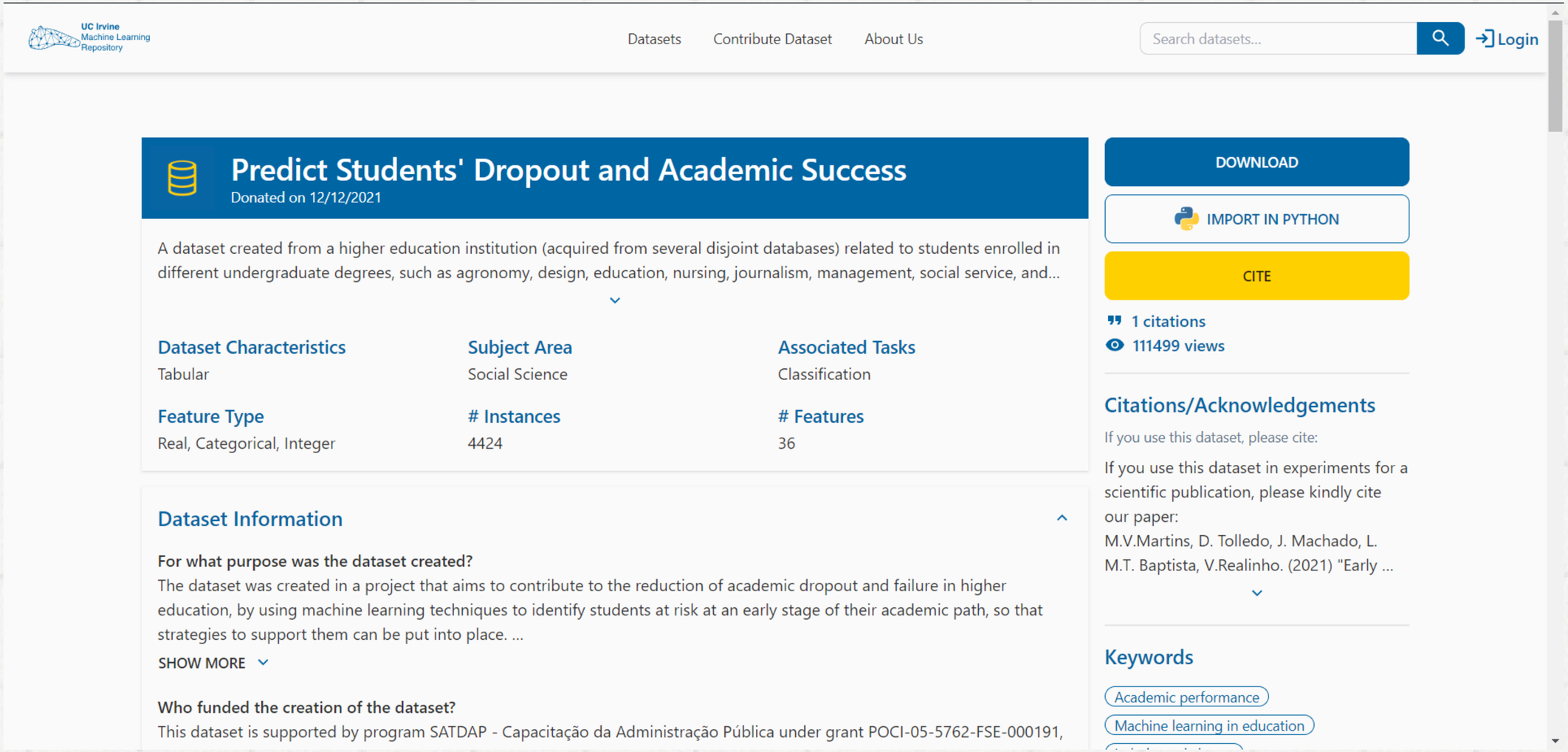
Data Science Project

Predict Students Dropout and Academic Success



Predict Students Dropout and Academic Success

Creators

Valentim Realinho
vrealinho@
Instituto Politécnico de Portalegre

Mónica Vieira Martins
mvmartins@ipportalegre.pt
Instituto Politécnico de Portalegre

Jorge Machado
jmachado@ipportalegre.pt
Instituto Politécnico de Portalegre

Luís Baptista
lmtb@ipportalegre.pt
Instituto Politécnico de Portalegre

Infomation about dataset

1) Marital status :

- 1 – single
- 2 – married
- 3 – widower
- 4 – divorced
- 5 – facto union
- 6 – legally separated

2) Application mode:

- 1 - 1st phase - general contingent
- 2 - Ordinance No. 612/93
- 5 - 1st phase - special contingent (Azores Island)
- 7 - Holders of other higher courses
- 10 - Ordinance No. 854-B/99
- 15 - International student (bachelor)
- 16 - 1st phase - special contingent (Madeira Island)
- 17 - 2nd phase - general contingent
- 18 - 3rd phase - general contingent
- 26 - Ordinance No. 533-A/99, item b2) (Different Plan)
- 27 - Ordinance No. 533-A/99, item b3 (Other Institution)
- 39 - Over 23 years old
- 42 - Transfer
- 43 - Change of course
- 44 - Technological specialization diploma holders
- 51 - Change of institution/course
- 53 - Short cycle diploma holders
- 57 - Change of institution/course (International)

3) Application order:

- 0 - first choice
- 9 - last choice

4) Course:

- 33 - Biofuel Production Technologies
- 171 - Animation and Multimedia Design
- 8014 - Social Service (evening attendance)
- 9003 - Agronomy
- 9070 - Communication Design
- 9085 - Veterinary Nursing
- 9119 - Informatics Engineering
- 9130 - Equinculture
- 9147 - Management
- 9238 - Social Service
- 9254 - Tourism
- 9500 - Nursing
- 9556 - Oral Hygiene
- 9670 - Advertising and Marketing Management
- 9773 - Journalism and Communication
- 9853 - Basic Education
- 9991 - Management (evening attendance)

5) Daytime/evening attendance:

- 1 – daytime
- 0 - evening

6) Previous qualification:

- 1 - Secondary education
- 2 - Higher education - bachelor's degree
- 3 - Higher education - degree
- 4 - Higher education - master's
- 5 - Higher education - doctorate
- 6 - Frequency of higher education
- 9 - 12th year of schooling - not completed
- 10 - 11th year of schooling - not completed
- 12 - Other - 11th year of schooling
- 14 - 10th year of schooling
- 15 - 10th year of schooling - not completed
- 19 - Basic education 3rd cycle (9th/10th/11th year) or equiv.
- 38 - Basic education 2nd cycle (6th/7th/8th year) or equiv.
- 39 - Technological specialization course
- 40 - Higher education - degree (1st cycle)
- 42 - Professional higher technical course
- 43 - Higher education - master (2nd cycle)

7) Previous qualification (grade):

- Grade of previous qualification (between 0 and 200)

8) Nacionality:

- 1 - Portuguese
- 2 - German
- 6 - Spanish;
- 11 - Italian;
- 13 - Dutch
- 14 - English
- 17 - Lithuanian
- 21 - Angolan
- 22 - Cape Verdean
- 24 - Guinean
- 25 - Mozambican
- 26 - Santomean
- 32 - Turkish
- 41 - Brazilian
- 62 - Romanian
- 100 - Moldova (Republic of)
- 101 - Mexican
- 103 - Ukrainian
- 105 - Russian
- 108 - Cuban
- 109 - Colombian

9) Mother's and 10) Father's qualification:

- 1 - Secondary Education - 12th Year of Schooling or Eq.
- 2 - Higher Education - Bachelor's Degree
- 3 - Higher Education - Degree
- 4 - Higher Education - Master's
- 5 - Higher Education - Doctorate
- 6 - Frequency of Higher Education
- 9 - 12th Year of Schooling - Not Completed
- 10 - 11th Year of Schooling - Not Completed
- 11 - 7th Year (Old)
- 12 - Other - 11th Year of Schooling
- 13 - 2nd year complementary high school course
- 14 - 10th Year of Schooling
- 18 - General commerce course
- 19 - Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv.
- 20 - Complementary High School Course
- 22 - Technical-professional course
- 25 - Complementary High School Course - not concluded
- 26 - 7th year of schooling
- 27 - 2nd cycle of the general high school course
- 29 - 9th Year of Schooling - Not Completed
- 30 - 8th year of schooling
- 34 - Unknown
- 35 - Can't read or write
- 36 - Can read without having a 4th year of schooling
- 37 - Basic education 1st cycle (4th/5th year) or equiv.
- 38 - Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv.
- 39 - Technological specialization course
- 40 - Higher education - degree (1st cycle)
- 41 - Specialized higher studies course
- 42 - Professional higher technical course
- 43 - Higher Education - Master (2nd cycle)
- 44 - Higher Education - Doctorate (3rd cycle)

11) Mother's and 12) Father's occupation:

- 0 - Student
- 1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers
- 2 - Specialists in Intellectual and Scientific Activities
- 3 - Intermediate Level Technicians and Professions
- 4 - Administrative staff
- 5 - Personal Services, Security and Safety Workers and Sellers
- 6 - Farmers and Skilled Workers in Agriculture, Fisheries and Forestry
- 7 - Skilled Workers in Industry, Construction and Craftsmen
- 8 - Installation and Machine Operators and Assembly Workers
- 9 - Unskilled Workers
- 10 - Armed Forces Professions
- 90 - Other Situation
- 99 - (blank)
- 101 - Armed Forces Officers
- 102 - Armed Forces Sergeants
- 103 - Other Armed Forces personnel
- 112 - Directors of administrative and commercial services
- 114 - Hotel, catering, trade and other services directors
- 121 - Specialists in the physical sciences, mathematics, engineering and related techniques
- 122 - Health professionals
- 123 - teachers
- 124 - Specialists in finance, accounting, administrative organization, public and commercial relations
- 131 - Intermediate level science and engineering technicians and professions
- 132 - Technicians and professionals, of intermediate level of health
- 134 - Intermediate level technicians from legal, social, sports, cultural and similar services
- 135 - Information and communication technology technicians
- 141 - Office workers, secretaries in general and data processing operators
- 143 - Data, accounting, statistical, financial services and registry-related operators
- 144 - Other administrative support staff
- 151 - personal service workers
- 152 - sellers
- 153 - Personal care workers and the like
- 154 - Protection and security services personnel
- 161 - Market-oriented farmers and skilled agricultural and animal production workers
- 163 - Farmers, livestock keepers, fishermen, hunters and gatherers, subsistence
- 171 - Skilled construction workers and the like, except electricians
- 172 - Skilled workers in metallurgy, metalworking and similar
- 174 - Skilled workers in electricity and electronics
- 175 - Workers in food processing, woodworking, clothing and other industries and crafts
- 181 - Fixed plant and machine operators
- 182 - assembly workers
- 183 - Vehicle drivers and mobile equipment operators
- 192 - Unskilled workers in agriculture, animal production, fisheries and forestry
- 193 - Unskilled workers in extractive industry, construction, manufacturing and transport
- 194 - Meal preparation assistants
- 195 - Street vendors (except food) and street service providers

Infomation about dataset

13) Admission grade:

- Admission grade (between 0 and 200)

14) Displaced, 15) Educational special needs, 16) Debtor, 17) Tuition fees up to date, 19) Scholarship holder, 21) International

- 1 – yes
- 0 – no

18) Gender

- 1 – male
- 0 – female

20) Age at enrollment

- Age of studend at enrollment

22) Curricular units 1st sem (credited)

- Number of curricular units credited in the 1st semester

23) Curricular units 1st sem (enrolled)

- Number of curricular units enrolled in the 1st semester

24) Curricular units 1st sem (evaluations)

- Number of evaluations to curricular units in the 1st semester

25) Curricular units 1st sem (approved)

- Number of evaluations to units approved in the 1st semester

26) Curricular units 1st sem (grade)

- Grade average in the 1st semester (between 0 and 20)

27) Curricular units 1st sem (without evaluations)

- Number of curricular units without evalutions in the 1st semester

28) Curricular units 2nd sem (credited)

- Number of curricular units credited in the 2nd semester

29) Curricular units 2nd sem (enrolled)

- Number of curricular units enrolled in the 2nd semester

30) Curricular units 2nd sem (evaluations)

- Number of evaluations to curricular units in the 2nd semester

31) Curricular units 2nd sem (approved)

- Number of curricular units approved in the 2nd semester

32) Curricular units 2nd sem (grade)

- Grade average in the 2nd semester (between 0 and 20)

33) Curricular units 2nd sem (without evaluations)

- Number of curricular units without evalutions in the 1st semester

34) Unemployment rate

- Unemployment rate (%)

35) Inflation rate

- Inflation rate (%)

36) GDP

- GDP

37) Target

- Target. The problem is formulated as a three category classification task (dropout, enrolled, and graduate) at the end of the normal duration of the course

หัวข้อที่ 1

แนวโน้ม คะแนนสอบเข้า และ ผลการเรียนเฉลี่ย ของนักเรียนที่ลาออกเทียบกับนักเรียนที่สำเร็จการศึกษาในแต่ละสาขาวิชา

คะแนนสอบเข้าเฉลี่ย

นักเรียนที่ลาออก

Average admission grade for dropout students: Course	
Advertising and Marketing Management	119.994737
Agronomy	127.439535
Animation and Multimedia Design	131.412195
Basic Education	122.414118
Biofuel Production Technologies	116.262500
Communication Design	125.176471
Equinculture	130.600000
Informatics Engineering	124.690217
Journalism and Communication	126.460396
Management	120.524627
Management (evening attendance)	128.525000
Nursing	126.516949
Oral Hygiene	124.806061
Social Service	123.187692
Social Service (evening attendance)	119.826761
Tourism	119.627083
Veterinary Nursing	128.983333

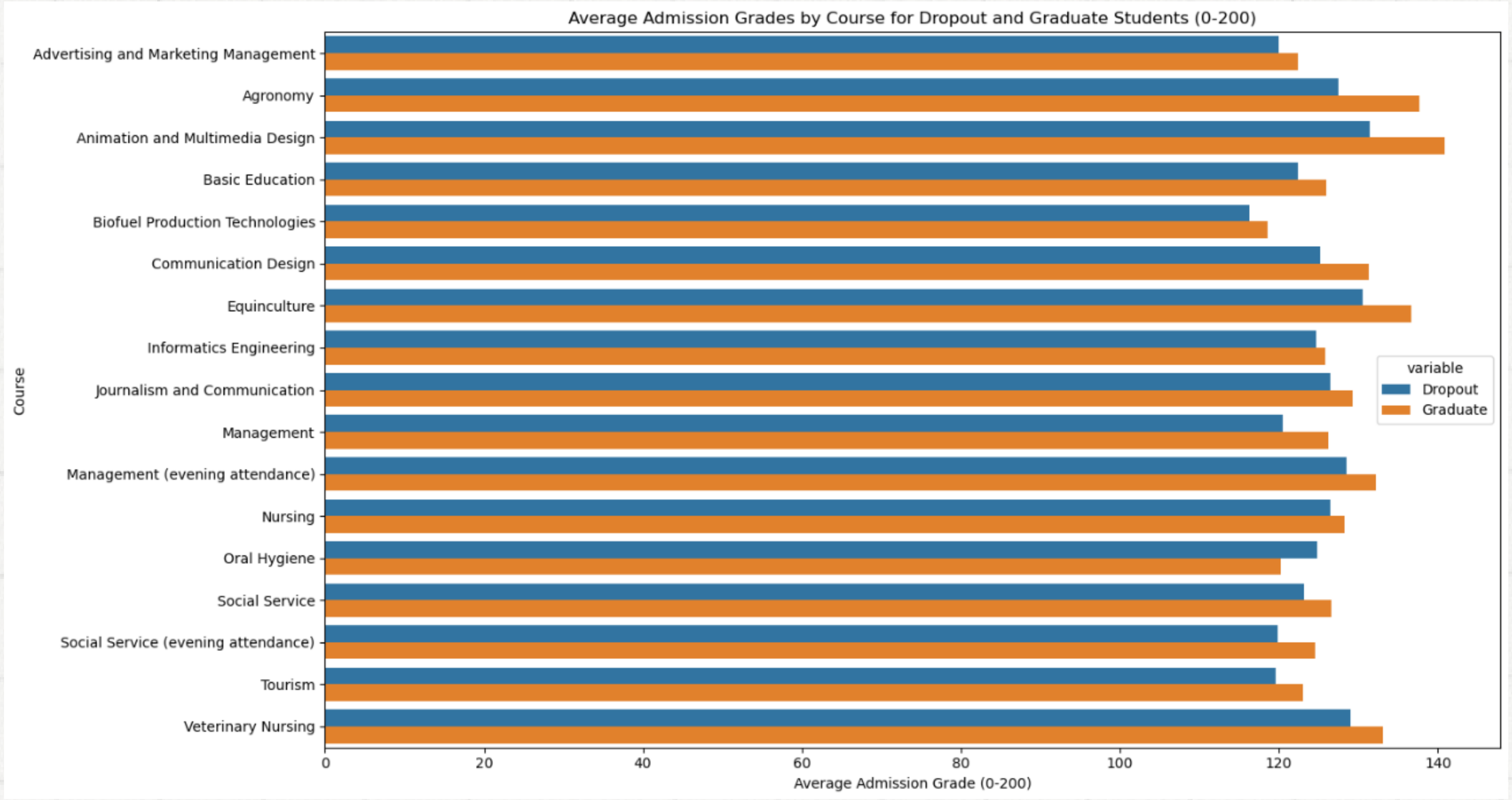
Average admission grade dropout: 124.96136523574945

นักเรียนที่จบการศึกษา

Average admission grade for graduate students: Course	
Advertising and Marketing Management	122.415200
Agronomy	137.687356
Animation and Multimedia Design	140.815625
Basic Education	125.947368
Biofuel Production Technologies	118.600000
Communication Design	131.251880
Equinculture	136.695238
Informatics Engineering	125.885714
Journalism and Communication	129.245408
Management	126.233333
Management (evening attendance)	132.238462
Nursing	128.254562
Oral Hygiene	120.291667
Social Service	126.620968
Social Service (evening attendance)	124.573171
Tourism	123.059130
Veterinary Nursing	133.098837

Average admission grade graduate: 128.79443186962428

คะแนนสอบเข้าเฉลี่ย



ผลการเรียนเฉลี่ย

ผลการเรียนเฉลี่ยในแต่ละสาขาวิชาของ นักเรียนที่ลาออก

เทอม1

Course	Curricular units 1st sem (grade)
Advertising and Marketing Management	8.559737
Agronomy	7.182419
Animation and Multimedia Design	0.738485
Basic Education	9.639806
Biofuel Production Technologies	7.910038
Communication Design	9.192994
Equinculture	7.814403
Informatics Engineering	6.466949
Journalism and Communication	8.764875
Management	7.203374
Management (evening attendance)	6.354282
Nursing	7.745677
Oral Hygiene	8.548247
Social Service	7.033700
Social Service (evening attendance)	6.553199
Tourism	7.359995
Veterinary Nursing	8.050040

Curricular units 1st sem (grade)	7.256656

เทอม2

Course	Curricular units 2nd sem (grade)
Advertising and Marketing Management	6.417845
Agronomy	5.457184
Animation and Multimedia Design	0.730820
Basic Education	7.982395
Biofuel Production Technologies	7.093750
Communication Design	6.696125
Equinculture	6.819017
Informatics Engineering	6.313859
Journalism and Communication	6.583616
Management	5.867219
Management (evening attendance)	6.221624
Nursing	6.534007
Oral Hygiene	4.711616
Social Service	5.698864
Social Service (evening attendance)	4.481433
Tourism	5.835938
Veterinary Nursing	6.465093

Curricular units 2nd sem (grade)	5.899339

ผลการเรียนเฉลี่ย

ผลการเรียนเฉลี่ยในแต่ละสาขาวิชาของ นักเรียนที่จบการศึกษา

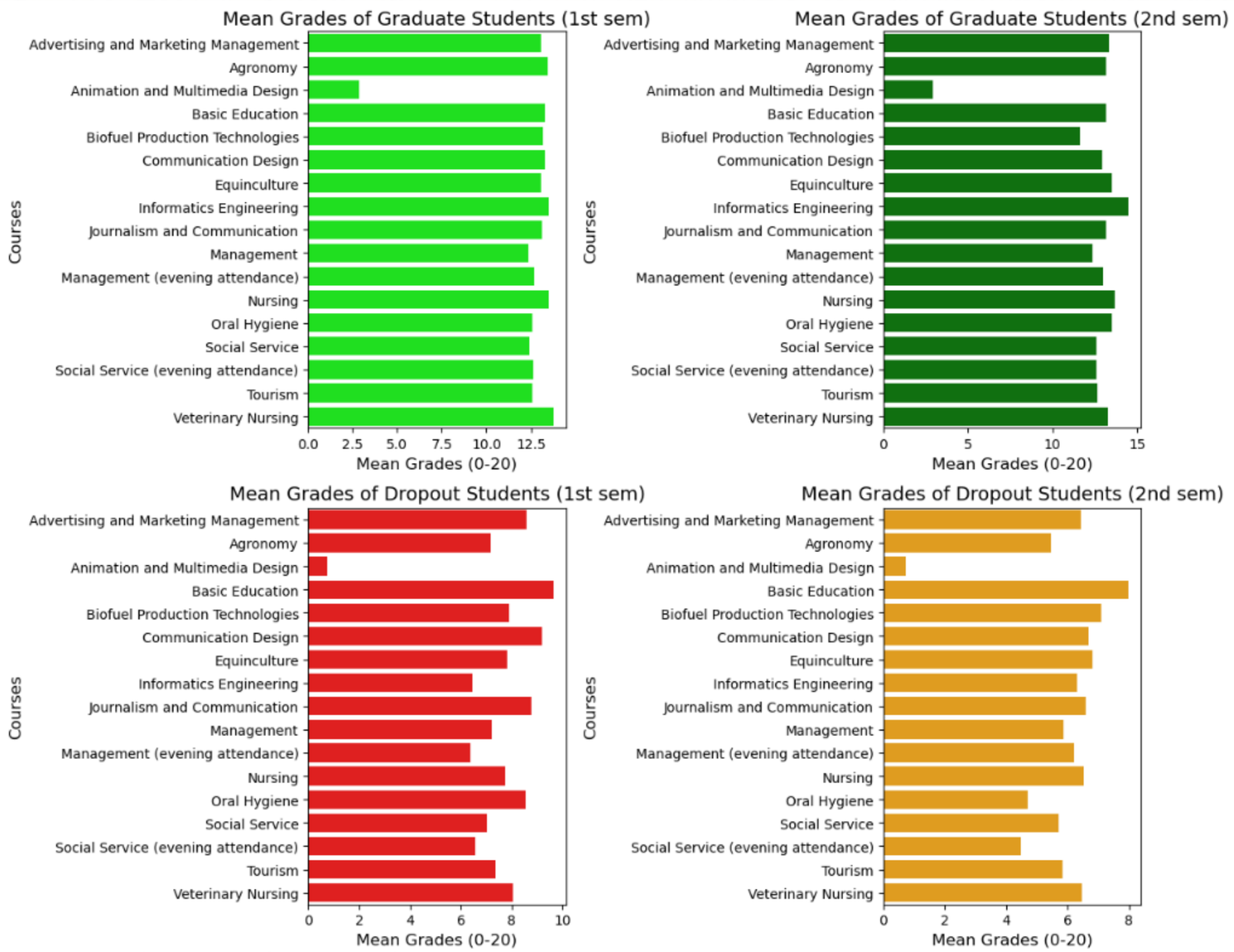
เทอม1

Course	Curricular units 1st sem (grade)
Advertising and Marketing Management	13.073914
Agronomy	13.427467
Animation and Multimedia Design	2.832257
Basic Education	13.305618
Biofuel Production Technologies	13.181818
Communication Design	13.264593
Equinculture	13.062354
Informatics Engineering	13.492857
Journalism and Communication	13.151397
Management	12.358313
Management (evening attendance)	12.672532
Nursing	13.504966
Oral Hygiene	12.576842
Social Service	12.445086
Social Service (evening attendance)	12.646194
Tourism	12.577087
Veterinary Nursing	13.772682
Curricular units 1st sem (grade)	12.643655

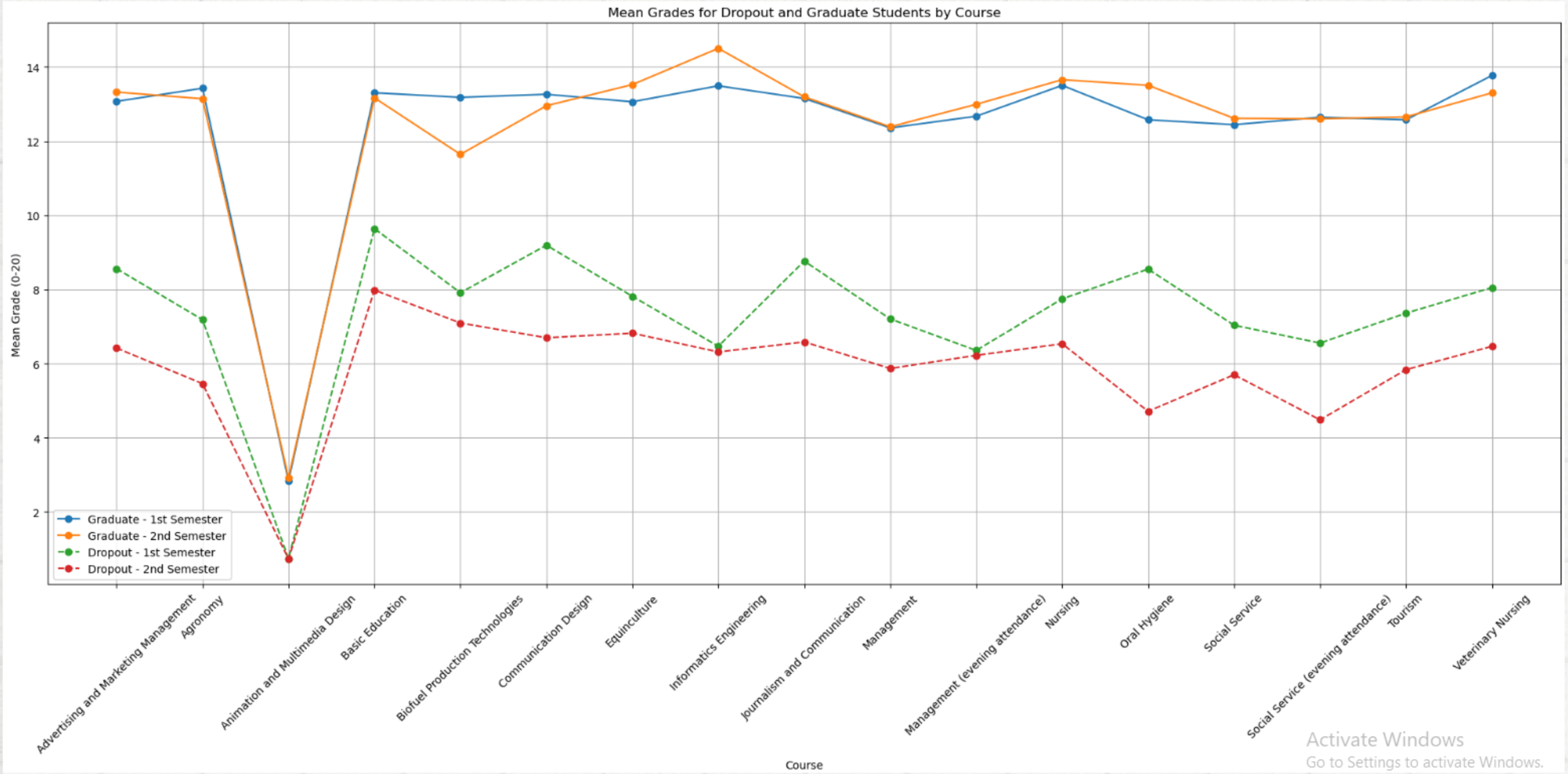
เทอม2

Course	Curricular units 2nd sem (grade)
Advertising and Marketing Management	13.323565
Agronomy	13.144639
Animation and Multimedia Design	2.917942
Basic Education	13.168130
Biofuel Production Technologies	11.642857
Communication Design	12.955402
Equinculture	13.525283
Informatics Engineering	14.503912
Journalism and Communication	13.195196
Management	12.388127
Management (evening attendance)	12.993746
Nursing	13.658520
Oral Hygiene	13.506804
Social Service	12.615322
Social Service (evening attendance)	12.602444
Tourism	12.655231
Veterinary Nursing	13.303611
Curricular units 2nd sem (grade)	12.697276

ผลการเรียนเฉลี่ย



ผลการเรียนเฉลี่ย



หัวข้อที่ 2
การทำโมเดล Classification

เทรนโมเดล KNN

```
knn = KNeighborsClassifier( n_neighbors= clf.best_params_['n_neighbors'], weights=clf.best_params_['weights'])
knn.fit(X_train, y_train)
y_predict = knn.predict(X_test)
```

Metric

```
print('accuracy: ', accuracy_score(y_test, y_predict)) # ความแม่นย่าของโมเดลเป็นเท่าไหร่
print('recall:   ', recall_score(y_test, y_predict)) # จากค่า graduate ทั้งหมดโมเดลทำนายทำนายได้อย่างถูกต้องเท่าไหร่
print('precision:', precision_score(y_test, y_predict)) # จากค่า prediction ของ graduate ทั้งหมดโมเดลทำนายทำนายได้อย่างถูกต้องเท่าไหร่
print('r2:       ', r2_score(y_test, y_predict)) # โมเดลอธิบายความแปรปรวนของข้อมูลได้เท่าไหร่
```

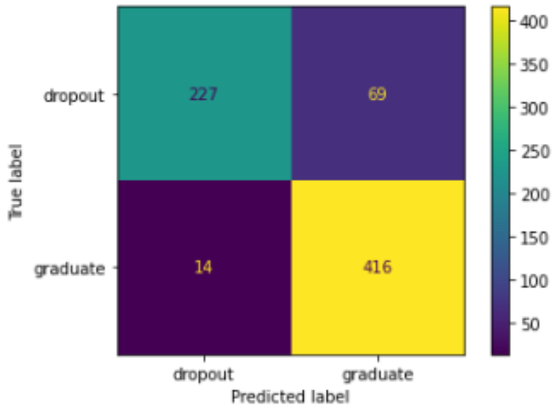
```
accuracy:  0.8856749311294766
recall:    0.9674418604651163
precision: 0.8577319587628865
r2:        0.526571338780641
```

โมเดลมีความแม่นย่าที่ 88% และโมเดลสามารถอธิบายความแปรปรวนได้ที่ 52%

Confusion Matrix

```
cm = confusion_matrix(y_test, y_predict)
graph = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=['dropout', 'graduate'])

graph.plot()
plt.show()
```



วัดความ Overfitting / Underfitting

```
y_predict = knn.predict(X_test)
print('accuracy on testing set: ', accuracy_score(y_test, y_predict))
y_predict = knn.predict(X_train)
print('accuracy on training set:', accuracy_score(y_train, y_predict)) # accuracy on training set
```

```
accuracy on testing set:  0.8856749311294766
accuracy on training set: 0.8966942148760331
```

คะแนนความแม่นยำของการทำนายจาก training set สูงกว่า testing set เล็กน้อย

หัวข้อที่ 3
Hypothesis testing

หัวข้อ 3

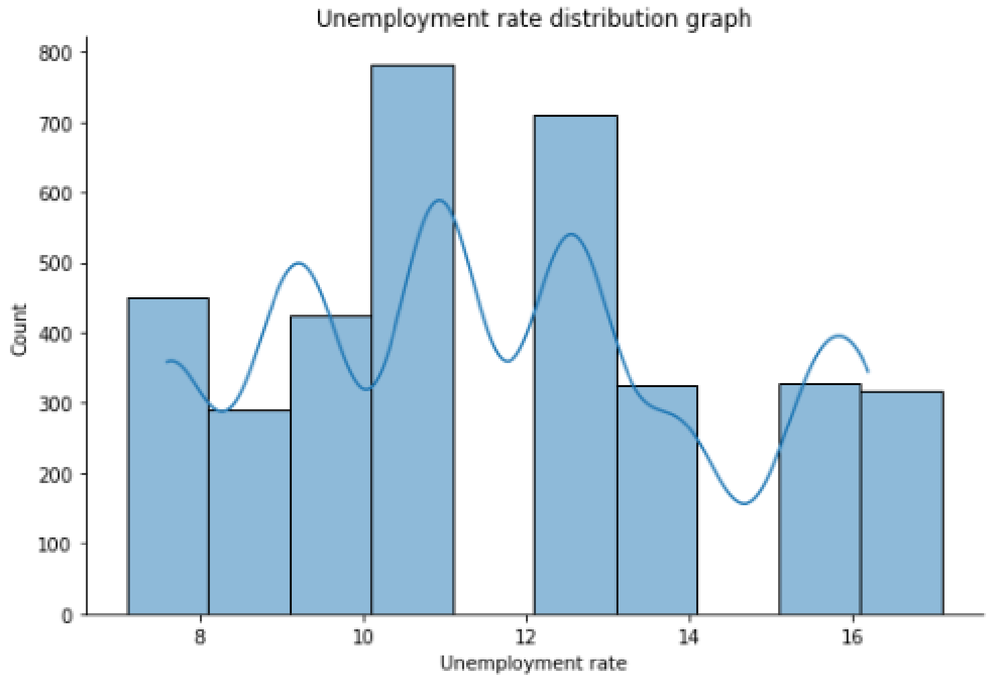
ทดสอบสมมุติฐานอัตราการว่างงานมีความสัมพันธ์เชิงเส้นกับการเรียนไม่จบหรือไม่ที่ระดับนัยสำคัญ 0.05

H0 : อัตราการว่างงานไม่มีความสัมพันธ์เชิงเส้นกับการเรียนไม่จบ

H1 : อัตราการว่างงานมีความสัมพันธ์เชิงเส้นกับการเรียนไม่จบ

```
import seaborn as sns
import matplotlib.pyplot as plt
```

```
sns.displot(df, x='Unemployment rate', kde=True, discrete=True, aspect=1.5)
plt.title("Unemployment rate distribution graph")
plt.show()
```



```python
from scipy import stats
```

```python
x = df['Unemployment rate']
y = df['Target']
rho, p_val = stats.spearmanr(x, y) # ข้อมูลเป็นประชากรจึงใช้ spearman
print('')
print(f'rho     : {rho}')    # ค่าสัมประสิทธิ์สหสัมพันธ์
print(f'p-value : {p_val}')  # ค่า p-value
```

```
rho     : 0.00842922527232026
p-value : 0.6116703699132648
```

จากค่า p-value 0.61 ซึ่งมากกว่าค่านัยสำคัญ 0.05 จึงยอมรับ H0 นั่นคืออัตราการว่างงานไม่มีความสัมพันธ์เชิงเส้นกับการเรียนไม่จบเป็นจริง

Member

นายณกรณ์ บุญประสงค์ 6510405458
นายธนวัฒน์ อุกฤษฎ์อัสดร 6510405563

Thank you very much!