

**PROJECT BASED LAB REPORT**

**On**

**Predict Book Reviews**

**Submitted in partial fulfilment of the**

**Requirements for the award of the Degree of**

**Bachelor of Technology**

**In**

**Computer science and Engineering**

**Under the esteemed guidance of**

**Dr.Md.Ismail**

**BE,Mtech,Phd**

**By**

|  |  |
| --- | --- |
| **STUDENT ID** | **STUDENT NAME** |
| **M. Sai Teja Reddy** | **170030832** |
| **M. V. S. Pavan Kalyan** | **170030860** |

**(DST-FIST Sponsored Department)**

**K L EDUCATION FOUNDATION**

**Green Fields, Vaddeswaram, Guntur District-522 502**

**2019-2020**

**K L EDUCATION FOUNDATION**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**(DST-FIST Sponsored Department)**

****

**CERTIFICATEEEEE**

This is to certify that this project based lab report entitled **“predict book reviews** “ is a bonafide work done by M.Sai Teja Reddy 170030832, M.V.S. Pavan Kalyan 170030860 in the course 17CS3159 Data Warehousing and Mining in partial fulfilment of the requirements for the award of Degree in Bachelor of Technology in **COMPUTER SCIENCE AND ENGNEERING** during the Odd Semester of Academic year 2019-2020.

**Faculty in Charge Head of the Department**

**K L EDUCATION FOUNDATION**

**DEPT OF COMPUTER SCINCE AND ENGINEERING**

**(DST-FIST Sponsored Department)**

****

**DECLARATION**

We hereby declare that this project based lab report entitled **“predict book reviews”** has been prepared by us in the course 17CS3159 Data Warehousing and Mining in partial fulfilment of the requirement for the award of degree bachelor of technology in **COMPUTER SCIENCE AND ENGINEERING** during the Odd Semester of the academic year 2019-2020.We also declare that this project-based lab report is of our own effort and it has not been submitted to any other university for the award of any degree.

**Date:**

**place:**

**ACKNOWLEDGEMENT**

Our sincere thanks to **Dr.Md.Ismail** in the Lab for her outstanding support throughout the project for the successful completion of the work.

We express our gratitude to **Dr. Vijaya Sri Kompalli,** Course Co-ordinator for **17CS3159 Data Warehousing and Mining** course in the Computer Science and Engineering Department for providing us with adequate planning and support and means by which we can complete this project-based Lab.

We express our gratitude to **Mr.V. HARIKIRAN,** Head of the Department forcomputer science and Engineering for providing us with adequate facilities, ways and means by which we can complete this project-based Lab.

We would like to place on record the deep sense of gratitude to the honourable Vice Chancellor, K L University for providing the necessary facilities to carry the project-based Lab.

Last but not the least, we thank all Teaching and Non-Teaching Staff of our department and especially our classmates and our friends for their support in the completion of our project-based Lab.

**M. Sai Teja Reddy 170030832**

**M.V.S. Pavan Kalyan 170030860**

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **CHAPTERS** | **PAGE NO** |
| ABSTRACT | 7 |
| CHAPTER 1: INTRODUCTION |  |
| * 1. INTRODUCTION | 8 |
| * 1. PROBLEM DEFINITION | 8 |
| * 1. SCOPE | 8 |
| * 1. PURPOSE | 9 |
| 1.5 PROBLEM AND EXISTING TECHNOLOGY | 9 |
| 1.6 PRPOSED SYSTEM | 9 |
|  |  |
| CHAPTER 2: REQIUREMENTS & ANALYSIS |  |
| 2.1 PLATFORM REQUIREMENTS | 10 |
| 2.2 MODULE DESCRIPTION | 11 |
|  |  |
| CHAPTER 3: DESIGN & IMPLEMENTATION |  |
|  |  |
| 3.1 ALGORITHMS | 12 |
| 3.2 PSEUDO CODE | 13 |
| CHAPTER 4: SCREENSHOTS | 17 |
|  |  |
| CHAPTER 5: CONCLUSION | 22 |
|  |  |
| CHAPTER 6: REFERENCES | 22 |

**LIST OF FIGURES AND TABLES**

|  |  |  |
| --- | --- | --- |
| **FIGURES AND TABLES** | | **PAGE NO** |
| Tab2.1. | | 10 |
|  | |  |
|  | |  |
|  | |  |
|  | |  |
|  | |  |
|  | |  |
|  | |  |
|  |
|  | |  |
|  | |  |
|  | |  |
|  | |  |
|  | |  |
|  | |  |
|  | |  |

**Abstract**

Over the years, several statistical tools have been used to analyze and predict book reviews from different point of view. One of the biggest challenges for data mining Today is to predict the paths of books through the data mining process. Successful book review prediction in early course stage depends on many factors. Data mining techniques could be used for this kind of job. Data mining techniques are widely used in educational field to find new hidden patterns from books’s data. The hidden patterns that are discovered can be used to understand the problem arise in the educational field. Data Mining (DM), or Knowledge Discovery in Databases (KDD), is an approach to discover useful information from large amount of data. Data mining techniques apply various methods in order to discover and extract patterns from stored data Based on collected books information, different data mining techniques need to be used. For the purpose of this project AZURE data mining software is used for the prediction of final book reviews based on parameters in the given dataset. The dataset contains information about different kinds of books and ratings based on review. Book review data is used for test dataset.

**1.INTRODUCTION**

A data warehouse is a technique for collecting and managing data from varied sources to provide meaningful business insights. It is a blend of technologies and components which allows the strategic use of data.

Data mining is looking for hidden, valid, and potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected/ previously unknown relationships amongst the data.

* 1. **Introduction:**

Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we develop the software that can learn how to classify the data items into groups. For example, we can apply classification in the application that given all records of customer who left the company, predict who will probably leave the company in a future period. In this case, we divide the records of employees into two groups that named “leave” and “stay”. And then we can ask our data mining software to classify the employees into separate groups.

**1.2 Problem Definition**

In simple classifier model, a simple count of positive and negative data points will define overall positive or negative sets. There is a problem with this. For example, in the case of words in sentences, “Great” and “Good” both are positive words. But “Great” has a higher impact than “Good”. We need to train our model to weigh identified data points.

Classifier model boundaries can be a simple line to separate “positive” from “negative” outcome to more complex hyperplane to separate multiple groups.

**1.3 Scope**

Predictive processing (PP) is the name for a class of theories in cognitive neuroscience which present the prospect of unifying various accounts of perception, action, and very many other ordinary and pathological cognitive phenomena. Roughly, the central idea is that both perception and action can be explained in terms of a mechanism whose sole function is the minimisation of surprise. There are an astounding range of ordinary and pathological cognitive phenomena that have possible explanations expressed in the theoretical vocabulary of PP, and the number of these continues to grow.

**1.4 Purpose**

**Book reviews** can be brief or long. They can critique and/or summarize the **book**. They can be written by readers or professional **book** reviewers. Reader **reviews** tend to be more personal, focusing on the individual reader's experience while reading the **book**.

**1.5 Problem and Existing Technology**

At present situation there is huge amount of data in world , which we cant even think or imagine .among this large data reaching customers satisfaction and needs individually is very difficult . considering every single customer and know the status of relation with the company is impossible. This situation may leads to companies fall down.the existing technology is manual prediction of the relation with company but is inaccurate and some times totally mislead the company.

**1.6 Proposed System**

we propose a review rating prediction method based on user context and product context by incorporating user information and product information into review texts. Our method firstly models the user context information of reviews, and then models the product context information of reviews. Finally, a review rating prediction method that is based on user context and product context is proposed. Our method consists of three main parts.

**Objectives of Proposed System:**

\*Easily predict the performance of student.

\*Huge data can be predict very easily.

\*High accuracy

**2.REQUIREMENTS**

**2.1. Platform Requirements**

|  |  |  |
| --- | --- | --- |
| Hardware/Software | Hardware / Software element | Specification /version |
| Hardware | Processor | I3 |
| RAM | 8GB |
| Hard Disk | 1TB |
| Software | OS |  |
| Python IDE  Microsoft Azure |

**Tab 2.1. Platform Requirements**

**2.2. Modules Description**

In this project we have two sub modules which are listed below.

**2.2.1:Loading dataset**

\*importing the required packages

\*load dataset

\*preprocessing techniques

\*graphs

\*selecting required features and target

**2.2.2 :Classification techniques**

**\***knn classifier

\*decision tree classifier

**3.DESIGN AND IMPLEMENTATION**

**3.1 Algorithms:**

**Knn-classification:**

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition. choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee. Cross-validation is another way to respectively determine a good K value by using an independent dataset to validate the K value. Historically, the optimal K for most datasets has been between 3-10.

**Decision tree classifier:**

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and

their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.A decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

**3.2Source Code**

**Importing packages:**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model\_selection import train\_test\_split

from sklearn.externals.six import StringIO

from IPython.display import Image

from sklearn.tree import export\_graphviz

import pydotplus

**Importing dataset:**

data=pd.read\_csv("C:\\Users\\mylaptop\\Desktop\\book.csv")

data.head()

**KNN-CLASSIFICATION:**

from sklearn.neighbors import KNeighborsClassifier

feature\_cols = ['Review\_1', 'Review\_2', 'Review\_3', 'Review\_4']

k=['class']

X = data1[feature\_cols]

y = data1[k]

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.1, random\_state=1)

knn = KNeighborsClassifier(n\_neighbors=5)

knn.fit(X\_train, y\_train)

y\_pred = knn.predict(X\_test)

print('predicted output in k nearest neighbours:',y\_pred)

print("Accuracy:",metrics.accuracy\_score(y\_test, y\_pred))

**DECISION TREE CLASSIFIER:**

feature\_cols = ['Review\_1', 'Review\_2', 'Review\_3', 'Review\_4']

k=['class']

X = data[feature\_cols]

y = data[k]

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.1, random\_state=1)

clf = DecisionTreeClassifier(criterion="entropy", max\_depth=3)

clf = clf.fit(X\_train,y\_train)

y\_pred = clf.predict(X\_test)

print("predected output is:",y\_pred)

print("Accuracy:",metrics.accuracy\_score(y\_test, y\_pred))

-------------------------------------------------------------

/\*for visualizing tree\*/

from sklearn.externals.six import StringIO

from IPython.display import Image

from sklearn.tree import export\_graphviz

import pydotplus

dot\_data = StringIO()

export\_graphviz(clf, out\_file=dot\_data,

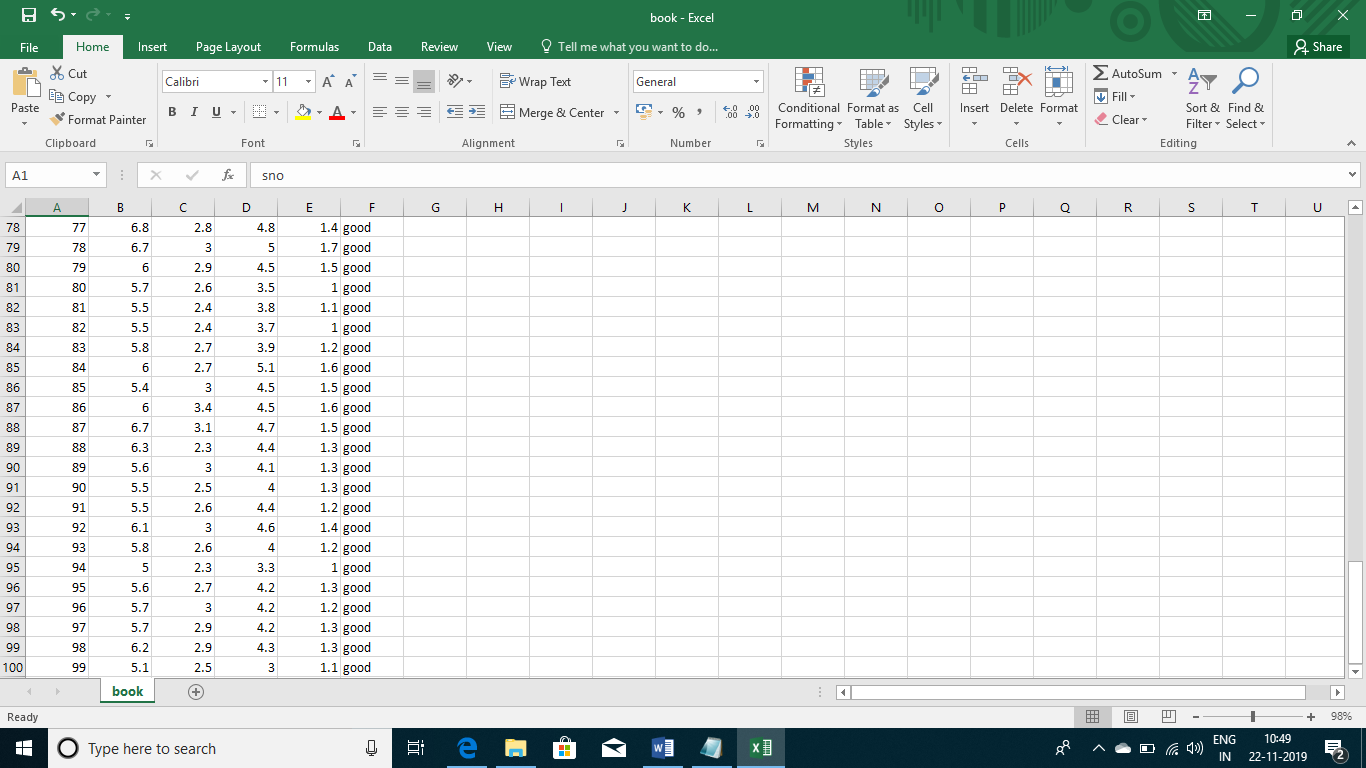
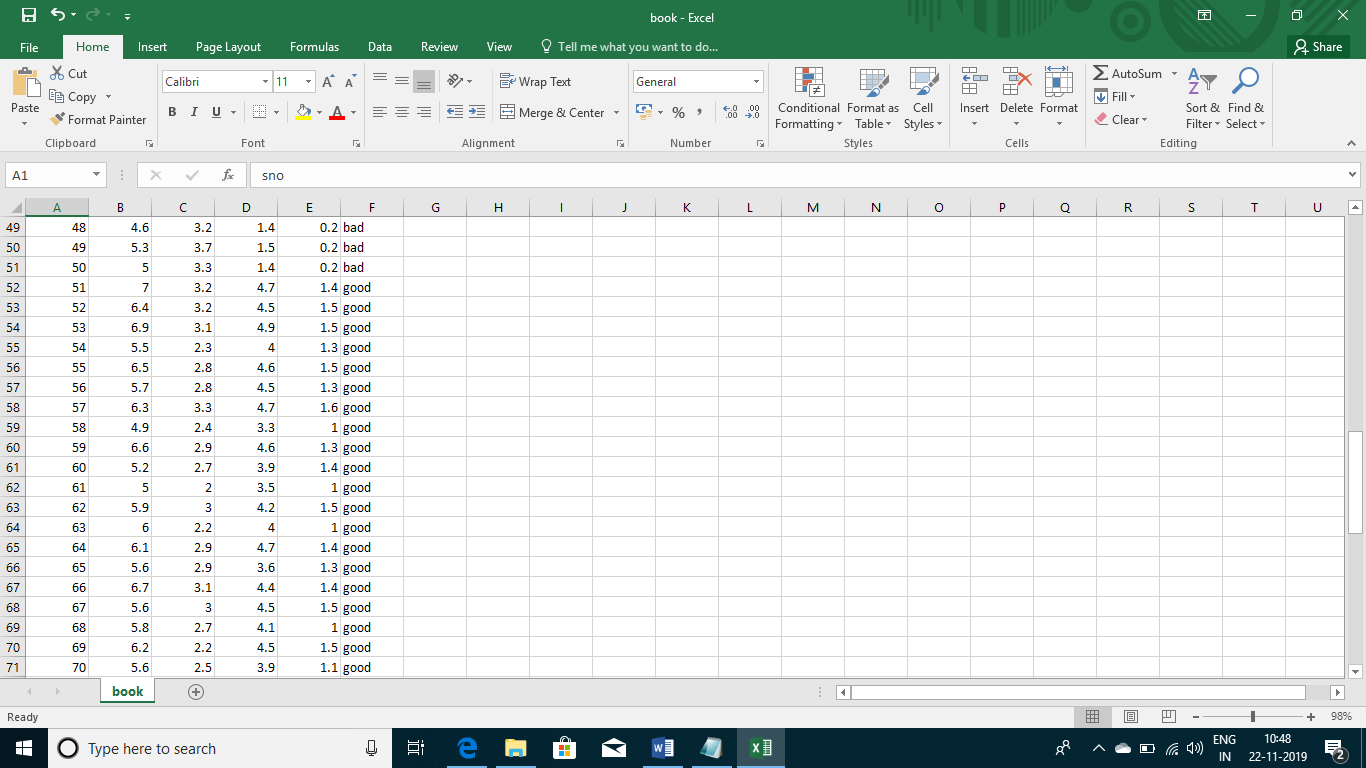
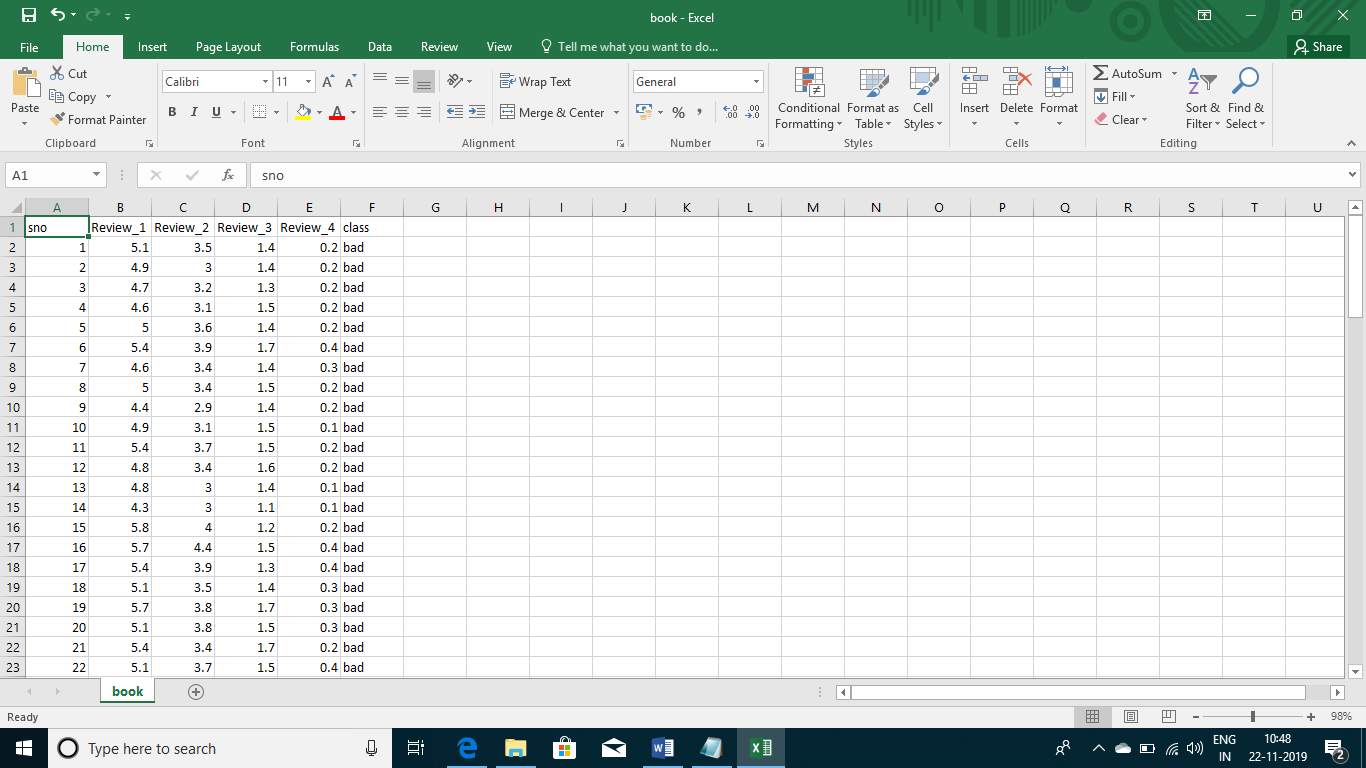
filled=True, rounded=True,

special\_characters=True, feature\_names = feature\_cols,class\_names=['bad','good'])

graph = pydotplus.graph\_from\_dot\_data(dot\_data.getvalue())

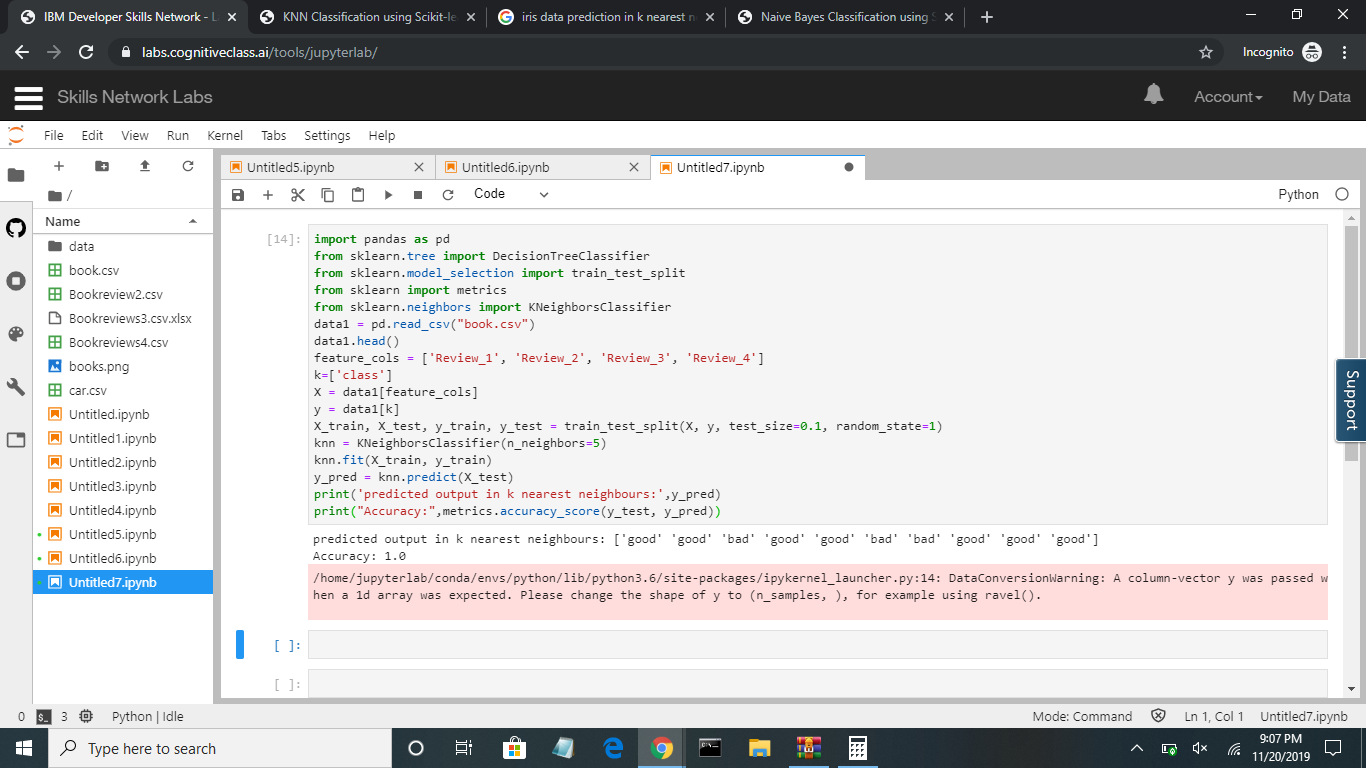
graph.write\_png('books.png')

Image(graph.create\_png())

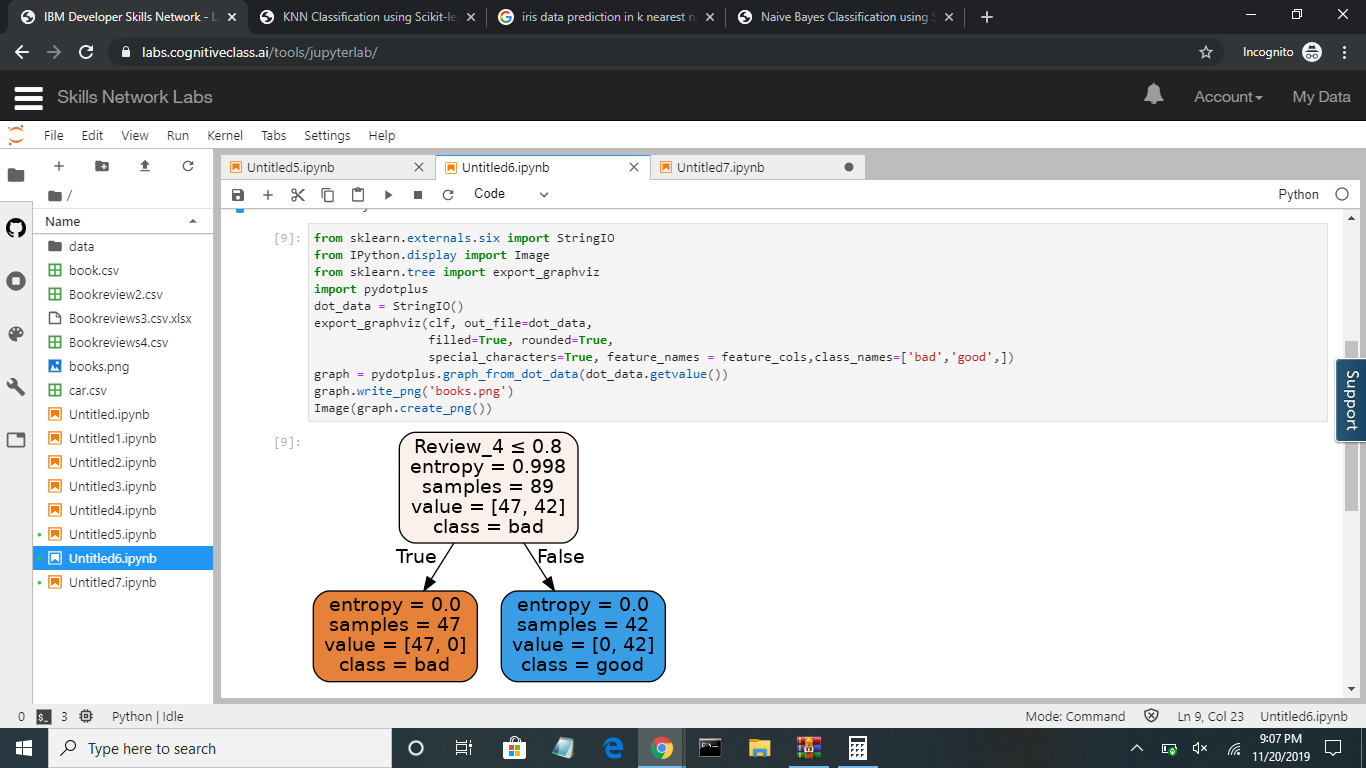
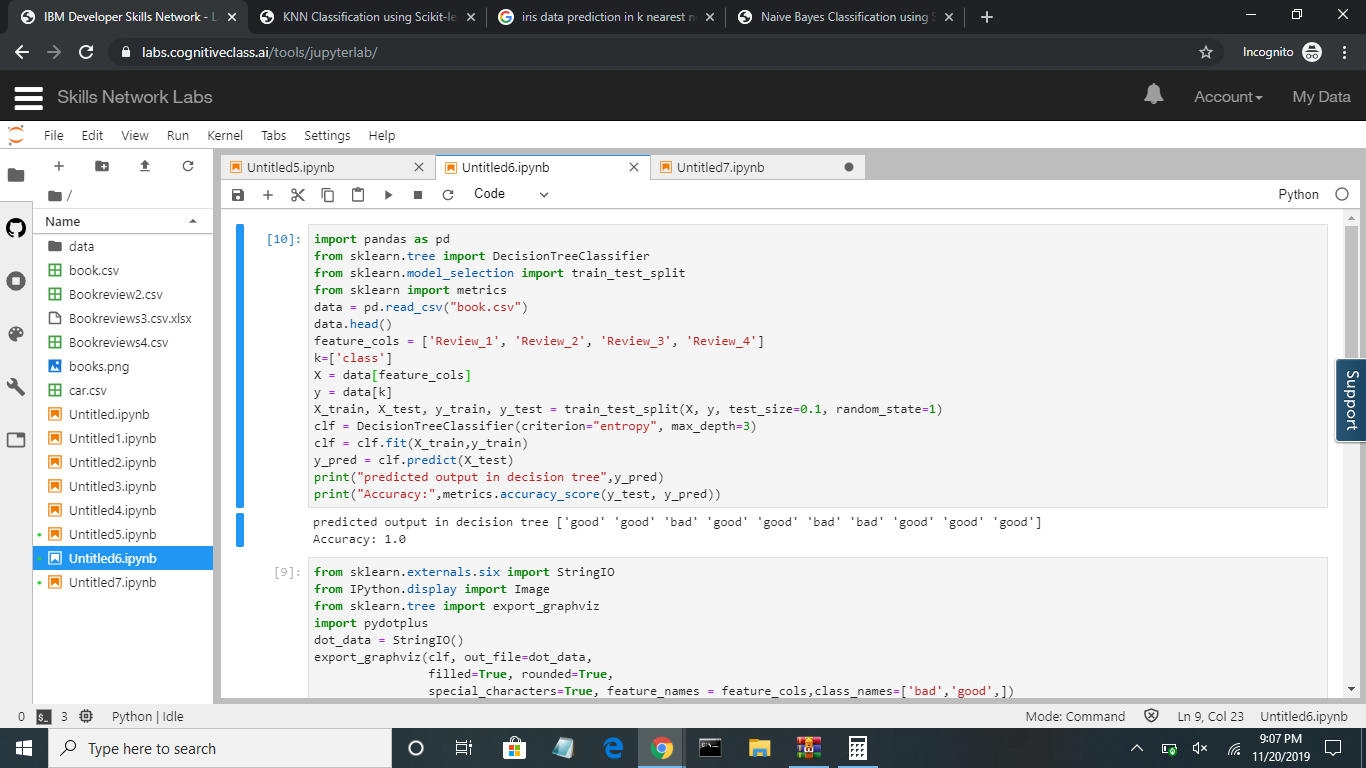
**DATASET : csv file**

**4****.SCREENSHOTS**

**MODULE-1:**

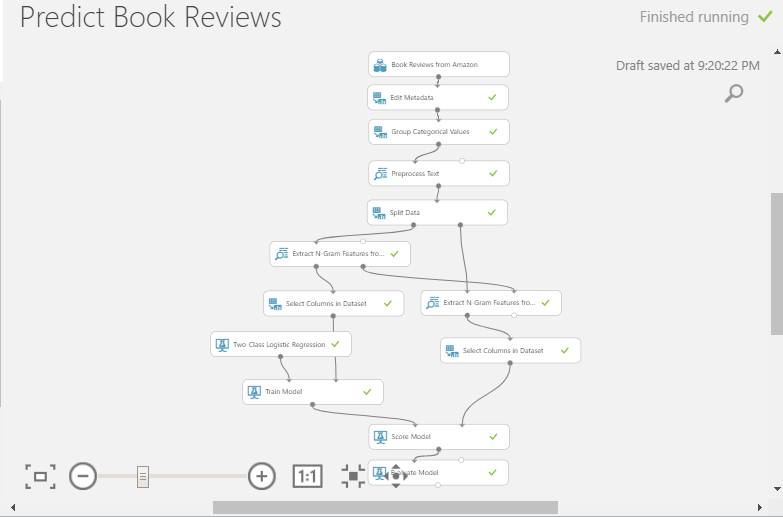
****

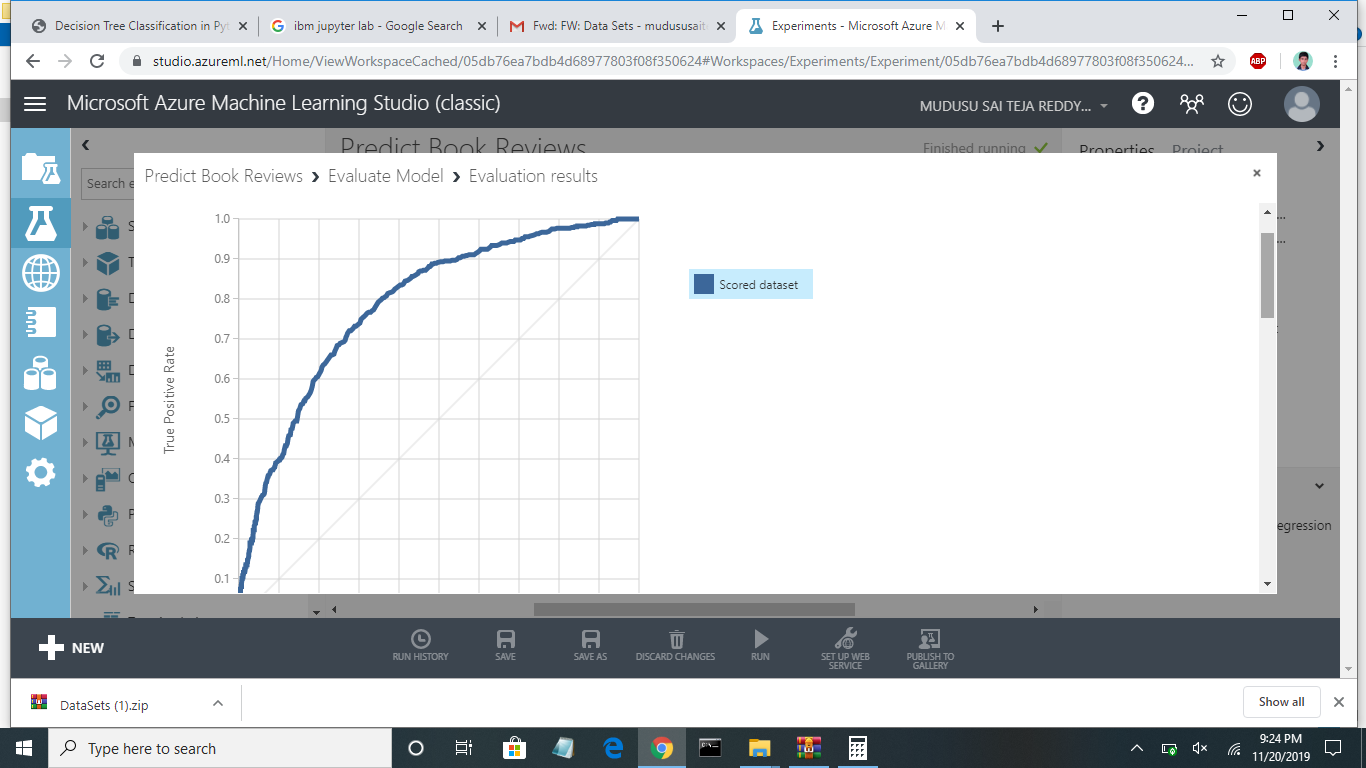
**MODULE-2:**

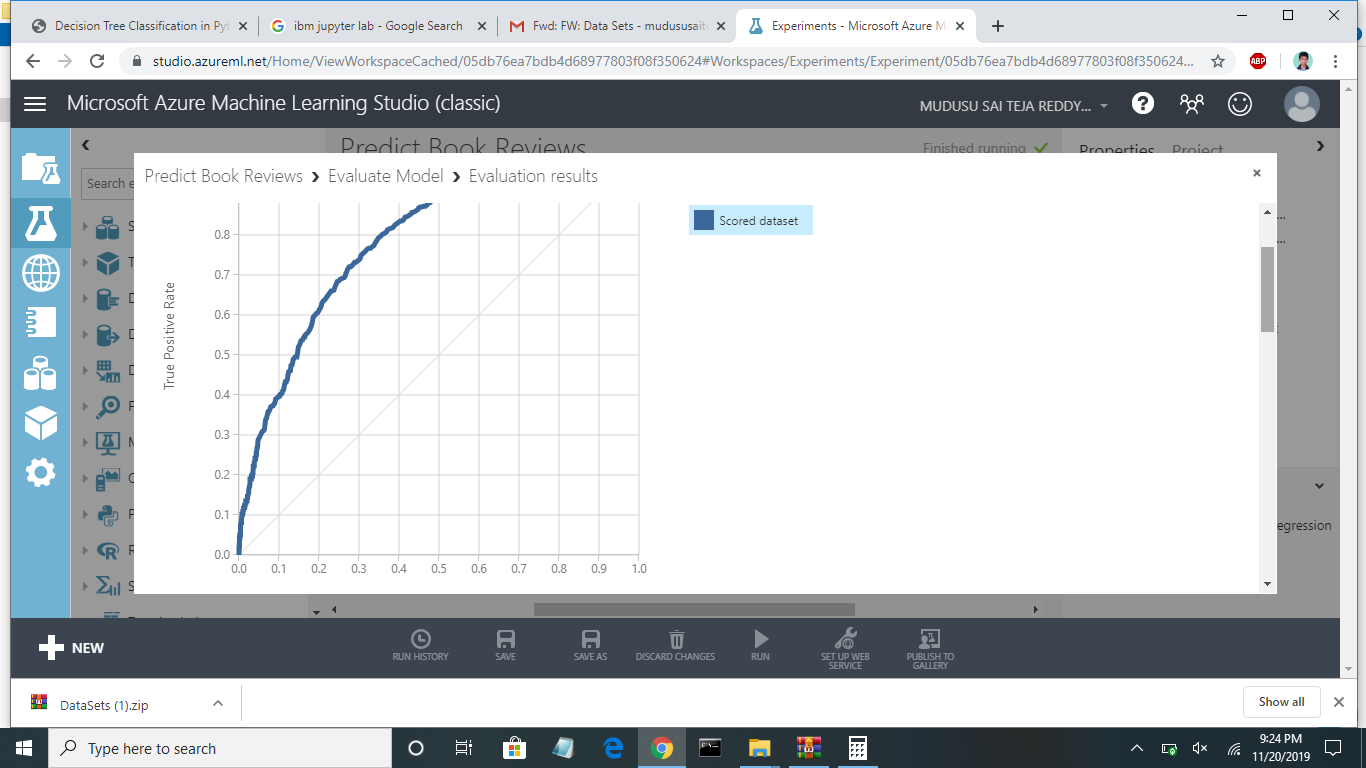
****

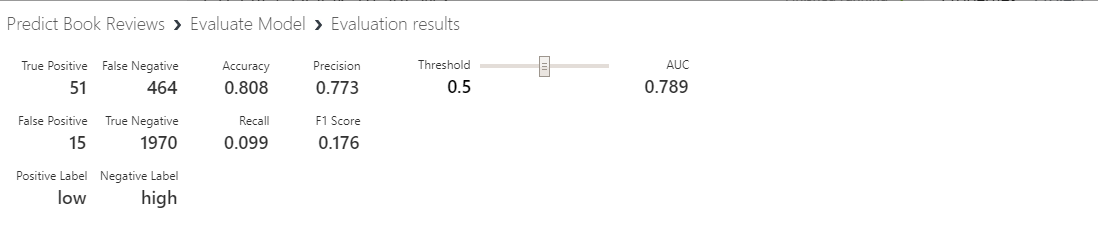
**MODULE-3:**

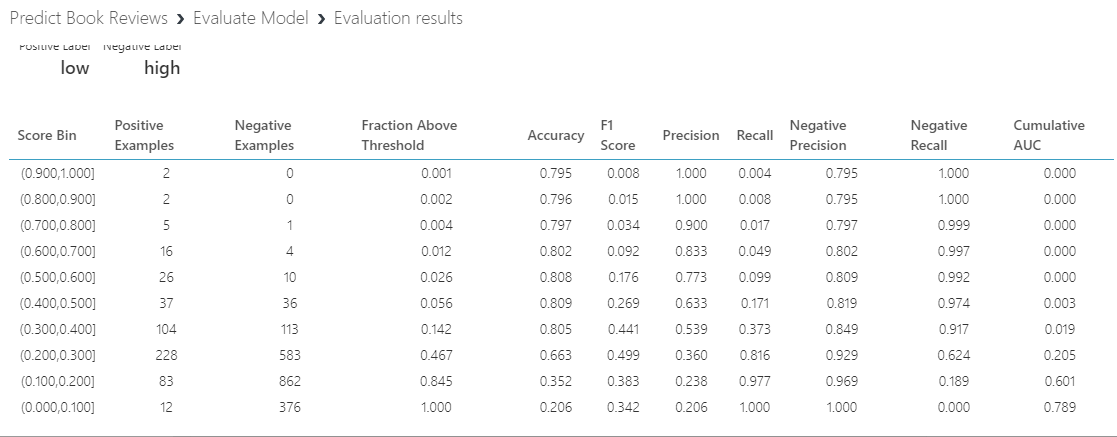
**BY APPLYING AZURE MODELING WE GET THE BELOW PREDICTIONS ON BOOKS**

****

****

****

****

****

**5.CONCLUSION**

After applying different classification techniques like k-Nearest Neighbors classification and decision tree classifier .

\*k-nearest neighbors gives 76% accuracy.

\*Decision tree classifier gives 72% accuracy.

After considering all the accuracies of the classifications random forest classifier gives the maximum accuracy that is 80%.

**6.REFERENCES**

1. <https://gallery.azure.ai/Experiment/Predict-Book-Reviews-1>
2. <https://bookjelly.com/book-review-prediction-machines/>
3. <https://www.guru99.com/machine-learning-tutorial.html>
4. <https://towardsdatascience.com/>