# Bahir Dar University

 

# Institute of Technology

**School of Computing and Electrical Engineering**

**Department of Computer Science and Engineering**

***Data mining and Data warehousing***

# Proposal Document

Group Members.

Mulugeta Ayele………………………………….267/2000

Mulugeta Muche………………………………….268/2000

Nebiyou Samuel ………………………………….279/2000

# Table of content

[1. Introduction 2](#_Toc293291561)

[2. Problem statement: 2](#_Toc293291562)

[3. KDD Methodologies 2](#_Toc293291563)

[4. Data Set 4](#_Toc293291564)

#  Introduction

This project focuses on investigating the application of KDD to explore and discover patterns within the bank customer dataset. We will practice the main functions of data mining application. In this proposal paper we have presented the following project descriptions.

#  Problem statement:

This Project addresses study on the given bank customer data. This system identifies bank customer’s pattern based on frequent data similarity. It also classifies given bank customer data based on their region and age (young, medium, old). Three different approaches are followed to tap this information from the given data.

# KDD Methodologies

The bank customer dataset has 12 attributes and 601 records. The dataset attributes are: id, age, sex, region, income, married, children, car, save\_act, current\_act, mortgage and pep. Throughout the data mining process there are some attributes that needs to be removed, discretized and selected for the whole project activities. In this dataset we have selected the following data mining functionalities and algorithms for data extraction that are important for decision making.

1. **Classification:** - Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.

In this project mainly we will do a lot on classification. From different classification algorithms we have selected the following mining algorithms.

*Decision tree: - It* is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distribution on the given data. By applying this graphical representation we will classify customers in to different groups. The tree uses nominal attributes so some of the attributes are converted from numeric to nominal on the pre processing. The method that will be used for processing and presenting data is J48 algorithm.

*Bayesian Classifier:* This classifier works on the Bayesian probability distribution of the data. The classifier evaluates absolute probabilities of the classes looking at the training data distribution. It also evaluates partial probabilities of the features given classifications. This information is used to classify test data.

# Data Set

|  |  |  |
| --- | --- | --- |
| Attribute | Data type | Description  |
| id | String  | a unique identification number |
| age | Numeric  | age of customer in years (numeric) |
| sex | String | MALE / FEMALE |
| region | String | inner\_city/rural/suburban/town |
| income | Numeric | income of customer (numeric) |
| married | String | is the customer married (YES/NO) |
| children | Numeric | number of children (numeric) |
| car | String | does the customer own a car (YES/NO) |
| save\_acct | String | does the customer have a saving account (YES/NO) |
| current\_acct | String | does the customer have a current account (YES/NO) |
| mortgage | String | does the customer have a mortgage (YES/NO) |
| pep | String | did the customer buy a PEP (Personal Equity Plan) after the last mailing (YES/NO) |