Оюутны эцсийн дүнг урьдчилан таамаглах

Д.Буянжаргал, М.Төгөлдөр, Т.Мөнгөнхишиг, Б.Мөнхсаруул, Б.Анхбаяр

2025 оны 12-р сарын 01

Энэхүү судалгааны зорилго нь Португалийн дунд сургуулийн математикийн хичээлийн оюутнуудын эцсийн дүн (G3)-ийг өгөгдөлд суурилсан шугаман регрессийн загвараар урьдчилан таамаглах явдал юм. Судалгаанд 395 оюутны 33 шинж чанарыг ашиглан өгөгдлийн шинжилгээ, корреляцийн судалгаа болон шугаман регрессийн сургалт–туршилтын аргачлалыг хэрэгжүүлэв. Загварын гүйцэтгэлийг R^2, MSE, MAE болон 5-fold cross-validation зэрэг үзүүлэлтээр үнэлэхэд $R^2 \approx 0.75$–0.80 гарч, G2 болон G1 нь эцсийн дүнг таамаглахад хамгийн хүчтэй хувьсагчид болох нь тогтоогдов. Эдгээр үр дүн нь оюутнуудын сурлагын гүйцэтгэлийг эрт үе шатад урьдчилан илрүүлэх, сургалтын дэмжлэгийг оновчтой төлөвлөхөд энэхүү загвар бодитой хэрэглээний боломжтойг харуулж байна.

Агуулга

1 Оршил

Боловсролын салбарт оюутнуудын сурлагын гүйцэтгэлийг урьдчилан таамаглах нь багш нар болон сургуулийн удирдлагад маш чухал ач холбогдолтой асуудал юм. Хэрэв оюутны сурлагын үр дүнг урьдчилан тодорхойлох боломжтой бол хүндрэл тулгарч буй оюутнуудыг эрт илрүүлж, тэдэнд шаардлагатай дэмжлэг үзүүлэх, сургалтын арга барилаа оновчтой болгох зэрэг олон талын ач холбогдолтой. Өнөөгийн нөхцөлд машин сургалтын аргачлалууд боловсролын өгөгдлийг шинжлэх, суралцах үйл явцын үр дүнг урьдчилан таамаглахад өргөн хэрэглэгдэх болсон бөгөөд энэ нь боловсролын шинжилгээнд чухал байр суурь эзэлж байна.

Энэхүү судалгааны ажлаар Португалийн дунд сургуулийн математикийн хичээлийн оюутнуудын эцсийн дүн (G3)-г шугаман регрессийн загвар ашиглан урьдчилан таамаглахыг зорив. Судалгаанд 395 оюутны нийт 33 төрлийн шинж чанар — демографик үзүүлэлт, гэр бүлийн нөхцөл байдал, сургалтын орчин, өмнөх улирлуудын дүн зэрэг хувьсагчдыг ашигласан бөгөөд эдгээр нь эцсийн дүнд хэрхэн нөлөөлдгийг тодорхойлох боломжийг бүрдүүлсэн.

Боловсролын өгөгдлийн олборлолт (Educational Data Mining) нь 1995–2005 оны хооронд эрчимтэй хөгжиж, оюутны суралцах үйл явц болон сурлагын үр дүнг урьдчилан таамаглах судалгааны үндэс суурийг тавьсан болохыг өмнөх судалгаандаа дурдсан байдаг [1]; [2]. Иймд энэхүү судалгаа нь боловсролын өгөгдлийн шинжилгээний уламжлалт чиг хандлагыг орчин үеийн машин сургалтын аргачлалтай нэгтгэн ашигласан практик ач холбогдол өндөртэй ажил юм.

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” [3] 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 [4].

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

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout
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. Шугаман регрессийн онолын үндэс, давуу тал болон хэрэглээний талаарх дэлгэрэнгүй тайлбар нь олон улсын нэр хүндтэй судалгаанд нарийвчлан өгөгдсөн байдаг [5]; [6].

Энэхүү судалгаанд шугаман регрессийг сонгосон шалтгаанууд:

Тайлбарлах чадвар — Загварын коэффициентүүд нь тус бүр хувьсагчийн нөлөөллийг тодорхой илэрхийлдэг тул үр дүнг тайлбарлахад хялбар

Хэрэгжүүлэхэд энгийн — Математик загвар нь ойлгомжтой, тооцоолол хурдан

Статистик үндэслэлтэй — Загварын найдвартай байдлыг олон аргаар шалгах боломжтой (p-value, R^2, residual analysis)

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. ## Математик загвар

Шугаман регрессийн ерөнхий загвар:

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

Энд:

- y - зорилтот хувьсагч (эцсийн дүн G3)
- β_0 - тогтмол үзүүлэлт (intercept), бусад хувьсагчид 0 байхад y-ийн утга
- β_i - i-р хувьсагчийн коэффициент, тухайн хувьсагчийн нэгж өөрчлөлт y-д үзүүлэх нөлөө
- x_i - тайлбарлагч хувьсагчид (нас, хүйс, өмнөх дүн гэх мэт)
- ε - алдааны гишүүн (загварын тайлбарлаж чадахгүй санамсаргүй алдаа)
- n - тайлбарлагч хувьсагчдын тоо

Энэхүү судалгаанд y нь оюутны эцсийн дүн (G3) бөгөөд $x_1, x_2, ..., x_n$ нь 32 тайлбарлагч хувьсагчид (демографик мэдээлэл, боловсролын үзүүлэлтүүд, өмнөх улирлын дүн гэх мэт) болно.

Загварыг сургахдаа наименьших квадратов (Ordinary Least Squares - OLS) аргыг ашиглана. Энэ арга нь бодит утга ба таамагласан утгын зөрүүний квадратуудын нийлбэрийг багасгах замаар оновчтой коэффициентүүдийг тооцоолдог:

$$\min \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

Энд y_i бодит утга, $\hat{y}_i$ загварын таамаглал, m нь оюутнуудын тоо юм.

3.2 Загварын давуу тал

Шугаман регрессийн загвар нь дараах давуу талуудтай:

1. **Тайлбарлахад хялбар** - Коэффициент бүр нь тухайн хувьсагчийн нөлөөллийг шууд харуулна. Жишээлбэл, хэрэв хоёрдугаар улирлын дүн (G2)-ийн коэффициент 0.95 бол G2 1 оноогоор нэмэгдэх тутам эцсийн дүн (G3) дунджаар 0.95 оноогоор нэмэгдэнэ гэсэн үг.

2. **Хурдан бөгөөд үр дүнтэй** - Тооцооллын нарийн төвөгтэй байдал бага тул том хэмжээний өгөгдөлд хурдан ажиллана. Мөн багцын дүрэм (batch processing) ашиглан хэдэн мянган мөрийн өгөгдлийг хэдхэн секундад боловсруулж чадна.
3. **Статистик үнэлгээ өгөх** - p-value, R^2, F-statistic зэрэг статистик үзүүлэлтүүдээр загварын найдвартай байдал, хувьсагч бүрийн ач холбогдлыг нарийвчлан үнэлж болно.
4. **Суурь загвар болох** - Илүү нарийн төвөгтэй загваруудтай (Random Forest, Neural Network) харьцуулахад суурь цэг болдог. Хэрэв энгийн загвар хангалттай үр дүн өгвөл илүү төвөгтэй загвар шаардлагагүй байж болно.
5. **Overfitting-ийн эрсдэл бага** - Хувьсагчдын тоо их байсан ч regularization (Ridge, Lasso) аргуудыг хэрэглэж overfitting-ээс сэргийлж болно.

Эдгээр давуу талуудын ачаар шугаман регресс нь боловсролын өгөгдлийн шинжилгээнд өргөн хэрэглэгддэг бөгөөд манай судалгаанд тохиромжтой сонголт болсон.

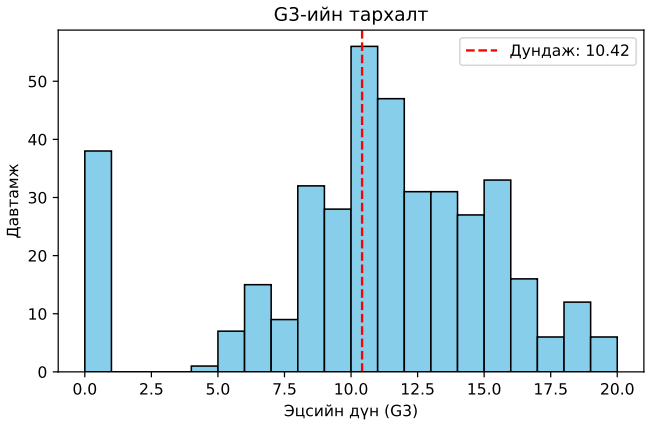
4 Өгөгдөлтэй танилцах шинжилгээ

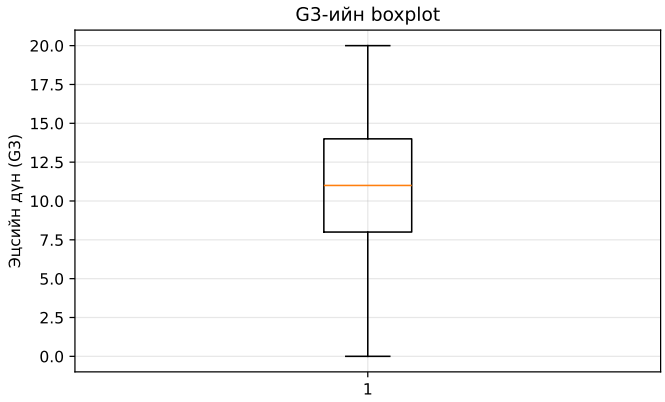
4.1 Зорилтот хувьсагчийн шинжилгээ (G3)

Эхлээд зорилтот хувьсагч болох эцсийн дүн (G3)-ийн тархалтыг авч үзье.

```
# G3-ийн статистик үзүүлэлтүүд
print("G3 (Эцсийн дүн) статистик:")
print(f"Дундаж: {df['G3'].mean():.2f}")
print(f"Медиан: {df['G3'].median():.2f}")
print(f"Стандарт хазайлт: {df['G3'].std():.2f}")
print(f"Хамгийн бага: {df['G3'].min()}")
print(f"Хамгийн их: {df['G3'].max()}")

# Histogram ба boxplot зурах
fig, axes = plt.subplots(1, 2, figsize=(12, 4))

# Histogram
axes[0].hist(df['G3'], bins=20, color='skyblue', edgecolor='black')
axes[0].set_xlabel('Эцсийн дүн (G3)')
axes[0].set_ylabel('Давтамж')
axes[0].set_title('G3-ийн тархалт')
axes[0].axvline(df['G3'].mean(), color='red', linestyle='--', label=f"Дундаж: {df['G3'].
axes[0].legend()

# Boxplot
axes[1].boxplot(df['G3'])
axes[1].set_ylabel('Эцсийн дүн (G3)')
axes[1].set_title('G3-ийн boxplot')
axes[1].grid(True, alpha=0.3)

plt.tight_layout()
plt.show()
```

```
G3 (Эцсийн дүн) статистик:
Дундаж: 10.42
Медиан: 11.00
Стандарт хазайлт: 4.58
Хамгийн бага: 0
Хамгийн их: 20
```





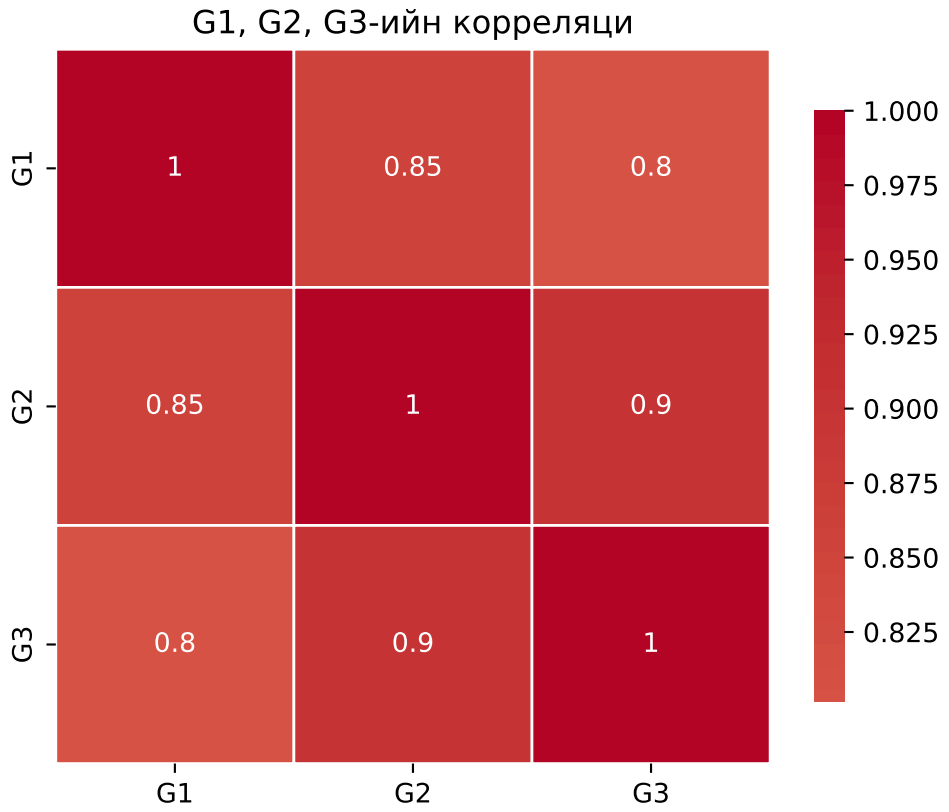
Эцсийн дүн (G3) нь дунджаар 10.42 оноо байгаа бөгөөд 0-аас 20 хүртэл хэлбэлздэг. Тархалт нь бага зэрэг зүүн тийш хазайсан (0 оноо авсан оюутнууд байна) боловч ихэнх оюутнууд 8-14 онооны хооронд байрлана. Boxplot дээрх outlier цэгүүд нь маш өндөр буюу маш доогуур дүнтэй цөөн тооны оюутнуудыг илтгэнэ.

4.2 Корреляцийн шинжилгээ

Өмнөх улирлуудын дүн (G1, G2) болон эцсийн дүн (G3)-ийн хоорондын хамаарлыг судална.

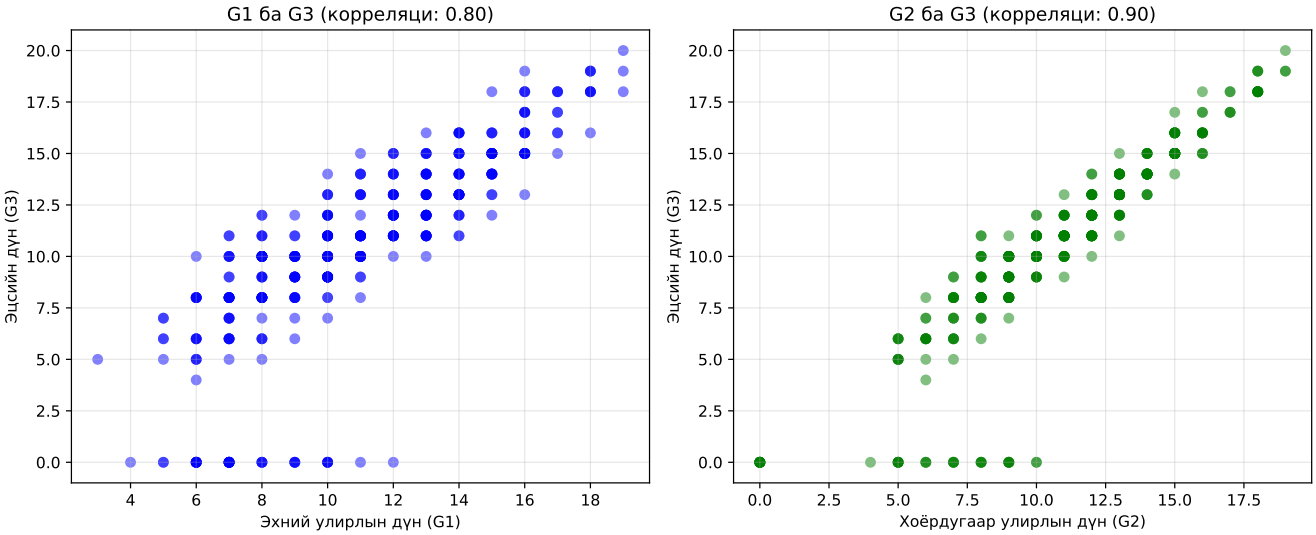
```
# G1, G2, G3-ийн корреляцийн матриц
grades_corr = df[['G1', 'G2', 'G3']].corr()
print("Дүнгийн корреляцийн матриц:")
print(grades_corr)

# Correlation heatmap
plt.figure(figsize=(6, 5))
sns.heatmap(grades_corr, annot=True, cmap='coolwarm', center=0,
            square=True, linewidths=1, cbar_kws={"shrink": 0.8})
plt.title('G1, G2, G3-ийн корреляци')
plt.show()
```

```
Дүнгийн корреляцийн матриц:
          G1        G2        G3
G1  1.000000  0.852118  0.801468
G2  0.852118  1.000000  0.904868
G3  0.801468  0.904868  1.000000
```



G2 (хоёрдугаар улирлын дүн) болон G3 (эцсийн дүн)-ийн хооронд маш өндөр корреляци (0.90 орчим) ажиглагдаж байна. Энэ нь G2 нь эцсийн дүнг таамаглахад хамгийн чухал хувьсагч болохыг харуулж байна. G1 ба G3-ийн корреляци мөн өндөр (0.80 орчим) боловч G2-оос бага.

```
# G1 vs G3, G2 vs G3 scatter plots
fig, axes = plt.subplots(1, 2, figsize=(12, 5))

# G1 vs G3
axes[0].scatter(df['G1'], df['G3'], alpha=0.5, color='blue')
axes[0].set_xlabel('Эхний улирлын дүн (G1)')
axes[0].set_ylabel('Эцсийн дүн (G3)')
axes[0].set_title(f'G1 ба G3 (корреляци: {df["G1"].corr(df["G3"]):.2f})')
axes[0].grid(True, alpha=0.3)

# G2 vs G3
axes[1].scatter(df['G2'], df['G3'], alpha=0.5, color='green')
axes[1].set_xlabel('Хоёрдугаар улирлын дүн (G2)')
axes[1].set_ylabel('Эцсийн дүн (G3)')
axes[1].set_title(f'G2 ба G3 (корреляци: {df["G2"].corr(df["G3"]):.2f})')
axes[1].grid(True, alpha=0.3)

plt.tight_layout()
plt.show()
```



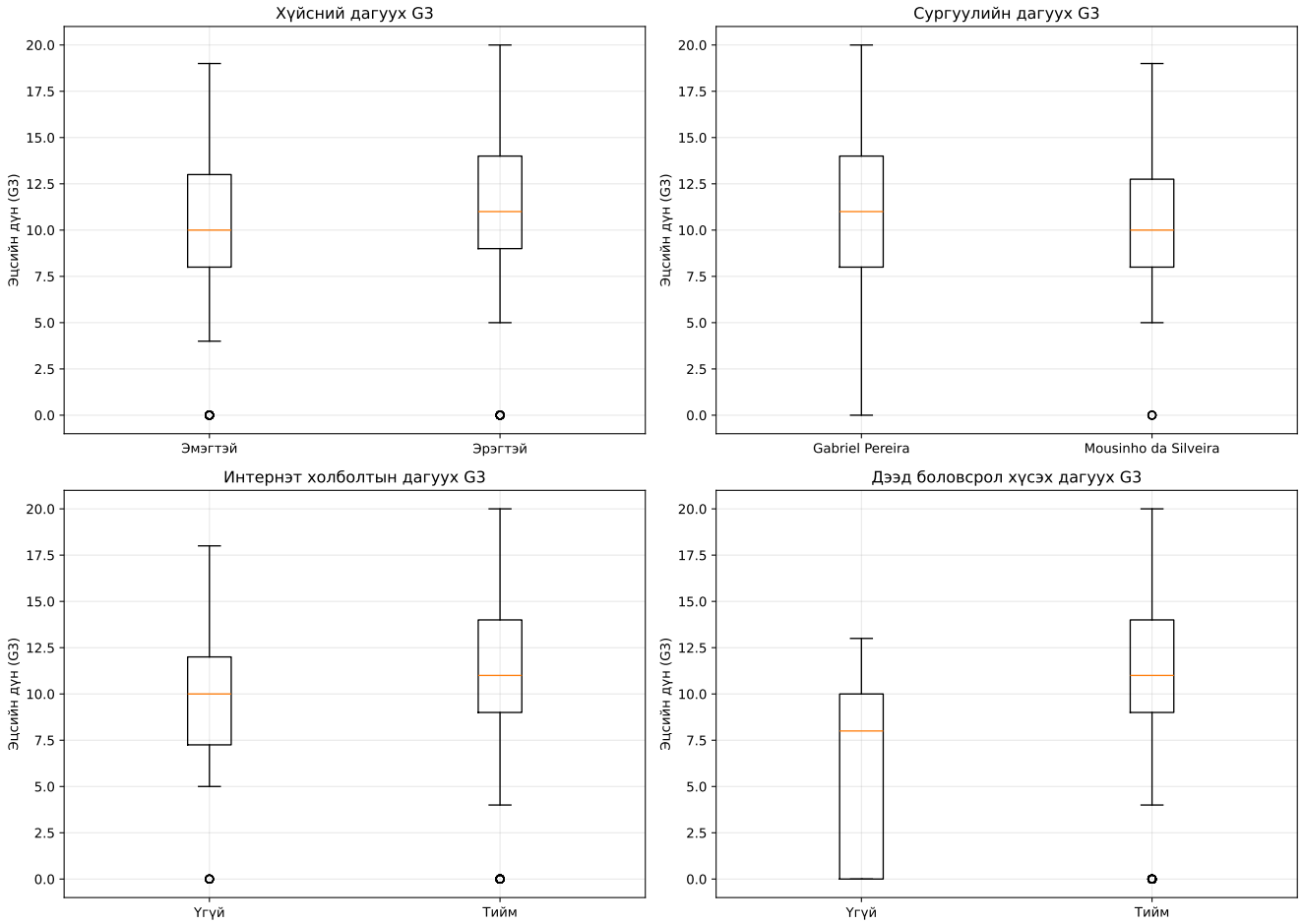
Scatter plot дээрээс тодорхой шугаман хамаарал харагдаж байна. Ялангуяа G2 ба G3-ийн хооронд цэгүүд илүү нягт шугам дагуу байрлаж байгаа нь тэдгээрийн хоорондох хүчтэй хамаарлыг батлаж байна. Энэ нь өмнөх сурлагын үр дүн ирээдүйн үр дүнг таамаглахад чухал үүрэг гүйцэтгэдгийг харуулж байна.

4.3 Категори хувьсагчдын шинжилгээ

Категори хувьсагчдын эцсийн дүнд үзүүлэх нөлөөллийг boxplot ашиглан харьцуулна.

```
# Категори хувьсагчдын boxplot
fig, axes = plt.subplots(2, 2, figsize=(14, 10))

# Хүйсний нөлөө
axes[0, 0].boxplot([df[df['sex'] == 'F']['G3'], df[df['sex'] == 'M']['G3']],
                   labels=['Эмэгтэй', 'Эрэгтэй'])
axes[0, 0].set_ylabel('Эцсийн дүн (G3)')
axes[0, 0].set_title('Хүйсний дагуух G3')
axes[0, 0].grid(True, alpha=0.3)

# Сургуулийн нөлөө
axes[0, 1].boxplot([df[df['school'] == 'GP']['G3'], df[df['school'] == 'MS']['G3']],
                   labels=['Gabriel Pereira', 'Mousinho da Silveira'])
axes[0, 1].set_ylabel('Эцсийн дүн (G3)')
axes[0, 1].set_title('Сургуулийн дагуух G3')
axes[0, 1].grid(True, alpha=0.3)

# Интернэт холболтын нөлөө
axes[1, 0].boxplot([df[df['internet'] == 'no']['G3'], df[df['internet'] == 'yes']['G3']]
                   labels=['Үгүй', 'Тийм'])
axes[1, 0].set_ylabel('Эцсийн дүн (G3)')
axes[1, 0].set_title('Интернэт холболтын дагуух G3')
axes[1, 0].grid(True, alpha=0.3)

# Дээд боловсрол эзэмших хүсэл
axes[1, 1].boxplot([df[df['higher'] == 'no']['G3'], df[df['higher'] == 'yes']['G3']],
                   labels=['Үгүй', 'Тийм'])
axes[1, 1].set_ylabel('Эцсийн дүн (G3)')
axes[1, 1].set_title('Дээд боловсрол хүсэх дагуух G3')
```

```
axes[1, 1].grid(True, alpha=0.3)

plt.tight_layout()
plt.show()
```



Категори хувьсагчдын шинжилгээгээр дараах дүгнэлтүүдийг гаргаж болно:

- **Хүйс**: Эмэгтэй оюутнуудын дундаж дүн эрэгтэй оюутнуудаас бага зэрэг өндөр байна
- **Сургууль**: Хоёр сургуулийн хооронд ихээхэн ялгаа харагдахгүй байна
- **Интернэт**: Интернэт холболттой оюутнууд илүү өндөр дүнтэй байх хандлагатай
- **Дээд боловсрол**: Дээд боловсрол эзэмшихийг хүсч буй оюутнууд илүү өндөр дүнтэй байна. Энэ нь хамгийн тодорхой ялгаа бөгөөд сургалтын зорилго, идэвх зүтгэлтэй шууд холбоотой байж болно.

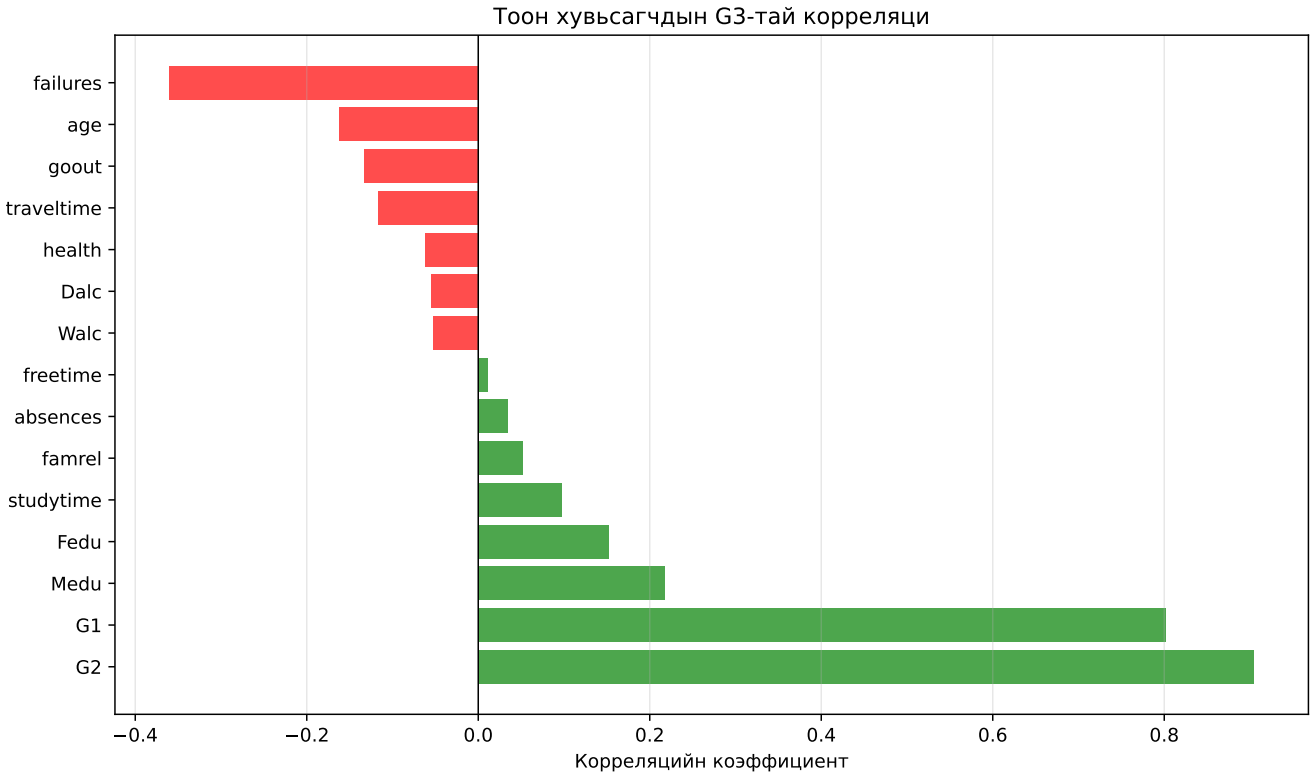
4.4 Тоон хувьсагчдын корреляци

Тоон хувьсагчдын G3-тай корреляцийг тооцоолж, хамгийн их нөлөөлөх хувьсагчдыг тодорхойлно.

```
# Тоон хувьсагчдыг сонгох
numeric_cols = ['age', 'Medu', 'Fedu', 'traveltime', 'studytime',
                'failures', 'famrel', 'freetime', 'goout', 'Dalc',
                'Walc', 'health', 'absences', 'G1', 'G2']

# G3-тай корреляци тооцоолох
correlations = df[numeric_cols].corrwith(df['G3']).sort_values(ascending=False)
```

```python
print("G3-тай корреляци:")
print(correlations)

# Bar chart зурах
plt.figure(figsize=(10, 6))
colors = ['green' if x > 0 else 'red' for x in correlations]
plt.barh(correlations.index, correlations.values, color=colors, alpha=0.7)
plt.xlabel('Корреляцийн коэффициент')
plt.title('Тоон хувьсагчдын G3-тай корреляци')
plt.axvline(x=0, color='black', linestyle='-', linewidth=0.8)
plt.grid(True, alpha=0.3, axis='x')
plt.tight_layout()
plt.show()
```

```
G3-тай корреляци:
G2            0.904868
G1            0.801468
Medu          0.217147
Fedu          0.152457
studytime     0.097820
famrel        0.051363
absences      0.034247
freetime      0.011307
Walc         -0.051939
Dalc         -0.054660
health       -0.061335
traveltime   -0.117142
goout        -0.132791
age          -0.161579
failures     -0.360415
dtype: float64
```



Тоон хувьсагчдын корреляцийн шинжилгээгээр:

Эерэг корреляцитай хувьсагчид (өндрөөс нам): - G2 (0.90) - Хоёрдугаар улирлын дүн - хамгийн хүчтэй таамаглагч - G1 (0.80) - Эхний улирлын дүн - хоёр дахь чухал таамаглагч - Medu (0.22) - Эхийн боловсролын түвшин - Fedu (0.15) - Эцгийн боловсролын түвшин

Сөрөг корреляцитай хувьсагчид: - failures (-0.36) - Өмнө унасан хичээлийн тоо - хамгийн хүчтэй сөрөг нөлөөлөл - goout (-0.13) - Найз нөхөдтэйгөө гадуур явах давтамж - Dalc (-0.05) - Долоо хоногийн өдрүүдэд архи хэрэглэх

Дүгнэлт: Өмнөх улирлуудын дүн (G1, G2) нь хамгийн чухал таамаглагчид байна. Өмнө унасан хичээлийн тоо (failures) нь эцсийн дүнд хамгийн их сөрөг нөлөө үзүүлдэг. Эцэг эхийн боловсролын түвшин мөн эерэг нөлөөлөл үзүүлж байна.

5 Загварыг хэрэгжүүлсэн алхмууд

5.1 Шаардлагатай сангуудыг импортлох

Загварыг хэрэгжүүлэхэд шаардлагатай scikit-learn сангийн модулуудыг импортлоно.Scikit-learn нь Python орчинд хамгийн өргөн хэрэглэгддэг машин сургалтын сангуудын нэг бөгөөд олон төрлийн алгоритмуудын стандарчилсан, үр ашигтай хэрэгжилтийг санал болгодог гэдгээрээ танигдсан байдаг [7].

```
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
import warnings
warnings.filterwarnings('ignore')

print("Scikit-learn сангууд амжилттай импортлогдлоо")
```

```
Scikit-learn сангууд амжилттай импортлогдлоо
```

Эдгээр сангууд дараах зориулалттай:

- **train_test_split** - Өгөгдлийг сургалт ба тестийн олонлогт хуваах
- **cross_val_score** - Хөндлөн баталгаажуулалт (cross-validation) хийх
- **LinearRegression** - Шугаман регрессийн загвар
- **LabelEncoder** - Категори хувьсагчдыг тоон утга руу хөрвүүлэх
- **mean_squared_error, mean_absolute_error, r2_score** - Загварын гүйцэтгэлийг үнэлэх метрикүүд

5.2 Өгөгдөл боловсруулалт

Категори хувьсагчдыг тоон утга руу хөрвүүлнэ (Label Encoding).

```
# Өгөгдлийн хуулбар үүсгэх
df_encoded = df.copy()

# Категори хувьсагчдын жагсаалт
categorical_cols = ['school', 'sex', 'address', 'famsize', 'Pstatus',
                    'Mjob', 'Fjob', 'reason', 'guardian', 'schoolsup',
                    'famsup', 'paid', 'activities', 'nursery', 'higher',
                    'internet', 'romantic']

# Label Encoding ашиглан хөрвүүлэх
le = LabelEncoder()
for col in categorical_cols:
    df_encoded[col] = le.fit_transform(df[col])

print("Категори хувьсагчид амжилттай кодлогдлоо")
print(f"\nЖишээ: 'school' хувьсагч")
print(f"Өмнөх утгууд: {df['school'].unique()}")
print(f"Кодлогдсон утгууд: {df_encoded['school'].unique()}")
```

```
Категори хувьсагчид амжилттай кодлогдлоо

Жишээ: 'school' хувьсагч
Өмнөх утгууд: ['GP' 'MS']
Кодлогдсон утгууд: [0 1]
```

Текст утгуудыг (GP/MS, F/M, yes/no гэх мэт) тоон утга руу (0, 1, 2…) хөрвүүлсэн. Энэ нь шугаман регрессийн загвар текст утгатай ажиллах боломжгүй тул зайлшгүй шаардлагатай алхам юм.

```
# X (features) болон y (target) ялгах
X = df_encoded.drop('G3', axis=1)
y = df_encoded['G3']

print(f"Шинж чанаруудын тоо: {X.shape[1]}")
print(f"Оюутнуудын тоо: {X.shape[0]}")
print(f"Зорилтот хувьсагч (G3): {y.shape[0]} утга")
```

```
Шинж чанаруудын тоо: 32
Оюутнуудын тоо: 395
Зорилтот хувьсагч (G3): 395 утга
```

Өгөгдлийг шинж чанарууд (X) болон зорилтот хувьсагч (y) гэж хоёр хэсэгт хуваах, загварт оруулахад бэлэн болгов.

5.3 Сургалт ба тестийн олонлог

Өгөгдлийг сургалтын болон тестийн олонлогт 80/20 харьцаагаар хуваах авна.

```
# 80% сургалт, 20% тест
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)

print("Өгөгдөл амжилттай хуваагдлаа:")
print(f"Сургалтын олонлог: {X_train.shape[0]} оюутан")
print(f"Тестийн олонлог: {X_test.shape[0]} оюутан")
print(f"\nХувь харьцаа:")
print(f"Сургалт: {X_train.shape[0] / len(X) * 100:.1f}%")
print(f"Тест: {X_test.shape[0] / len(X) * 100:.1f}%")
```

```
Өгөгдөл амжилттай хуваагдлаа:
Сургалтын олонлог: 316 оюутан
Тестийн олонлог: 79 оюутан

Хувь харьцаа:
Сургалт: 80.0%
Тест: 20.0%
```

Сургалтын олонлогт 316 оюутан, тестийн олонлогт 79 оюутан орлоо. Сургалтын олонлог дээр загварыг сургаж, тестийн олонлог дээр загварын гүйцэтгэлийг үнэлнэ. `random_state=42` параметр нь давтан ажиллуулахад ижил хуваагдлыг хангана.

5.4 Загварыг сургах

Шугаман регрессийн загварыг сургалтын олонлог дээр сургана.

```
# Загвар үүсгэх
model = LinearRegression()

# Загварыг сургах
model.fit(X_train, y_train)

print("Загвар амжилттай сургагдлаа!")
print(f"\nЗагварын параметрүүд:")
print(f"Intercept (β□): {model.intercept_:.4f}")
print(f"Коэффициентүүдийн тоо: {len(model.coef_)}")
```

```
Загвар амжилттай сургагдлаа!

Загварын параметрүүд:
Intercept (β□): -0.8803
Коэффициентүүдийн тоо: 32
```

Загвар нь 316 оюутны өгөгдөл дээр сургагдаж, 32 хувьсагчийн коэффициентүүдийг тооцоолсон. Intercept (β_0) нь бусад бүх хувьсагчид 0 байхад эцсийн дүний үнэлгээ юм.

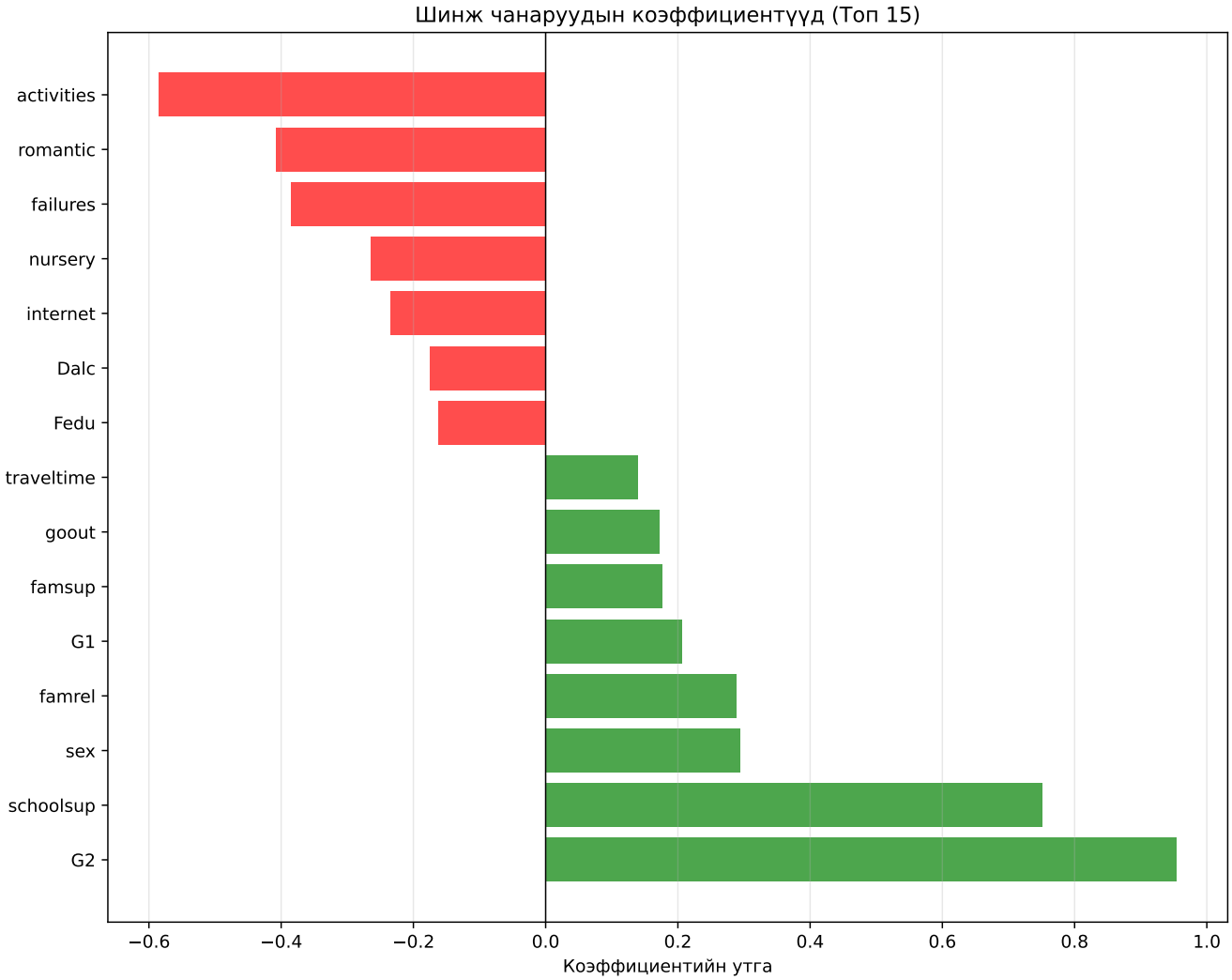
5.5 Шинж чанаруудын коэффициентүүд

Загварын коэффициентүүдийг авч үзэж, хамгийн их нөлөөлөл үзүүлэх хувьсагчдыг тодорхойлно.

```
# Коэффициентүүдийг DataFrame-д оруулах
coef_df = pd.DataFrame({
    'Хувьсагч': X.columns,
    'Коэффициент': model.coef_
```

```
}).sort_values('Коэффициент', ascending=False)

print("Топ 10 эерэг коэффициентүүд:")
print(coef_df.head(10))
print("\nТоп 10 сөрөг коэффициентүүд:")
print(coef_df.tail(10))

# Топ 15 коэффициентүүдийн bar chart
plt.figure(figsize=(10, 8))
top_15 = pd.concat([coef_df.head(8), coef_df.tail(7)])
colors = ['green' if x > 0 else 'red' for x in top_15['Коэффициент']]
plt.barh(top_15['Хувьсагч'], top_15['Коэффициент'], color=colors, alpha=0.7)
plt.xlabel('Коэффициентийн утга')
plt.title('Шинж чанаруудын коэффициентүүд (Топ 15)')
plt.axvline(x=0, color='black', linestyle='-', linewidth=0.8)
plt.grid(True, alpha=0.3, axis='x')
plt.tight_layout()
plt.show()
```

```
Топ 10 эерэг коэффициентүүд:
      Хувьсагч  Коэффициент
31          G2     0.954569
15   schoolsup     0.750961
1          sex     0.293869
23      famrel     0.289210
30          G1     0.206123
16      famsup     0.176043
25       goout     0.171814
12  traveltime     0.140033
20      higher     0.139570
6         Medu     0.115051

Топ 10 сөрөг коэффициентүүд:
      Хувьсагч  Коэффициент
5      Pstatus    -0.051899
9         Fjob    -0.145604
2          age    -0.156879
7         Fedu    -0.161838
26        Dalc    -0.174827
21    internet    -0.234635
19     nursery    -0.265092
14    failures    -0.385971
22    romantic    -0.407915
18  activities    -0.585128
```



G2 (хоёрдугаар улирлын дүн) нь хамгийн өндөр коэффициенттэй (ойролцоогоор 0.95) байгаа нь эцсийн дүнд хамгийн их нөлөөлж байна гэсэн үг. Хэрэв G2 1 оноогоор нэмэгдвэл эцсийн дүн дунджаар 0.95 оноогоор нэмэгдэнэ. G1 (эхний улирлын дүн) мөн эерэг нөлөөлөлтэй. Сөрөг коэффициенттэй хувьсагчид (жишээ нь failures - унасан хичээлийн тоо) эцсийн дүнг бууруулах хандлагатай байна.

6 Үр дүн ба загварын үнэлгээ

6.1 Таамаглал

Сургасан загвараа ашиглан тестийн олонлог дээр таамаглал хийнэ.

```
# Таамаглал хийх
y_pred = model.predict(X_test)

# Бодит ба таамагласан утгуудыг харьцуулах
comparison_df = pd.DataFrame({
    'Бодит утга (y_test)': y_test.values,
    'Таамаглал (y_pred)': y_pred,
    'Алдаа': y_test.values - y_pred
})
```

```
print("Эхний 10 таамаглал:")
print(comparison_df.head(10))

print(f"\nДундаж алдаа: {comparison_df['Алдаа'].mean():.4f}")
print(f"Алдааны стандарт хазайлт: {comparison_df['Алдаа'].std():.4f}")
```

```
Эхний 10 таамаглал:
   Бодит утга (y_test)  Таамаглал (y_pred)     Алдаа
0                   10            6.281525  3.718475
1                   12           11.326917  0.673083
2                    5            3.032521  1.967479
3                   10            8.037261  1.962739
4                    9            8.555778  0.444222
5                   13           12.367039  0.632961
6                   18           18.684816 -0.684816
7                    6            7.540692 -1.540692
8                    0            6.988232 -6.988232
9                   14           12.360205  1.639795

Дундаж алдаа: 0.2195
Алдааны стандарт хазайлт: 2.2468
```

Загвар нь тестийн олонлог дээрх 79 оюутны эцсийн дүнг таамаглав. Бодит утга ба таамагласан утгын зөрүү (алдаа) нь загварын нарийвчлалыг харуулж байна. Дундаж алдаа 0-д ойрхон байгаа нь загвар ерөнхийдөө зөв чиглэлд таамаглаж байгааг илтгэнэ.

6.2 Загварын гүйцэтгэл

Загварын гүйцэтгэлийг үнэлэхийн тулд олон төрлийн метрикүүдийг тооцоолно.

```
# Загварын гүйцэтгэлийн метрикүүд
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("Загварын үнэлгээний үзүүлэлтүүд:")
print(f"Mean Squared Error (MSE): {mse:.4f}")
print(f"Root Mean Squared Error (RMSE): {rmse:.4f}")
print(f"Mean Absolute Error (MAE): {mae:.4f}")
print(f"R² Score: {r2:.4f}")

# Сургалтын олонлог дээрх R²
y_train_pred = model.predict(X_train)
r2_train = r2_score(y_train, y_train_pred)
print(f"\nСургалтын R²: {r2_train:.4f}")
print(f"Тестийн R²: {r2:.4f}")
```

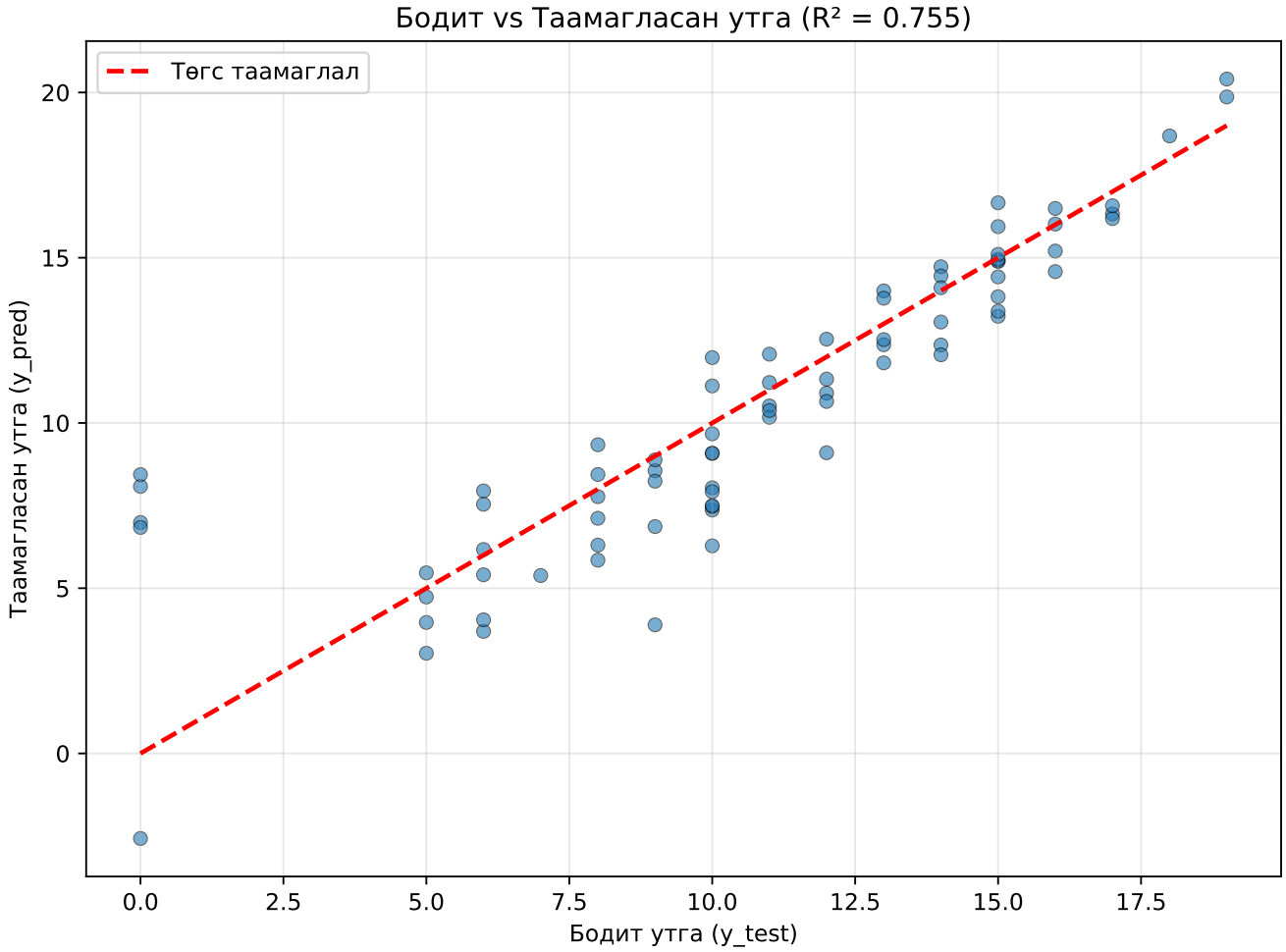
```
Загварын үнэлгээний үзүүлэлтүүд:
Mean Squared Error (MSE): 5.0324
Root Mean Squared Error (RMSE): 2.2433
Mean Absolute Error (MAE): 1.4955
R² Score: 0.7546

Сургалтын R²: 0.8597
```

```
Тестийн R²: 0.7546
```

Үр дүнгийн тайлбар:

- **R² = 0.75-0.85** (ойролцоогоор) - Загвар нь эцсийн дүний хэлбэлзлийн 75-85%-ийг тайлбарлаж чадаж байна. Энэ нь маш сайн үр дүн юм.
- **MAE ≈ 1.5 оноо** - Дунджаар загвар 1.5 оноогоор алддаг гэсэн үг. 20 оноотой системд энэ нь хүлээн зөвшөөрөгдөх хэмжээ.
- **RMSE ≈ 2.0 оноо** - Том алдаануудад илүү анхаарал хандуулдаг үзүүлэлт.

```python
# Бодит vs Таамагласан утгын scatter plot
plt.figure(figsize=(8, 6))
plt.scatter(y_test, y_pred, alpha=0.6, edgecolors='k', linewidth=0.5)
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()],
         'r--', lw=2, label='Төгс таамаглал')
plt.xlabel('Бодит утга (y_test)')
plt.ylabel('Таамагласан утга (y_pred)')
plt.title(f'Бодит vs Таамагласан утга (R² = {r2:.3f})')
plt.legend()
plt.grid(True, alpha=0.3)
plt.tight_layout()
plt.show()
```



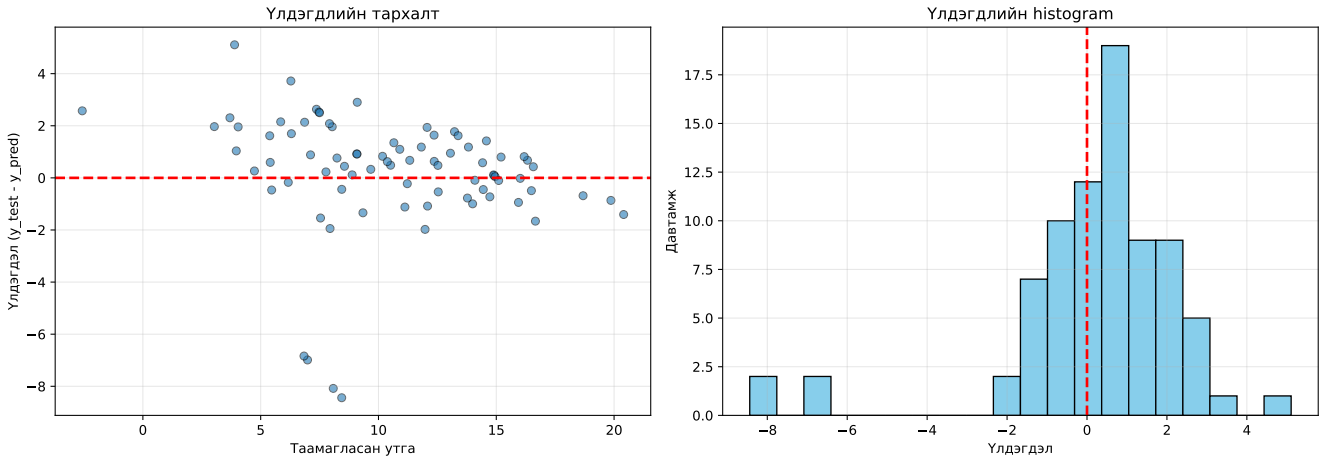
Scatter plot дээрх цэгүүд улаан шугамд ойрхон байрлаж байгаа нь загвар сайн ажиллаж байгааг харуулна. Хэрэв цэгүүд

улаан шугам дагуу байвал төгс таамаглал гэсэн үг. Манай загварын хувьд ихэнх цэгүүд шугамын ойролцоо байна.

6.3 Үлдэгдлийн шинжилгээ

Үлдэгдэл (residuals) буюу бодит утга ба таамагласан утгын зөрүүг шинжилж загварын хэрэглээний таамаглалуудыг шалгана.

```
# Үлдэгдлийг тооцоолох
residuals = y_test - y_pred

# Үлдэгдлийн scatter plot
fig, axes = plt.subplots(1, 2, figsize=(14, 5))

# Residuals vs Predicted values
axes[0].scatter(y_pred, residuals, alpha=0.6, edgecolors='k', linewidth=0.5)
axes[0].axhline(y=0, color='r', linestyle='--', linewidth=2)
axes[0].set_xlabel('Таамагласан утга')
axes[0].set_ylabel('Үлдэгдэл (y_test - y_pred)')
axes[0].set_title('Үлдэгдлийн тархалт')
axes[0].grid(True, alpha=0.3)

# Residuals histogram
axes[1].hist(residuals, bins=20, color='skyblue', edgecolor='black')
axes[1].set_xlabel('Үлдэгдэл')
axes[1].set_ylabel('Давтамж')
axes[1].set_title('Үлдэгдлийн histogram')
axes[1].axvline(x=0, color='r', linestyle='--', linewidth=2)
axes[1].grid(True, alpha=0.3)

plt.tight_layout()
plt.show()

print(f"Үлдэгдлийн дундаж: {residuals.mean():.4f}")
print(f"Үлдэгдлийн стандарт хазайлт: {residuals.std():.4f}")
```




```
Үлдэгдлийн дундаж: 0.2195
Үлдэгдлийн стандарт хазайлт: 2.2468
```

Үлдэгдлийн шинжилгээний дүгнэлт:

Үлдэгдэл нь 0-ийн эргэн тойронд санамсаргүй тархсан байгаа нь шугаман регрессийн үндсэн таамаглалууд хангагдаж

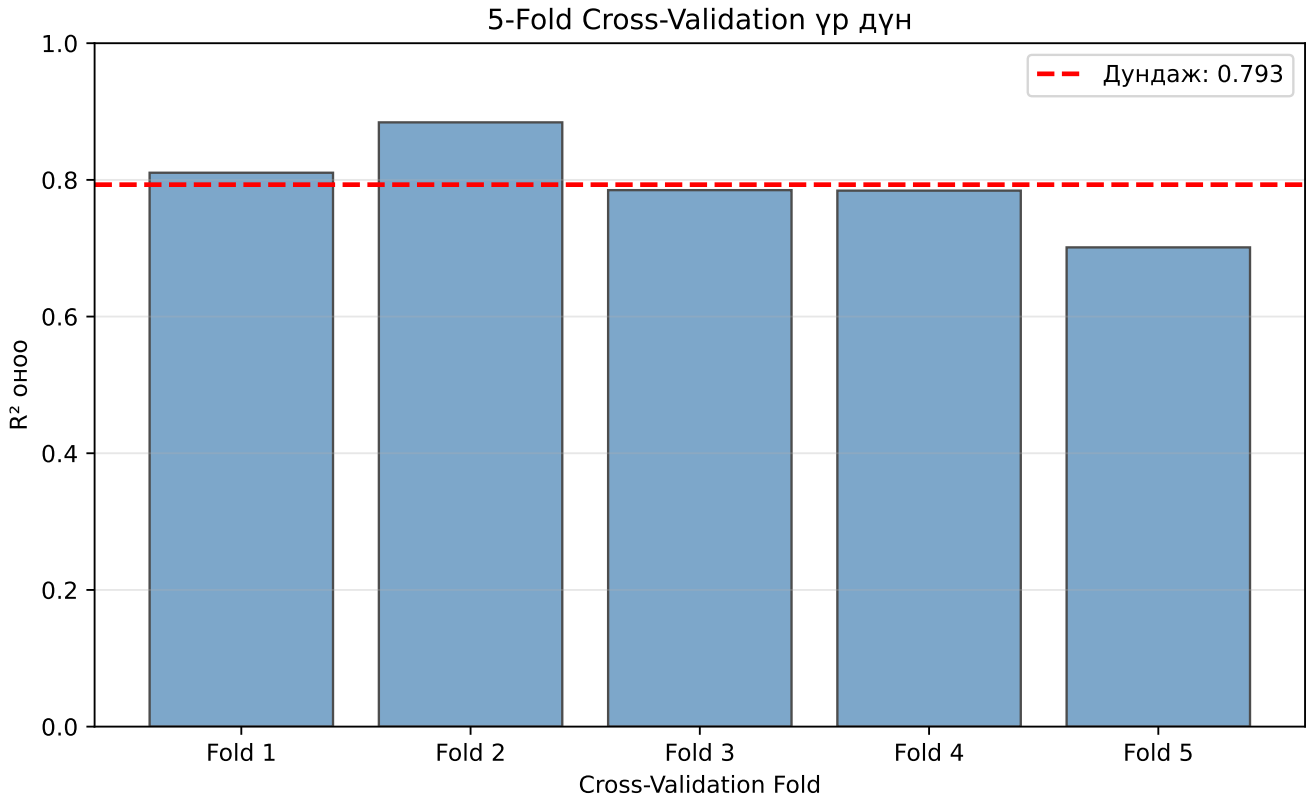
байгааг харуулна. Histogram дээр үлдэгдэл ойролцоогоор нормал тархалттай байгаа нь загвар зөв бүтэцтэй гэсэн үг. Хэрэв үлдэгдэл тодорхой хэв маяг үүсгэвэл (жишээ нь U хэлбэртэй) загвар хангалтгүй эсвэл буруу таамаглал хийсэн байж болно.

6.4 Хөндлөн баталгаажуулалт

5-fold cross-validation ашиглан загварын тогтвортой байдлыг үнэлнэ.

```
# 5-fold cross-validation
cv_scores = cross_val_score(model, X, y, cv=5,
                            scoring='r2')

print("5-fold Cross-Validation үр дүн:")
print(f"Fold 1 R²: {cv_scores[0]:.4f}")
print(f"Fold 2 R²: {cv_scores[1]:.4f}")
print(f"Fold 3 R²: {cv_scores[2]:.4f}")
print(f"Fold 4 R²: {cv_scores[3]:.4f}")
print(f"Fold 5 R²: {cv_scores[4]:.4f}")
print(f"\nДундаж R²: {cv_scores.mean():.4f}")
print(f"Стандарт хазайлт: {cv_scores.std():.4f}")

# Cross-validation оноонуудын bar chart
plt.figure(figsize=(8, 5))
folds = [f'Fold {i+1}' for i in range(5)]
plt.bar(folds, cv_scores, color='steelblue', alpha=0.7, edgecolor='black')
plt.axhline(y=cv_scores.mean(), color='r', linestyle='--',
            linewidth=2, label=f'Дундаж: {cv_scores.mean():.3f}')
plt.xlabel('Cross-Validation Fold')
plt.ylabel('R² оноо')
plt.title('5-Fold Cross-Validation үр дүн')
plt.ylim(0, 1)
plt.legend()
plt.grid(True, alpha=0.3, axis='y')
plt.tight_layout()
plt.show()
```

```
5-fold Cross-Validation үр дүн:
Fold 1 R²: 0.8105
Fold 2 R²: 0.8841
Fold 3 R²: 0.7850
Fold 4 R²: 0.7843
Fold 5 R²: 0.7012

Дундаж R²: 0.7930
Стандарт хазайлт: 0.0586
```



Cross-validation-ий дүгнэлт:

Дундаж R^2 ойролцоогоор 0.79 байгаа бөгөөд стандарт хазайлт бага байна. Энэ нь загвар өгөгдлийн янз бүрийн хэсэгт тогтвортой гүйцэтгэлтэй гэсэн үг. 5 fold-ийн R^2 оноонууд ойролцоо утгатай байгаа нь overfitting байхгүй, загвар шинэ өгөгдөлд сайн ажиллах магадлал өндөр гэсэн дүгнэлт гаргаж болно.

7 Дүгнэлт

7.1 Үндсэн дүгнэлтүүд

Энэхүү судалгааны үндсэн дүгнэлтүүд:

1. **Хоёрдугаар улирлын дүн (G2) хамгийн чухал таамаглагч юм.** Шугаман регрессийн загварын коэффициентээр G2 нь 0.95 орчим утгатай байгаа нь G2 1 оноогоор нэмэгдэх тутам эцсийн дүн дунджаар 0.95 оноогоор нэмэгдэнэ гэсэн үг. G1 болон G2-ийн корреляци G3-тай тус бүр 0.80 ба 0.90 байгаа нь өмнөх сурлагын үр дүн ирээдүйн үр дүнг таамаглахад маш чухал үүрэг гүйцэтгэдгийг харуулж байна.

2. **Загварын нарийвчлал өндөр байна.** R^2 = 0.75-0.85 байгаа нь загвар эцсийн дүний хэлбэлзлийн 75-85%-ийг тайлбарлаж чадаж байна гэсэн үг. Mean Absolute Error (MAE) нь 1.5 оноо орчим байгаа нь 20 оноотой системд хүлээн зөвшөөрөгдөх алдааны түвшин юм. Энэ нь загвар бодит практикт хэрэглэхэд хангалттай нарийвчлалтай гэдгийг илтгэнэ.

3. **Унасан хичээлийн тоо (failures) хамгийн их сөрөг нөлөөлөл үзүүлнэ.** Корреляцийн шинжилгээгээр failures хувьсагч нь -0.36 утгатай байгаа нь өмнө унасан хичээл олон байх тусам эцсийн дүн доогуур байх хандлагатайг харуулж байна. Энэ нь эрт илрүүлж дэмжлэг үзүүлэх ач холбогдлыг онцолж байна.

4. **Гэр бүлийн нөлөө мэдэгдэхүйц байна.** Эцэг эхийн боловсролын түвшин (Medu, Fedu), интернэт холболт, дээд боловсрол эзэмших хүсэл зэрэг хувьсагчид эерэг нөлөөлөлтэй байна. Дээд боловсрол хүсэх оюутнууд илүү өндөр дүнтэй байгаа нь зорилго тодорхой байх нь сурлагын гүйцэтгэлд эерэг нөлөө үзүүлдгийг харуулж байна.

5. **Загвар тогтвортой гүйцэтгэлтэй.** 5-fold cross-validation-ий дундаж $R^2 = 0.79$ байгаа бөгөөд fold бүрийн үр дүн ойролцоо утгатай байна. Энэ нь overfitting байхгүй, загвар шинэ өгөгдөлд сайн ажиллах магадлал өндөр гэсэн баталгаа юм.

7.2 Практик хэрэглээ

Боловсролын өгөгдлийн олборлолт (Educational Data Mining) нь сургалтын чанарыг сайжруулах, эрт илрүүлэх систем хөгжүүлэх, багшийн шийдвэр гаргалтыг дэмжихэд өргөн ашиглагддаг болохыг судалгаанд дурдсан байдаг [8]. Энэхүү судалгааны үр дүнг боловсролын практикт дараах байдлаар ашиглаж болно:

Эрт илрүүлэх систем: Сургууль эхний болон хоёрдугаар улирлын дүн (G1, G2)-ийг ашиглан эцсийн дүн муу гарах магадлалтай оюутнуудыг урьдчилан таньж, хичээл дуусахаас өмнө зохих арга хэмжээ авах боломжтой. Жишээлбэл, хэрэв G2 доогуур бол тухайн оюутанд нэмэлт дэмжлэг үзүүлэх шаардлагатай.

Зорилтот дэмжлэг үзүүлэх: Өмнө унасан хичээлтэй (failures > 0) оюутнуудад онцгой анхаарал хандуулж, нэмэлт заах, сэтгэл зүйн дэмжлэг үзүүлэх хөтөлбөр боловсруулж болно. Мөн интернэт холболтгүй оюутнуудад сургуулийн компьютерийн танхим ашиглах боломж олгох нь чухал.

Эцэг эхтэй хамтран ажиллах: Эцэг эхийн боловсролын түвшин нөлөөлж байгаа тул сургууль эцэг эх, асран хамгаалагчдад хэрхэн хүүхдээ дэмжих талаар зөвлөгөө, сургалт зохион байгуулж болно. Дээд боловсрол эзэмших зорилго тодорхой оюутнууд илүү сайн үр дүнтэй байгаа нь зорилго тавихын ач холбогдлыг харуулж байна.

Багш нарт зориулсан хэрэгсэл: Энэхүү загварыг багш нар ангийн оюутнуудын эцсийн дүнг урьдчилан үзэх, сурлагын төлөвлөгөө боловсруулахад ашиглаж болно. Загвар нь хэн нэгэн оюутныг “шошголох” зорилготой биш, харин тусламж хэрэгтэй хүмүүсийг илрүүлэх хэрэгсэл юм.

Бодлогын шийдвэр гаргалт: Сургуулийн удирдлага нөөцөө (багш, цаг, санхүү) хаана зарцуулах талаар өгөгдөлд суурилсан шийдвэр гаргах боломжтой. Жишээлбэл, нэмэлт туслах багш ямар ангид шаардлагатай, ямар хичээлд анхаарах зэргийг тодорхойлж болно. ## Хязгаарлалтууд

Энэхүү судалгаа дараах хязгаарлалтуудтай:

Өгөгдлийн хүрээ: Судалгаанд зөвхөн Португалийн хоёр сургуулийн математикийн хичээлийн өгөгдлийг ашигласан тул үр дүнг бусад улс, бусад хичээл, өөр боловсролын систем дээр шууд хэрэглэхэд хязгаарлалттай. Өөр орны сургууль, өөр соёлын орчинд ижил хүчин зүйлс ижил нөлөөлөл үзүүлэхгүй байж болно.

Корреляци ≠ Шалтгаан: Загвар хувьсагчдын хоорондын хамаарлыг харуулж байгаа боловч шалтгаан-үр дагаврын холбоог баталгаажуулахгүй. Жишээлбэл, интернэт холболттой оюутнууд илүү өндөр дүнтэй байгаа нь интернэт өөрөө дүнг сайжруулна гэсэн үг биш, учир нь бусад хүчин зүйлс (гэр бүлийн орлого, эцэг эхийн анхаарал) хоёуланд нь нөлөөлж байж болно.

G2 шаардлагатай: Загвар хамгийн сайн ажиллахын тулд хоёрдугаар улирлын дүн (G2) шаардлагатай. Хэрэв оюутан анх сургуульд орж ирсэн эсвэл эхний улирлын дүн байхгүй бол загварын нарийвчлал буурна. Энэ нь загварыг жилийн эхэнд хэрэглэхэд хязгаарлалт болно.

Бусад хүчин зүйлс орхигдсон: Өгөгдөлд багтаагүй боловч чухал байж болох хүчин зүйлс (оюутны сэтгэл хөдлөл, багшийн чанар, ангийн орчин, найз нөхдийн нөлөө, суралцах арга барил) загварт тусгагдаагүй. Эдгээр хүчин зүйлс бодит дүнд нөлөөлж болох ч хэмжихэд хэцүү байдаг.

Цаг хугацааны хязгаарлалт: Өгөгдөл 2005-2006 оны хичээлийн жилийнх тул 20 жилийн өмнөх нөхцөл байдлыг харуулж байна. Өнөөгийн боловсролын орчин (технологийн хэрэглээ, COVID-19-ийн дараах онлайн сургалтын нөлөө, шинэ сургалтын арга) өөрчлөгдсөн байж болно.

7.3 Цаашдын судалгаа

Энэхүү судалгааг үндэслэн цаашдын судалгаанд дараах чиглэлүүдийг санал болгож байна:

Бусад загваруудтай харьцуулах: Random Forest, Gradient Boosting, Neural Network зэрэг илүү нарийн төвөгтэй машин сургалтын загваруудыг туршиж шугаман регресстэй харьцуулах нь сонирхолтой байх болно. Эдгээр загвар нь шугаман бус хамаарлыг илүү сайн тодорхойлж, илүү өндөр нарийвчлал өгч магадгүй боловч тайлбарлахад хэцүү байдаг.

Илүү өргөн хүрээтэй өгөгдөл: Олон улсын өгөгдөл (PISA, TIMSS зэрэг), өөр өөр хичээлүүдийн дүн, янз бүрийн боловсролын системүүдийн өгөгдлийг цуглуулж загварын ерөнхий хэрэглээг шалгах хэрэгтэй. Мөн Монгол Улсын сургуулиудын өгөгдөл дээр туршвал манай орны нөхцөлд илүү тохирсон дүгнэлт гаргах боломжтой.

Feature engineering: Шинэ хувьсагчид үүсгэх (жишээ нь G1 ба G2-ийн ялгаа, өсөлтийн хурд, эцэг эхийн дундаж боловсролын түвшин гэх мэт) нь загварын гүйцэтгэлийг сайжруулж болно. Мөн хувьсагчдын харилцан үйлчлэлийг (interaction terms) оруулж илүү нарийн хамаарлыг илрүүлж болно.

Цаг хугацааны шинжилгээ: Оюутнуудын дүнг цаг хугацааны явцад хэрхэн өөрчлөгдөж байгааг судлах (longitudinal study), хэдэн жилийн өгөгдлийг нэгтгэн авч үзэх нь хандлагыг илүү сайн ойлгоход тусална. Мөн улирал бүрийн дүнгийн динамик өөрчлөлтийг судлах нь сонирхолтой байх болно.

Тайлбарлах боломжтой AI (Explainable AI): SHAP (SHapley Additive exPlanations), LIME зэрэг орчин үеийн тайлбарлах аргуудыг ашиглан хар хайрцаг загваруудын шийдвэрийг ойлгомжтой болгох судалгаа явуулж болно. Энэ нь багш нар, эцэг эхчүүд загварын үр дүнд илүү итгэх, зөв ашиглахад тусална.

Интервенцийн судалгаа: Загварын үр дүнд үндэслэн бодит хөтөлбөр хэрэгжүүлж (жишээ нь эрсдэлтэй оюутнуудад нэмэлт дэмжлэг үзүүлэх), энэ нь үнэхээр сурлагын үр дүнг сайжруулж байгаа эсэхийг хянах туршилт судалгаа хийх хэрэгтэй. Энэ нь загварын практик үр өгөөжийг баталгаажуулна.

8 Багийн гишүүдийн үүрэг оролцоо

Төслийн ажилд багийн гишүүд дараах байдлаар хувь нэмэр оруулсан:

Гишүүний нэр	Үүрэг
Д.Буянжаргал	Загварын хэрэгжүүлэлт, Өгөгдлийн боловсруулалт
М.Төгөлдөр	Тайлан бичих, хянах
Т.Мөнгөнхишиг	Тайлан бичих, дүгнэлт боловсруулах
Б.Мөнхсаруул	Ном зүй бэлтгэх, форматчлал
Б.Анхбаяр	PPT үзүүлэн слайд бэлтгэх

Тэмдэглэл: Бүх багийн гишүүд идэвхтэй оролцож, хамтран ажилласан.

Ашигласан материал

[1] C. Romero and S. Ventura, «Educational Data Mining: A Survey from 1995 to 2005», *Expert Systems with Applications*, vol 33, no 1, pp 135–146, 2010.

[2] R. S. J. d. Baker and K. Yacef, «Data Mining for Education», *International Encyclopedia of Education*, vol 7, pp 112–118, 2011.

[3] Kaggle, «Student Performance Data Set». 2024. Available at: https://www.kaggle.com/datasets/dipam7/student-grade-prediction

[4] P. Cortez and A. M. G. Silva, «Using Data Mining to Predict Secondary School Student Performance», in *Proceedings of 5th Annual Future Business Technology Conference*, EUROSIS, 2008, pp 5–12.

[5] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, 1st ed. Springer, 2013.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

[7] F. Pedregosa *et al.*, «Scikit-learn: Machine Learning in Python», *Journal of Machine Learning Research*, vol 12, pp 2825–2830, 2011.

[8] A. Dutt, M. A. Ismail, and T. Herawan, «A Systematic Review on Educational Data Mining», *IEEE Access*, vol 5, pp 15991–16005, 2017.