Car Evaluation Using Machine Learning

Md Sayeedur Rahman	**170104111**
Md Tanvir Hasan	**170104115**
Muhtasim Sajat	**170104145**

Project Report

Course ID: CSE 4214

Course Name: Pattern Recognition Lab

Semester: Fall 2020

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

September 2021

Car Evaluation Using Machine Learning

Submitted by

Md Sayeedur Rahman	**170104111**
Md Tanvir Hasan	**170104115**
Muhtasim Sajat	**170104145**

Submitted To

Faisal Muhammad Shah, Associate Professor

Farzad Ahmed, Lecturer

Md. Tanvir Rouf Shawon, Lecturer

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

September 2021

ABSTRACT

Automobiles are an inextricable element of our daily life. There are many types of automobiles made by various manufacturers; as a result, customers must make a selection.When a person contemplates purchasing a car, there are a number of factors that might affect his or her decision on the type of vehicle to purchase. A buyer's or driver's decision is influenced by a variety of variables, including price, safety, and the car's luxury and space.The automobile assessment database includes important structured information on the car's characteristics that everyone should look at before making a decision. This data-set is labeled according to the PRICE, COMFORT, and SAFETY requirements.The goal of this study is to identify the decision-making process, identify vehicle factors such as car pricing value, and utilize various variables to determine which cars are excellent and which are acceptable, based on the target value's unaccepted values.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Everyone, especially first-time buyers or those who are unfamiliar with how the automotive industry works, should understand the concept of making a decision on a car purchase. We need a car as a mode of transportation in general, but when we add pleasure to it, we tend to forget that we shouldn't underestimate it.Classifying a good automobile from a good to a bad one is generally done physically with the help of a car sales agent who encourages us to acquire this along these lines or from the conclusion of our relatives and friends who have already encountered vehicle problems. It would have been preferable to have a gadget that could assess automobile characteristics and determine if the vehicle is an X or a Y vehicle. If such a gadget exists, there should be no hesitation in purchasing a vehicle.In today's world, the vehicle sales agent is always encouraging us to buy this car or not. We may or may not realize it, but we are essentially neglecting variables that might benefit us financially, comfortably, and safely in the long term.We process the data in this assignment by exploring the variables' relationships with the attributes and modeling the data using various classification models, including K nearest neighbor and Decision trees, in terms of their best set of parameters for each case and performance on the car evaluation data set.

Chapter 2

Literature Reviews

This research [1] compares seven categorical variable encoding techniques to be utilized with Artificial Neural Networks for classification on a categorical dataset. For training, the Car Evaluation dataset provided by UCI is utilized. When compared to data pre-processed by the other approaches, the data encoded using Sum Coding and Backward Difference Coding techniques provides the best accuracy.

In this work [2], they propose a deep learning-based technique for modeling and predicting sequential decisions made by designers in the context of system design. The combination of the function-behavior-structure model for design process characterization and the long short term memory unit model for deep leaning is at the core of this method. This method [2] is presented in a solar energy system design case study, and the accuracy of its predictions is tested against various frequently used models for sequential design decisions, including the Markov Chain model, Hidden Markov Chain model, and random sequence generation model. The findings show that the suggested method outperforms existing standard models. This means that throughout a system design project, designers are extremely likely to rely on both short-term and long-term memories of previous design decisions to guide their decision-making in the future. As long as the sequential design action data is accessible, our technique may be used in a variety of additional design environments.

In their paper, [3] they examine the detection of misleading information and its significance in decision-making in online materials. The goal is to figure out how alternative approaches may be utilized to solve the problem. This study [3] divides the four forms of misleading information that circulates on social media into four categories. They also go through four deep learning and eight machine learning approaches for detecting fake information. The findings of the paper [3] will give insight into the many forms of false information, associated detection strategies, and the link between false information and decision making for the researchers.Previous research in the subject of fake information detection offered a literature overview. However, by offering precise responses to the stated study questions, they

conducted a comprehensive literature review. As a result, their contribution to the area is unique since this sort of research has never been done before.

Chapter 3

Data Collection & Processing

3.1 Data Collection

For this project, the Car Evaluation Dataset was chosen from the UCI Machine Learning library. There are 1727 instances and 6 characteristics in this collection. To read the automobile assessment data set from our system disk, we're loading the appropriate pandas modules.

3.2 Data Preprocessing

Before proceeding with the module analysis, the dataset from the UCI repository must be cleansed and of standard quality.

Missing values and extreme values, known as outliers, are common in data sets, and these numbers might affect out tests and even cause modules to fail. It is preferable to eliminate any outliers and fill in the missing data with values that are close to them. We don't have any missing values or any type of outliers in our dataset.

	Buying	Maint_cost	Doors	Persons	Lug_boot	Safety	Over_all_score
0	vhigh	vhigh	2	2	small	low	unacc
1	vhigh	vhigh	2	2	small	med	unacc
2	vhigh	vhigh	2	2	small	high	unacc
3	vhigh	vhigh	2	2	med	low	unacc
4	vhigh	vhigh	2	2	med	med	unacc
5	vhigh	vhigh	2	2	med	high	unacc
6	vhigh	vhigh	2	2	big	low	unacc
7	vhigh	vhigh	2	2	big	med	unacc
8	vhigh	vhigh	2	2	big	high	unacc
9	vhigh	vhigh	2	4	small	low	unacc

Figure 3.1: Sample(10 rows) of our Dataset

According to the above result, finding missing values is a simple process, but deciding how to handle missing values is more challenging. Missing values in categorical data sets are not a problem since we can consider them as NA, but missing values in numerical variables would complicate our analysis. Data cleansing is completed before we begin.

It's a good idea to examine the dimensiona of our dataset before commencing the analysis by looking at the share and description of the variables.

3.3 Encoding

Our dataset is in human-readable form. So to make it understandable to computer, dataset needs to be encoded.

Categorical Encoding

Categorical Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

	Buying	Maint_cost	Doors	Persons	Lug_boot	Safety
0	1	1	1	1	1	1
1	1	1	1	1	1	2
2	1	1	1	1	1	3
3	1	1	1	1	2	1
4	1	1	1	1	2	2
...	...	...	...	...	...	...
1723	4	4	4	3	2	2
1724	4	4	4	3	2	3
1725	4	4	4	3	3	1
1726	4	4	4	3	3	2
1727	4	4	4	3	3	3

1728 rows × 6 columns

Figure 3.2: After Encoding of our Dataset.

3.4 Data Distribution

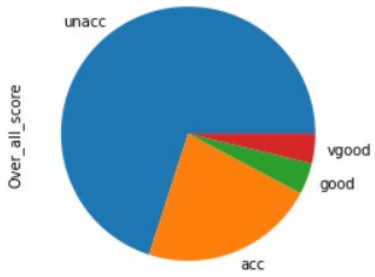

Figure 3.3: Before sampling

From the above chart We can see that our dataset has imbalance distribution of data. So before training our dataset, It needs to be balanced.

Sampling

Data sampling refers to statistical methods for selecting observations from the domain with the objective of estimating a population parameter. Whereas data resampling refers to methods for economically using a collected dataset to improve the estimate of the population parameter and help to quantify the uncertainty of the estimate.

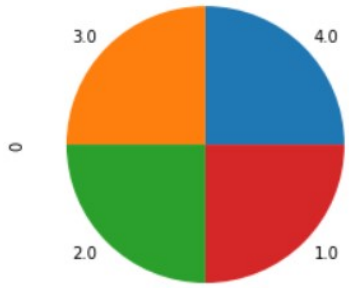

Figure 3.4: After sampling

From the above figure we can see that after applying sampling our dataset is now balanced.

Chapter 4

Methodology

The experiment is conducted out utilizing the Random Forest Classifier , Support Vector Machine and Decision trees classifier models. The goal of this experiment is to figure out which classifier is best for our data set in terms of categorizing the trained and tested sets, as well as making predictions based on the training data. The experiment's complete process is listed below.

4.1 Support Vector Machine

The Support Vector Machine, or SVM, is a linear model that may be used to solve classification and regression issues. It can handle both linear and nonlinear problems and is useful for a wide range of applications. SVM is a basic concept: The method divides the data into classes by drawing a line or hyperplane.

4.2 Random Forest

A random forest is a machine learning approach for solving classification and regression issues. It makes use of ensemble learning, which is a technique for solving difficult problems by combining many classifiers. Many decision trees make up a random forest algorithm.

4.3 Decision Tree Classifier

A decision tree is a module that employs a tree-like graph or a module containing decision conditions and probable outcomes. It's one way to show an algorithm that's entirely made

up of conditional control statements.

Each internal node has a flowchart-like structure that is conditional on each attribute, with each branch representing the condition's conclusion and each leaf node representing a class table. The categorization criteria follows a top-down method from the root to the leaf . When a sub node is split into several sub nodes, it is referred to as a decision node. When a node can't be divided any further into subnodes

4.4 K-Fold Cross Validation

In the field of machine learning cross validation is widely used. Original dataset is partitioned amoung k equal size subsamples. From K subsamples, one of the subsamples is takenas the validation data for the testing model and the remaining k-1 subsamples are used astraining data. In our project work we used 10 fold cross validation. The advantage of this procedure is that each sample is given the opportunity to be used.

4.5 Methodology Diagram

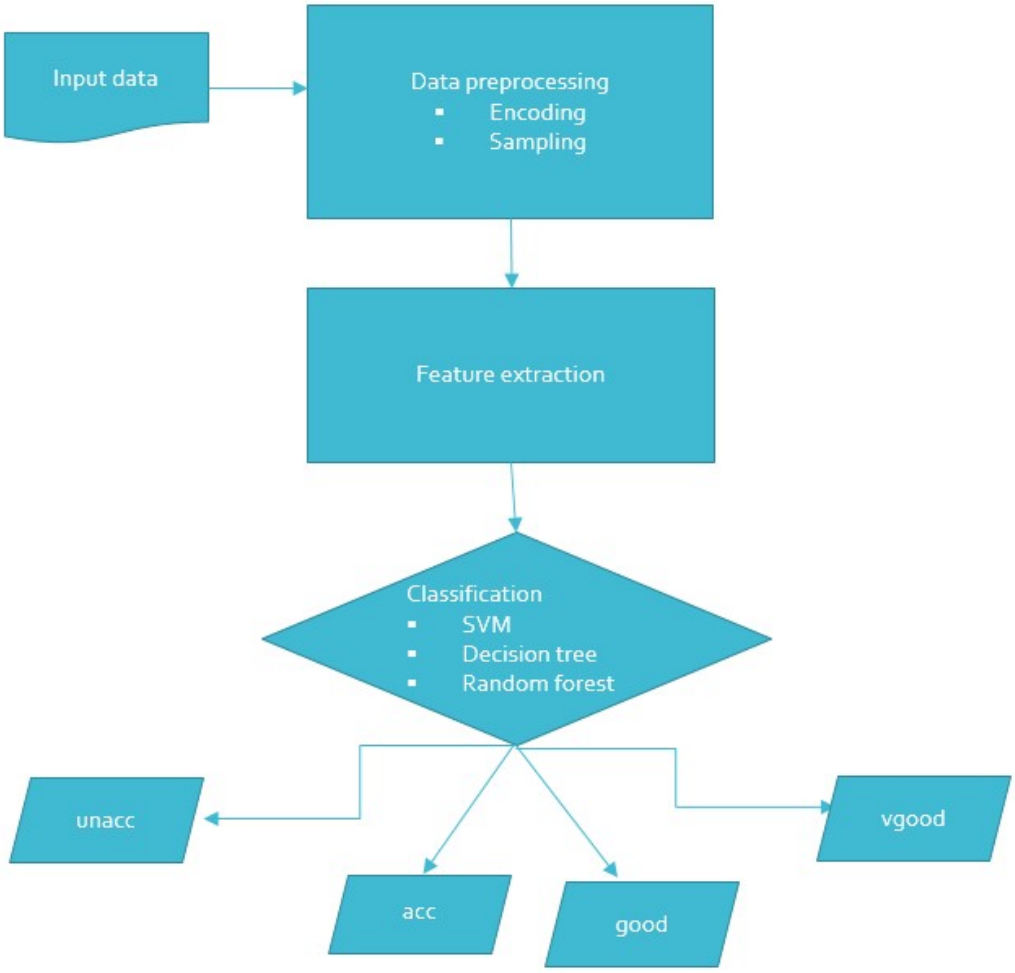

Figure 4.1: Methodology

Chapter 5

Experiments and Results

The presentation of the results is based on following model analysis. We used Support vector machine, Random forest and Decision tree algorithms for the classification purpose by using built in sklearn library. We have performed hyperparameter tuning approach using GridsearchCV. We optimized each algorithm by finding the best possible hyperparamter. The best accuracy obtained by each algorithm according to their best parameters are given bellow.

Table 5.1: Machine Learning based models Performance metrics

Models	**Accuracy**	**Best Parameters**
SVM	96.9628%	C=10,gamma=0.55 kernel=rbf
Random Forest	93.3388%	Max-depth=6, max features=sqrt, min leaf=1
Decision tree	98.3678%	Criterion=entropy, maxdepth=11

5.1 Support Vector Machine

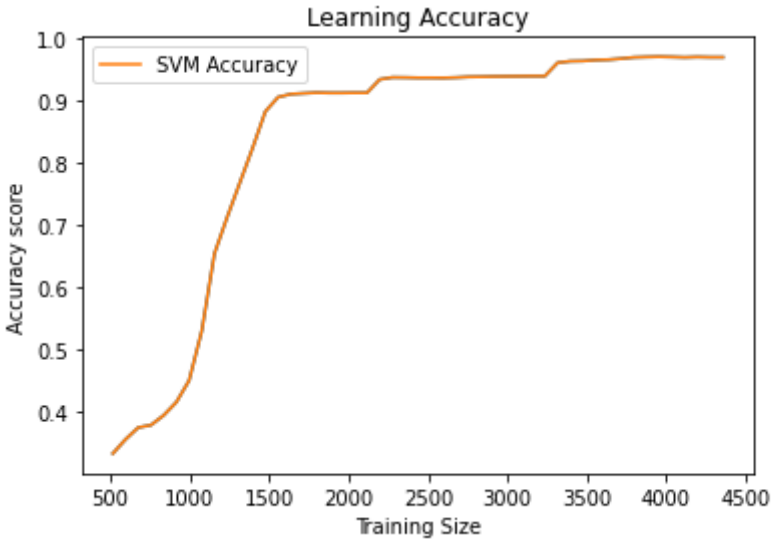

Figure 5.1: Learning Accuracy for Support Vector Machine Model

Confusion matrix

Below is the output of the 80-20 split of the data set confusion metrics for our SVM model.

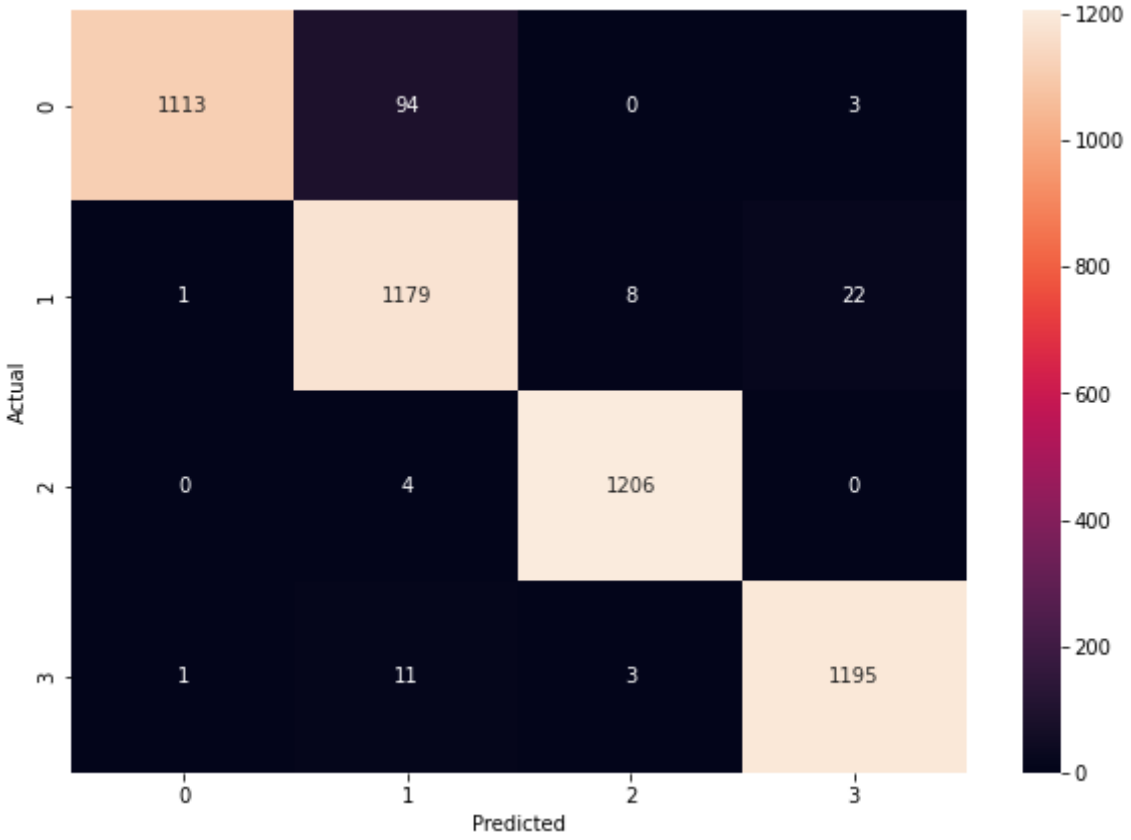

Figure 5.2: Confusion Matrix for Support Vector Machine Model

5.2 Random Forest Classifier

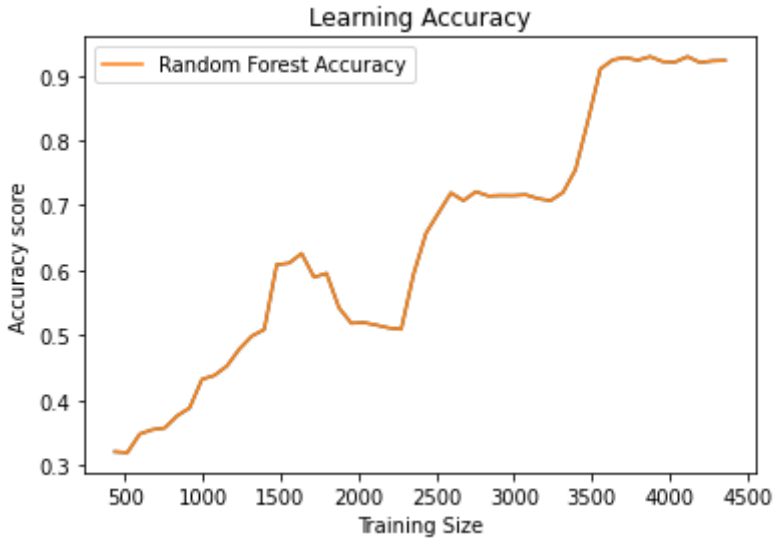

Figure 5.3: Learning Accuracy for Random Forest Classifier Model

Confusion matrix

Below is the output of the 80-20 split of the data set confusion metrics for our random forest model.

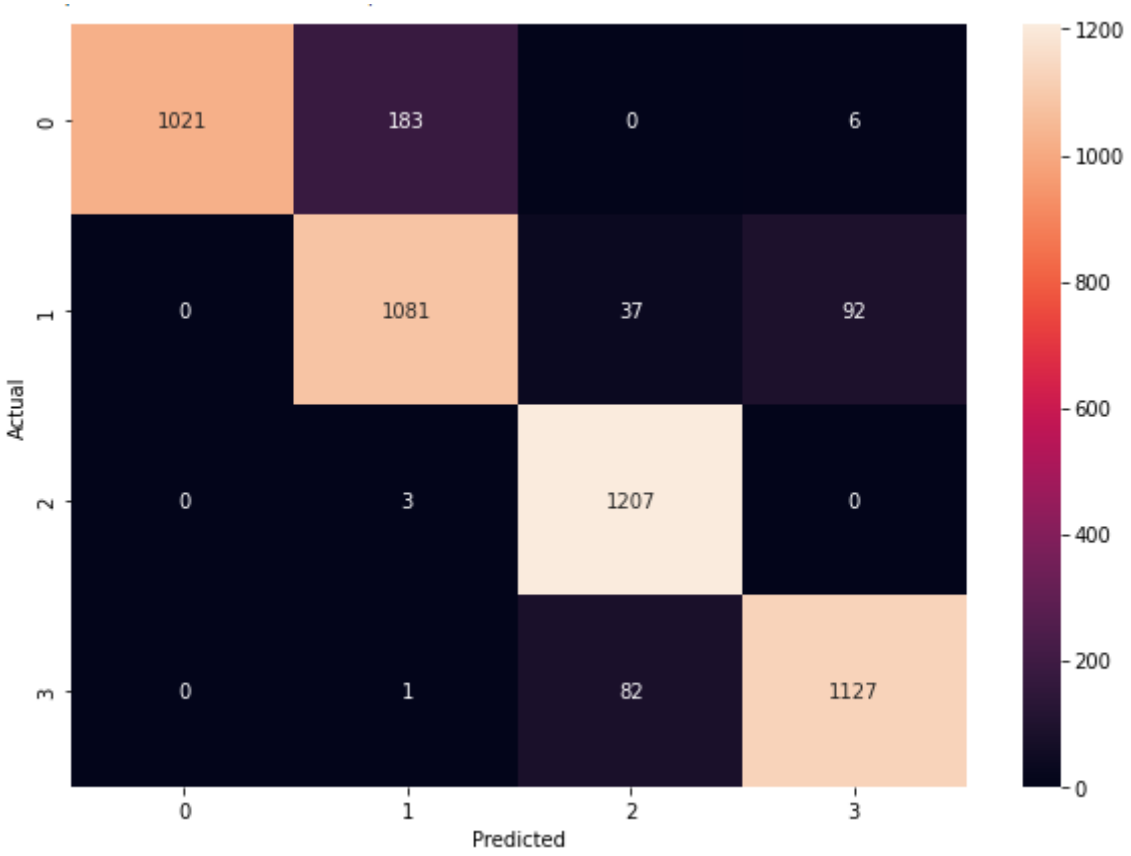

Figure 5.4: Confusion Matrix of Random Forest Classifier Model

5.3 Decision Tree Classifier

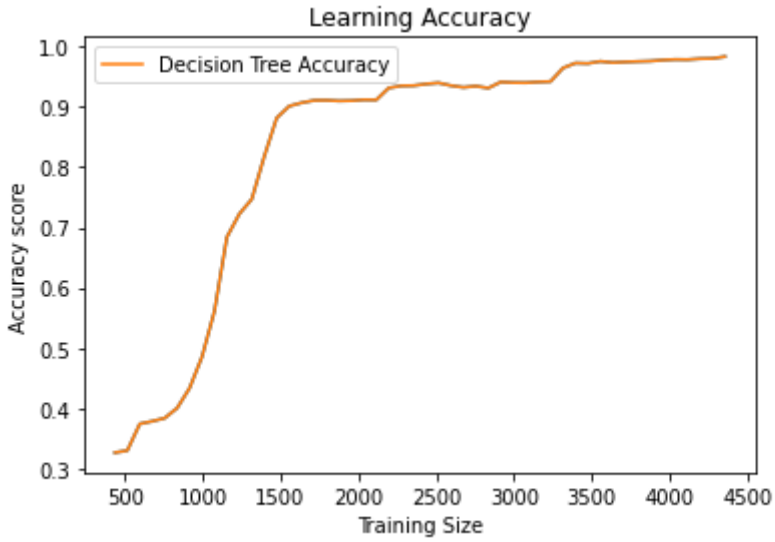

Figure 5.5: Learning Accuracy for Decision Tree Classifier

Confusion matrix

Below is the output of an 80-20 split of the data set confusion metrics for our decision tree model.

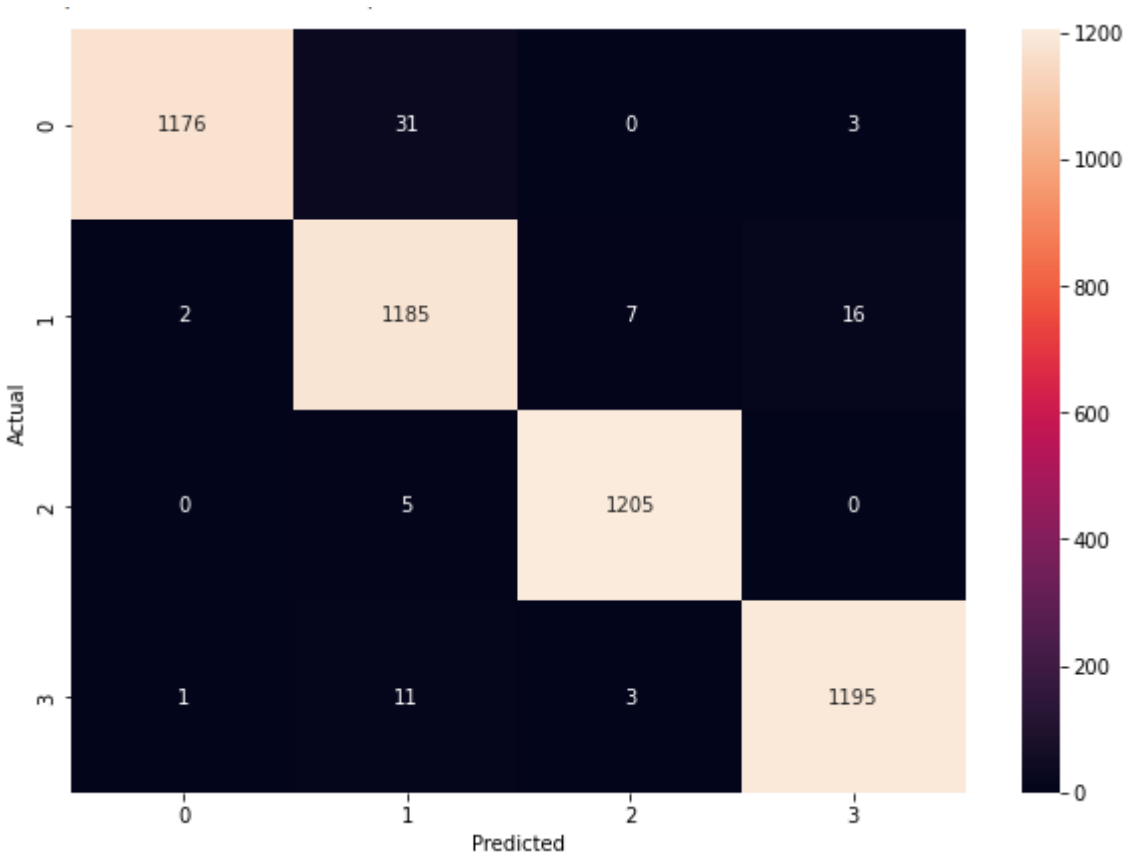

Figure 5.6: Confusion Matrix for Decision Tree Classifier

Chapter 6

Future Work and Conclusion

This paper demonstrated and compared the classification accuracy of different machine learning classification models applied to categorical data encoded using categorical encoding techniques. The aim of this study was to find out the most accurate model for the used car dataset. According to prediction results, the decision tree algorithm performs best, with an accuracy of around 98.36%.

We will try to train our model with a larger dataset in the future, and we'll try to apply deep neural network algorithms to our dataset.

References

[1] K. Potdar, T. S. Pardawala, and C. D. Pai, "A comparative study of categorical variable encoding techniques for neural network classifiers," *International journal of computer applications*, vol. 175, no. 4, pp. 7–9, 2017.

[2] M. H. Rahman, C. Xie, and Z. Sha, "A deep learning based approach to predict sequential design decisions," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 59179, p. V001T02A029, American Society of Mechanical Engineers, 2019.

[3] A. Habib, M. Z. Asghar, A. Khan, A. Habib, and A. Khan, "False information detection in online content and its role in decision making: a systematic literature review," *Social Network Analysis and Mining*, vol. 9, no. 1, pp. 1–20, 2019.

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This project report was generated on Wednesday 6th October, 2021 at 6:05pm.