Decision Tree models are an effective means of data classification, which is of paramount importance in data mining. They are perfect for many categorization jobs because of how easily they can be used, how effectively they handle different sorts of data, and how interpretable they are. Insights provided by Decision Trees enable operational efficiency and strategic success, whether for decision making, pattern recognition, risk management, or improving the customer experience. Decision Trees are foundational to modern data-driven techniques due to their enormous benefits in data classification, despite constraints such as data quality, shifting patterns, and the requirement for ongoing model upgrades.

**Real-World Applications of Data Classification:**

Data classification is a powerful tool that may improve decision-making and operational efficiency in many different industries. Three notable practical uses include identifying spam emails, predicting loan payments, and analyzing online shoppers. All these uses highlight the usefulness and adaptability of categorization models for improving efficiency and resolving difficult issues.

**1. Email Spam Detection:**

**a. Overview:** The identification of spam emails is among the oldest and most prevalent uses of data classification. The main objective is to protect users from dangers by removing unwanted and frequently harmful emails from their inboxes.

**b. Classification Techniques:** Machine learning methods are commonly used by spam detection systems to determine if an email is spam or not [5]. To begin, you will need a labeled dataset that includes both spam and legitimate email instances to train your model.

**The Application of Decision Trees:**

**Feature Selection:** Select aspects from emails by looking for things like the subject line, sender's email address, frequency of links, and the presence of specific keywords.

**Tree Construction:** Using an analysis of these features, the decision tree algorithm builds a tree in which a decision rule is represented by each branch and a feature is represented by each node.

**Classification:** To categorize incoming emails as "spam" or "not spam," they are compared to the decision tree.

For instance:

Consider an email as "spam" if it has more than two links and the term "free" in it.

If not, categorize as "not spam".

**2. Loan Payment Prediction:**

**a. Overview:**The banking sector relies heavily on loan payment prediction applications, such as classification models, to determine the probability of a borrower's payment default [6]. Because of this, banks are better able to control their exposure to risk and make educated loan selections.

**b. Classification Techniques:**Data from previous loan applicants, both those who defaulted and those who did not, is used to build predictive models.

**The Application of Decision Trees:**

**Feature Selection:** Collect information on the applicant, including their credit score, income, work history, loan amount, and amount of debt they currently owe.

**Tree Construction:** These features are used by the decision tree method to build a model that forecasts the chance of a default.

**Classification:** The decision tree is used to evaluate each new loan application and forecast the result (default or no default).

For instance:

Identify the applicant as "high risk" (likely to default) if their credit score is less than 600 and their debt-to-income ratio is high.

Identify as "low risk" (unlikely to default) if the debt-to-income ratio is modest and the credit score is higher than 700.

**3. E-commerce Customer Analysis:**

**a. Overview:**Data classification helps businesses in the e-commerce sector understand their clients better, create targeted marketing campaigns, and monitor consumer behavior [7]. Improved client satisfaction and increased business growth are two outcomes of this application.

**b. Classification Techniques:**Based on demographics, purchase habits, and personal preferences, customer data is divided into various categories.

**The Application of Decision Trees:**

**Feature Selection:** Gather information on past transactions, browsing habits, frequency of purchases, average transaction value, and client demographics.

**Tree Construction:** Using these features, the decision tree algorithm builds a model that divides clients into various groups.

**Classification:** Based on their data, each customer is categorized into a group (such as "high-value customer," "bargain shopper," or "infrequent buyer").

For instance:

Consider a customer to be a "high-value customer" if they routinely shop and buy high-value things.

Identify a customer as a "bargain shopper" if they only make purchases during sales events and spend very little money.

Decision trees assist in data-driven decision making in each of these applications by decomposing complicated datasets into more manageable, easily understood rules. This improves the capacity to correctly classify fresh data using patterns that have been learned.

**b) Challenges in Designing Data Classification:**

Designing data classification systems involves several challenges:

1. **Data Quality:** Input data quality has a major bearing on how well the categorization model works. Unreliable forecasts may result from data that is either skewed, noisy, or incomplete. The first and most important stage in constructing a model is to ensure that the data is of high quality, clean, and representative.
2. **Feature Selection:** A major step in improving classification accuracy is determining which features (attributes) are most important. Adding unnecessary or duplicate features might increase the model's complexity and decrease its performance. To overcome this obstacle, many employ dimensionality reduction, correlation analysis, and feature importance.
3. **Model Selection:** Choosing the right algorithm for categorization is critical. The data type and the nature of the problem dictate the algorithm to be used, as each approach has its own set of advantages and disadvantages. Some common algorithms are neural networks, k-Nearest Neighbors, decision trees, and support vector machines.
4. **Overfitting and Underfitting:** When a model memorizes all the features, including outliers and noise, from its training data, something called overfitting happens, and the model is unable to generalize successfully to new data. In the case of underfitting, the model's simplicity prevents it from capturing the true patterns present in the data. It is essential to construct robust models by balancing these two concerns.
5. **Class Imbalance:** The distribution of class labels is sometimes uneven in real-world datasets, with some classes being underrepresented in comparison to others. [8] Due to this imbalance, the model may perform poorly in the minority class while favoring the dominant one. This problem can be reduced with the use of resampling, cost-sensitive learning, and synthetic data synthesis (e.g., SMOTE (Synthetic Minority Over Sampling Technique)).
6. **Interpretability:** People commonly refer to certain classification models, like deep neural networks, as "black boxes" because of the difficulty in understanding how they make decisions. This can be an issue in fields like healthcare and finance where comprehending the reasoning behind a choice is vital. However, decision trees may not necessarily yield the best accuracy, but they are easier to understand and work with [9].
7. **Scalability:** The amount of computing power needed to train and run classification models is directly proportional to the data volume. A major obstacle is making sure the model can handle big datasets without slowing down.
8. **Deployment:** It takes extensive planning to integrate a trained model into an operating setting. This involves thinking about things like system latency, making predictions in real-time, and keeping the model's performance up as new data comes in.

**TASK 2: Dataset and Tool Evaluation**

**Motivation to Select the Simple Loan Classification Dataset**

To build and assess a classification model, the Simple Loan Classification Dataset is a great option for many reasons. For the practical issue of loan acceptance prediction, this dataset offers a wide variety of demographic and financial characteristics. Listed below are the primary reasons for using this dataset:

**1. Relevance to Financial Decision-Making:**

**Loan Approval Prediction:**

For financial institutions, the approval of loans is a crucial decision-making process. Precise forecasts can reduce risks, expedite the loan approval procedure, and guarantee that credit is granted to borrowers who are most likely to repay it.

The dataset contains key characteristics that financial companies frequently use to evaluate an applicant's creditworthiness, including age, income, and credit score. The dataset is therefore extremely important to financial decision-making in the actual world.

**2. Diverse and Informative Features:**

**Comprehensive Demographic Information:**

**Age:** Borrowing approval trends by age group can be better understood with an eye toward the age distribution.

**Gender:** By looking at gender, we can be sure that lending methods are fair and that there are no prejudices in the approval processes.

**Occupation and Education Level:** Important aspects in determining a loan applicant's financial stability and stability include their occupation and degree of education.

**Marital Status:** One factor that can be used to predict loan outcomes is marital status. This is because marital status can impact a person's financial stability and accountability.

**Financial Indicators:**

**Income:** The capacity to repay a loan is directly related to the applicant's annual income. Loan approval rates tend to be higher for borrowers with higher incomes.

**Credit Score:** To approve a loan, this is among the most important criteria. A person's credit report is a summary of their financial history and how creditworthy they are.

**3. Practical Application and Real-World Utility:**

**Risk Management:**

The insights collected from this dataset can help banks and other financial organizations better manage risk. Lenders can improve their default prediction and decision-making capabilities by constructing a strong classification model [10].

**Regulatory Compliance:**

The dataset guarantees transparent and justifiable loan approval procedures, which promotes compliance with regulatory criteria. It can be easier to record and defend loan decisions before regulatory agencies if one is aware of the components that go into the process.

**Customer Relationship Management:**

Financial organizations can enhance customer happiness by precisely forecasting loan approval. Accurate and prompt predictions facilitate speedier application processing, improving client satisfaction.

**4. Opportunities for Advanced Data Analysis and Model Building:**

**Feature Engineering:** There are many features in this dataset. For instance, converting income to log scale, calculating ratios of income to loan amount, or combining occupation and education level to create indicators of socioeconomic position.

**Model Interpretability:** Decision trees provide great interpretability and can be used with this dataset. It is simpler for financial analysts and other stakeholders to comprehend the model's decision-making process, which facilitates acting on and placing trust in the model's forecasts.

**Balanced Dataset:** A well-balanced dataset guarantees that the model can effectively learn to distinguish between the two groups of loan applications—approved and refused. This aids in developing a model that predicts loan approval and denial.

**5. Addressing Societal and Ethical Considerations:**

**Fairness and Bias Detection:**

The collection enables the identification and rectification of biases in the loan approval procedure by including characteristics like gender, marital status, and educational attainment. It is possible to audit models to make sure they don't unjustly discriminate against any certain group.

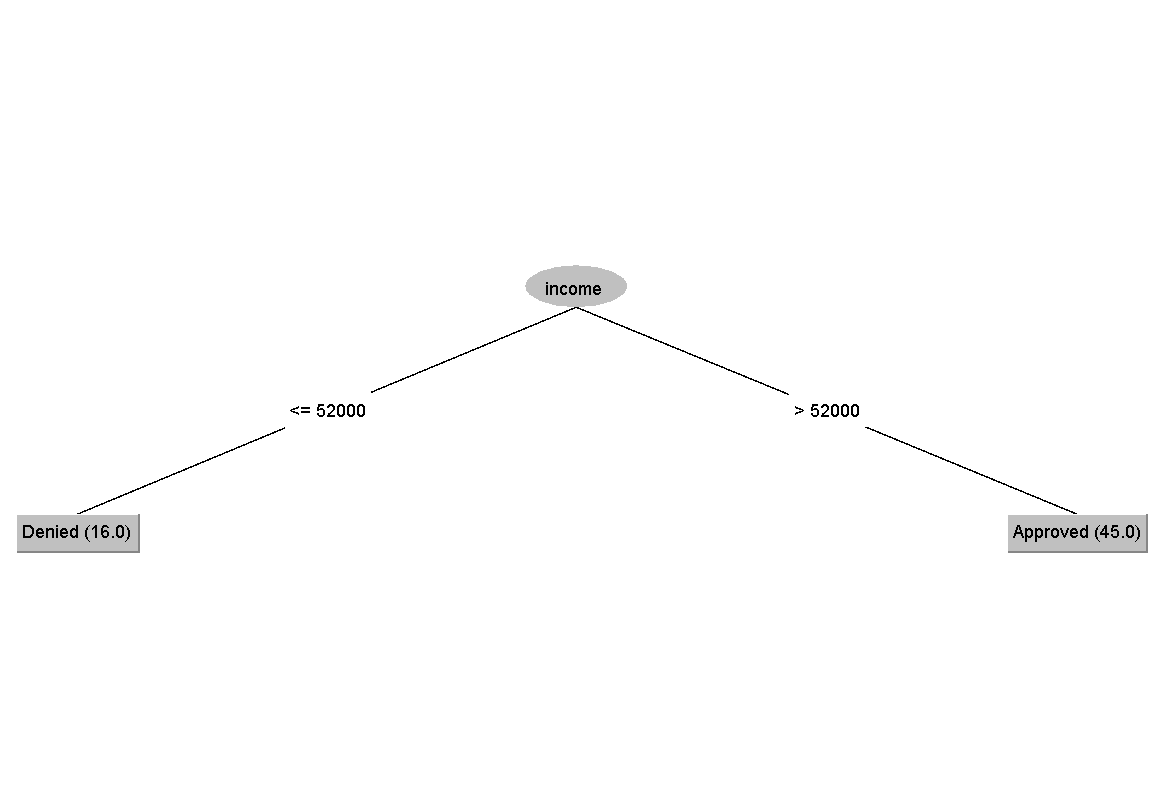
**Socio-Economic Impact:**

By comprehending the elements that contribute to loan acceptance, financial institutions can develop lending policies that are more inclusive. For example, seeing that borrowers from particular categories are routinely turned down for loans could result in focused financial literacy and assistance initiatives.

Because of its relevance to financial decision-making, variety of informative features, practical application, opportunities for advanced data analysis, and ability to address ethical and societal considerations, the Simple Loan Classification Dataset is a good choice for developing a classification model. Financial institutions can guarantee fair lending practices, strengthen their risk management plans, and expedite the loan approval process by utilizing this dataset. This makes it a useful tool for financial analysts and data scientists who want to develop reliable, understandable, and moral models for predicting loan acceptance.

Selecting the appropriate dataset is a crucial step in any machine learning project. The motivation behind this selection revolves around several key factors that ensure the dataset is suitable for the intended analysis and objectives. The rationale behind choosing the right dataset is to make sure that it is relevant to the problem, balanced, scalable, interpretable, comparable, and of excellent quality. It also needs to be of adequate size for instructional reasons. The Iris dataset is a perfect example of a choice that is perfect for developing and accessing decision tree models in an educational setting because of its balance, applicability, ease of use, and simplicity.

Figure 1: Decision Tree using WEKA



### **Justification for Using WEKA**

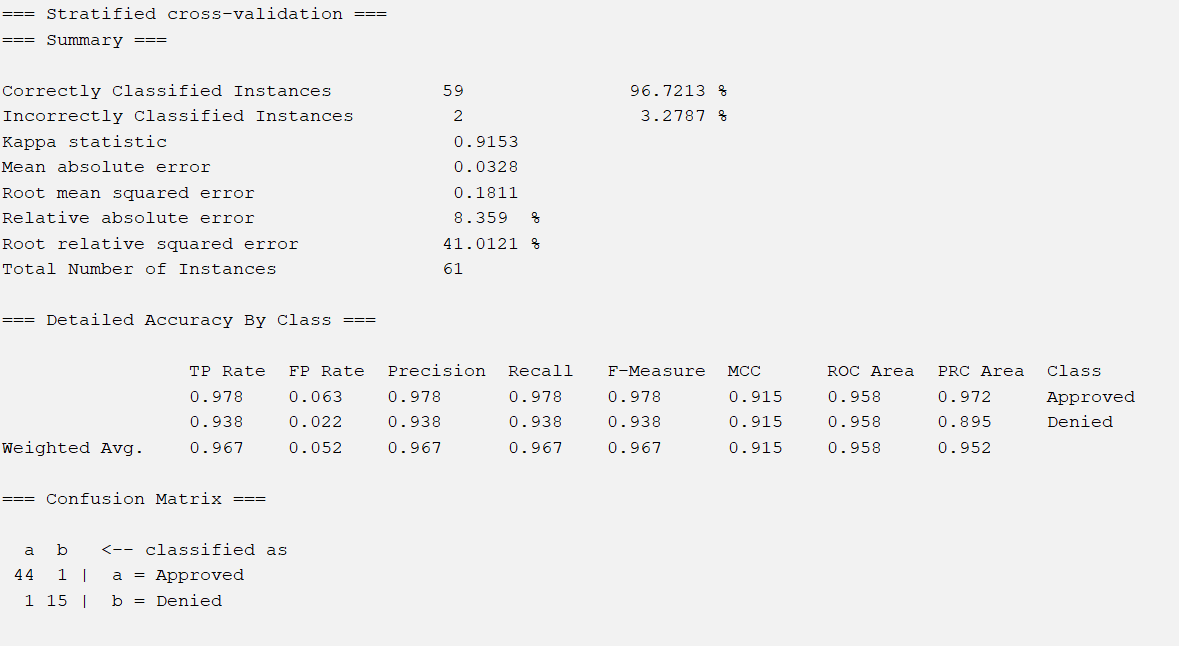
WEKA's strong feature set, ease of use, and extensive support for both novice and expert users make it a worthy pick. WEKA is a perfect tool for the dataset because of its speedy preprocessing capabilities, decision tree algorithm application, and result visualization. Furthermore, testing with various settings and algorithms is made easier by WEKA's graphical user interface (GUI), which promotes a more user-friendly method of machine learning.

**Task 3:**

**Results**

Figure 2 shows the results obtained from WEKA using decision tree.

Figure 2: WEKA Results



Since this is a decision tree model, we retrieved the five evaluations metric as below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Accuracy | Precision | F1-Score | Recall | AUC |
| 96.7% | 96.7% | 96.7% | 96.7% | 95.8% |

Below is interpretation for Confusion Matrix of Decision Tree:

True Positives (TP): 44

True Negatives (TN): 15

False Positives (FP): 1

False Negatives (FN): 1

Based on the results, the decision tree model correctly classifies 59 instances out of 61 instances, achieving a high overall accuracy of 96.72%. The confusion matrix reveals that the model has a low false positive rate and false negative rate. This means the misclassification rate of the model is low. The precision and recall are high as well. In short, the decision tree model performs well in classifying loan applications, effectively distinguishing between approved and denied instances with high accuracy, precision, recall, F1-Score and AUC.

**Summary:**

Data classification is a supervised learning method used to classify data into predefined categories. It involves training a model using labeled data to classify new data. Decision tree models are a prominent and effective technique in data classification, providing clear and explainable decisions for various applications.

Designing a data classification system involves multiple challenges. This paper chooses a simple loan classification dataset because it is relevant to financial decision-making, has diverse and informative characteristics, and can solve social and ethical issues through data analysis. At the same time, it helps financial institutions strengthen risk management, regulatory compliance, and customer relationship management.

WEKA is used because of its powerful functions, ease of use, and support for novices and experts. It provides fast preprocessing, decision tree algorithm application, and result visualization, making it an ideal tool for datasets. Based on the results of using WEKA, the decision tree model achieved high accuracy (96.72%), precision, recall, f1 score, and AUC when classifying loan applications, indicating high classification efficiency and low misclassification rate.

**Potential Future Application:**

Data classification can be widely used in various fields because it can increase efficiency, improve decision-making, and increase overall operational efficiency. For example: In the financial field, these models can enhance risk management by improving credit scoring systems and loan approval processes, ensuring that borrowers are assessed more accurately and fairly. In the healthcare field, patient data can be analyzed to predict potential health problems and recommend personalized treatment plans. In the field of cybersecurity, these models can detect anomalies and potential threats by classifying network activities, thereby strengthening security measures.

**References:**

[1] Data, “What is Data Classification in Machine Learning?,” *Aya Data*, May 29, 2024. <https://www.ayadata.ai/blog-posts/data-classification-in-machine-learning/>

[2] S. H. S. Nor, S. Ismail, and B. W. Yap, “Personal bankruptcy prediction using decision tree model,” *Journal of Economics Finance and Administrative Science/Journal of Economics, Finance and Administrative Science*, vol. 24, no. 47, pp. 157–170, Mar. 2019, doi: 10.1108/jefas-08-2018-0076.

[3] “How credit scores can improve risk assessment,” May 15, 2024. <https://www.comarch.com/finance/articles/how-credit-scores-can-improve-risk-assessment/>

[4] Z. He, C. Sheng, Y. Liu, and Q. Zou, “Instance-Based Classification Through Hypothesis Testing,” *IEEE Access*, vol. 9, pp. 17485–17494, Jan. 2021, doi: 10.1109/access.2021.3053778.

[5] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, “Machine learning for email spam filtering: review, approaches and open research problems,” *Heliyon*, vol. 5, no. 6, p. e01802, Jun. 2019, doi: 10.1016/j.heliyon.2019.e01802.

[6] N. Uddin, Md. K. U. Ahamed, M. A. Uddin, Md. M. Islam, Md. A. Talukder, and S. Aryal, “An ensemble machine learning based bank loan approval predictions system with a smart application,” *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 327–339, Jun. 2023, doi: 10.1016/j.ijcce.2023.09.001.

[7] Z. Oberemok, “Data-Driven Marketing for E-Commerce: Benefits, Trends, Examples,” *Claspo.io*, Jan. 17, 2024. <https://claspo.io/blog/data-driven-marketing-for-e-commerce-benefits-trends-examples/>

[8] O. Yenigün, “Handling Class Imbalance in Machine Learning - Okan Yenigün - Medium,” *Medium*, May 06, 2024. [Online]. Available: <https://medium.com/@okanyenigun/handling-class-imbalance-in-machine-learning-cb1473e825ce>

[9] GeeksforGeeks, “How Decision Tree Depth Impact on the Accuracy,” *GeeksforGeeks*, May 24, 2024. <https://www.geeksforgeeks.org/how-decision-tree-depth-impact-on-the-accuracy/>

[10] G. Barcia, “Unveiling the Future of Lending: Predictive Analytics Revolutionize Loan Default Prediction,” Jul. 26, 2023. <https://www.linkedin.com/pulse/unveiling-future-lending-predictive-analytics-loan-default-barcia/>