Оюутны эцсийн дүнг урьдчилан таамаглах

Багийн гишүүдийн нэрс

2025 оны 12-р сарын 4

Энэхүү төслийн ажлын зорилго нь Португалийн дунд сургуулийн математикийн хичээлийн оюутнуудын эцсийн дүн (G3)-ийг урьдчилан таамаглахад шугаман регрессийн загварыг ашиглан статистикийн арга зүйг практикт хэрэглэх юм. Өгөгдлийн олонлогт 395 оюутны 33 шинж чанарын мэдээлэл багтсан бөгөөд загварын үр дүнг үнэлсэн. Үр дүнгээр оюутнуудын хоёр дахь үеийн дүн (G2) нь эцсийн дүнг урьдчилан таамаглахад хамгийн чухал хүчин зүйл бөгөөд загварын R^2 үзүүлэлт 0.75 байв.

Агуулга

1 Оршил

Боловсролын салбарт оюутнуудын сурлагын гүйцэтгэлийг урьдчилан таамаглах нь багш нар болон сургуулийн удирдлагад маш чухал ач холбогдолтой асуудал юм. Сурлагын үр дүнг өмнө нь мэдэх боломжтой бол хүндрэл тулгарч буй оюутнуудыг эрт илрүүлж, тэдэнд зохих дэмжлэг үзүүлэх, сургалтын арга барилаа сайжруулах боломж бүрдэнэ. Машин сургалтын аргууд өнөө үед боловсролын өгөгдлийг шинжлэх, ирээдүйн үр дүнг таамаглахад өргөн хэрэглэгддэг болсон.

Энэхүү судалгаа нь Португалийн дунд сургуулийн математикийн хичээлийн оюутнуудын эцсийн дүнг (G3) урьдчилан таамаглахад шугаман регрессийн загварыг ашиглан хийгдсэн. Судалгаанд 395 оюутны 33 янзын шинж чанар буюу демографик мэдээлэл, гэр бүлийн нөхцөл байдал, өмнөх сурлагын дүн зэрэг хувьсагчдыг ашигласан.

1.1 Төслийн ажлын зорилго

Төслийн ажлын үндсэн зорилгууд:

1. Оюутнуудын эцсийн дүнд хамгийн их нөлөөлөх хүчин зүйлсийг тодорхойлох
2. Шугаман регрессийн загварыг ашиглан математикийн эцсийн дүнг урьдчилан таамаглах
3. Загварын үр дүнг статистикийн аргуудаар үнэлэх
4. Боловсролын салбарт практикт хэрэглэх боломжтой санал дүгнэлт гаргах

2 Өгөгдлийн эх сурвалж ба тайлбар

2.1 Эх сурвалж

Энэхүү судалгаанд ашигласан өгөгдлийг Kaggle платформын “Student Performance Data Set” [1] эх сурвалжаас авсан болно. Уг өгөгдлийн олонлог нь Португалийн хоёр дунд сургуулийн (Gabriel Pereira ба Mousinho da Silveira) математикийн хичээлийн оюутнуудын 2005-2006 оны хичээлийн жилийн мэдээлэл юм. Өгөгдлийг Paulo Cortez ба Alice Silva нар цуглуулж, анх 2008 онд судалгаандаа ашигласан [2].

Өгөгдлийн олонлог нь нийт 395 оюутны мэдээллийг агуулж байгаа бөгөөд тус бүрт 33 шинж чанар (feature) байна. Эдгээр шинж чанарууд нь демографик мэдээлэл, нийгэм-эдийн засгийн үзүүлэлтүүд, гэр бүлийн нөхцөл байдал, сургуулийн дэмжлэг, өмнөх улирлуудын дүн зэрэг өргөн хүрээний хувьсагчдыг хамарна.

2.2 Өгөгдлийн тайлбар

Өгөгдлийн олонлог нь 395 оюутны 33 шинж чанарын мэдээллийг агуулна. Эдгээр шинж чанаруудыг дараах бүлэгт хуваан авч үзнэ:

Демографик мэдээлэл:

- `school` - Сургууль (Gabriel Pereira эсвэл Mousinho da Silveira)
- `sex` - Хүйс (эрэгтэй/эмэгтэй)
- `age` - Нас (15-22 нас)
- `address` - Амьдрах газар (хот/хөдөө)
- `famsize` - Гэр бүлийн хэмжээ (3-аас их эсвэл бага)
- `Pstatus` - Эцэг эхийн хамт амьдрах эсэх

Боловсролын мэдээлэл:

- `Medu` - Эхийн боловсролын түвшин (0-4, 0=үгүй, 4=дээд боловсрол)
- `Fedu` - Эцгийн боловсролын түвшин (0-4)
- `studytime` - Долоо хоногт суралцахад зарцуулах цаг (1: <2 цаг, 2: 2-5 цаг, 3: 5-10 цаг, 4: >10 цаг)
- `failures` - Өмнө унасан хичээлийн тоо (0-4)
- `schoolsup` - Сургуулийн нэмэлт дэмжлэг авсан эсэх
- `higher` - Дээд боловсрол эзэмшихийг хүсч байгаа эсэх
- `internet` - Гэртээ интернэт холболт байгаа эсэх
- `absences` - Хичээл тасалсан тоо (0-93)

Сурлагын дүн:

- `G1` - Эхний улирлын дүн (0-20 оноо)
- `G2` - Хоёрдугаар улирлын дүн (0-20 оноо)
- `G3` - Эцсийн дүн (0-20 оноо) - **зорилтот хувьсагч**

Судалгааны зорилтот хувьсагч нь `G3` буюу эцсийн дүн бөгөөд энэ нь жилийн эцсийн үнэлгээ юм. Бусад хувьсагчдыг ашиглан энэ дүнг урьдчилан таамаглахыг зорино.

2.3 Өгөгдлийг уншиж танилцах

Эхлээд шаардлагатай сангуудыг импортлож, өгөгдлийг уншина.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Өгөгдлийг уншиж авах
df = pd.read_csv('student-mat.csv')

# Өгөгдлийн хэмжээ
print(f"Өгөгдлийн хэмжээ: {df.shape[0]} мөр, {df.shape[1]} багана")
```

```
Өгөгдлийн хэмжээ: 395 мөр, 33 багана
```

Өгөгдөл амжилттай ачаалагдсан. Нийт 395 оюутны 33 шинж чанарын мэдээлэл байна.

Өгөгдлийн эхний хэдэн мөрийг харцгаая:

```
df.head()
```

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1

Өгөгдлийн статистик үзүүлэлтүүдийг авч үзье:

```
df.describe()
```

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout
count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000
mean	16.696203	2.749367	2.521519	1.448101	2.035443	0.334177	3.944304	3.235443	3.108861
std	1.276043	1.094735	1.088201	0.697505	0.839240	0.743651	0.896659	0.998862	1.113278
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000
25%	16.000000	2.000000	2.000000	1.000000	1.000000	0.000000	4.000000	3.000000	2.000000
50%	17.000000	3.000000	2.000000	1.000000	2.000000	0.000000	4.000000	3.000000	3.000000
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.000000	5.000000	4.000000	4.000000
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000	5.000000	5.000000	5.000000

Дутуу утгатай мөрүүдийг шалгацгаая:

```
print("Дутуу утгын тоо:")
print(df.isnull().sum())
```

```
Дутуу утгын тоо:
school        0
sex           0
age           0
address       0
famsize       0
Pstatus       0
Medu          0
Fedu          0
Mjob          0
Fjob          0
reason        0
guardian      0
traveltime    0
studytime     0
failures      0
schoolsup     0
famsup        0
paid          0
activities    0
nursery       0
higher        0
internet      0
romantic      0
famrel        0
freetime      0
goout         0
Dalc          0
Walc          0
health        0
absences      0
G1            0
G2            0
G3            0
dtype: int64
```

Өгөгдөлд дутуу утга байхгүй тул нөхөх шаардлагагүй болно. Бүх 395 оюутны бүрэн мэдээлэлтэй байна.

3 Хэрэглэсэн арга загварын танилцуулга

3.1 Шугаман регресс

Шугаман регресс нь хамгийн өргөн хэрэглэгддэг статистикийн загварчлалын аргуудын нэг бөгөөд хоёр ба түүнээс дээш хувьсагчдын хоорондын шугаман хамаарлыг тодорхойлоход ашиглагдана. Энэ аргын үндсэн зорилго нь тайлбарлагч хувьсагчдын (features) утгуудаас хамааруулан зорилтот хувьсагчийн (target) утгыг таамаглах явдал юм.

Энэхүү судалгаанд шугаман регрессийг сонгосон шалтгаанууд:

1. **Тайлбарлах чадвар** - Загварын коэффициентүүд нь тус бүр хувьсагчийн нөлөөллийг тодорхой илэрхийлдэг тул үр дүнг тайлбарлахад хялбар
2. **Хэрэгжүүлэхэд энгийн** - Математик загвар нь ойлгомжтой, тооцоолол хурдан

3. **Статистик үндэслэлтэй** - Загварын найдвартай байдлыг олон аргаар шалгах боломжтой (p-value, R^2, residual analysis)
4. **Бусад загвартай харьцуулах суурь загвар** - Илүү нарийн төвөгтэй загваруудын (neural network, random forest) гүйцэтгэлийг харьцуулахад суурь цэг болдог

Оюутны эцсийн дүнг таамаглахад шугаман регресс тохиромжтой, учир нь өмнөх улирлуудын дүн (G1, G2) болон бусад хувьсагчид эцсийн дүнтэй (G3) шугаман хамаарал үүсгэдэг нь өгөгдлийн шинжилгээгээр тогтоогдсон.

3.2 Математик загвар

Шугаман регрессийн ерөнхий загвар:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$$

Энд:

- [y тайлбар]
- [β_0 тайлбар]
- [β_i тайлбар]
- [x_i тайлбар]
- [ε тайлбар]

3.3 Загварын давуу тал

[Коэффициент ойлгогдох, хэрэгжүүлэхэд хялбар, хурдан, статистик үндэслэлтэй гэх мэт]

4 Өгөгдөлтэй танилцах шинжилгээ

4.1 Зорилтот хувьсагчийн шинжилгээ (G3)

```
# G3 статистик (дундаж, медиан, std, min, max)
# Histogram + boxplot зурах
```

[G3 тархалтын тайлбар]

4.2 Корреляцийн шинжилгээ

```
# G1, G2, G3 корреляцийн матриц
# Heatmap зурах
```

[G2-G3 корреляци өндөр тухай]

```
# G1 vs G3, G2 vs G3 scatter plots
```

[Шугаман хамаарал харагдаж байна]

4.3 Категори хувьсагчдын шинжилгээ

```
# sex, school, internet, higher гэх мэтийн boxplot-ууд
```

[Дээд боловсрол хүсч буй оюутнууд илүү өндөр дүнтэй]

4.4 Тоон хувьсагчдын корреляци

```
# Тоон хувьсагчдын G3-тай корреляци
# Bar chart
```

[failures хамгийн сөрөг корреляцитай]

5 Загварыг хэрэгжүүлсэн алхмууд

5.1 Шаардлагатай сангуудыг импортлох

```
# sklearn импортлох
# train_test_split, LinearRegression, LabelEncoder
# mean_squared_error, r2_score гэх мэт
```

[Эдгээр сангууд юунд ашиглагдах]

5.2 Өгөгдөл боловсруулалт

```
# Категори хувьсагчдыг кодлох (Label Encoding)
```

[Текст утгуудыг тоонд хөрвүүлсэн]

```
# X (features) болон y (target) ялгах
```

5.3 Сургалт ба тестийн олонлог

```
# train_test_split 80/20
```

[316 сургалт, 79 тест]

5.4 Загварыг сургах

```
# LinearRegression()
# model.fit()
```

5.5 Шинж чанаруудын коэффициентүүд

```
# Коэффициентүүдийг dataframe-д оруулах
# Bar chart зурах
# Топ 10 харуулах
```

[G2 хамгийн өндөр коэффициенттэй (0.95)]

6 Үр дүн ба загварын үнэлгээ

6.1 Таамаглал

```
# y_pred = model.predict()
# Comparison table (бодит vs таамаглал)
```

6.2 Загварын гүйцэтгэл

```
# MSE, RMSE, MAE, R² тооцоолох
# Үр дүнг хэвлэх
```

[R^2 = 0.75, MAE = 1.5 тайлбар]

```
# Бодит vs таамаглал scatter plot
```

[Улаан шугамд ойрхон = сайн таамаглал]

6.3 Үлдэгдлийн шинжилгээ

```
# Residual scatter plot
# Residual histogram
```

[Үлдэгдэл нормал тархалттай = таамаглал хангагдсан]

6.4 Хөндлөн баталгаажуулалт

```
# 5-fold cross-validation
# R² оноонуудыг харуулах
```

[Дундаж R^2 = 0.79, тогтвортой]

7 Дүгнэлт

7.1 Үндсэн дүгнэлтүүд

Төслийн ажлын үндсэн дүгнэлтүүд:

1. **[G2-ийн тухай дүгнэлт]**
2. **[Загварын нарийвчлалын тухай дүгнэлт]**
3. **[Failures-ийн сөрөг нөлөөний тухай]**
4. **[Бусад хүчин зүйлсийн тухай]**
5. **[Cross-validation-ий тухай]**

7.2 Практик хэрэглээ

[Хэрхэн хэрэглэж болох - эрт таних, багш дэмжлэг гэх мэт]

7.3 Хязгаарлалтууд

[Зөвхөн Португалийн өгөгдөл, корреляци≠шалтгаан, G2 шаардлагатай]

7.4 Цаашдын судалгаа

[Бусад загвар туршиж үзэх, feature engineering, илүү их өгөгдөл]

8 Багийн гишүүдийн үүрэг оролцоо

Гишүүний нэр	Үүрэг	Хувь нэмэр
[Нэр 1]	[Үүрэг 1]	25%
[Нэр 2]	[Үүрэг 2]	30%
[Нэр 3]	[Үүрэг 3]	25%
[Нэр 4]	[Үүрэг 4]	20%

Тэмдэглэл: [Бүх гишүүд идэвхтэй оролцсон]

Ашигласан материал

[1] Kaggle, «Student Grade Prediction Dataset». 2024. Available at: https://www.kaggle.com/datasets/dipam7/student-grade-prediction

[2] P. Cortez and A. Silva, «Using Data Mining to Predict Secondary School Student Performance», *EUROSIS*, 2008.