Student Category Classification by Their Academics Performance

Using K Nearest Neighbours Classifier

Submitted To

Dr. Mohammad Shoyaib
Professor
IIT, University of Dhaka
&
Kishan Kumar Ganguly
Lecturer
IIT, University of Dhaka

Submitted By

Shezan Al-Mahmud
BSSE 1023

Submission Date : 5 September 2021

Table of Contents

1 : Introduction

Educational quality is compulsory in the development of each country. The data in the education domain is increasing day by day. The data collected from students are usually used for making simple queries for decision making. But most of the data remains unused due to complexity and large volume data sets. Therefore, analyzing this huge amount of educational data is a great interest to predict student performance.

One of the biggest challenges is to improve the quality of the educational processes so as to enhance student's performance. Instructors can update their teaching methodology to fulfill the requirement of poor performance students and can provide additional guidance to deserving students. The prediction results might help students develop a good understanding of how well or bad they would perform and then can take steps accordingly. Increasing student retention is a long-term target of any educational institutions around the globe. There are many positive impacts of increased retention such as increased college reputation, ranking and better job opportunities for alumni etc.

2 : Scope of the Project

The scope of this project is strictly limited to using K Nearest Neighbours (KNN) algorithm as the one and only classification technique to predict student academics performance. We aspired to build the best possible machine learning model using KNN, but did not consider any other classification algorithm.

The students are classified into three intervals based on their academic performance. (i) Low-Level, (ii) Middle-Level and (iii) High-Level.

3 : Objectives

Academic achievement is a big concern for academic institutions all over the world. The main objective of this project was to build a fully working tool for student category classification by their academics performance. Another objective was to improve the machine learning model used in this project to have a higher accuracy in the prediction. The prediction results might help students develop a good understanding of how well or bad they would perform and then can take steps accordingly. Instructors can update their teaching methodology to fulfill the requirement of poor performance students and can provide additional guidance to deserving students.

4 : Methodology

The procedure has mainly-

1. Read xAPI-Edu-Data.csv file and
2. Implement KNN algorithm to find the nearest ones.

The machine learning model used in this tool was carefully built using dataset preprocessing, machine learning classifier and evaluation process. The program uses the K nearest neighbor algorithm with the value of **K** to classify the data points. I have run the number of fold = 5 cross-validation.

4.1 : Description of the Dataset

The dataset consists of 305 males and 175 females. The students come from different origins such as 179 students are from Kuwait, 172 students are from Jordan, 28 students from Palestine, 22 students are from Iraq, 17 students from Lebanon, 12 students from Tunis, 11 students from Saudi Arabia, 9 students from Egypt, 7 students from Syria, 6 students from USA, Iran and Libya, 4 students from Morocco and one student from Venezuela. The dataset is collected through two educational semesters: 245 student records are collected during the first semester and 235 student records are collected during the second semester. The data set includes also the school attendance feature such as the students are classified into two categories based on their absence days: 191 students exceed 7 absence days and 289 students their absence days under 7. This dataset includes also a new category of features; this feature is parent parturition in the educational process. Parent participation features have two sub features: Parent Answering Survey and Parent School Satisfaction. There are 270 of the parents who answered the survey and 210 are not, 292 of the parents are satisfied with the school and 188 are not.

The dataset is properly explained below -

Number of Instances : 480

Number of Attributes : 16

Attribute Description :

1 : Gender - student's gender (nominal: 'Male' or 'Female’)

2 : Nationality- student's nationality (nominal:’ Kuwait’,’ Lebanon’,’ Egypt’,’ SaudiArabia’,’ USA’,’ Jordan’,’

Venezuela’,’ Iran’,’ Tunis’,’ Morocco’,’ Syria’,’ Palestine’,’ Iraq’,’ Libya’)

3 : Place of birth- student's Place of birth (nominal:’ Kuwait’,’ Lebanon’,’ Egypt’,’ SaudiArabia’,’ USA’,’ Jordan’,’

Venezuela’,’ Iran’,’ Tunis’,’ Morocco’,’ Syria’,’ Palestine’,’ Iraq’,’ Libya’)

4 : Educational Stages- educational level student belongs (nominal: ‘lower level’,’MiddleSchool’,’HighSchool’)

5 : Grade Levels- grade student belongs (nominal: ‘G-01’, ‘G-02’, ‘G-03’, ‘G-04’, ‘G-05’, ‘G-06’, ‘G-07’, ‘G-08’, ‘G-09’, ‘G-10’, ‘G-11’, ‘G-12 ‘)

6 : Section ID- classroom student belongs (nominal:’A’,’B’,’C’)

7 : Topic- course topic (nominal:’ English’,’ Spanish’, ‘French’,’ Arabic’,’ IT’,’ Math’,’ Chemistry’, ‘Biology’, ‘Science’,’ History’,’ Quran’,’ Geology’)

8 : Semester- school year semester (nominal:’ First’,’ Second’)

9 : Parent responsible for student (nominal:’mom’,’father’)

10 : Raised hand- how many times the student raises his/her hand on classroom (numeric:0-100)

11 : Visited resources- how many times the student visits a course content(numeric:0-100)

12 : Viewing announcements-how many times the student checks the new announcements(numeric:0-100)

13 : Discussion groups- how many times the student participate on discussion groups (numeric:0-100)

14 : Parent Answering Survey- parent answered the surveys which are provided from school or not

(nominal:’Yes’,’No’)

15 : Parent School Satisfaction- the Degree of parent satisfaction from school(nominal:’Yes’,’No’)

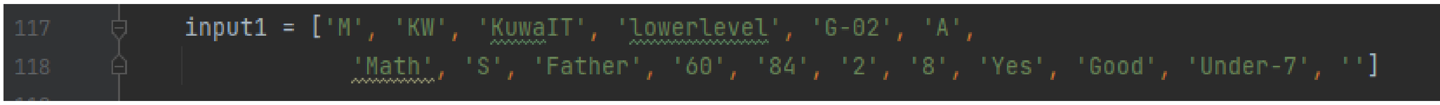
16 : Student Absence Days-the number of absence days for each student (nominal: above-7, under-7)

4.2 : Machine Learning Classifier Used

K Nearest Neighbours (KNN) classifier was used to build the machine learning model for this project. Different variations of KNN were experimented with to have better results. I tried weighted KNN for implementation. No machine learning library was used to implement the classifier used in this project.

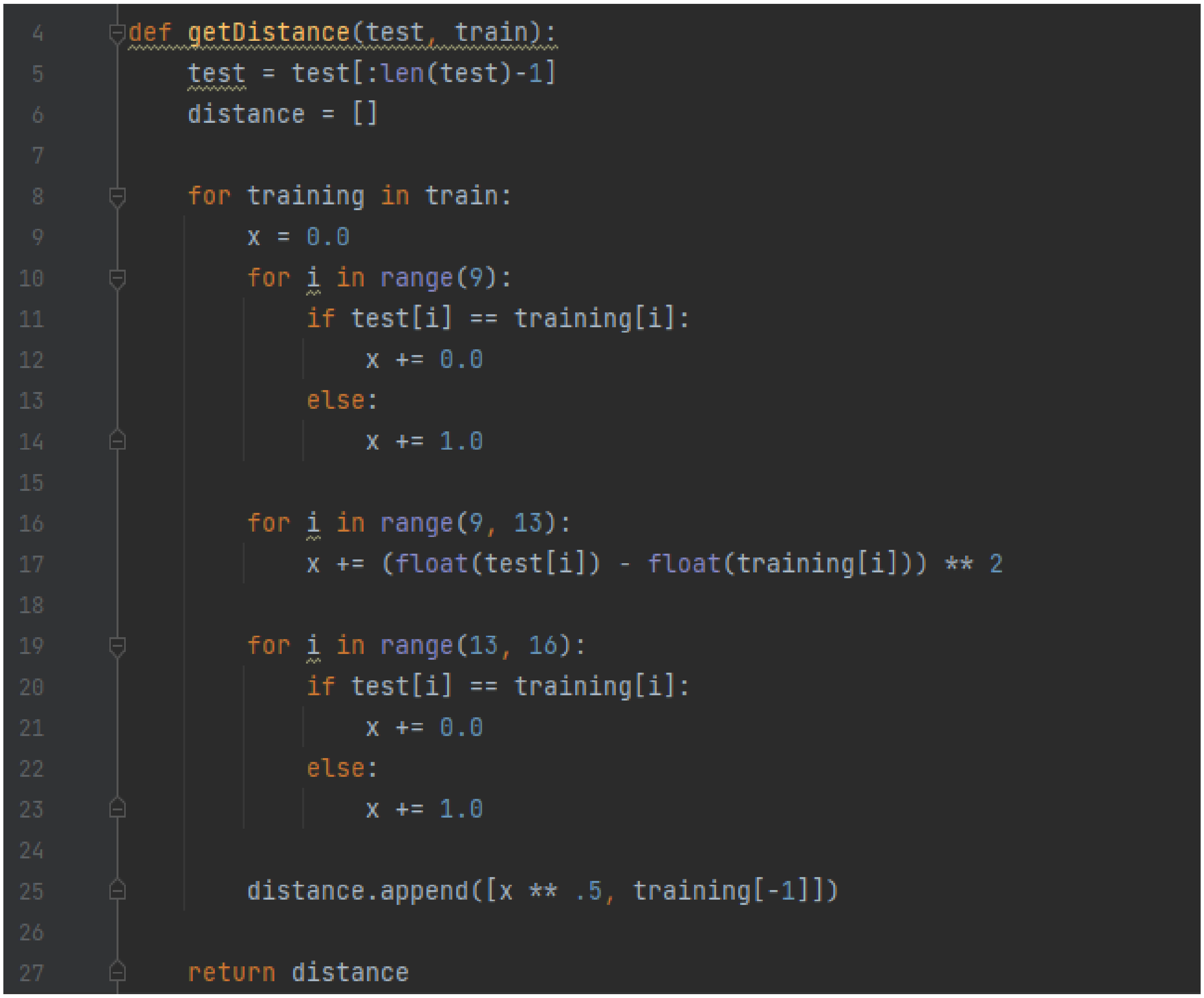
4.3 Euclidean Distance Based Weighted KNN

I calculated the euclidean distance between the input data and others data and sorted the distances. -



```
input1 = ['M', 'KW', 'KuwaIT', 'lowerlevel', 'G-02', 'A',
              'Math', 'S', 'Father', '60', '84', '2', '8', 'Yes', 'Good', 'Under-7', '']
```

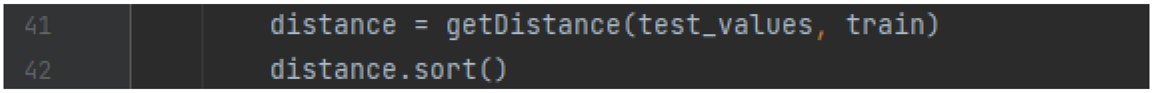
Figure-1 : Input data



```python
def getDistance(test, train):
    test = test[:len(test)-1]
    distance = []

    for training in train:
        x = 0.0
        for i in range(9):
            if test[i] == training[i]:
                x += 0.0
            else:
                x += 1.0

        for i in range(9, 13):
            x += (float(test[i]) - float(training[i])) ** 2

        for i in range(13, 16):
            if test[i] == training[i]:
                x += 0.0
            else:
                x += 1.0

        distance.append([x ** .5, training[-1]])

    return distance
```

Figure-2 : Calculated the distance

Note : Here, 1st 9 attributes and 14th-16th attributes are string type. That's why, if test data and training data are the same then it's value will be **0.0** for distance otherwise it's value will be **1.0**



```python
        distance = getDistance(test_values, train)
        distance.sort()
```

Figure-3 : After calculating distances, sorted the distances

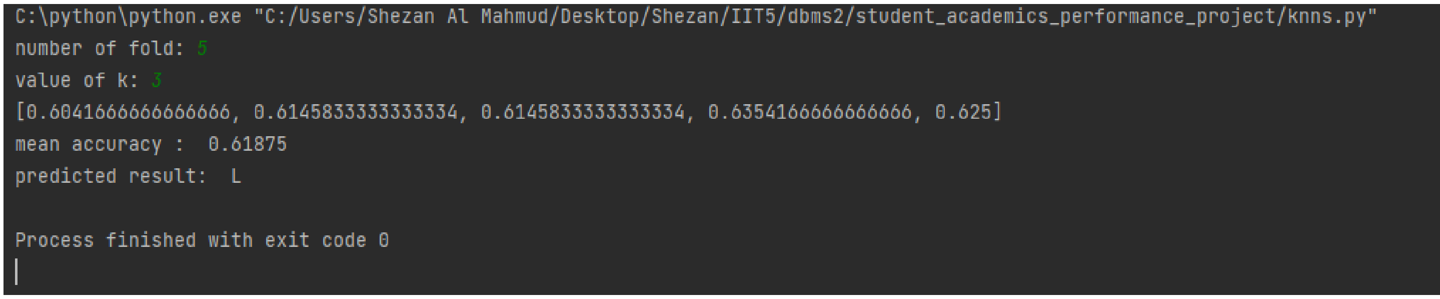
5 : Results

The program uses the K nearest neighbor algorithm with the value of **K** to classify the data points. I have run the number of fold = 5 cross-validation.

The students are classified into three intervals based on their academic performance.

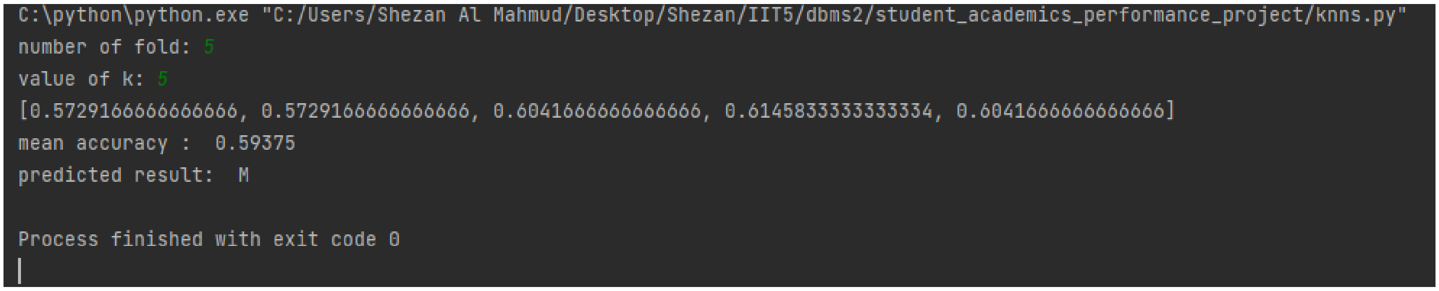
i. Low-Level (L)
ii. Middle-Level (M)
iii. High-Level (H)

Sample results are given below for the various value of **K** :



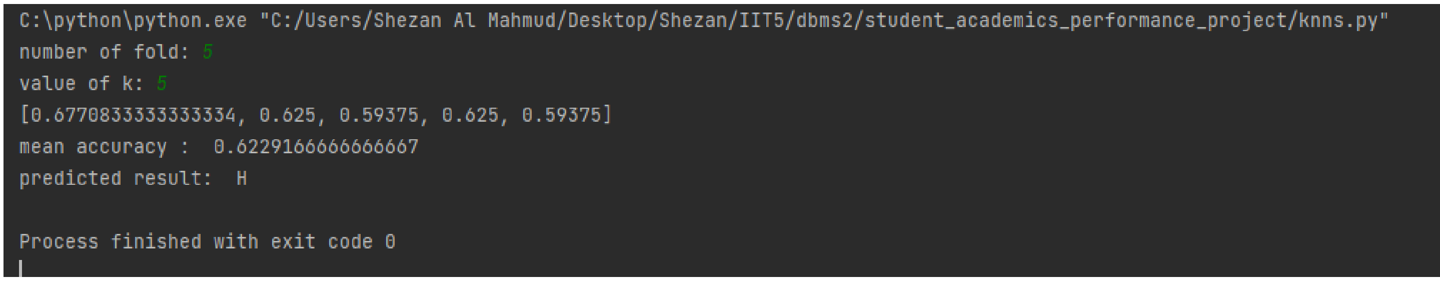
```
C:\python\python.exe "C:/Users/Shezan Al Mahmud/Desktop/Shezan/IIT5/dbms2/student_academics_performance_project/knns.py"
number of fold: 5
value of k: 3
[0.6041666666666666, 0.6145833333333334, 0.6145833333333334, 0.6354166666666666, 0.625]
mean accuracy :  0.61875
predicted result:  L

Process finished with exit code 0
```

Figure-4 : Output-1



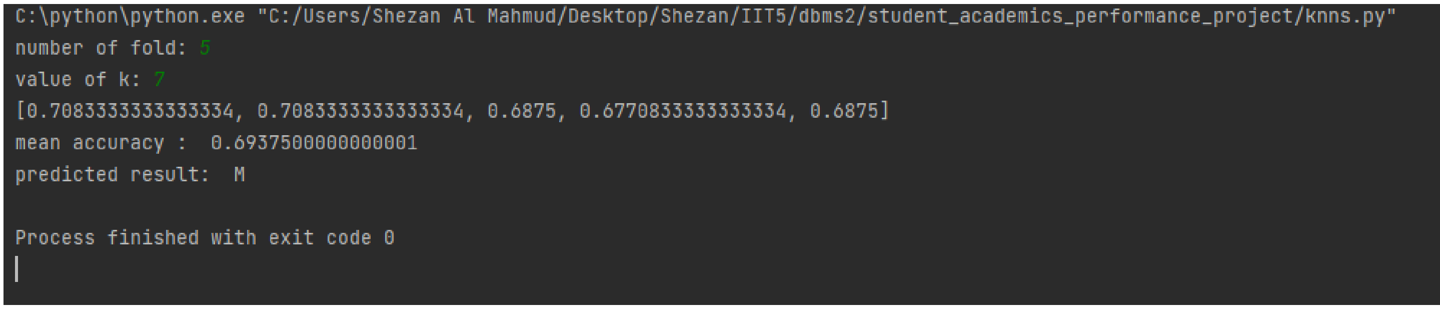
```
C:\python\python.exe "C:/Users/Shezan Al Mahmud/Desktop/Shezan/IIT5/dbms2/student_academics_performance_project/knns.py"
number of fold: 5
value of k: 5
[0.5729166666666666, 0.5729166666666666, 0.6041666666666666, 0.6145833333333334, 0.6041666666666666]
mean accuracy :  0.59375
predicted result:  M

Process finished with exit code 0
```

Figure-5 : Output-2



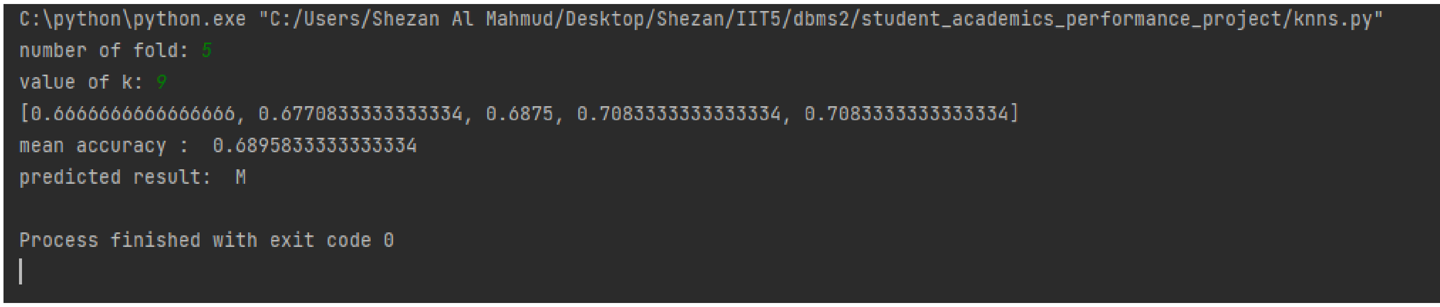
```
C:\python\python.exe "C:/Users/Shezan Al Mahmud/Desktop/Shezan/IIT5/dbms2/student_academics_performance_project/knns.py"
number of fold: 5
value of k: 5
[0.6770833333333334, 0.625, 0.59375, 0.625, 0.59375]
mean accuracy :  0.6229166666666667
predicted result:  H

Process finished with exit code 0
```

Figure-6 : Output-3



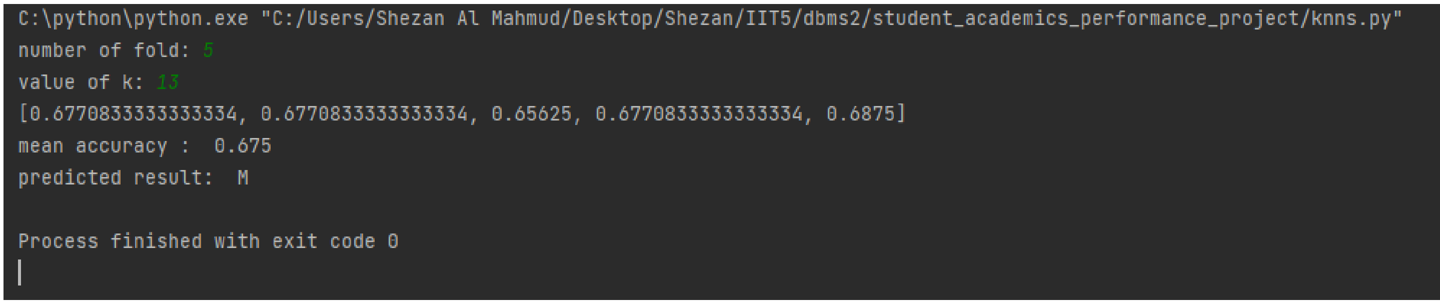
```
C:\python\python.exe "C:/Users/Shezan Al Mahmud/Desktop/Shezan/IIT5/dbms2/student_academics_performance_project/knns.py"
number of fold: 5
value of k: 7
[0.7083333333333334, 0.7083333333333334, 0.6875, 0.6770833333333334, 0.6875]
mean accuracy :  0.6937500000000001
predicted result:  M

Process finished with exit code 0
```

Figure-7 : Output-4



```
C:\python\python.exe "C:/Users/Shezan Al Mahmud/Desktop/Shezan/IIT5/dbms2/student_academics_performance_project/knns.py"
number of fold: 5
value of k: 9
[0.6666666666666666, 0.6770833333333334, 0.6875, 0.7083333333333334, 0.7083333333333334]
mean accuracy :  0.6895833333333334
predicted result:  M

Process finished with exit code 0
```

Figure-8 : Output-5



```
C:\python\python.exe "C:/Users/Shezan Al Mahmud/Desktop/Shezan/IIT5/dbms2/student_academics_performance_project/knns.py"
number of fold: 5
value of k: 13
[0.6770833333333334, 0.6770833333333334, 0.65625, 0.6770833333333334, 0.6875]
mean accuracy :  0.675
predicted result:  M

Process finished with exit code 0
```

Figure-9 : Output-6

6 : Tools and Technology

- Python (Language)
- PyCharm (Python IDE)

7 : Conclusion

The accurate student academic performance prediction model is demanded of every educational institute nowadays. But to resolve the data quality issues in the student perfor-mance prediction model is often the biggest challenge. This project presented a student performance prediction model based on the K Nearest Neighbours Classifier. In this project, I intended to build a useful tool for Student Category Classification by student's academics performance. It finished obtaining an average accuracy of 63%.

8 : Github Link

https://github.com/shezan7/Student-Classification-by-KNN

9. References

Dataset :
https://archive.ics.uci.edu/ml/datasets/Student+Academics+Performance?fbclid=IwAR29tiuhNhgK0RHBOfIcP0SFvks-eMl0Zs1cWM0Nv-SYXyVVuLFmX7FKQRI#