# Table of contents



# 1 Оршил

Энэ төслийн зорилго нь хувь хүний жилийн орлого **$50,000-аас дээш** ($> 50\mathbf{K}$) эсэхийг урьдчилан таамаглах логистик регрессийн загвар боловсруулах явдал юм. Үүний тулд бид Adult Income өгөгдлийн санг ашиглаж, хоёртын ангиллын асуудлыг дотооддоо хэрэгжүүлсэн градиент бууруулалтын аргаар шийдвэрлэсэн.

Манай загвар дараах онцлогуудтай:

- Гараас хэрэгжүүлсэн логистик регресс (градиент бууруулалт, L2 регуляризаци, сурах хурдны бууралт)
- Sklearn Pipeline ашигласан урьдчилсан боловсруулалт (стандартчилал, one-hot кодчилол)
- Ангиудын тэнцвэргүй байдлыг харгалзан `class_weight="balanced"` ашигласан
- Сургалт--баталгаажуулалтын тусдаа салгалт болон тогтвортой турших боломжтой бүтэц

Эдгээр нь загварыг илүү ойлгомжтой, давтагдахуйц, мөн тогтвортой болгодог. Дараагийн хэсэгт өгөгдөл, онцлогууд, болон сонгосон аргачлалуудыг дэлгэрэнгүй тайлбарлана.

---

# 2 Өгөгдөл

## 2.1 Өгөгдлийн эх үүсвэр ба зорилго

Adult Income Dataset нь АНУ-ын 1994 оны Хүн амын тооллогын түүвэр мэдээлэлд тулгуурладаг. Нас, боловсрол, мэргэжил, гэрлэлтийн байдал, хүйс гэх мэт нийгэм-эдийн засгийн үзүүлэлтүүд багтдаг.

Зорилтот хувьсагч:

- `income_>50K`
  - $0 \rightarrow \leq 50\text{K}$
  - $1 \rightarrow > 50\text{K}$

Нийт ойролцоогоор 44,000 мөртэй. Үүнээс 80% нь сургалтад (35,165), 20% нь баталгаажуулалтад (8,792) ашиглагдсан.

## 2.2 Ангиудын тэнцвэргүй байдал

Өгөгдлийн гол сорилт бол зорилтот ангиудын тэнцвэргүй байдал юм.

- $\leq 50\text{K} \rightarrow$ **76%**
- $> 50\text{K} \rightarrow$ **24%**

Иймээс:

- Энгийн ``бүгдийг $\leq 50\text{K}$ гэж таамагладаг'' загвар хүртэл ≈**76%accuracy** үзүүлж чадна
- Гэхдээ энэ нь хэрэгцээгүй тул **Recall**, **Precision**, **F1-score** илүү чухал үзүүлэлтүүд болдог

Тиймээс бид зөвхөн accuracy бус, ангиудын тэнцвэртэй үзүүлэлтийг үндсэн үнэлгээнд ашигласан.

## 2.3 Онцлогийн сонголт

Өгөгдөл анх олон тооны баганатай бөгөөд зарим нь хоорондоо их давхцдаг.
Загварыг удирдахад амархан байлгах үүднээс онцлогуудыг дараах байдлаар сонгов.

### 2.3.1 Ашигласан онцлогууд (9)

**Тоон (5):** `age`, `educational-num`, `capital-gain`, `capital-loss`, `hours-per-week`

**Категори (4):** `education`, `marital-status`, `occupation`, `gender`

### 2.3.2 Хассан онцлогууд

`workclass`, `fnlwgt`, `relationship`, `race`, `native-country` гэх мэт ---
Эдгээрийг хассан гол шалтгаанууд:

- зарим нь давхацдаг (`education` vs `educational-num`)
- загварын гүйцэтгэлд мэдэгдэхүйц хувь нэмэр оруулахгүй
- шаардлагагүй шуугиан үүсгэх хандлагатай

Зорилго нь **тэнцвэртэй, ойлгомжтой, илүүдэлгүй** онцлогийн багц гаргах явдал байв.

## 2.4 Өгөгдөл хуваах стратеги

Бид өгөгдлийг стратифик хуваалт ашиглан **80/20** харьцаагаар салгасан.

Үр дүн:

- Ангиудын харьцаа ижил хадгалагдсан
- Туршилт бүрт тогтвортой үр дүн гардаг
- Pipeline-д урьдчилсан боловсруулалт болон сургалт логик дараалаар холбогдох боломжтой

## 2.5 Урьдчилсан боловсруулалт

### 2.5.1 Тоон онцлогууд

`StandardScaler` ашиглаж:

- дундаж: $\mu = 0$
- стандарт хазайлт: $\sigma = 1$

Ингэснээр градиент бууруулалтын тогтвортой байдал сайжирдаг.

### 2.5.2 Категори онцлогууд

`OneHotEncoder(handle_unknown='ignore')` ашигласан. Жишээ нь:
`education=Bachelors` → `education_Bachelors=1`, бусад 0.

**Классын жин:**
`class_weight='balanced'` нь цөөн ангийг илүү чухалчилж, L2 loss-д илүү жин өгдөг.

**Гиперпараметрүүдийн ерөнхий утгууд:**

- learning_rate ≈ 0.1
- max_iter ≈ 800
- reg_lambda = 1e-4
- lr_decay = 1e-4
- threshold = 0.5

Эдгээр нь тогтвортой, илүү тэнцвэртэй сургалтыг хангадаг.

---

## 3 Ном зүй / References