Data Mining Final Project Prospectus

Team X: Razan Alshehri, Xinyu Cai, Zihan Xie, Ningyuan Xiong

***Abstract*— This prospectus summarizes the project goals and expectations and includes a summary of what had done so far. In addition, it includes a work-assigned table and and a detailed timeline about what we are going to accomplish.**

I. INTRODUCTION

Decision trees are very common and are widely used in the field of data mining. They are simple, yet very effective classification methods. The underlying concept of decision trees is recursive partitioning where the decision tree algorithm recursively partition data into different subsets, or nodes, until all the partitioned data have the same target variable. Sometimes, partitioning stops when further splitting does not add further value, or information, to the predictions of the target variable.

II. DATASET

The 4 datasets we chose are the German Credit Risk Dataset, Red Wine Quality Dataset, USA Cars Dataset, and Adult Income Dataset. The German Credit Risk Dataset classifies people described by a set of attributes as good or bad credit risks. The Red Wine Quality Dataset is related to red and white vinho verde wine samples that model wine quality based on physicochemical tests. USA Cars Dataset includes information about 28 brands of clean and used vehicles for sale in the U.S. The Adult Income Dataset describes annual income results from various factors. Within these 3 data sets, the German Credit Risk Dataset, Red Wine Quality Dataset, and the Adult Income Dataset have categorical target variables, while the USA Cars Dataset has a continuous target variable.

III. EXPECTATIONS AND GOALS

This project looks at four different decision tree algorithms that are used for classification purposes. These algorithms are Iterative Dichotomiser 3 (ID3), C4.5, Classification And Regression Tree (CART), and Chi-square automatic interaction detection (CHAID). The project examines the mathematical approach of each algorithm in determining which feature offers the most informational gain and where partitioning should occur. In addition, it looks at how the four chosen algorithms differ from the others and the advantages and disadvantages of each. Finally, the project will introduce four different data sets to test the performance of all four algorithms on each data set and the differences in partitioning decisions and metrics used. If time permits, additional exploration of more data sets and potentially trying to test different performance measures besides accuracy.

More specifically, the Classification And Regression Tree (CART) algorithm is the basic decision tree algorithm, but unlike the other three algorithms, the CART algorithm is used for both regression and classification decision trees. The implementation of CART is very similar to C4.5 but it splits variables based on numeric values. It uses Gini impurity to measure how often a randomly chosen element from the set would be incorrectly labeled. In this final project, we will use the CART algorithm to construct both a regression and classification decision tree on four datasets.

ID3 algorithm is a classification algorithm that builds a decision tree using a greedy approach by selecting the best attribute that yields maximum Information Gain or minimum Entropy. The project will discuss the mathematical approach to ID3. Since ID3 is the precursor to the C4.5 algorithm, the project will compare the ID3 algorithm to C4.5 in great detail.

C4.5 can only deal with categorical target variables, so binning the continuous target variable into categorical is necessary. Also, some pruning techniques will be used to prevent overfitting problems.

CHAID uses a chi-square measurement metric to find out the most important feature and apply this recursively until the sub-datasets have a single decision. CHAID uses multiway splits by default and prevents overfitting problems where a node is only split if a significance criterion is fulfilled.

TABLE I
WORK ASSIGNED TABLE

Name	Algorithm	Dataset
Razen Alshehri	CHAID	Adult Income Dataset [1]
Xinyu Cai	ID3	German Credit Risk Dataset [2]
Ningyuan Xiong	CART	Red Wine Quality Dataset [3]
Zihan Xie	C4.5	USA Cars Dataset [4]

IV. PROGRESS

Up to now, we have finalized the expectations and goals for the project and found four data sets suitable for building decision trees. We have assigned each algorithm to each group member, and each of us has already started to work on our algorithms. In addition, we finished doing EDA and have a general understanding of the data sets. The timeline of what we are going to do next has been constructed according to everyone's schedule.

TABLE II

TIMELINE TABLE

Tasks Need To Be Done	Due Date
EDA and Data Preprocessing	11/15/2021
Learn the assigned DT algorithm	11/16/2021
Algorithm implementation	11/18/2021
Learn performance metrics e.g. Recall	11/21/2021
Discuss four algorithm's similarities and differences	11/22/2021
Make presentation PPT	11/26/2021
Final Essay Start	
Motivation,introduction and literature	11/30/2021
Theoretical and results parts	12/4/2021
Combine the write up and revise	12/5/2021

REFERENCES

[1] https://www.kaggle.com/wenruliu/adult-income-dataset?select=adult.csv

[2] https://www.kaggle.com/kabure/german-credit-data-with-risk

[3] https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

[4] https://www.kaggle.com/doaaalsenani/usa-cers-dataset