***A Synopsis for***

## Project Based learning, semester-5

***Submitted by***

**Chetan Malhotra (1413101098)**



**Under supervision of**

**Ms. Sandeep Kaur**

## SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

## GALGOTIAS UNIVERSITY

## GREATER NOIDA, UTTAR PRADESH, INDIA

## **SEPTEMBER,** 2016

**PROJECT TOPIC- *Implementing Decision tree algorithms and building a predictive model that can solve real world problems.***

**CONTENTS**

1. OBJECTIVE

2. INTRODUCTION

3. SCOPE OF THE PROJECT

4. METHODOLOGY, IMPLEMENTATION AND METRICS

5.TECHNOLOGIES USED

6.REFERENCES

**OBJECTIVE**

* To study what Decision trees are.
* To study what Decision tress are capable of doing.
* To study the the creation and implementation of Decision trees.
* To study about various Decision tree algorithms and compare them on the basis of their running time, entropies and the accuracy of their supervised learning.
* To create a model that predicts the value of a target variable by earning simple decision rules inferred from the data features.

**INTRODUCTION**

***->About:***

A **decision tree** is a decision support tool that uses a tree-like graphor [model](https://en.wikipedia.org/wiki/Causal_model)of decisions and their possible consequences, including [chance](https://en.wikipedia.org/wiki/Probability)event outcomes, resource costs, and [utility](https://en.wikipedia.org/wiki/Utility). Decision trees are commonly used in [operations research](https://en.wikipedia.org/wiki/Operations_research), specifically in [decision analysis](https://en.wikipedia.org/wiki/Decision_analysis), to help identify a strategy most likely to reach a [goal](https://en.wikipedia.org/wiki/Goal), but are also a popular tool in machine learning.

**Decision tree learning** uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modeling approaches used in [statistics](https://en.wikipedia.org/wiki/Statistics), data mining and machine learning.

***->A decision tree consists of 3 types of nodes:***

1.Decision node – commonly represented by squares.

2.Chance nodes – represented by circles.

3.End nodes-represented by triangles.

***->Advantages of Decision trees:***

1.Are simple to understand and interpret.

2.Can be combined with other decision techniques.

3.Have value even with little and hard data-set.

**THEREFORE, Decision Trees must be the best modeling tool for prediction and decision making because of their tolerance to missing data, their,acceptance of different data types, and their robustness to assumptions about the input distributions.**

**SCOPE OF THE PROJECT**

Decision making is not at all a new science, but it is always working to find new approaches, algorithms and techniques to make the target decisions more precise, efficient and reliable, no matter how complex the data set is.

Decision making is used almost everywhere, by large organizations for risk management, for learning trends, etc; by an individual who wants to set up a new business, but is confused between a number of options.

**This is where Decision tree and its algorithms come into play as Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.**

Thus predictive models developed by implementing decision trees are able to handle multi-output problems and the cost of using the model(predicting data) is logarithmic in the number of data points used to train the tree.

***Applications of predictive models built with decision trees:***

* **Remote Sensing:** Remote sensing has been a strong application area for pattern recognition work on decision trees.
* **Physics:** Decision trees have been used for the detection of physical particles.
* **Financial analysis:** Use of CART (a Decision tree algorithm) for asserting the attractiveness of buy-writes and investments.
* **Software development:** Regression trees (and back propagation networks) were recently used to estimate the development effort of a given software module in, where it is argued that machine learning methods compare favorably with traditional methods.
* **Astronomy:** Astronomy has been an active domain for using automated classification techniques. Use of decision trees for filtering noise from Hubble Space Telescope images was reported recently. Decision trees have helped in star-galaxy classification, determining galaxy counts and discovering quasars.
* **Object recognition:** Tree based classification has been used recently for recognizing three dimensional objects, for high level vision and also text recognition.

**METHODOLOGY, IMPLEMENTATIONS AND METRICS**

Algorithms for constructing decision trees usually work top-down, by choosing a variable at each step that best splits the set of items. Different algorithms use different metrics for measuring "best". These generally measure the homogeneity of the target variable within the subsets. These metrics are applied to each candidate subset, and the resulting values are combined (e.g., averaged) to provide a measure of the quality of the split.

***HOW TO CREATE A TREE?***

There are a number of algorithms to build and implement a Decision tree about which I have described in this section later. ***But inventing a new algorithm for creating a tree is easy. Here is one:***

1.**Selection of a splitting variable.**

2.**Number of branches:** This algorithm always splits a node into two branches so as to avoid having to decide what an appropriate number of branches would be.

3.**Elusive best splits:** This split finds the split on the selected variable that maximizes some measure of separation of target values.

4.**Recursive partitioning:** This step splits the entire data set, the root node, into two nodes. Each node is then split into more nodes, and so on. Each node in its turn is considered in isolation.

5.**Stopping or pruning:** The size of tree may be the most important single determinant of quality, more important, perhaps, than creating good individual splits. Trees that are too small do not define the data well whereas trees that are too large have leaves with too little data to make any reliable predictions.

6.**Multivariate splits:** Here, the algorithm uses a single input to define a splitting rule.

7.**Missing values:** In this step, the simplest strategy is to regard a missing value as a special non-missing value. ***VARIOUS ALGORITHMS THAT WILL BE USED TO GENERATE DECISION TREES FROM A DATA SET***

* **ID3**: (Iterative Dichotomiser 3) was developed in 1986 by Ross Quinlan. The algorithm creates a multi-way tree, finding for each node (i.e. in a greedy manner) the categorical feature that will yield the largest information gain for categorical targets. Trees are grown to their maximum size and then a pruning step is usually applied to improve the ability of the tree to generalize to unseen data.
* **C4.5:** is the successor to ID3 and removed the restriction that features must be categorical by dynamically defining a discrete attribute (based on numerical variables) that partitions the continuous attribute value into a discrete set of intervals. C4.5 converts the trained trees (i.e. the output of the ID3 algorithm) into sets of if-then rules. These accuracy of each rule is then evaluated to determine the order in which they should be applied. Pruning is done by removing a rule’s precondition if the accuracy of the rule improves without it.
* **C5.0:** is Quinlan’s latest version release under a proprietary license. It uses less memory and builds smaller rule-sets than C4.5 while being more accurate.
* **CART:** (Classification and Regression Trees) is very similar to C4.5, but it differs in that it supports numerical target variables (regression) and does not compute rule sets. CART constructs binary trees using the feature and threshold that yield the largest information gain at each node.

**TECHNOLOGIES USED**

* **Python**: as the programming language for implementing all the algorithms.
* Various machine learning and scientific python libraries like scikit-learn, NumPy and SciPy.
* **Flask:** Flask is on of the many frameworks of python used for server-side scripting.
* **R:** **R** is a programming language and software environment for statistical computing. It will bes used for entropy calculation in this project.
* **OPENSHIFT AND TRAVIS CI:** OPENSHIFT and TRAVIS CI will provide PaaS for hosting and testing my web app on a cloud server.
* **GITHUB:** for source control.
* **HTML, CSS , JAVASCRIPT:** for developing the User interface .
* **Common file extensions that will be generated during the project:** .py, .R, .dat, .csv, .html, .css, .js
* **The project will be developed on UBUNTU but everyone will be able to access it over the world wide web, thanks to Cloud computing.**

**REFERENCES**

* [https://en.wikipedia.org](https://en.wikipedia.org/)
* <http://scikit-learn.org/stable/modules/tree.html>
* <https://github.com/>
* Decision Trees for Predictive Modeling by Padraic G. Neville SAS Institute Inc. 4 August 1999
* <http://www.patricklamle.com/>

**To view and download the soft copy of this project synopsis , please visit:**

https://github.com/debuggermalhotra/Decision-tree-2016/blob/master/Project%20synopsis.docx

*THANK YOU*