Pattern Recognition based on Hobby(Reading)

Md. Shakhawat Hossain Mridha[1*] **and Farzana Sharmin Mou**[1+]

[1]Ahsanullah University of Science and Technology, Computer Science and Engineering, Dhaka, 1208, Bangladesh
[*]shakhawathossainmridha@gmail.com(15-01-04-067)
[+]farzanamou773@gmail.com(15-01-04-076)
[1]these authors contributed equally to this work

ABSTRACT

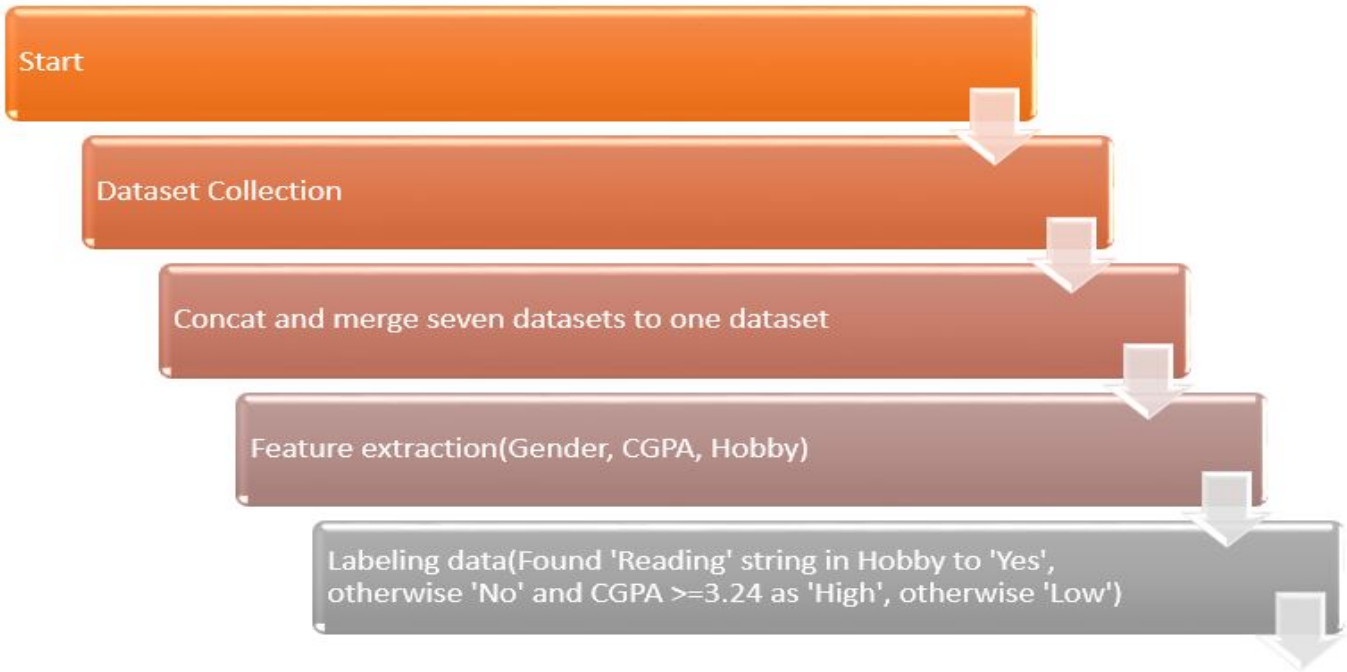
Hobbies are an important aspect of leisure time for many people, bringing both balance and joy to normally stressful and responsibility-driven lives. While certain hobbies offer relaxation, others provide outlets for creative expression or physical fitness. Personality type certainly influences what kinds of hobbies people engage in. Here we tried to find out a group of people who are interested in reading.

Introduction

Hobbies reflect the inner most desires of people, help them fulfill their unmet needs and make them feel special. But people can be doing the same exact hobby for completely different reasons. Because of that, it is very difficult to specify a person's hobby. So, we tried to find out a pattern of a group of people who likes to read.
We worked with two features(Gender, CGPA) and a category(Hobby) which was our ground truth. We worked with these features because those people are rich in knowledge who read. It is not obvious that he scores a good mark but there is a high probability. Then we pre-processed data and using KNN algorithm we predicted new people's hobby and found out a correlation.

Methods

Initially, we had 7 datasets, where 6 of them had same columns with different rows and other one had all rows with new coloumns. So, to solve this problem we had to follow some steps described below:



Step to solution:

1. Collected seven datasets
2. Concat and merge dataset to one
3. Feature extraction
4. Labeling data
5. KNN Algorithm
6. Correlation
7. Prediction

Background

Algorithm types
Pattern recognition systems are in many cases trained from labeled "training" data (supervised learning), but when no labeled data are available other algorithms can be used to discover previously unknown patterns (unsupervised learning). Machine learning is the common term for supervised learning methods and originates from artificial intelligence, whereas KDD and data mining have a larger focus on unsupervised methods and stronger connection to business use.

KNN algorithm
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 already in the beginning of 1970's as a non-parametric technique.

Scikit
Scikit-learn (formerly scikits.learn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Matplotlib
Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits.
Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc., with just a few lines of code.
For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.

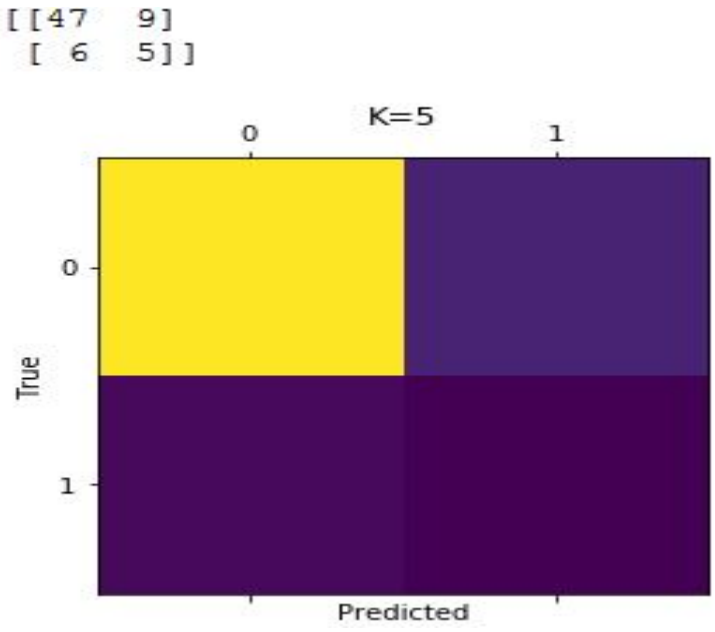
Results

Here we extracted two types of results:

1. Confusion matrix plot

2. Plotting correlation and predict

Confusion matrix plot: Here we worked with 97 samples, where 30 samples were for train set and 67 were for test set. When we ran KNN algorithm with K=5, then we got 77.6% accuracy. And got this confusion matrix with plotting like this below.

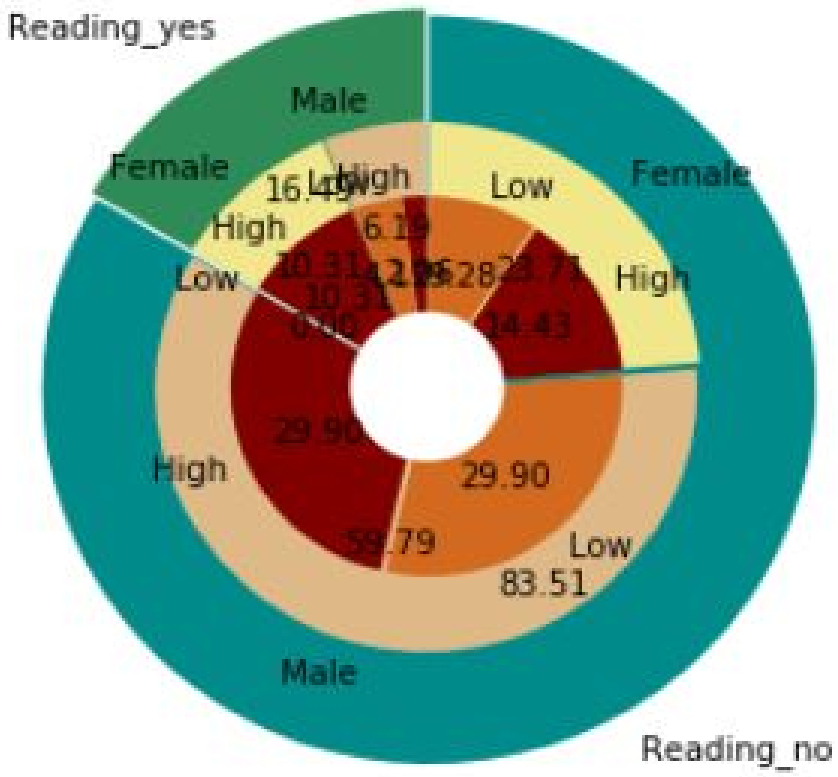


Here we can see that, from test set 56 samples were supposed to be predicted as 0, but our solution got 47 as 0, other as 9. On the other hand, 11 samples were supposed to be 1, where our solution predicted 6 as 1, others as 0. Then we plotted this conception in a graph.

Plotting correlation and predict:
Here we grouped by Hobby, Gender and CGPA, then counted sample per group. Here is the table of correlation:

-	-	-	Count
Hobby	Gender	CGPA high or low	-
no	Female	High	14
-	-	Low	9
-	Male	High	29
-	-	Low	29
yes	Female	High	10
-	Male	High	2
-	-	Low	4

Here is the plot of correlation:



It shows that, female with high CGPA are most likely to read books.

Discussion

Here, we tried to find out a pattern. In future, we will add more features to make it more justifiable.