САНХҮҮ ЭДИЙН ЗАСГИЙН ИХ СУРГУУЛЬ
ЭКОНОМИКСИЙН ТЭНХИМ

БОРЖИГИН
Дарханбаатарын ЭРДЭНЭТӨГӨЛДӨР

ТӨГСӨГЧДИЙН ХӨДӨЛМӨР ЭРХЛЭЛТИЙГ ӨГӨГДӨЛ ОЛБОРЛОЛТ ДАХЬ АНГИЛЛЫН АЛГОРИТМД ҮНДЭСЛЭН ТААМАГЛАХ НЬ

Мэргэжлийн индекс
031101

Эдийн засгийн ухааны бакалаврын
зэрэг горилсон дипломын төсөл

Удирдсан
П.Гантөмөр /Ph.D/

Улаанбаатар. 2023

САНХҮҮ ЭДИЙН ЗАСГИЙН ИХ СУРГУУЛЬ
ЭКОНОМИКСИЙН ТЭНХИМ

БОРЖИГИН
Дарханбаатарын ЭРДЭНЭТӨГӨЛДӨР

ТӨГСӨГЧДИЙН ХӨДӨЛМӨР ЭРХЛЭЛТИЙГ ӨГӨГДӨЛ ОЛБОРЛОЛТ ДАХЬ АНГИЛЛЫН АЛГОРИТМД ҮНДЭСЛЭН ТААМАГЛАХ НЬ

Мэргэжлийн индекс
031101

Эдийн засгийн ухааны бакалаврын
зэрэг горилсон дипломын төсөл

Удирдагч: . . . . . П.Гантөмөр /Ph.D/

Шүүмжлэгч: . . . . . Д.Мөнхцэцэг /MA/

Улаанбаатар. 2023

ГАРЧГИЙН ТОВЬЁОГ

ХҮСНЭГТЭН МЭДЭЭЛЛИЙН ЖАГСААЛТ

ЗУРГАН МЭДЭЭЛЛИЙН ЖАГСААЛТ

ХАВСРАЛТУУДЫН ЖАГСААЛТ

ТОВЧИЛСОН ҮГС, НЭР ТОМЪЁОНЫ ТАЙЛБАР

ҮСХ	Үндэсний Статистикийн Хороо
СЭЗИС	Санхүү Эдийн Засгийн Их Сургууль
МУИС	Монгол Улсын Их Сургууль
ХНХСИ	Хөдөлмөр Нийгмийн Хамгааллын Судалгааны Институц
БУС	Байгалийн ухааны сургууль
НУС	Нийгмийн ухааны сургууль
ХУС	Хүмүүнлэгийн ухааны сургууль
БС	Бизнесийн сургууль
ОУХНУС	Олон улсын харилцааа, нийгмийн ухааны сургууль
ХЗС.	Хууль зүйн сургууль
ХШУИС.	Хэрэглээний шинжлэх ухаан, инженерчлэлийн сургууль
ЗС	Завхан сургууль
ЭС	Эрдэнэт сургууль

ТОВЧ ХУРААНГУЙ

Энэхүү дипломын ажлаар орчин үед тренд болоод буй өгөгдөл олборлолт буюу Data Mining-ийн аргуудаар төгсөгчдийн хөдөлмөр эрхлэлтийн байдлыг урьдчилан таамаглах ангиллын загвар үүсгэж, хамгийн өндөр нарийвчлал бүхий аргыг тодорхойлох билээ. Мөн түүнчлэн, төгсөгчдийн ажилгүйдлийн шалтгаан болон хөдөлмөрийн зах зээлд нийцэхүйц мэргэжилтнийг бэлдэхэд шаардлагатай хүчин зүйлсийг өгөгдөл олборлолтын аргаар тодорхойлох юм. Судалгааны үндсэн арга зүйд Нэйви Бэйес, К-Хамгийн ойрын хөрш, Шийдвэрийн мод зэрэг ангиллын алгоритм байх ба загвараа баталгаажуулахын тулд МУИС-ийн 2018-2019 оны төгсөгчдийн мэдээллийг ашиглан гүйцэтгэсэн болно. Төгсөгчдийн ерөнхий мэдээллийг агуулсан 12 хувьсагчдаар хөдөлмөр эрхлэлтийг урьдчилан таамаглах ангиллын загварыг бий болгохдоо суралцах болон үнэлгээний гэсэн хоёр үе шатаар ангиллын загварыг бий болгосон эцэст нь гүйцэтгэлийн үнэн зөв байдал (accuracy), алдаа (error), нарийвчлал (precision), санах ой (recall) зэрэг янз бүрийн параметрүүд дээр үндэслэн үнэлсэн билээ. Судалгааны үр дүнд, сурлагын голч дүн ямар нэг байдлаар ажил хөдөлмөр эрхлэхэд нөлөө үзүүлдэг ба хүчин зүйл хамаарлын шинжилгээгээр 0.9 буюу өндөр хамааралтай байв. Харин голч оноо бага байх тусам ажил эрхлэлтийн хувь хэмжээ буурахад нөлөө үзүүлдэг гэсэн хамаарал ажиглагдсан. Мөн тухайн төгсөгчийн мэргэжил ямар газар ажиллахтай эерэг хамааралтай буюу 0.6 байна. Үндсэн арга зүйн хэрэгжүүлэлтийн үр дүнг харвал Нэйви Бэйес 91.76%, К-Хамгийн ойрын хөрш 98.64%, Шийдвэрийн мод алгоритмын ангиллын загвар 99.61% буюу хамгийн өндөр нарийвчлалтай байна. Тиймээс Шийдвэрийн модны алгоритм нь төгсөгчдийн хөдөлмөр эрхлэлтийн байдлыг урьдчилан таамаглахад илүү тохиромжтой гэсэн үр дүн гарсан юм.

Эдийн засгийн бүтээлийн сэтгүүлийн ангиллын индекс: C52, C55

Түлхүүр үгс: Өгөгдөл олборлолт, К-Хамгийн ойрын хөрш, Шийдвэрийн мод, Нэйви Бэйес, Ангиллын алгоритм

ОРШИЛ

Өгөгдөл олборлолт нь асар их хэмжээний өгөгдөлд хурдан дүн шинжилгээ хийх чадвартай тул орчин үед бүх төрлийн салбарт хэрэглэгдсээр байна. Тухайлбал, өгөгдөл олборлолт нь ихэвчлэн бизнест оновчтой стратеги сонгох болон алдагдсан боломжийн зардлыг бууруулах зэргээр ашиглагдаж байсан ч өнөө үед судлаачид, төрийн байгууллага гэх мэт өргөн хүрээнд ашиглагдаж үнэ цэнэтэй хөрөнгөд тооцогддог болжээ (Investopedia.com). Энэхүү судалгааны ажлын хүрээнд боловсролын салбарын өгөгдөл олборлолт, түүний ач холбогдолын талаар авч үзэх ба энэ нь сургалтын чанар болон оюутны чанарын гүйцэтгэлийг сайжруулах мөн хөдөлмөрийн зах зээлд нийцэхүйц төгсөгч бэлдэхэд чухал ач холбогдолтой юм. Валеро, Ван Ренен нар 2018 онд их дээд сургуулийн эдийн засагт үзүүлэх нөлөөг судалж үр дүнд нь нэг хүнд ногдох ДНБ болон их дээд сургуулиудын хооронд хүчтэй бөгөөд эерэг хамаарлыг олж тогтоосон ба тухайн бүс нутагт их дээд сургуулийн тоо 10%-иар нэмэгдсэн нь нэг хүнд ногдох ДНБ-ий 0.4%-иар өссөнтэй холбоотой гэж тайлбарласан байв (Valero, 2018). Энэ нь боловсролын эдийн засагт үзүүлэх нөлөө нь эдийн засагч болон улс төрчдийн гол сэдэв болохыг илтгэнэ. Учир нь боловсрол нь улс орны бүтээмжийг нэмэгдүүлэх төдийгүй төгсөгчийг сайн чанарын ажлын байртай болох боломжийг нэмэгдүүлэх хүчин зүйлсийн гол цөм нь бөгөөд ажиллах хүчний боловсрол, ур чадварын түвшин өндөр байх нь ажилгүйдлээс хамгаалах, эмзэг хөдөлмөр эрхлэлт буюу түр зуурын ажил эрхлэлтээс тодорхой хэмжээгээр хамгаалдгийг тогтоосон байна (Sparreboom 2014). Нөгөө талаас, ажил эрхлэлт нь ядуу өрхүүдийг орлоготой болгож, бараа үйлчилгээний дотоод эрэлтийг өсгөж, нийт өсөлтийг идэвхжүүлдэг. Тиймдээ ч аливаа улс орны хамгийн чухал асуудлуудын нэг нь хөдөлмөр эрхлэлтийг хэрхэн нэмэгдүүлэх, хэнд ямар бодлого хэрэгжүүлэх вэ гэдгийг шийдвэрлэх явдал юм. Түүнчлэн ихэнх эмпирик судалгаа болон эдийн засагчдын онолд улс орны бүтээмжийг дээшлүүлэх, хүмүүсийн ажил эрхлэлтийг нэмэгдүүлэх замаар амьдралын чанарт нөлөөлөх гол хүчин зүйл бол боловсрол хэмээн нотолсон байна.

Нэмж дурдахад боловсрол, мэдлэг, ур чадвар, хүмүүн капитал нь капитал, хөдөлмөрөөс гадна үйлдвэрлэлийн гол хүчин зүйлсийн нэг. Үүнээс үзвэл их сургууль төгсөгчид буюу шинээр хөдөлмөрийн зах зээлд орогсдын талаар судлах нь дараа үеийн төгсөгчдийн чадварыг сайжруулах, хөдөлмөр эрхлэлт, тэдний чадварыг нэмэгдүүлэх боломж бүрдэнэ. Монгол Улсын хувьд их, дээд сургууль төгсөгчдийн тоо 2021 оны байдлаар нийт 58343 хүн буйгаас 79% ажилтай, 1.5% ажилгүй, 20% нь ажиллах хүчнээс гадуур байгаа нь дийлэнх нь ажил хайх итгэлээ алдсан буйг илтгэнэ. Энэ нь манай улсын нийт төгсөгчдийн дөрвөн хүн тутмын нэг нь их сургуулиа төгссөн хэдий ч ажилгүй байгааг харуулж байна (БСШУЯ). Одоогийн байдлаар оюутны хөдөлмөр эрхлэлт, ахиц дэвшилд дүн шинжилгээ хийх, хянах тогтолцоо манай улсад байхгүй (YCX) ба үүний шалтгаан нь шинжилгээ хийх арга зүйн хувьд дутмаг, орчин үеийн өгөгдөл олборлох арга ашигладаггүйтэй холбоотой юм. Хэдийгээр, манай улс арга зүйн хувьд дутмаг хэдий ч тогтвортой хөгжлийн зорилтын хүрээнд төрөөс хөдөлмөр эрхлэлтийн талаар баримтлах бодлогын 1.3.6-д “... ажилгүйдлийн түвшин хөдөлмөрийн насны залуучуудын дунд ихсэж байгаа тул их дээд сургууль, коллеж, мэргэжлийн боловсрол, сургалтын

байгууллагыг төгсөгчдөд чиглэсэн хөдөлмөр эрхлэлтийн бодлогыг хэрэгжүүлэх шаардлагатай байна" хэмээн тусгасан ажээ. Энэ нь цаашид боловсролын салбар сайжрах, орчин үеийн оновчтой арга, арга зүйг ашиглах үндэс тавигдсан гэж харж байна. Учир нь оюутнуудын хөдөлмөр эрхлэлтийн хандлагыг урьдчилан таамаглах үр дүнтэй аргыг олох нь маш чухал бөгөөд өгөгдөл олборлолт, машин сургалтын аргуудыг судалгааны ажилд ашиглах нь орчин үеийн судлаачдад өндөр ач холбогдолтой болно. Дээрх судалгааны асуудлын хүрээнд дараах асуултуудыг дэвшүүллээ. Үүнд:

- Асуулт 1: Их, дээд сургууль төгсөгчид ажлын байртай болоход гол нөлөөлдөг хүчин зүйлс юу вэ?
- Асуулт 2: Оюутны голч төгсөгчдийг ажилд ороход нөлөөлдөг үү?

Энэхүү судалгааны зорилго нь өгөгдөл олборлолт дахь ангиллын алгоритм ашиглан төгсөгчдийн хөдөлмөр эрхлэлтийн загварыг үнэлэх ба Нэйви Бэйес, К-Хамгийн ойрын хөрш, Шийдвэрийн мод зэрэг хэд хэдэн аргыг харьцуулах билээ. Үүсгэсэн загварыг баталгаажуулахын тулд МУИС-ийг 2018-2019 онд төгссөн төгсөгчдийн мэдээллийг ашиглан гүйцэтгэсэн ба уг загварыг салбар сургууль тус бүрээр төгсөгчдийг ажилтай эсвэл ажилгүй байгаа эсэхийг таамаглахад ашигласан. Өнөөгийн хөдөлмөрийн зах зээлийн өөрчлөлтийн үйл явцтай уялдуулан төгсөгчдийн хөдөлмөр эрхлэлтийн төлөв байдлыг гаргах, ажилгүйдлийн шалтгаан, нөхцөлийг тодорхойлох, их сургуулиас бэлтгэгдэж буй мэргэжилтний ур чадвар хөдөлмөрийн зах зээлд нийцэж буй эсэх, төгсөгчдийн хөдөлмөр эрхлэлтэд тулгамдаж буй хүндрэл бэрхшээл зэргийг тодорхойлоход оршино. Их дээд, сургуулийн төгсөгчдийн тоо жил бүр нэмэгдсээр байх тул төгсөгчид хөдөлмөрийн зах зээлд амжилттай гархын тулд илүү их өрсөлдөөнтэй тулгарсаар байна. Энэ нь боловсролын байгууллагууд төгсөгчдөө хөдөлмөрийн зах зээлд нэвтрэх хангалттай ур чадвараар хангахад нь туслах шаардлага тулгарж буйг илтгэнэ. Уг судалгааны ажлын хүрээнд өгөгдөл олборлолт дахь ангиллын алгоритмын онол аргазүйг судалж, эзэмшин алгоритм хоорондын ялгарах онцлог давуу талуудыг онолын болон эмпирик судалгаанд тулгуурлан тодорхойлох зорилтыг дэвшүүлсэн болно. Дээрх асуудлын хүрээнд дараах таамаглалыг урдчилан тавьж байна. Үүнд:

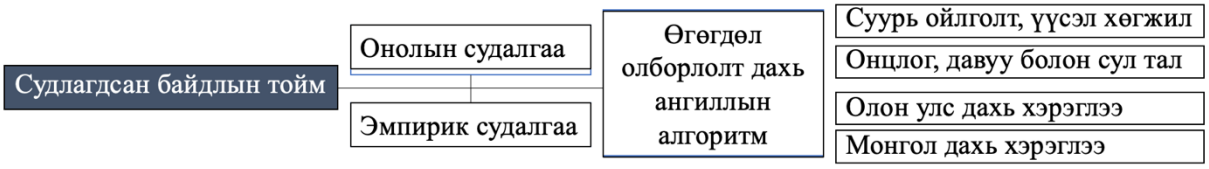
Таамаглал: *Шийдвэрийн модны алгоритм нь хүлээн зөвшөөрөгдөх түвшинд оюутнуудын хөдөлмөр эрхлэлтийг урьдчилан таамаглахад илүү тохиромжтой.*

Шийдвэрийн мод алгоритм нь чанарын болон тоон өгөгдөлд ангилал үүсгэх боломжтой тул бусад өгөгдөл олборлолтын алгоритм болон шугаман регрессийн аргаас давуу талтай билээ. Мөн өгөгдөл дахь хоосон утга (Null) шийдвэрийн модонд нөлөөлөхгүй бөгөөд ихэнх эмпирик судалгаанд хамгийн сайн нарийвчлалтай ангилагч хэмээн дүгнэсэн байдаг. Үүний нэг жишээ нь Малайз улсын судлаач Сапаат нар нь 2017 онд, Ван Нурул судлаач нь 2019 онд энэ төрлийн судалгаанд Шийдвэрийн мод, Нэйви Бэйес, К-Хамгийн ойрын хөрш зэрэг алгоритмыг ашигласан ба уг алгоритмыг илүү нарийвчлалтай таамагладаг гэж нотолсон байна (Sapaat, 2017

I БҮЛЭГ. СУДЛАГДСАН БАЙДАЛ

Энэхүү бүлэгт өгөгдөл олборлолтыг ашиглан оюутны хөдөлмөр эрхлэлтийг урьдчилан таамаглах талаарх эмпирик судалгаа, мөн эдгээр ойлголтуудад хамааралтай онолын судалгаануудыг хийнэ. Ном зүйн тойм хийх ерөнхий төлөвлөгөөг Зураг I-1 дээр харуулж байна (дэлгэрэнгүйг Хавсралт 1-ээс харна уу).

Зураг II.1 Судлагдсан байдал хийх оюуны зураглал



Эх сурвалж: Судлаачийн зураглал

Судлагдсан байдлын судалгаагаар өгөгдөл олборлолт, түүний ангиллын алгоритм болоод төгссөн оюутны хөдөлмөр эрхлэлт, амжилтад нөлөөлөгч хүчин зүйлс зэрэг агуулгын хүрээнд онолын болон эмпирик судалгааны мэдээллийг цуглуулсан билээ.

1.1 Онолын судлагдсан байдал

1.1.1 Өгөгдөл олборлолтын суурь ойлголт, үүсэл хөгжил

Өгөгдөл олборлолтыг “Data mining”, зарим тохиолдолд “Knowledge Discovery from Databases” гэж нэрлэдэг ба гол зорилго нь өгөгдлөөс мэдээлэл гарган авах түүнийгээ цаашид ашиглах боломжтой мэдлэг болгон хувиргах явдал юм. Уг шинжлэх ухааны гол ойлголтуудыг дараах Хүснэгт I.1-т харуулж байна. Эдийн засагт газар, капитал, хөдөлмөр нь үндсэн орц болдог бол сүүлийн жилүүдэд өгөгдөл нь чухал орцуудын нэг болжээ. Учир нь өгөгдлийг ашиглан шаардлагатай мэдлэгийг бий болгосноор илүү үр дүнтэй стратеги боловсруулах, зардлыг багасгаж ашгийг нэмэгдүүлэх зэрэг давуу талтай юм (Yan.C, 2019).

Хүснэгт II.1 Өгөгдөл олборлолтын тухай

Англи нэршил	Монгол нэршил	Тайлбар	Жишээ
Data	Өгөгдөл	Хадгалагдсан баримтууд, тоо, текст	метадата
Information	Мэдээлэл	Өгөгдлийн зүй тогтол, хоорондын хамаарал	ямар бүтээгдэхүүн, хэзээ зарагдаж байна
Knowledge	Мэдлэг	Мэдээллийг ашиглан ирээдүйн талаар мэдэх	ямар бараанд урамшуулал үзүүлэхээ тодорхойлох

Эх сурвалж: Н.Баттүшиг, 2017

Хүн төрөлхтөн өгөгдөл, мэдээллийг боловсруулах тодорхой үе шатуудыг алхан өнөө үед хүрч ирсэн ажээ. Өгөгдөл олборлолт нь 2000 оноос эрчимтэй хөгжиж эхэлсэн хэдий ч тэртээх 1960 онд өгөгдлийг цуглуулах, өнгөрсөнд хандсан тогтмол өгөгдөл дээр ажиллаж байсан нь Зураг 1.2-оос харагдаж байна. Ийнхүү бид томоохон өгөгдлийн сангуудаас урьд нь мэдэгдэж байгаагүй тодорхой зүй тогтол бүхий мэдээллийг ялгах (patterns),

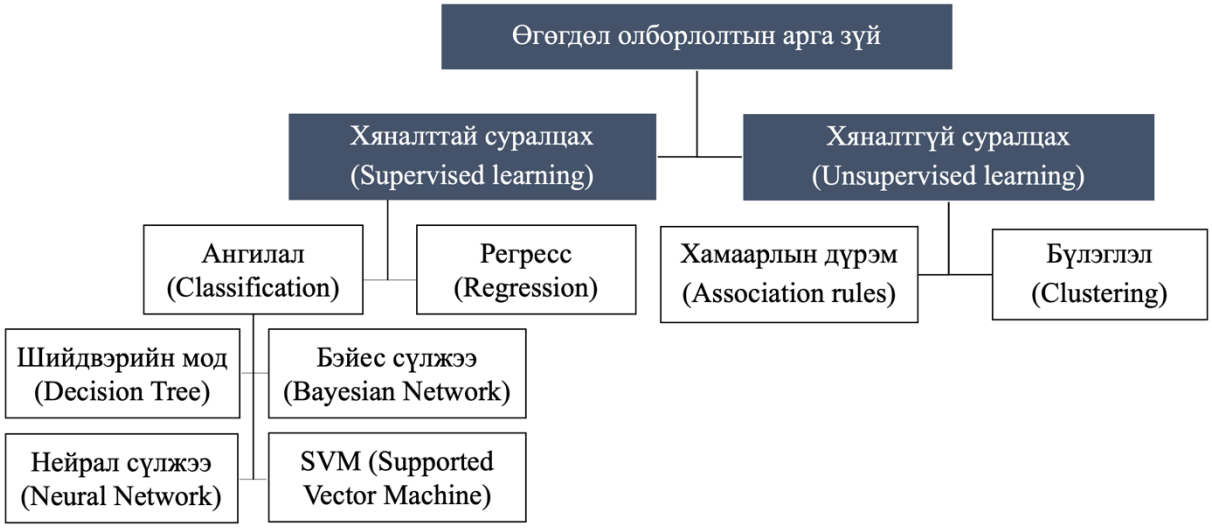
ирээдүйг таамаглах (predict) боломжийг олсон юм. Мөн өгөгдөл олборлолт нь шийдвэр гаргалтад шаардлагатай өгөгдлийн хэв маягийг илрүүлэх үйл ажиллагааг хэлдэг аж (Han.J, 2011).

Зураг II.2 Өгөгдөл олборлолтын түүхэн үе шат

Эх сурвалж: (Paidi.A, 2012)

Нөгөө талаас, өгөгдөл олборлолт нь өгөгдлийн сан, хиймэл оюун ухаан, машин сургалт, статистик гэх мэт салбаруудыг багтаасан олон талт салбар ухаан (Lemos, 2003) төдийгүй мэдээллийн технологийн хувьслын нэгээхэн үр дүн билээ. Уг ойлголтын арга зүйн төрлүүдийг Зураг I.3-аас харж болно. Өгөгдөл олборлолт нь хяналттай суралцах, хяналтгүй суралцах гэсэн 2 төрөлтэй ба гол ялгаа нь өгөгдсөн даалгаврыг хэрхэн гүйцэтгэхийг гаднаас заана эсвэл хиймэл оюун ухаан өөрөө сурах явдал юм.

Зураг II.3 Өгөгдөл олборлолтын төрлүүд



Эх сурвалж: (Laura Garach, 2014)

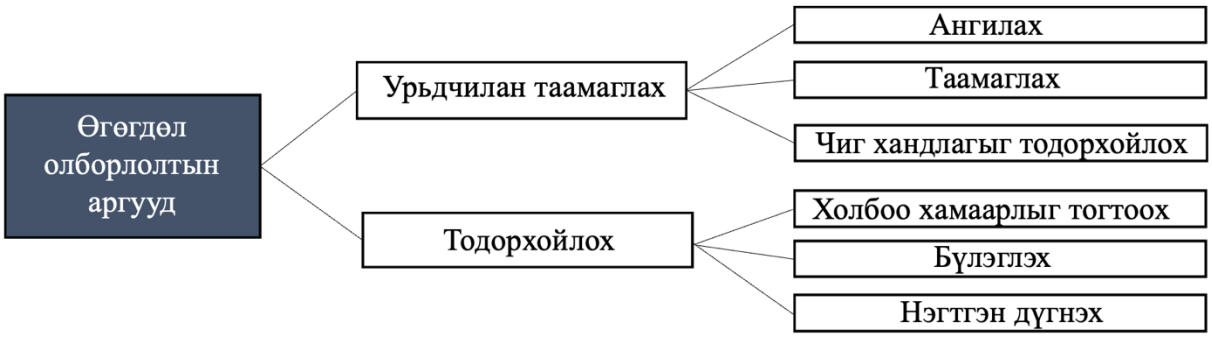
Мэдлэгийг их хэмжээний өгөгдлөөс олборлохын тулд тодорхой аргууд буюу машин сургалтын төрөл бүрийн алгоритмуудыг ашигладаг. Өгөгдөл олборлолтын алгоритм гэдэг нь өгөгдлөөс олборлолтын загвар үүсгэх бүлэг тооцооллуудыг нэрлэдэг байна. Түгээмэл хэрэглэгддэг үндсэн алгоритмуудын онцлог шинж чанарыг дурдвал:

- **Ангилал:** Энэ алгоритм нь хяналттай суралцах төрөлд багтдаг буюу төгсгөлөг тооны утгуудаас таамагладаг (Элдэв-Очир.Б, 2019). Өгөгдлийг ялгаатай ангиудад хуваахад ашиглагддаг. Түгээмэл хэрэглэгддэг алгоритмууд нь Нэйви Бэйес, Шийдвэрийн мод зэрэг алгоритмууд ордог.

- **Регресс:** Уг алгоритм нь таамаглалд бий болсон алдаануудыг тооцоолон илүү боловсронгуй болгодог бөгөөд хувьсагч хоорондын харилцааг загварчлахад чиглэгддэг.
- **Хамаарлын дүрэм:** Өгөгдөлд орших хувьсагч хоорондын ажиглагдсан холбоо хамаарлыг тайлбарлах дүрмийг бий болгодог. Өөрөөр хэлбэл өгөгдлийн дунд байх ялгаатай шинж чанаруудын хоорондын хамаарлыг хайж олно.
- **Бүлэглэл:** Энэ алгоритм нь ихэнх судалгаанд “Кластер” хэмээн нэршсэн байдаг. Төсөөтэй шинж чанар бүхий объектуудыг бүлэглэх ба регресстэй мөн адил асуудлыг тодорхойлдгоороо онцлогтой.

Эдгээрээс үзвэл өгөгдөл олборлолтын алгоритмууд нь судлаачид дунд таамаглал хийх, түүнд дүн шинжилгээ хийх чиглэлд илүү өргөн хэрэглэгддэг болох нь харагдаж буй бөгөөд хамгийн түгээмэл хэрэглэгддэг нь бүлэглэл болон ангиллын алгоритмууд аж. Бүлэглэлийн алгоритмууд нь нийт өгөгдөлд байгаа бүхий л зүй тогтлуудыг шинжилдэг бол ангилалтын алгоритмууд нь зөвхөн урьдчилан тодорхойлсон хамааран хувьсагчдад нөлөөлөх зүй тогтлуудыг шинжилдэг байна (Wu.X, 2008). Мөн эдгээр аргуудыг урьдчилан таамаглах болон тодорхойлох гэсэн хоёр хэсэгт хувааж болох ажээ (Б. Энхтуул, 2020). Үүнийг Зураг I.4-т дүрслэн харууллаа.

Зураг II.4 Өгөгдөл олборлолтын аргууд



Эх сурвалж: (Б.Энхтуул, 2020)

1.1.2 Шийдвэрийн мод ангилагч алгоритм /Decision Tree/

Шийдвэрийн мод нь өгөгдөлд орших элементүүдийн бодит үнэ цэнэ дээр суурилсан шийдвэр гаргалтын загварыг бий болгодог индукцийн арга юм. Уг алгоритмыг 1963 онд Висконсин Мэдисоны их сургуулийн Статистикийн тэнхим зохиож байжээ. Ийнхүү 1966 онд Познан технологийн их сургууль хүний оюун ухааныг судалсан нь анхны судалгаанд ашиглагдсан тохиолдол байв. Ийнхүү шийдвэрийн модны алгоритмыг 1974 оноос эхлэн статистикийн ухааны профессорууд болох Лео Брейман, Чарльз Стоун, Стэнфордын Жером Фридман, Ричард Олшен нар ангилал ба регрессийн мод CART-ийг боловсруулснаар шийдвэрийн мод алгоритмыг илүү боловсронгуй болгосон юм. Гэвч шинэ үзэл баримтлал гарч ирсэн буюу 1986 онд Жон Росс Куинлан нь CART болон бусад алгоритмууд нь асуулт бүрд зөвхөн хоёрхон хариултай тул алгоритмын бүтцийг сайжруулах ёстой гэж үзсэний үндсэн дээр C4.5 алгоритм бий болж зарим зангилааг нэмж өгснөөр өгөгдөл олборлолтын шилдэг 10 алгоритмын нэг болсон ажээ (Springer LNCS, 2008). Энэхүү алгоритмын гол санаа нь шийдвэрийн эрсдэл болон үр ашгийг

үнэлэх боломжийг олгох ба шийдвэр гаргалтаас үүдэлтэй учирч болох эрсдэлийг бууруулах зорилготой байдаг байна.

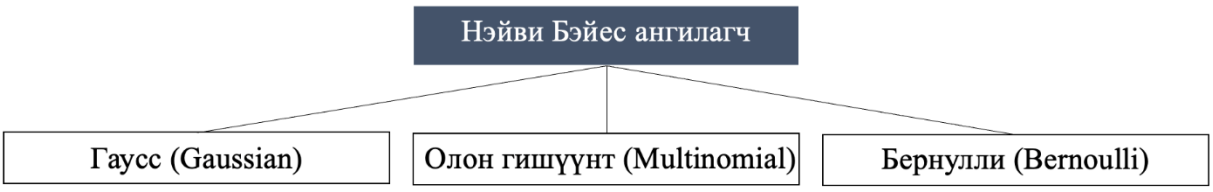
- **Шийдвэрийн мод алгоритмын онцлог ба давуу тал**

Шийдвэрийн мод нь төрөл бүрийн үйл ажиллагааны чиглэл болон шийдвэр гаргалтад туслах хамгийн шилдэг алгоритм юм. Уг алгоритмын гол давуу тал нь үр дүн нь ойлгомжтойгоос гадна тайлбарлахад хялбар байдаг байна. Мөн урьдчилан өгөгдлийг бэлтгэх болон өгөгдлийг стандартчилах шаардлагагүй. Харин сул тал нь хэт их хэмжээний өгөгдлийн олонлог нь шийдвэрийн модыг хэт том болгож эцсийн шийдвэр гарах үед алдаа гарах боломжийг нэмэгдүүлдэг ажээ.

1.1.3 Нэйви Бэйес ангилагч алгоритм /Naïve Bayes/

Нэйви Бэйесийн алгоритм нь Бэйесийн теорем дээр суурилсан нөхцөлт магадлалт ангилагч юм. Уг алгоритм нь анх XVIII зууны II-р хагаст бий болсон ба бином тархалт дахь магадлалын параметрийн тархалтыг хэрхэн тооцоолох талаар судалж байсан Томас Бэйесийн нэрээр нэрлэгджээ. Мөн түүнчлэн статистик болон компьютерын шинжлэх ухааны ном, зохиолд Нэйви Бэйесийн загваруудыг энгийн Бэйес (Simple Bayes), бие даасан Бэйес (Independence Bayes) гэх мэт янз бүрээр нэрлэсэн байдаг. Нэйви Бэйес ангилагч нь маш өндөр чадлын цар хүрээтэй (highly scalable) учраас шугаман параметрүүдийг шаарддаг, чанарын хувьсагчтай илүү сайн ажилладаг байна. Мөн уг алгоритм нь бие даасан хувьсагчид статистикийн хувьд ч бие даасан байдаг хэмээн тайлбарладаг ба өгөгдлийн хэмжээ их байх тусам Нэйви Бэйес алгоритм тохиромжтой байдаг. Ижил тооны янз бүрийн тархалт дээр суурилан Нэйви Бэйес алгоритмыг Гаусс, Олон гишүүнт, Бернулли хэмээн ангилах ба үүнийг Зураг I.5-д дүрслэн харууллаа.

Зураг II.5 Нэйви Бэйес ангиллын алгоритмын төрлүүд



Эх сурвалж: (Б.Энхтуул, 2020)

- **Нэйви Бэйес алгоритмын урьдач нөхцөл**

Хувьсагчид нь бие даасан (Independent) эсвэл ижил (Equal) байна гэж үздэг байна. Энэ нь ямар нэг хос шинж чанараас хамаарахгүй буюу энгийнээр машины өнгө нь түүний төрөлд хамааралгүй гэж үздэг. Харин ижил гэдэг нь хувьсагчид үр дүнд ижил нөлөө үзүүлнэ гэх санааг илэрхийлдэг.

- **Нэйви Бэйес ангиллын алгоритмын онцлог ба давуу тал**

Энэхүү алгоритм нь магадлалыг тооцоолоход хялбар болон давталт хийх шаардлагагүй байдаг давуу талтай билээ. Өөрөөр хэлбэл хэрэгжүүлэхэд хялбар бөгөөд хамгийн хурдан ангиллыг таамагладаг байна. Харин сул тал нь өгөгдлийн олонлог дахь чанарын хувьсагч нь ажиглагдаагүй утгатай бол уг загвар нь тэг магадлал хэмээн үзэж таамаглах боломжгүй болдог ба үүнийг тэг давтамж гэж нэрлэдэг.

1.1.4 К-Хамгийн ойрын хөрш ангилагч алгоритм /K-Nearest Neighbor/

К-Хамгийн ойрын хөрш алгоритм нь 1951 онд Эвелин Фикс болон Жозеф Ходжес нарын хөгжүүлсэн параметрийн бус ангилагч алгоритм юм. Тус алгоритмыг ангилал эсвэл регрессийн аль алинд ашиглах боломжтой ба параметрийн бус гэдэг нь үндсэн өгөгдөл болон түүний тархалтын талаар ямар ч таамаглал гаргахгүй гэсэн үг билээ. 1960 оноос тус алгоритм нь хамгийн түгээмэл хэрэглэгддэг статистик аргуудын нэг болсон ажээ. Мөн К-Хамгийн ойрын хөрш алгоритмыг залхуу суралцагчид буюу “Lazy learner” гэж нэрлэдэг. Учир нь хяналттай суралцах бүлгийн алгоритм болхийнхоо хувьд боломжит бүх тохиолдлуудыг хадгалж, ижил төстэй байдлыг зайны функц дээр үндэслэн шинэ тохиолдлуудыг ангилдаг байна. К-Хамгийн ойрын хөрш нь машин сургалтын алгоритмуудын нэг бөгөөд эрүүл мэндийн салбараас эхлээд эдийн засаг, санхүүгийн салбаруудад өргөнөөр ашиглаж байна.

- **К-Хамгийн ойрын хөрш алгоритмын онцлог ба давуу тал**

К-Хамгийн ойрын хөрш алгоритмын ижил төстэй байдлыг хэмжигч зайны функцэд Евклидийн зай, Манхэттэн зай, Минковвскийн зай юм. Эдгээр гурван зайны функц нь зөвхөн тасралтгүй хувьсагчдад ашигладаг ба чанарын хувьсагчтай үед Хаммингийн зайг ашиглана. Зайны функцийг өгөгдлийн олонлогоос шууд тооцоолох нэг томоохон дутагдал нь хувьсагчид өөр өөр хэмжүүртэй эсвэл тоон болон чанарын хувьсагчдын холимог байх явдал юм (Sallay, 2012). Жишээлбэл, хэрэв нэг хувьсагч нь жилийн орлогод тулгуурласан төгрөгөөр, нөгөө нь насаар тооцсон бол орлого нь тооцоолсон зайд илүү их нөлөөлнө. Тиймээс стандартчилал ашиглан үүнийг шийддэг байна.

1.2 Өгөгдөл олборлолтын ангиллын алгоритм ашигласан эмпирик ажлууд

Өгөгдөл олборлолтыг дэлхий дахинаа санхүү, эдийн засаг, бизнес болон инженерчлэл, анагаах ухаан, аж үйлдвэр, хөдөө аж ахуй зэрэг салбарт өргөнөөр ашигладаг бөгөөд ирээдүйн нөхцөл байдлыг урьдчилан таамаглахын тулд боловсролын салбарт ашиглаж байна. Энэхүү хэсэгт гадаад, дотоодын судлаачдын хийсэн эмпирик ажлуудыг танилцуулах болно.

1.2.1 Олон улсад хийгдсэн эмпирик судалгаа

“Ангиллын алгоритм ашиглан оюутны гүйцэтгэлийг таамаглах” тухай судалгааг Дорина Кабекчиева (2012) нь Болгар улсын их сургуулийн 2007-2009 онуудад элсэн орсон 10,330 оюутнаас хүйс, төрсөн он, газар, төгссөн ахлах сургууль гэх мэтээр 20 хувьсагч бүхий асуулга авч өгөгдлийн олонлогоо бүрдүүлжээ. Тус судалгааны ажлын зорилго нь их сургуулийн удирдлагад өгөгдөл олборлолтын хэрэглээний ач холбогдлыг таниулах, их дээд сургуулийн элсэлтийн компанит ажлыг үр дүнтэй явуулахад хувь нэмэр оруулах, хамгийн өндөр дүнтэй оюутнуудыг сургуульдаа татах боломжийг олгох байлаа. Судалгааны арга зүйн хувьд OneR-ийн дүрэм, К-Хамгийн ойрын хөрш, Шийдвэрийн мод, Нэйви Бэйес, Нейрал сүлжээ зэрэг ангиллын алгоритмыг ашиглан гүйцэтгэсэн. Үр дүнд нь эдгээр алгоритмын таамаглах чадвар 52%-67% хооронд хэлбэлзсэн байна. Тиймээс оюутнуудын гүйцэтгэлийг таамаглахад тохиромжгүй байх магадлалтай гэж үзжээ. Илүү нарийвчлалтай үр дүнд хүрэхийн тулд өгөгдлийн багцыг өөрчлөх, ангиллын

алгоритмын хувьсагчийг тохируулах гэх мэт их сургуулийн өгөгдөл олборлох төслийг хэрэгжүүлэх цаашдын чиглэлийг тодорхойлох хэрэгтэй гэсэн байв. (Dorina Kabakchieva, 2012)

Өгөгдөл олборлолт нь их хэмжээний өгөгдлийг хурдан шинжлэх чадвартай тул төрөл бүрийн салбарт хэрэглэгдсээр буй билээ. Бангсук Жантаван (2013) нь "Төгсөгчдийн ажлын байрыг урьдчилан таамаглах ангиллын загвар бий болгох нь" судалгаанд ангиллын алгоритмыг ашиглан төгсөгчдийн хөдөлмөр эрхлэлтийн загварыг бий болгох мөн Нэйви Бэйесийн арга, Шийдвэрийн модны арга зэрэг өгөгдөл олборлох хэд хэдэн аргыг харьцуулах зорилготой юм. Уг судлаач нь алгоритмын үүсгэсэн ангиллын загвараа баталгаажуулахын тулд Тайландын Маежо их сургууль төгсөөд 12 сар болсон төгсөгчдийн мэдээлэлд үндэслэн уг төгсөгч нь ажилтай, ажилгүй байсан эсвэл тодорхойгүй нөхцөл байдалд буй эсэхийг таамаглажээ. Үр дүнд нь их, дээд сургуулиуд төгсөгчдөө хөдөлмөрийн зах зээлд нэвтрэхэд хангалттай ур чадвараар хангах боломжийг бүрдүүлэхэд Нэйви Бэйесийн арга нь хамгийн өндөр нарийвчлалтай буюу 99.77% хүрсэн бол Шийдвэрийн модны арга нь 98.31% хүрчээ. (Bangsuk Jantawan, 2013)

Дээрх судлаачтай ижил зорилгоор Аззиати Бинти (2016) судлаач нь "Шинээр төгсөгчдийн ажил эрхлэлтийг өгөгдөл олборлолтын хяналттай болон хяналтгүй суралцах аргууд ашиглан таамаглах нь" сэдэвтэйгээр бүтээжээ. Уг судалгаанд К-Хамгийн ойрын хөрш, Нэйви Бэйес, Шийдвэрийн мод зэрэг хэд хэдэн алгоритм ашигласан ба улсын сургууль төгсөөд 6 сар болж буй төгсөгчдийн ажил эрхлэлтийг байдлыг таамаглах хамгийн сайн загварыг олох нь гол зорилго байв. Судалгааны үр дүнд төгсөгчид ажилд орох, цаашид суралцах, ур чадвараа дээшлүүлэх, ажилд орох гэж буй болон ажилгүй байгаа эсэхээс үл хамааран нөлөөлж буй шинж чанаруудыг тодорхойлж хамгийн өндөр нарийвчлалтай загварыг К-Хамгийн ойрын хөрш алгоритм 97.78% гаргажээ. Энэхүү судалгаанаас харвал уг алгоритм нь тус судалгааны ажилд хамгийн хүлээлт үүсгэсэн арга болоод байна. (Azziati Binti, 2016)

Мөн энэ төрлийн судалгааг Тажул Мифта (2016) судлаач Нэйви Бэйес, Логистик регресс, К-Хамгийн ойрын хөрш, Шийдвэрийн мод зэрэг алгоритмыг ашиглан гүйцэтгэжээ. Түүний судалгааны зорилго нь оюутан сургуулиа төгссөний дараа хувийн эсвэл төрийн албанд орох, ажилгүй байх, үргэлжлүүлэн сурах эсэхийг таамаглах байв. Хүн ам зүйн хүчин зүйл болон гадаад, дотоод шинж чанаруудыг ашигласан ба үр дүнд нь Шийдвэрийн модны алгоритм хамгийн өндөр нарийвчлалтай байсан юм. (Tajul Mifta, 2016)

Ажилгүйдэл нь дэлхий даяар тулгамдсан асуудал болоод буй бөгөөд энэ төрлийн судалгаа сүүлийн 10 гаруй жилүүдэд эрчимтэй өсжээ. Тиймээс их, дээд сургууль төгсөгчдийн хөдөлмөрийн зах зээлийг судлах нь ажилгүйдлийн асуудалд онцлон анхаарах чухал элементүүдийн нэг юм. "Төгсөгчдийн ажил эрхлэлтийг ангиллын алгоритм ашиглан таамаглах нь" нэртэй судалгааг Залинда Отман (2017) бүтээсэн ба ажилд ороход голч дүнгээс гадна өөр ямар төрлийн шинж чанар нөлөөлж буй эсэхийг судлахыг зорьжээ. Ингэхдээ, Шийдвэрийн мод, SVM, J48 зэрэг алгоритмуудыг ашигласан бөгөөд Малайзын мэргэжлийн коллежийн 633 оюутнаар өгөгдлийн олонлогоо бүрдүүлсэн байв. Үр дүнд нь Шийдвэрийн мод алгоритм нь 66.48% нарийвчлалтай гарч

нас, мэргэжлийн дадлага, гэр бүлийн байдал, англи хэлний мэдлэг зэрэг шинж чанар нь ажилд ороход нөлөөлж буйг нотолсон байна. Энэхүү судалгаа нь Малайзын дээд боловсролын байгууллагуудад төгсөгчдөө хөдөлмөрийн зах зээлд орохоос өмнө шаардлагатай ур чадварт бэлтгэхэд туслах хэмжээний ач холбогдолтой судалгаа юм. (Zalinda Othman, 2017)

Сургуулиа төгсөөд шууд ажлын байртай болох нь оюутнуудын гол зорилго билээ. Тиймээс энэ төрлийн судалгааг Малайз улсын их дээд сургууль төгсөгчдийн ажилд орох тухай таамаглалыг Ван Нурул (2019) хийсэн ба нь хөдөлмөр эрхлэлтийн байдлыг ажил эрхлэлтийн хүчин зүйл дээр үндэслэн урьдчилан таамаглажээ. Өгөгдлөө 2015-2018 оны нийт 375507 төгсөгчдийн мэдээллийг цуглуулсан ба янз бүрийн чанарын хувьсагчдыг задлахын тулд хяналттай болон хяналтгүй суралцах алгоритмыг ашиглан шинжилгээ хийжээ. Үр дүнд нь Шийдвэрийн мод алгоритм нь хамгийн өндөр нарийвчлалтай байсан ба К-Кластер алгоритм нь оюутны сэтгэл ханамжийг тодорхойлох 7 бүлэг үүсгэсэн бөгөөд хамгийн алдартай бүлэглэл нь оюутнууд ур чадвартаа сэтгэл хангалуун байсан ч сургалтын хөтөлбөртөө сэтгэл хангалуун бус төгсөгчдөөс бүрдсэн байна. Эдгээр өгөгдөл олборлолтын аргууд нь төгсөгчдийн ажил эрхлэлтийн байдлыг хянах, ажлын байртай болж чадах хүчин зүйлсийг тодорхойлох боломжийг бүрдүүлнэ хэмээн үзжээ. (Wan Nurul, 2019)

Их өгөгдлийг сүүлийн жилүүдэд эрчимтэй олборлож шаардлагатай мэдлэгийг бий болгох нь дэлхий дахинд тренд болоод буй. Юуанг Ванг (2022) энэхүү их өгөгдлийг ашиглан мэргэжил сонголтыг урьдчилан таамаглах судалгааг хийжээ. Уг судалгаанд коллеж төгссөн 18000 оюутны мэдээлэлд үндэслэн Машин сургалт түүний дотор Шийдвэрийн мод, Санамсаргүй ой (Random Forest), XGB (Extreme Gradient Boost) зэрэг аргуудыг хэрэгжүүлсэн ба бие даасан шинж чанаруудын ач холбогдлыг шинжлэхийн тулд Шаплигийн нэмэлт тайлбарыг (SHAP) ашигласан. Үр дүнд нь XGB нь 89.1% нарийвчлалтай байсан тул мэргэжил сонголтыг сайн таамаглана гэж үзжээ. Мөн коллеж төгссөний дараа ажилд орох, ажилгүй байх, мэргэжил дээшлүүлэх зэрэг хоорондын харилцан үйлчлэлийг судалсны үндсэнд голч дүн нь харьцангуй их нөлөө үзүүлдэг бөгөөд дээд боловсролын байгууллагуудын үйл ажиллагаа төлөвлөлт, хэрэгжүүлэхэд туслах ач холбогдол бүхий судалгаа болжээ. (Yuang Wang, 2022)

1.2.2 Монгол Улсад хийгдсэн эмпирик судалгаа

Оюутны хөдөлмөр эрхлэлт сэдвийн хүрээнд судлаач М.Сэлэнгэ, Д.Энхзул нар 2018 онд оюутны хөдөлмөр эрхлэлтийн шалтгаан, үр нөлөө болон тэдгээрт нөлөөлж буй хүчин зүйлсийг судласан ба онцлох давуу тал нь цалин хэрэглээнд хүртээмжтэй байдаг эсэх, хамгийн олдоц ихтэй ажлын байр зэргийг тодорхойлжээ. Судалгааны үр дүнд, 80% нь хэрэглээгээ хангах зорилгоор хөдөлмөр эрхэлдэг ба өрхийн дундаж зардал 1% өсөхөд оюутны хөдөлмөр эрхлэлт 2.06%-аар өсж буй үр дүн Логистик регрессийн шинжилгээгээр гарчээ. (М.Сэлэнгэ, Д.Энхзул, 2018)

Ажилгүйдэл, ядуурал зэрэг нь өнөөгийн Монгол Улсын маань чухал асуудлуудын нэг болоод буй. Хөдөлмөр нийгмийн хамгааллын яамнаас төгсөгчдийн хөдөлмөр эрхлэлтийн судалгааны 2016 оны үр дүнгээс харвал улсын хэмжээнд нийт төгсөгчдийн 61.6% нь

ажил эрхэлж буй бөгөөд ажил хайж байгаа төгсөгчид ажлын байрны сонголт хийхдээ өсөн хөгжих боломж болон ажил мэргэжлийн ирээдүй, цалин, хөдөлмөрийн нөхцөл, аюулгүй ажиллагааг илүү чухалчилж авч үздэг байна. Ажил хайх явцад тулгардаг гол бэрхшээлүүд нь ажлын байрны болон ажлын байрны мэдээллийн хомсдол, ажил олгогч танил тал хардаг зэрэг асуудал байна. ҮСХ-ийн 2019 онд гаргасан судалгаанд МСҮТ төгсөгчдийн 62%, Их, дээд сургууль төгсөгчдийн 76% нь мэргэжлээрээ ажиллаж буй статистик мэдээг гаргажээ. Мөн онд МУИС өөрийн сургуулийн төгсөгчдийн мэдээлэлд үндэслэн тоон болон чанарын арга зүйг ашиглан тандалтын судалгаа хийсэн ба үр дүнд нь тухайн төгсөгч ажилд ороход хамгийн их хувийг хувийн зан чанар, төгссөн сургууль, ерөнхий ур чадвар нь хамгийн их хамааралтай байв.

1.2.3 Төгсөгчдийн ажил эрхлэлтэд нөлөөлөх хүчин зүйлсийн судалгаа

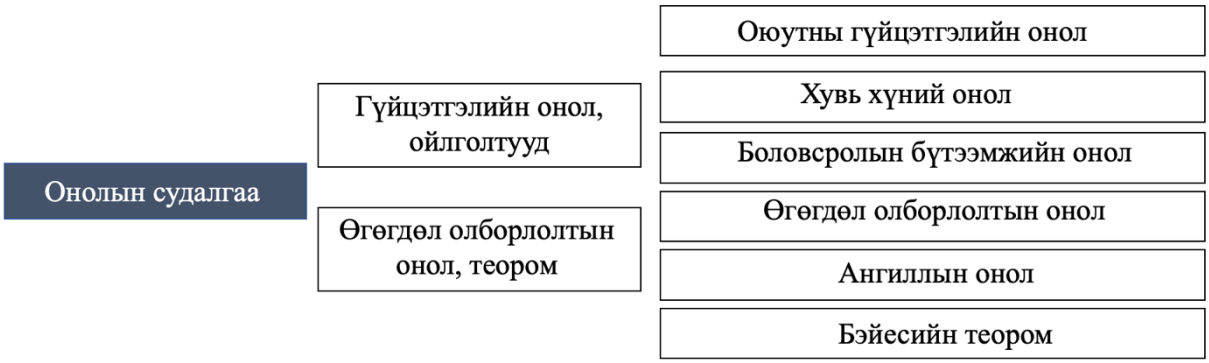
Оюутны сурлага болон цаашдын амьдралд хүртэх амжилтын гол суурь нь их сургуульд эзэмшсэн боловсрол билээ. Тиймээс Нигери улсын судлаач Олэдебину Току (2018) нь үүнд нөлөөлөгч хүчин зүйлсийн талаарх судалгааг хийжээ. Энэхүү судалгаанд тус улсын 480 оюутныг санамсаргүй байдлаар сонгон нийгэм-эдийн засгийн онцлог, гэр бүлийн гарал үүсэл, багшийн ур чадвар, сургуулийн орчин зэрэг хүчин зүйлсүүд дээр үнэлгээ хийсний үндсэнд эцэг, эхийн гарал үүсэл, сургуулийн орчин, багшийн ур чадвар хамгийн их нөлөөлдгийг тодорхойлсон юм. Уг судлаачийн санал болгож буй зөвлөмж нь сургуульд санхүүжилт авах, сургалтын орчныг сайжруулах гэсэн байв. (Oladebinu Toku, 2018)

Сурлагын амжилт, голч дүн нь тухайн төгсөгчийг ажилд авах гэж буй хүний хардаг чухал элементүүдийн нэг ажээ. Степен Карл (2021) оюутнуудын гүйцэтгэлийг сайжруулах зорилгоор түүнд нөлөөлөгч хүчин зүйлсийг судалсан байна. Уг судлаач нь тоон судалгааны арга ашигласан ба онлайнаар судалгааны өгөгдлийг цуглуулжээ. Энэхүү судалгааны үр дүнд сургуулийн сургалтын орчин нь тухайн оюутны сурлагад хамгийн өндөр нөлөөлдөг хэмээн дүгнэжээ. Мөн сургуулийн дотоод болон гадаад хүчин зүйлүүд болох эцэг эхийн хүмүүжлийн хэв маяг, оюутны онцлог, интернэт хэрэглээний түвшин, багшийн ур чадвар зэрэг нь нөлөөлсөн байна. (Stephen Karl, 2021)

II БҮЛЭГ. ОНОЛЫН УХАГДАХУУН БА ЗАГВАР

Энэхүү хэсэгт дипломын ажлын суурилж буй онол ба загварын талаар авч үзэх болно. Тухайн оюутан төгсөөд шууд ажлын байртай болж буй нь түүний сурлагын чанар буюу гүйцэтгэлтэй холбоотой юм. Тиймээс оюутны академик гүйцэтгэл, хувь хүний онолууд болон өгөгдөл олборлолттой холбоотой онолуудыг судалсан билээ. Онолын ухагдахуун хийх ерөнхий төлөвлөгөөг Зураг II.1 дээр харуулж байна.

Зураг II.1 Онолын судалгаа хийх оюун зураглал



Эх сурвалж: (Судлаачийн зураглал)

2.1 Оюутны гүйцэтгэлийн онол

2.1.1 Хувь хүний онол (Фридман, Розенман)

Фридман, Розенман нарын хувь хүний онол нь 1974 онд гарсан ба хувь хүний хэв шинжийг нийт дөрвөн төрөлд хуваан авч үзжээ. Уг онол нь хувь хүний тодорхой ангиллуудыг судалж, дараа нь тухайн ангилалд үндэслэн хувь хүмүүсийг бүлэглэдэг байна. Хувь хүний хэв шинжийн төрлүүд нь дараах байдлаар байна. Үүнд:

- А төрлийн зан чанар- Энэ төрлийн хүмүүс нь маш их урам зоригтой бөгөөд цаг хугацаа дуусч буй мэт үргэлж яарч байдаг аж. А хэлбэрийн зан чанартай хүмүүс нь өрсөлдөх чадвартай, яаруу, тэвчээргүй, дайсагнасан, түрэмгий, сэтгэл түгшсэн төлөвд байдаг бөгөөд энэ төрлийн хүмүүс нь цусны даралт ихсэх, зүрх судасны өвчин тусах магадлалтай байна.
- В төрлийн зан чанар- Уг төрөлд багтах хүмүүс нь А төрлийн зан чанарын яг эсрэг хүмүүс буюу тайван, тэвчээртэй байдаг аж. В хэлбэрийн зан чанартай хүмүүс ямар нэг ажлыг хойшлуулах хандлагатай байдаг бөгөөд тайван, бүтээлч тэвчээртэй хүмүүс байдаг.
- С төрлийн зан чанар- Ийм зан чанартай хүмүүс нарийн ширийн зүйлд дуртай, аливаа зүйлс хэрхэн явагддагийг мэдэхийн тулд цаг заваа зарцуулдаг. Тэд эвтэй, бусдад халамжтай, тэвчээртэй хүмүүс байдаг байна.
- D төрлийн зан чанар- Энэ төрлийн зан чанартай хүмүүс нь гутранги үзэлтэй, уур хилэн ихтэй, сөрөг хандлагатай байдаг.

Фридман, Розенман нарын хувь хүний зан чанарын онол нь сонирхол татахуйц боловч хэтэрхий энгийн онол юм. Энэ онолын талаар маш их шүүмжлэл байдаг бөгөөд шүүмжлэлийн нэг чухал шалтгаан нь хүний зан үйл нь маш нарийн төвөгтэй бөгөөд цаг хугацаа, нөхцөл байдлаас хамааран харилцан адилгүй байдаг аж (Razia, 2022).

2.1.2 Уолбергсийн боловсролын бүтээмжийн онол

Уолбергсийн боловсролын бүтээмжийн онол нь 1981 онд гарсан бөгөөд 3000 гаруй судалгааг нэгтгэн дүгнэж, эмпирик байдлаар шалгагдсан цөөхөн онолуудын нэг юм (Diperna, 2002). Энэхүү онолын гол зорилго нь тухайн оюутны сургалтын амжилт, гүйцэтгэлд ямар хүчин зүйлс нөлөөлж буйг тодорхойлох аж. Ингэхдээ, 100 гаруй судалгаанд дүн шинжилгээ хийж, боловсрол судлаачид дунд санал асуулга явуулж сурлагын амжилтанд нөлөөлөгч 28 ангиллыг тодорхойлсон байна. Хамгийн өндөр нөлөө бүхий хувьсагчид нь 11 төрөл ба нийгэм болон сэтгэл хөдлөл зэргийг багтаасан ангийн удирдлага, эцэг эхийн дэмжлэг, сурагч багшийн харилцаа, сургуулийн соёл, ангийн уур амьсгал зэрэг оржээ. Харин хамгийн бага нөлөө бүхий хувьсагчдад байршил, сургуулийн бодлого, сургалтын төлөвлөгөө зэрэг орсон байна. Мөн түүнчлэн, оюутны онцлог шинж чанар буюу өмнөх амжилт, хандлага зэрэг нь шууд бус нөлөө үзүүлсэн дүгнэлтэд хүржээ.

2.1.3 Академик гүйцэтгэлийн онол

Оюутны академик гүйцэтгэлийн онол нь Уолбергсийн боловсролын бүтээмжийн онолоос эхлэлтэй боловч Фридман, Росеман нарын хувь хүний зан чанарын онолоос үүдэлтэй ажээ (Sophie, 2011). Энэхүү академик гүйцэтгэлийн онолыг судлаач Елгер нь 2007 онд боловсруулсан байна. Уг онолоор тухайн оюутны гүйцэтгэл болон гүйцэтгэлийн сайжруулалтыг тайлбарлахад ашиглах тогтолцоог бүрдүүлэх зорилготой бөгөөд зургаан үндсэн ойлголтыг онцолж өгсөн байна. Гүйцэтгэлийн өнөөгийн түвшин нь нөхцөл байдал, мэдлэгийн түвшин, ур чадварын түвшин, хувийн шинж чанар, хувийн хүчин зүйл, тогтмол хүчин зүйл гэсэн зургаан бүрэлдэхүүн хэсгээс бүхэлдээ хамаардаг гэж үзжээ. Гүйцэтгэлийг үр дүнтэй сайжруулахын тулд гурван аксиомыг санал болгожээ. Үүнд бүтээлч сэтгэлгээ, өөрийгөө хөгжүүлэх хүрээлэл бий болгох, сурсан зүйлсээ практикт ашиглаж дадлага хийх аж.

2.2 Өгөгдөл олборлолтын онол, теорем

2.2.1 Өгөгдөл олборлолтын онол

Өгөгдөл олборлолт нь машин сургалт, статистик болон мэдээллийн сангийн системийн огтлолцол дээрх аргуудыг агуулсан том өгөгдлийн багц дахь хэв маягийг задлах, илрүүлэх үйл явц бүхий компьютерийн шинжлэх ухаан, статистикийн салбар дундын салбар билээ. Энэхүү онолд дараах ойлголтууд багтна. Үүнд:

- Өгөгдлийн хэлбэр

Технологийн дэвшил нь их өгөгдлийг цуглуулах боломжийг олгосноороо асар их хэмжээний өгөгдөл хүн төрөлхтөнд бий болж байна. Энэхүү их өгөгдөл нь бүтэцлэгдсэн, хагас бүтэцлэгдсэн болон бүтэцлэгдээгүй гэсэн хэлбэртэй байна. Эдгээр хэлбэрийн

өгөгдлүүдийн гол ялгааг Зураг II.2-оос харж болно. Өөрөөр хэлбэл тодорхой загварын дагуу үүссэн эсвэл урьдчилан зохион байгуулагдаагүй, өгөгдлийн сан удирдах системээр үүссэн эсэхээс шалтгаалж хэлбэрээ тогтооно.

Зураг II.2 Өгөгдлийн хэлбэрүүд

Бүтэцлэгдсэн өгөгдөл

- Тодорхой загварын дагуу үүссэн, тогтмол шинжүүдтэй учир боловсруулахад хялбар

Хагас бүтэцлэгдсэн өгөгдөл

- Тодорхой хэмжээгээр зохион байгуулагдсан боловч өгөгдлийн сан удирдах системээр үүсгэгдээгүй

Бүтэцлэгдээгүй өгөгдөл

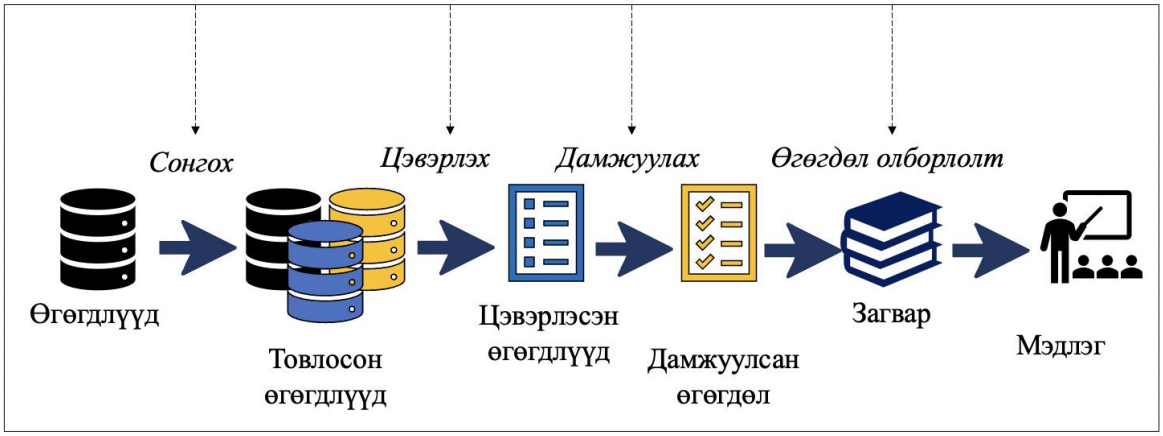
- Урьдчилан зохион байгуулагдаагүй, өгөгдлийн сан удирдах системээр үүсгэгдээгүй

Эх сурвалж: (Б.Энхтуул, 2020)

- Өгөгдөл олборлолтын үйл явц

Өгөгдөл олборлолтын гол зорилго нь өгөгдлөөс мэдээлэл гарган авах түүнийгээ цаашид ашиглах боломжтой мэдлэг болгон хувиргах билээ. Ийнхүү мэдлэг бий болгохын тулд 6 үе шатыг туулна. Үүнд: Бэлтэсэн өгөгдлүүдээс сонгож, товлосон өгөгдөл бий болгох, өгөгдлөө цэвэрлэсний дараа дамжуулж, загварыг бий болгосноор мэдлэг үүснэ. Дараах зурагт өгөгдөл олборлолтын үйл явцыг харуулж байна.

Зураг II.3 Өгөгдөл олборлолтын үйл явц



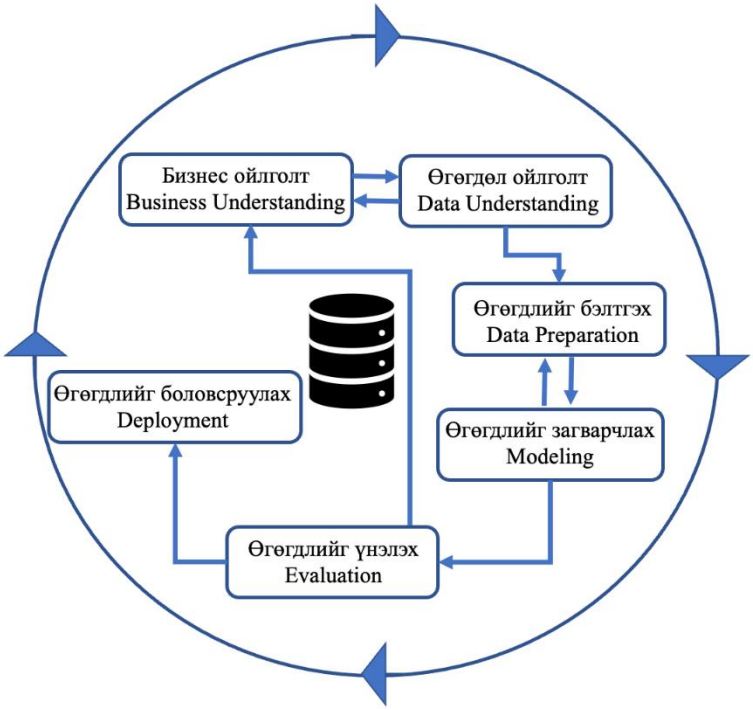
Эх сурвалж: (Jour, 2020)

Өгөгдөл олборлох үйл явц нь өгөгдлийн урьдчилсан боловсруулалт ба өгөгдөл олборлолт гэсэн хоёр хэсэгт хуваагддаг болох нь тодорхойлогдлоо. Өгөгдлийн урьдчилсан боловсруулалт нь өгөгдлийг цэвэрлэх, өгөгдлийг нэгтгэх, өгөгдлийг багасгах, өгөгдлийг хувиргах зэрэг орно. Өгөгдөл олборлолтын хэсэг нь өгөгдөл олборлох, хэв маягийн үнэлгээ, мэдээллийн мэдлэгийн дүрслэл гэж хэлж болно.

- **Олборлолтын загвар үүсгэх нь**

Олборлолтын загвар үүсгэхийн тулд алгоритм эхлээд туршилтын өгөгдөлд шинжилгээ хийж тодорхой хэв маяг, хандлагыг тодорхойлно. Алгоритм шинжилгээнийхээ үр дүнг олон давтах замаар олборлолтын загварын оновчтой параметрүүдийг тооцоолно. Үүний дараа уг загварыг цуглуулсан нийт өгөгдөлд биелүүлснээр үр ашигтай байж болох зүй тогтол, нарийвчилсан статистикуудыг илрүүлдэг байна. Өгөгдөл олборлолт гүйцэтгэхэд хамгийн түгээмэл хэрэглэгддэг процессийн загвар нь 1990 оноос өгөгдлийн шинжлэх ухаанд хэрэглэгдэж эхлэсэн Cross Industry Standard Process for Data Mining буюу CRISP-DM юм. Уг процессийн диаграмыг Зураг II.4-т дүрслэн харууллаа.

Зураг II.4 CRISP-DM процессийн диаграм



Эх сурвалж: (Azziaty,N, 2015)

CRISP-DM процесс нь нийт зургаан алхамаас бүрдэх ба үүнд: 1-т “Бизнес ойлголт” буюу төслийн зорилгыг бизнесийн талаас нь тодорхойлж түүнийгээ өгөгдөл олборлолтын асуудал болгон тавих. 2-т “Өгөгдлийг ойлгох” буюу өгөгдөлтэй танилцаж, чанарыг нь шалгах, таамаглал дэвшүүлэх боломж бүхий дэд хэсгүүдийг ялгах. 3-т “Өгөгдлийг бэлтгэх” буюу дараагийн шат болох өгөгдлийг загварчлахад шаардлагатай эцсийн өгөдлийн санг гарган авах. 4-т “Загвар байгуулах” буюу төрөл бүрийн загварчлалын арга техникүүдээс сонгон хэрэгжүүлж, тэдгээрийн параметрүүдийг оновчтой үр дүн гаргахаар тохируулна. 5-т “Үнэлэлт” буюу үүсгэсэн загварыг дахин шалгаж, турших, бизнес зорилгуудыг хэрхэн хангаж байгааг нягтлах. 6-т “Эцсийн хэрэглэгчдэд хүргэх” буюу олборлолтын загвар үүсгэснээр төсөл дуусдаггүй, түүнийг хэрэглэснээр гарсан үр дүнг хэрэглэгчдэд ойлгомжтой байдлаар хүргэх шаардлагатай байдаг байна.

2.2.2 Ангиллын онол

Ангилах гэдэг нь өгөгдсөн шинж х тус бүрийг урьдчилан тодорхойлсон ангилал у-д харгалзуулах үндсэн функц болох f(x,y)-ийг тодорхойлохын тулд суралцах үйл ажиллагаа билээ. Энэ нь өгөгдлийг зорилтот ангилал эсвэл ангиудад хуваах зорилготой өгөгдөл олборлох үйл явц юм. Ангилалтын алгоритм нь "Хяналттай суралцах" аргад багтах ба зөвхөн урьдчилан тодорхойлсон хамааран хувьсагчдад нөлөөлөх зүй тогтлуудыг шинжилдэг байна. Сургалтын өгөгдөлд орж байгаа өгөгдлүүд ямар ангилалд орох нь мэдэгдэж байх ёстой бөгөөд түүнийг ашиглан ангиллын загварыг байрлуулж, дараа нь уг загварыг зөв таниж байгаа эсэхийг туршилтын өгөгдлөөр шалгадаг. Мөн туршилтын өгөгдлийн ангилал мэдэгдэж байх ёстой юм. Учир нь туршилтын өгөгдлийг байршуулсан загварынхаа дагуу ангилаад өгөгдлийн өөрийнх нь мэдэгдэж байгаа ангилалтай хэд нь таарсан эсвэл ялгаатай буйгаас хамаарч үнэлгээ өгдөг билээ. Ангиллын загвар үнэн зөв ангилж байгаа эсэхийг туршилтын өгөгдөл, санамсаргүйгээр түүвэрлэх, солбиж турших гэсэн аргуудаар шалгадаг байна.

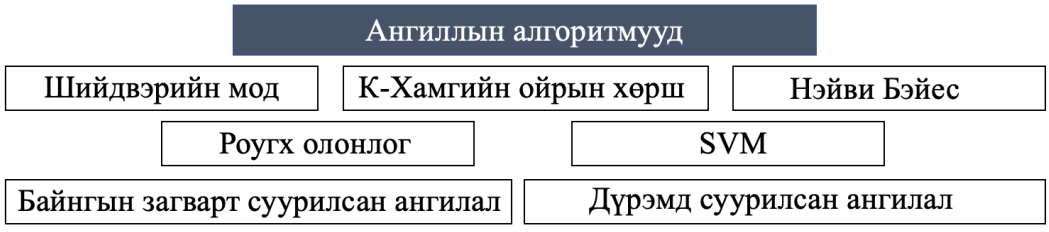
- Ангилах үйл явц

Ангилал нь сурах алхам болон ангилах алхам гэсэн хоёр шаттай үйл явц байх ба дараах байдлаар тодорхойлогдоно: **Сурах алхам:** Энэ нь ангиллын загварыг бий болгох алхам бөгөөд уг шатанд сургалтын өгөгдлийг ангиллын алгоритмаар шинжилдэг билээ. **Ангилах алхам:** Энэ нь өгөгдсөн өгөгдлийн ангиллын төрлийг урьдчилан таамаглахад загварыг ашигладаг алхам юм.

- Ангиллын алгоритмууд

Хүний нэг чухал үйл ажиллагаа бол нарийн, төвөгтэй үзэгдлүүдийг шинж чанарыг нь ашиглан ангилах явдал билээ. Үүнийг гүйцэтгэх үндсэн хэрэглүүр нь өгөгдөл олборлолт дахь ангиллын алгоритмууд болно. Classification буюу ангиллын алгоритмд багтах үндсэн аргуудыг Зураг II.5-д харуулж байна. Уг алгоритмын үндсэн алгоритмуудад: Шийдвэрийн модны арга, Нэйви Бэйесийн ангилал, Дүрэмд суурилсан ангилал, Вектор дэмжих машин (SVM), Ерөнхий шугаман загварууд, Бэйесийн ангилал, Буцах тархалтаар ангилах, К-Хамгийн ойрын хөрш зэрэг орно.

Зураг II.5 Ангиллын алгоритмын төрлүүд



Эх сурвалж: (Azziaty,N, 2015)

2.2.3 Бэйесийн теорем

Британи улсын математикч Томас Бэйесийн нэрээр нэрлэгдсэн Бэйесийн теорем нь XVIII зуунд үүссэн ба нөхцөлт магадлалыг тооцоолоход ашигладаг математикийн томьёо юм. Нөхцөлт магадлал гэдэг нь ижил, төстэй нөхцөл байдалд өмнөх үр дүнд тулгуурлан үр

дүн гаргадаг магадлал билээ. Бэйесийн теорем нь магадлалыг шинэчлэх буюу засах арга замыг өгдгөөрөө давуу талтай. Уг теоремыг мөн Бэйесийн дүрэм эсвэл Бэйесийн хууль гэж нэрлэдэг бөгөөд Бэйесийн статистикийн салбарын үндэс суурь болдог. Бэйесийн теорем нь ямар нэг үр дүн мэдэгдсэний дараа , түүнийг тусгасан нөхцөлт магадлалыг бодож олоход хэрэглэдэг ба уг теорем ёсооор P(A)>0 бол дараах томъёогоор илэрхийлэгдэнэ. Үүнд:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (1)$$

Энд, P(B)=B үзэгдэл явагдах магадлал
P(B|A)=A үзэгдэл явагдсаны дараа B үзэгдэл явагдах нөхцөлт магадлал

Бэйесийн теоремыг ойлгохын тулд магадлал болон нөхцөлт магадлалын талаар авч үзэх хэрэгтэй (Зураг II.6). Үүнд:

Зураг II.6 Бэйесийн теоремын үндсэн ойлголтууд

Эх сурвалж: (Sonner,N, 2020)

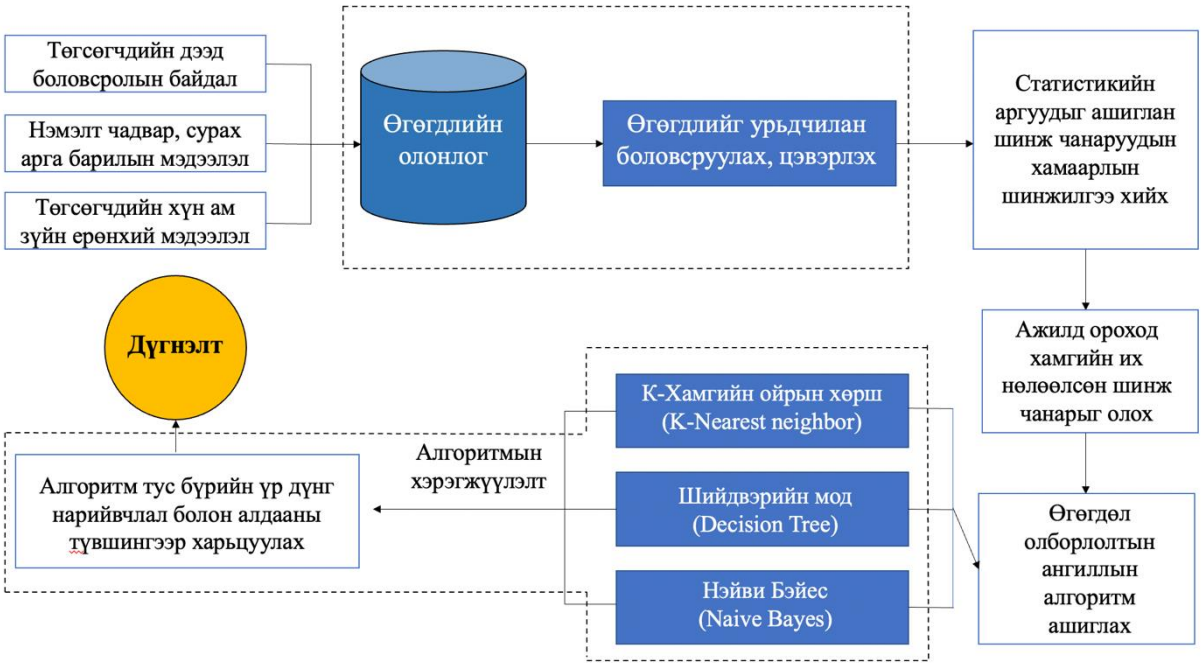
Магадлал гэдэг нь аливаа үйл явдал болох магадлалыг хэлдэг бөгөөд үргэлж 0-1 хооронд утгыг авна. Нөхцөлт магадлал гэдэг нь А үйл явдалтай холбоотой өөр үйл явдал аль хэдийн тохиолдсон тохиолдолд А үйл явдал болох магадлал юм. Мөн түүнчлэн, Бэйесийн теорем дээр суурилсан өгөгдөл олборлолтын алгоритмууд нь ангилал болон регресс гэх мэт Бэйесийн онолын элементийг ашигладаг билээ.

III БҮЛЭГ. ЭМПИРИК СУДАЛГААНЫ АРГАЗҮЙ

3.1 Судалгааны дизайн

Энэхүү судалгааны ажлын хүрээнд МУИС-ийг 2018-2019 оны хичээлийн жилд төгссөн төгсөгчдийн судалгаанд үндэслэн ангиллын загварын үнэлэх ба судалгааны дизайныг үндсэн арга зүй болон, өгөгдөлд тулгуурлан боловсруулсан болно (Зураг III.1-т дүрслэн харуулж байна).

Зураг III.1 Судалгааны дизайн



Эх сурвалж: Судлаачийн зураглал

Өмнөх хэсэгт дурдсаны дагуу төгсөгчдийн дээд боловсролын байдал, нэмэлт чадварын үнэлгээ болон хувийн мэдээллүүдээр өгөгдлийн олонлогоо бүрдүүлэх бөгөөд статистикийн арга ашиглан хувьсагч хоорондын хамаарлын шинжилгээ хийх юм. Энэ нь ажилд ороход хамгийн их нөлөөлсөн хувьсагчдыг илрүүлэх ач холбогдолтой бөгөөд өгөгдөл олборлолт дахь ангиллын алгоритмууд буюу К-Хамгийн ойрын хөрш, Шийдвэрийн мод, Нэйви Бэйес зэргийг R программ хангамж болон өгөгдөл олборлолтын WEKA software нь хамгийн тохиромжтой хэмээн санал болгосны дагуу ашиглан шинжилгээ хийсэн (Эдгээр ангиллын алгоритмыг дараагийн хэсэгт дэлгэрэнгүй тайлбарлана). Ийнхүү алгоритм тус бүр дээр үр дүн гарсны дараа нарийвчлал болон алдааны түвшин гэсэн 2 үзүүлэлт ашиглан алгоритмуудыг харьцуулж эцсийн дүгнэлтээ бичих юм. Эдгээрийг схем хэлбэрээр илэрхийлж, уг судалгааны ажлын ерөнхий дизайныг дээрх зурагт дүрсэлсэн билээ.

3.2 Загварыг үнэлэх статистик хэмжүүрүүд

Ангиллын алгоритмын үр дүнг Зөв ангилсан тохиолдол, Каппа статистик болон алдааны хэмжүүрүүдээр үнэлдэг байна. Хүснэгт III.1-д эдгээр үнэлгээний статистик хэмжүүрүүдийн математик тооцооллыг харуулсан болно. RMSE буюу дундаж квадрат алдааны язгуур нь дараах байдлаар загварын нарийвчлалыг үнэлнэ: RMSE < 10% үед маш сайн, RMSE 10% - 20% хооронд байвал сайн, RMSE 20% - 30% үед боломжийн, RMSE > 30% үед муу байна.

Хүснэгт III.1 Ангиллыг хэмжих үзүүлэлтүүд

Зөв ангилсан тохиолдол (Correctly Classified Instances)	CCI=(TP+TN/(TP+FP+FN+TN)
Буруу ангилсан тохиолдол (Incorrectly Classified Instances)	ICI=n-CCI
Каппа статистик (Kappa statistic)	$k=\frac{p_o-p_e}{1-p_e}$
Дундаж үнэмлэхүй алдаа (Mean absolute error)	$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i \rvert$
Дундаж квадрат алдааны язгуур (Root mean squared error)	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Харьцангуй үнэмлэхүй алдаа (Relative absolute error)	$\text{RAE} = \frac{\sum_{i=1}^{n}\lvert y_i - \hat{y}_i \rvert}{\sum_{i=1}^{n}\lvert y_i - \bar{y} \rvert}$
Харьцангуй квадрат алдааны язгуур (Root relative squared error)	$E_i = \sqrt{\frac{\sum_{j=1}^{n}(P_{(ij)} - T_j)^2}{\sum_{j=1}^{n}(T_j - \bar{T})^2}}$

Эх сурвалж: Albert,2011

Өгөгдөл олборлолтын ангиллын чанарыг хэмжихдээ Хүснэгт III.2-д буй хэмжүүрүүдийг ашиглана. Тухайлбал, Precision буюу нарийвчлал нь ангилагдсан үр дүнгийн хэд нь зөв болохыг илэрхийлдэг хувийн утга юм. Өөрөөр хэлбэл, нийт төгсөгчдийн хөдөлмөр эрхлэлтийн байдлыг зөв таамагласан бол 100%-ийн оноо өгөх болно. Recall буюу санах ой нь нийт түүврээс хэдэн ангиллыг зөв таамагласаныг харуулсан хувийн утга байна. F-Хэмжүүр нь дээрх хоёр хувийн утгын гармоник дундаж утга болно. Өгөгдөл олборлолтын ангиллын чанарыг хэмжих гол чанарын хэмжүүрүүд нь Precision буюу нарийвчлал болон Recall буюу эргэн дуудах байна.

Хүснэгт III.2 Чанарын хэмжүүрүүд

Томьёо	Тайлбар
$\boldsymbol{Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}}$	Эерэг ангилалд хамаарах эерэг ангийн таамаглалын тоог тодорхойлно
$\boldsymbol{Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}}$	Өгөгдлийн олонлог дахь бүх эерэг жишээнүүдийн эерэг ангиллын таамаглалын тоогоор илэрхийлнэ

$\boldsymbol{F-Measure=\frac{2\times Recall\times Precision}{Recall+Precision}}$	Нарийвчлал болон санах ойн асуудлыг тэнцвэржүүлж 1 оноо өгдөг
$\boldsymbol{MCC=\frac{(TP\times TN-FP\times FN)}{\sqrt{((TP+FP)(TP+FN)(TN+FP)(TN+FN))}}}$	Ангиллын чанарын хэмжүүр болгон машин сургалтанд ашиглагддаг

Эх сурвалж: (Albert,2011)

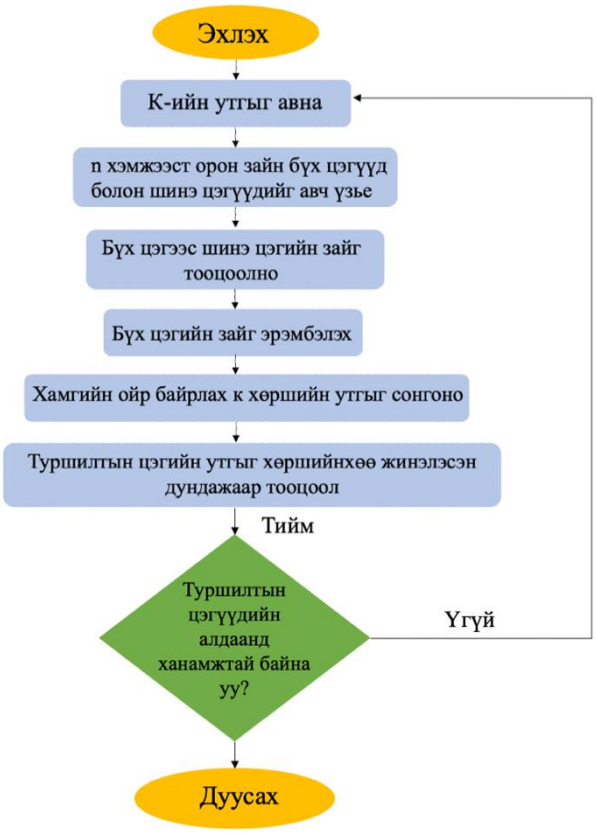
3.3 Судалгааны үндсэн арга, аргазүй

Өгөгдөл олборлолт дахь ангиллын алгоритмууд нь өгөгдлийн олонлогийг урьдчилан тодорхойлсон шинж чанараар нь ялгаж, бүлэглэх үйл явц билээ. Энэхүү хэсэгт судалгааны ажлын үндсэн арга зүй болох К-Хамгийн ойрын хөрш, Шийдвэрийн мод, Нэйви Бэйес зэрэг алгоритмуудыг дэлгэрэнгүй танилцуулах болно.

3.3.1 К-Хамгийн ойрын хөрш ангилагч алгоритм

К-Хамгийн ойрын хөрш алгоритмыг ангилал эсвэл регрессийн аль алинд ашиглах боломжтой параметрийн бус ангилагч юм. Энэ нь үндсэн өгөгдөл болон түүний тархалтын талаар ямар ч таамаглал гаргахгүй гэсэн утгыг илэрхийлнэ. К-Хамгийн ойрын хөрш нь машин сургалтын алгоритмуудын нэг бөгөөд эрүүл мэндийн салбараас эхлээд эдийн засаг, санхүүгийн салбаруудад өргөнөөр ашигладаг байна.

Зураг III.2 К-Хамгийн ойрын хөрш алгоритмын схем



Эх сурвалж: (Mohammed,2020)

К-Хамгийн ойрын хөрш алгоритмыг схем хэлбэрээр загварчилбал дээрх байдлаар дүрслэглдэнэ. Тус алгоритм нь боломжит бүх тохиолдлуудыг хадгалж, ижил төстэй байдлыг зайны функц дээр үндэслэн шинэ тохиолдлуудыг ангилдаг байна. К-Хамгийн ойрын хөрш алгоритмын ижил төстэй байдлыг хэмжигч зайны функцад дараах функцууд байх ба эдгээрээс хамгийн нийтлэг хэрэглэгддэг нь Евклидийн зай юм. Евклидийн зай, Манхэттэн зай болон Минковскийн зайны функц нь зөвхөн тасралтгүй хувьсагчдад ашигладаг ба чанарын хувьсагчтай үед Хаммингийн зайг (Hamming distance) ашиглана.

Хүснэгт III.3 Зайны функцууд

Зайны функц	**Томъёо**	**Давуу тал**	**Сул тал**
Евклидийн зай (Euclidean distance)	$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})\text{^}2}$	Тооцоолоход хялбар	Том хэмжээний өгөгдлийн олонлогт ажиллахад хэцүү
Манхэттэн зай (Manhattan distance)	$d(x,y) = \sum_{i=1}^{m} \lvert x_i - y_i \rvert$	Зайг зөв өнцгөөр хэмжинэ	Өндөр хэмжээст өгөгдөлд тохиромжгүй
Минковскийн зай (Minkowski distance)	$d(x,y) = \left(\sum_{i=1}^{k} (\lvert x_i - y_i \rvert)\text{^} \right)^{\frac{1}{q}}$	Зайг тооцоход диагналиар хэмждэггүй	Өндөр хэмжээст өгөгдөлд тохиромжгүй
Дундаж зай (Average distance)	$d_{ave} = \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - y_j)^2 \right)^{1/2}$	Том хэмжээний өгөгдлийн олонлогт сайн ажиллана	Хувьсагч нь зайны хэмжүүрт бие даан хувь нэмэр оруулдаг
Жинлэсэн Евклид (Weighted Euclidean)	$d_{we} = \left(\sum_{i=1}^{n} w_i (x_i - y_j)^2 \right)^{1/2}$	Чухал хувьсагчдын үр нөлөөг нэмэгдүүлдэг	Хувьсагч нь зайны хэмжүүрт бие даан хувь нэмэр оруулдаг
Хаммингийн зай (Hamming distance)	$D_H = \sum_{i=1}^{k} \lvert x_i - y_i \rvert$	Векторууд ижил урттай биш үед ашиглана	Векторууд тэнцүү үед ашиглагдахгүй

Эх сурвалж: (Rajesh,2020)

Зайны функцийг өгөгдлийн олонлогоос шууд тооцоолох нэг томоохон дутагдал нь хувьсагчид өөр өөр хэмжүүртэй эсвэл тоон болон чанарын хувьсагчдын холимог байх явдал юм (Sallay, 2012). Жишээлбэл, Хэрэв нэг хувьсагч нь жилийн орлогод тулгуурласан төгрөгөөр, нөгөө нь насаар тооцсон бол орлого нь тооцоолсон зайд илүү их нөлөөлнө. Тиймээс стандартчилал ашиглан үүнийг шийднэ. К-Хамгийн ойрын хөрш аргын стандарчилал дараах байдлаар хийгдэнэ:

$$X_s = \frac{X - mean}{s.d} \quad (2)$$

- k утга сонгох нь

Энэхүү алгоритмын k утга нь харьцуулж буй хөршүүдийн тоог илэрхийлэх ба харьцуулсан хамгийн ойрын k өгөгдлийн тоог харуулна. Өөрөөр хэлбэл, өгөгдлийн цэг бүрийн хувьд хамгийн ойрын k ажиглалтыг олж, дараа нь өгөгдлийн цэгийг олонлог гэж ангилдаг. Ихэвчлэн хамгийн ойрын k ажиглалтыг авч үзэж буй өгөгдлийн цэг хүртэлх Евклидийн хамгийн бага зайтай ажиглалт гэж тодорхойлдог. Жишээлбэл, хэрэв k = 3, мөн тодорхой өгөгдлийн цэгт хамгийн ойр байгаа гурван ажиглалт нь А, В, А ангилалд хамаарах бол алгоритм нь өгөгдлийн цэгийг А ангилалд ангилах болно.

3.3.2 Нэйви Бэйес ангилагч

Нэйви Бэйесийн алгоритм нь Бэйесийн теорем дээр суурилсан нөхцөлт магадлалт ангилагч билээ. Уг алгоритм нь шугаман параметрүүдийг шаарддаг, чанарын хувьсагчтай илүү сайн ажилладаг байна. Мөн бие даасан хувьсагчид статистикийн хувьд ч бие даасан байдаг хэмээн тайлбарладаг ба өгөгдлийн хэмжээ их байх тусам Нэйви Бэйес алгоритм тохиромжтой байдаг. Өөрөөр хэлбэл, уг ангилагч нь тухайн ангилал дахь онцлог шинж чанар нь бусад шинж чанараас хамааралгүй гэж таамагладаг байна. Уг таамаглалын нөхцөлт үл хамаарах анги гэж нэрлэнэ. Нөхцөлт магадлалын загвар нь n шинж чанарыг $x = (x_1, \dots, \dots x_n)$ векторoор илэрхийлэгдэнэ (Narashima Murty, 2011). Энэ тохиолдолд магадлалыг оноож өгдөг ба $p(C_k|x_1, \dots, \dots x_n)$ байна. Илүү уян хатан нөхцөлтэй байлгах зорилгоор Бэйсийн теоремыг ашиглан нөхцөлт магадлалыг дараах байдлаар задалж болно:

$$P(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \quad (3)$$

Энд, $p(C_k)p(x|C_k)$ нь ангийн шошгод өгөгдсөн онцлог шинж чанаруудын тодорхой хослолын магадлалыг хэлнэ. Энэ нь С-ээс хамаарахгүй x_i хувьсагчийн утгууд өгөгдсөн тул хуваагч нь тогтмол билээ. Тиймээс бутархайн хувиар нь хамтарсан магадлалын загвартай тэнцүү байна. Үүнийг дараах байдлаар бичнэ:

$$p(C_k, x_1, \dots, x_{\mathrm{n}}) \quad (4)$$

Нөхцөлт бие даасан байдал дахь урьдач нөхцөлийн дагуу, х хувьсагч нь C_K ангилалд хамааралгүй гэж үзвэл: $p(x_i|x_{i+1}, \dots, x_n, C_k) = p(x_i|C_k)$

Үүнээс нэгдсэн (joint) загварыг дараах байдлаар бичиж болно. Үүнд:

$$\begin{aligned} p(C_k|x_1, \dots, x_{\mathrm{n}}) &\propto p(C_k, x_1, \dots, x_{\mathrm{n}}) \\ &\propto p(C_k)p(x|C_k)p_2(x_2|C_k)\cdots \\ &\propto (C_k)\prod_{i=1}^{n} p(x_i|C_k) \end{aligned} \quad (5)$$

Энд, $\propto$ нь пропорциональ байдлыг илэрхийлнэ. Ийнхүү С ангиллын хувьсагч дээрх нөхцөлт тархалт нь: $p(C_k|x_1, \dots, \dots x_n) = \frac{1}{Z}p(C_k)\prod_{i=1}^{n} p(x_i|C_k)$ (6)

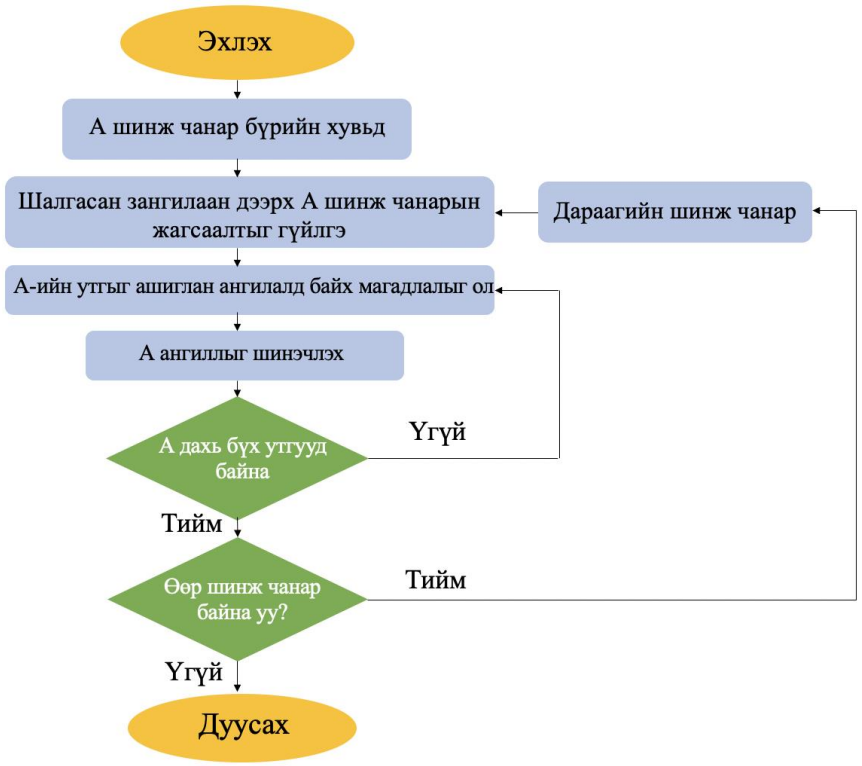
Энд, $Z = p(x) = \sum_k p(C_k)\, p(x|C_k)$ байна. Бэйсийн ангилагчаар дурын к-ийн хувьд у = $C_к$ гэсэн ангилал үүсгэх боломжтой бөгөөд томьёог доор үзүүлэв:

$$\hat{y} = \underset{k\in\{1,\dots K\}}{\operatorname{argmax}}\, p(C_k) \prod_{i=1}^{n} p(x_i\,|C_k) \qquad (7)$$

- Нэйви Бэйесийн алгоритм хэрхэн ажилладаг вэ?

Энэхүү ангилагч алгоритм нь дараах үе шатуудаар үзэгдлийн магадлалыг тооцно. Үүнд: 1-р шат: Ангиллын өгөгдсөн нөхцөлд приор магадлалыг тооцох. Приор магадлал гэдэг нь таамаглалаас үл хамаарах өгөгдлийн магадлалыг хэлнэ. 2-р шат: Анги бүрийн хувьд шинж чанар бүрээр боломжит магадлалыг олох. 3-р шат: Байесийн томьёонд эдгээр утгуудаа орлуулан постериор магадлалыг тооцох. Постериор магадлал гэдэг нь D өгөгдөл өгөгдсөн нөхцөл дэх h таамаглалын магадлал байна. 4-р шат: Өндөр магадлалтай ангид хамаарах орцууд өгөгддөг тул аль анги нь өндөр магадлалтай байгааг харах гэсэн үе шатууд орно. Нэйви Бэйес алгоритмыг схемээр загварчилбал дараах байдлаар байна.

Зураг III.3 Нэйви Бэйес алгоритмын схем



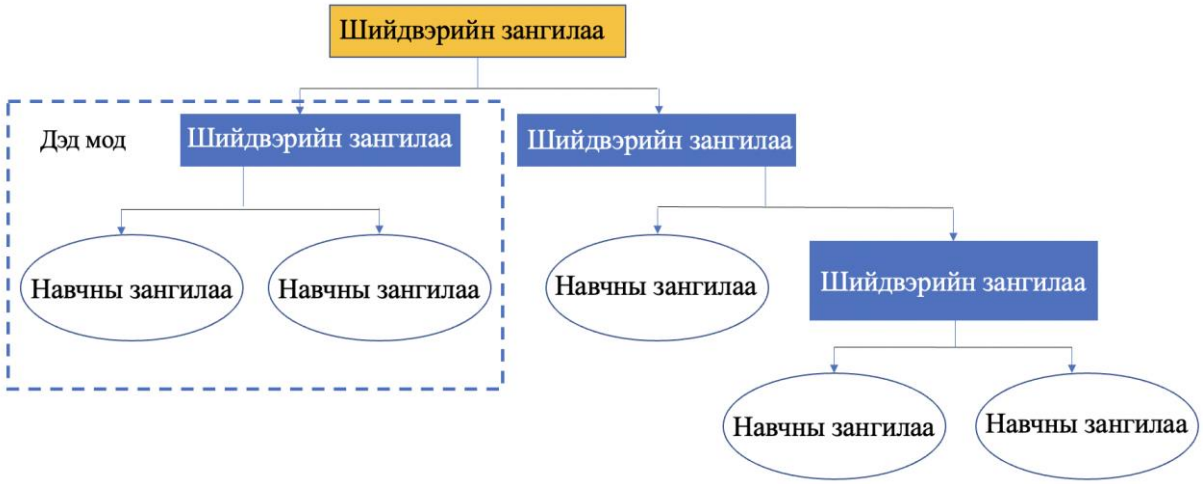
Эх сурвалж: (Sneha,2019)

3.3.3 Шийдвэрийн модны алгоритм

Шийдвэрийн мод алгоритмын гол санаа нь шийдвэрийн эрсдэл болон үр ашгийг үнэлэх боломжийг олгох ба шийдвэр гаргалтаас үүдэлтэй учирч болох эрсдэлийг бууруулах зорилготой байдаг байна. Шийдвэрийн модны арга нь өгөгдлийг рекурсив байдлаар дэд олонлогт хуваах явдал бөгөөд дотоод зангилаа бүр нь шинж чанарын туршилтыг буюу өгөгдлийн дэд олонлогийг агуулж, навчны зангилааг оруулаагүй тохиолдолд асуулт нь дэд олонлогийг хуваана, харин мөчир нь туршилтын үр дүнг, навчны зангилаа нь шинж

чанарыг илэрхийлдэг байна (Зураг III.4-д дүрслэн харууллаа). Өөрөөр хэлбэл хамгийн тохиромжтой таамаг бүхий шийдвэр гаргалт гарч ирэх хүртэл модны бүтцээр салбарлан байршина. Шийдвэрийн мод нь шийдвэрийн зангилаа, боломжийн зангилаа болон төгсгөлийн зангилаа гэсэн гурван төрлийн зангилаанаас бүрдэнэ. Шийдвэрийн зангилаа нь ихэвчлэн квадратаар дүрслэгддэг бол боломжийн зангилаа нь тойрог хэлбэрээр харин төгсгөлийн зангилаа нь гурвалжин дүрсээр дүрслэгддэг байна.

Зураг III.4 Шийдвэрийн модны бүтэц



Эх сурвалж: (Такавира Оливер, 2022)

Энэхүү судалгааны ажлын хүрээнд I-р бүлэгт дурдсанаар өгөгдөл олборлолтын шилдэг 10 алгоритмын нэг C4.5 алгоритмыг ашиглах билээ. Тус алгоритмыг 1986 онд Росс Куинлан бүтээсэн бөгөөд CART болон бусад алгоритмуудаас ялгаатай нь асуулт бүрд шинэ зангилаа нэмж өгснөөрөө давуу талтай юм (Springer LNCS, 2008). C4.5 алгоритм нь S эх олонлогоос эхэлдэг ба алгоритмын давталт бүрд S олонлогийн ашиглаагүй шинж чанарыг давтдаг. Тус алгоритмын гол санаа нь өмнө нь сонгогдоогүй байгаа шинж чанаруудыг харгалзан дэд олонлог бүр дээр давтагдах юм. Тэдгээр шинж чанарын энтроп H(S) болон мэдээллийн хожоог IG(S)-ийг тооцоолох ба дараах байдлаар байна. Үүнд:

- **Энтроп**: S эх олонлогийн санамсаргүй байдлын хэмжүүр юм. Тус алгоритм нь энтроп хамгийн бага байхыг илүүд үздэг байна.

$$H(S) = \sum_{x \in X} -p(x) log_2 p(x) \quad (8)$$

Энд, $H(S)$=0 үед S олонлог төгс ангилагдсан байна.
S- Энтропыг тооцож байгаа одоогийн өгөгдлийн олонлог
X- S ангиллын олонлог
p(x)- x ангиллын элементийн тоог S олонлогийн элементийн тоонд эзлэх хувь

Тус утга нь 0-ээс 1-ийн хооронд утга авч болох ба өгөгдлийн олонлог дахь түүвэр S-тэй нэг ангилалд хамаарах бол энтроп 0-тэй тэнцүү байна. Эсрэг тохиолдолд 1-тэй тэнцүү байна.

- **Мэдээллийн хожоо**: S олонлогийг A шинж чанарт хуваахаас өмнөх ба дараа хүртэлх энтропын зөрүүний хэмжүүр юм.

$$IG(S,A) = H(S) - \sum_{t \in T} p(t)H(t) = H(S) - H(S|A) \quad (9)$$

Энд, $H(S)$- S олонлогийн энтроп
T- S олонлогийг A атрибутаар хуваах замаар үүсгэсэн дэд олонлогууд
p(t)- S олонлогийн элементүүдийн тоонд эзлэх хувь
$H(T)$- t дэд олонлогийн энтроп

Тухайлбал, мэдээллийн хожоо нь өгөгдлийн олонлогийг шинж чанарт нь үндэслэн ангилсны дараа энтропийн өөрчлөлтийг хэмжих явдал юм. Шийдвэрийн модны алгоритм нь мэдээллийн хожоог хамгийн их байлгахыг хүсдэг бөгөөд энэ нь хамгийн их мэдээллийн хожоотой зангилаа нь эхлээд хуваагддагтай холбоотой аж.

IV БҮЛЭГ. ШИНЖИЛГЭЭНИЙ ХЭСЭГ

Өгөгдөл олборлолт нь их өгөгдөл буюу Big data дээр илүү сайн ажилдагаараа онцлог тул манай улсын хэмжээнд төгсөгчдийн мэдээллийг жил бүр цуглуулдаг болон түүврийн хэмжээ хангалттай эсэх зэргийг харгалзан үзэж МУИСургуулийн Оюутан, төгсөгчийн хэлтсээс явуулдаг "Төгсөгчдийн хөдөлмөр эрхлэлтийн судалгаа"-ийн өгөгдөлийг ашигласан болно. Тус судалгаа нь 2018-2019 оны хичээлийн жилд МУИСургууль төгсөгчдийн 40-өөс багагүй хувь буюу 1031 төгсөгчийг хамруулж, мөшгих (follow up) хэлбэрээр өгөгдлийн олонлогоо бүрдүүлсэн ба түүврийн төлөөлөх чадвар сайтай хийгддэгээрээ давуу талтай байв. Мөн энэхүү мэдээлэл нь оюутны хувийн мэдээлэл агуулсан байсан тул тус сургуулиас эрхлэн гаргадаг нэгдсэн тайлангаас шаардлагатай мэдээллүүдээ цуглуулсан болно.

4.1 Судалгааны түүвэрлэлт, бүтэц тоо хэмжээ

Уг судалгааны мэдээллийг цуглуулахдаа баримт бичгийн шинжилгээний арга болон телефон сурвалжлагын аргыг хэрэглэсэн байна. МУИС-г тус хичээлийн жилд 2512 оюутан 100 гаруй хөтөлбөрөөр суралцаж төгссөн ба тус мэдээлэлд тулгуурлан 40-өөс багагүй хувь тухайн түүвэрлэлтээр нийтдээ 1031 төгсөгч оролцжээ. Нийт өвөл болон хаврын төгсөлтийн харьцааг авч үзэхэд нийт 7 бүрэлдэхүүн сургуульд суралцаж, 22.5% нь өвөл, 77.5% хаврын улиралд тус тус төгссөн байна. Энэхүү төгссөн жилийн оюутны тоо хэмжээнд тулгуурлан 231 өвлийн төгсөгч, 800 хаврын төгсөгч оюутан судалгаанд хамрагдсан болно. Судалгаанд хамрагдах оюутнуудыг сонгохдоо төгсөгчийн нэрсийн жагсаалтад тулгуурлан, системчилсэн санамсаргүй түүвэрлэлийн ашигласан. Үүнд: N/n буюу N нь эх олонлогийн хэмжээ, n нь түүвэр олонлогийн хэмжээ ажээ. Шинжлэх ухааны сургуулийн хувьд N=1025, n=405 байх ба түүврийн алхам нь ойролцоогоор 3 буюу түүвэрлэлтийг хийхдээ нэрсийн жагсаалтаас эхний 3 төгсөгчийн нэрсээс аль нэгийг санамсаргүйгээр сонгож, уг зарчмаар түүвэрлэлт хийгджээ (МУИС, 2020). Ийнхүү тухайн түүвэрлэлт нь нийт эх олонлогийн 41%-ийг эзэлж байна. Дараах хүснэгтэд түүврийн тоо хэмжээг бүрэлдэхүүн сургууль тус бүр дээр харуулсан билээ.

Хүснэгт IV.1 Түүврийн тоо хэмжээ

МУИС-ийн бүрэлдэхүүн сургууль	Төлөвлөсөн түүвэрлэлт	Бодит түүвэрлэлт
Шинжлэх ухааны сургууль	405	444
Бизнесийн сургууль	222	246
Олон улсын харилцаа, нийтийн удирдлагын сургууль	49	36
Хууль зүйн сургууль	107	111
Хэрэглээний шинжлэх ухаан, инженерчлэлийн сургууль	144	131
Орхон салбар сургууль	65	34
Завхан салбар сургууль	33	29
НИЙТ	1025	1031

Эх сурвалж: Оюутан төгсөгчийн хэлтэс, МУИС

Ийнхүү судалгааны хувьсагчдыг сонгохдоо “Төгсөгчдийн хөдөлмөр эрхлэлтийг урьдчилан таамаглах ангиллын загварыг бий болгоход өгөгдөл олборлолтыг ашиглах ” нэртэй Ченг Па судлаачийн 2020 оны судалгааны ажил дээр тулгуурласан (Cheng Fa, 2020). Мөн судлагдсан байдлаас үзэхэд нийт судлаачдын ашиглаж буй хувьсагч нь нийтлэг, ижил хувьсагчид ашигласан байсан юм. Хүснэгт IV.2-т тус эмпирик ажлын хувьсагчдын товч тайлбарыг харуулж байна (Хавсралт 5-с дэлгэрэнгүй тайлбарыг харна уу)

Хүснэгт IV.2 Хувьсагчдын тайлбар

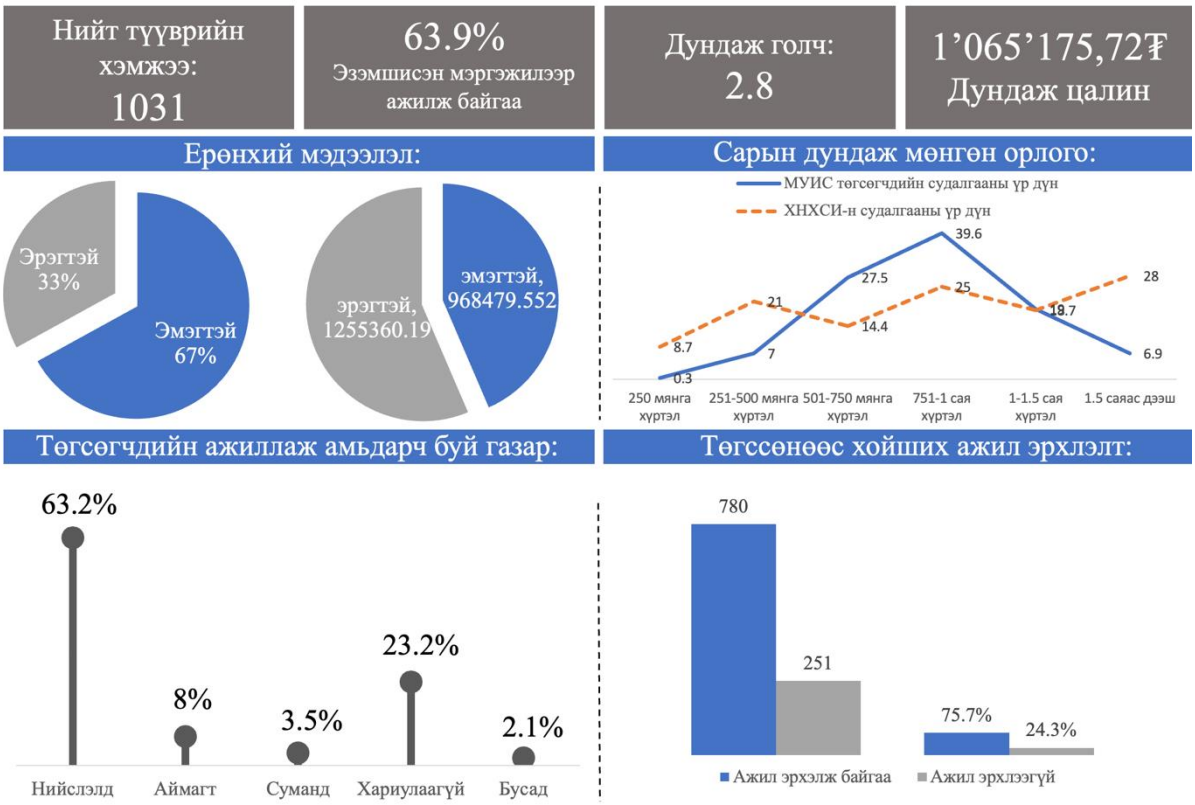
Хувьсагч	**Тэмдэглэгээ**	**Утга**	**Нэгж**	**Тайлбар**
Бүрэлдэхүүн сургууль	A_1	Чанарын	[1:7]	Нийт 7 төрлийн шинжлэх ухааны зорилго бүхий сургуулиуд
Салбар сургууль	A_2	Чанарын	[1:9]	Бүрэлдэхүүн сургууль тус бүрд үүссэн 9 салбар сургууль
Хөдөлмөрийн байдал	A_3	Чанарын	[1:2]	Ажилтай болон ажилгүй эсэх
Мэргэжил	A_4	Чанарын	[1:109]	Салбар сургууль бүрд бэлддэг 109 төрлийн мэргэжил
Ажил	A_5	Чанарын	[1:8]	Ямар төрлийн байгууллагад ажиллаж буй талаар мэдээлэл
Ажилд орсон хугацаа	A_6	Чанарын	[1:6]	Төгссөнөөс хойш хэдэн сарын хугацаанд ажилд орсон мэдээлэл
Мэргэжилээрээ ажиллаж буй эсэх	A_7	Чанарын	[1:3]	Мэргэжилээр болон мэргэжилээс өөр ажил хийж буй
Ажил хайх бэрхшээл	A_8	Чанарын	[1:11]	Дотоод болон гадаад хүчин зүйлүүд
Ажил хийхгүй байгаа шалтгаан	A_9	Чанарын	[1:14]	Дотоод болон гадаад хүчин зүйл шалтгаан
Голч	A_{10}	Тоон	1.0-4.0	Төгсөх үеийн голч
Хүйс	A_{11}	Чанарын	[1:2]	1- Эмэгтэй 2- Эрэгтэй
Сарын дундаж орлого	A_{12}	Тоон	[1:4]	100 мянгаас 5 саяын хооронд

Эх сурвалж: Судлаачийн тооцоолол

4.2 Өгөгдлийн танилцуулга

Судалгааны хүрээнд ажил эрхлэлтийн онцлог, орлогын ялгаатай байдал, ажилгүй байгаа шалтгаан, сурлагын голч, ажилгүй байгаа шалтгаан зэргийг хүйсийн онцлогтой харьцуулан, дүн шинжилгээ хийсэн болно. Судалгаанд хамрагдсан төгсөгчдийг хүйсийн хувьд авч үзэхэд 67% нь эмэгтэй, 33% нь эрэгтэй бөгөөд нийт эрэгтэйн 81.3% нь ажилтай бол нийт эмэгтэйн 72.3% нь ажил эрхэлж байна. Бүх салбар сургуулиудын голч дүнгийн дундаж нь 2.8 ба нийт төгсөгчдийн 63.9% нь эзэмшисэн мэргэжилээрээ ажиллаж байна. Судалгаанд хамрагдсан төгсөгчдийн дундаж цалин ₮1065175,72 бөгөөд ХНХСИ-с эрхлэн гаргадаг мөн үеийн Монгол Улсын нийт төгсөгчдийн дундаж орлогтой харьцуулбал улсын хэмжээнд төгсөгчдийн 25% нь 751мянгаас 1 сая хүртэлх цалин авч байгаа бол МУИС-ийн 39.6% нь тус цалинг авч буй нь дараах зургаас харагдах бөгөөд харьцангуй ялгаатай байна. Төгсөгчдийн сарын дундаж орлогыг эрэгтэй, эмэгтэй хүйсийн хувьд харьцуулж үзэхэд бага зэрэг ялгаа харагдаж байна. Тухайлбал эмэгтэйчүүдийн хувьд дунджаар ₮968.5 мянган төгрөгийг дундаж орлогыг сард олдог бол эрэгтэйчүүд сард дунджаар ₮1255.4 мянган төгрөгийн орлого буюу ₮286.9 мянган төгрөгөөр илүү орлоготой байна.

Зураг IV.1 Өгөгдлийн танилцуулга



Эх сурвалж: Судлаачийн тооцоолол

Мөн түүнчлэн төгсөгчдийн 63.2% нь нийслэлд амьдарч байгаа бөгөөд амьдарч буй байршлаас шалтгаалан орлогын түвшин ялгаатай буйг тодорхойлогч статистик хэсэгт дэлгэрэнгүй авч үзэх болно. Төгсөгчдийн 63.9% хувийн хэвшлийн байгууллагад ажиллаж байгаа бол 23.5% нь төрийн байгууллагад ажиллаж байв. Дараах хүснэгтэд

сургууль төгсөөд ажилд орсон хугацаа болон ажилд ороход нөлөөлсөн хүчин зүйлсийг харуулж байна.

Хүснэгт IV.3 Төгсөгчдийн хөдөлмөр эрхлэлтэд нөлөөлөгч хүчин зүйлс ба ажилд орсон хугацаа

Сургууль төгсөөд ажилд орсон хугацаа		**Ажилд орох хүчин зүйлс**	**Хүйсийн байдал**	
			эрэгтэй	эмэгтэй
Шууд ажилд орсон	14.1%	Сурлагын дүн	16%	17%
3 сарын дотор ажилд орсон	27.6%	Мэргэжлийн онцлог	18%	19%
4-7 сарын хугацаанд хүлээж байж орсон	11.4%	Төгссөн сургууль	39%	43%
8-12 сарын хугацаанд хүлээж байж орсон	9.6%	Гадаад хэлний мэдлэг	17%	23%
Нэг жил буюу түүнээс дээш хүлээж байж ажилд орсон	4.7%	Хувийн зан чанар	46%	42%
Нийт	67.4%	Ерөнхий ур чадвар	38%	27%
Хариулаагүй	32.6%	Мэргэжлийн дур сонирхол, чиг баримжаа	7%	10%
Нэгдсэн дүн	100%	Бусад хүчин зүйл (Танил гэх мэт)	4%	4%

Эх сурвалж: Судлаачийн тооцоолол

МУИС-ийг 2018-2019 оны хичээлийн жилд төгсөгчдөөс 14.1 хувь нь суралцах үедээ болон төгссөн даруйдаа ажил хөдөлмөр эрхэлсэн байна. Харин төгсөгчдийн 41% буюу хамгийн их хувь нь төгссөнөөс хойш 3 сарын дотор ажил хөдөлмөр эрхлэжээ. Мөн судалгааны үр дүнгээс үзэхэд төгсөгчдийн 17.% нь 4-7 сарын хугацаанд, 9.6% нь 8-12 сарын хугацаанд ажил хөдөлмөр эрхэлсэн бол, нийт төгсөгчдийн 6.9% нь нэг жил буюу түүнээс дээш хугацаанд хүлээж байж ажил хөдөлмөр эрхэлсэн байна. Төгсөгчдийн хувьд ажилд ороход нөлөөлсөн хүчин зүйлсийг дээрх байдлаар нэрлэсэн. Тухайлбал, ажилд ороход хувийн зан чанар 23.5%, төгссөн сургууль 22.7%, Ерөнхий ур чадвар 16.8% тус тус ихээхэн нөлөөлдөг эхний гурван хүчин зүйл болохыг төгсөгчид онцолсон байна. Хүйсийн байдлаар ажилд орох хүчин зүйлсийг харвал мөн адил эдгээр гурван хүчин зүйлийг тодотгожээ.

4.3 Түүврийн тодорхойлогч статистик

Тухайн жилд төгсөгчдийн голч оноо болон ажил эрхлэлт, сарын дундаж мөнгөн орлогын үзүүлэлтийг доорх хүснэгтэд харуулав. Ихэнх төгсөгчдийн хувьд сарын дундаж орлогоо

ойролцоогоор илэрхийлж хариулт өгсөн ба мөн хариулт өгөхөөс татгалзсан тохиолдол байсан болно. Ажлын байр ихэвчлэн нийслэл хотод байгаа нь төгсөгчдийн оршин суугаа газраа сонгоход нөлөөлөөд зогсохгүй, төвлөрөл шилжих хөдөлгөөний татах хүчин зүйлийн нэг болжээ. Мөн төгсөгч оюутнуудын хувьд харьцуулах шаардлагатай нэг үзүүлэлт бол тухайн суралцаж байх хугацааны сурлагын амжилт буюу сурлагын голч дүн билээ.

Хүснэгт IV.4 Төгсөгчдийн голч ба орлогын дундаж үзүүлэлт, сургууль тус бүрээр

Сургууль	N	Доод	Дээд	Дундаж	Хазайлт
ОУХНУС	36	2.4	3.8	3.2722	
	25	500000₮	8000000₮	1906400₮	2060802.59₮
НУС	262	2	3.9	2.99	
	47	100000₮	6000000₮	985106.38₮	821562.33₮
ХЗС	111	2.2	3.7	2.99	
	68	500000₮	10000000₮	1558161.76₮	1692407.08₮
ХУС	311	1.8	3.8	2.95	
	91	450000₮	7500000₮	963659.34₮	805095.788₮
БС	553	1.8	3.9	2.84	
	166	320000₮	2500000₮	1057789.16₮	341976.055₮
БУС	408	1.8	3.8	2.78	
	114	170000₮	10000000₮	919508.77₮	923617.36₮
ЗС	29	1.9	3.5	2.72	
	15	500000₮	9000000₮	1188666.67₮	2165660.13₮
ХШУИС	131	1.8	3.7	2.71	
	98	390000₮	2000000₮	891071.43₮	302690.896₮
ОС	159	1.6	4	2.67	
	2	500000₮	600000₮	550000₮	70710.678₮

Эх сурвалж: Судлаачийн тооцоолол

Орхон сургуулийн төгсөгчдийн голч хамгийн өндөр буюу 4.0 байгаа бол хамгийн бага голч дүн ХУС, БС, БУС зэрэг сургуулийн голчийн доод утга 1.8 байна. Дундаж утгаар авч үзвэл ОУХНУС сургууль хамгийн өндөр буюу 3.2 бөгөөд сарын орлого мөн ₮1,906,400 буюу хамгийн их утгатай байна.

Ажил эрхлэгчдийн хувьд ажлаа солих, өөрчлөх эсвэл ажлын байрандаа тогтвортой ажиллах нэгэн чухал үзүүлэлт нь цалин орлогын хэмжээ билээ. Хүснэгт IV.5-д ажил эрхлэлтийн байдал ба орлогын дундаж утгыг харуулж байна. Нийт 1031 түүврээс 616 төгсөгч сарын дундаж орлогын үзүүлэлтээ ойролцоогоор илэрхийлсэн байв. Ажил эрхэлж буй төгсөгчид сард дунджаар хамгийн багадаа 100 мянга, хамгийн ихдээ 10 сая хүртэл төгрөгийн орлого олдог байна. Харин ажил эрхлэдэггүй болон түр ажил эрхэлж буй 12 төгсөгчийн хувьд сард 450 мянгаас 2 сая хүртэл төгрөгийн орлогтой бөгөөд дундаж цалингийн хувьд авч үзвэл 1.6 дахин бага орлого олж байна.

Хүснэгт IV.5 Ажил эрхлэлтийн байдал ба орлогын хэмжээ

Та ажил эрхэлж байгаа юу?	N	Доод	Дээд	Дундаж	Хазайлт
Тийм	614	100000	10000000	1072671.01	993170.913
Үгүй	12	450000	2000000	681666.67	433628.105

Эх сурвалж: Судлаачийн тооцоолол

Хүснэгт IV.6-д буй Корреляцын шинжилгээний үр дүнгээс үзэхэд голч оноо ба ажил эрхлэлт эерэг хамааралтай буюу 0.9 байна. Тиймээс төгсөгчийн голч нь ажил эрхлэхтэй өндөр хамааралтайг илтгэж байна. Харин голч оноо бага байх тусам ажил эрхлэлтийн хувь хэмжээ буурахад нөлөө үзүүлдэг гэсэн хамаарал ажиглагдлаа. Мөн тухайн төгсөгчийн мэргэжил ямар газар ажиллахтай эерэг хамааралтай буюу 0.6 байна. Үүнээс сурлагын голч дүн ямар нэг байдлаар ажил хөдөлмөр эрхлэхэд нөлөө үзүүлдэг гэж үзнэ.

Хүснэгт IV.6 Түүврийн корреляцын шинжилгээ

	Хүйс	Ажил	Хугацаа	Голч	Мэргэжил	Салбар сургууль	Сургууль	Байршил	Мэргэжлээрээ ажиллах	Ажил хайх бэрхшээл	Ажил эрхлэлт
Хүйс	1.00	0.21	0.06	-0.10	-0.04	-0.18	-0.18	-0.18	-0.11	-0.08	-0.10
Ажил	0.21	1.00	-0.19	0.09	0.61	-0.15	-0.17	-0.11	-0.41	0.36	0.10
Хугацаа	0.06	-0.19	1.00	-0.77	-0.18	0.13	0.15	0.23	0.81	0.11	-0.77
Голч	-0.10	0.09	-0.77	1.00	0.10	0.17	0.15	0.00	-0.84	-0.14	0.92
Мэргэжил	-0.04	0.61	-0.18	0.10	1.00	-0.05	-0.08	0.05	-0.27	0.33	0.10
Салбар сургууль	-0.18	-0.15	0.13	0.17	-0.05	1.00	0.95	0.31	0.01	0.05	0.17
Сургууль	-0.18	-0.17	0.15	0.15	-0.08	0.95	1.00	0.26	0.03	0.02	0.14
байршил	-0.18	-0.11	0.23	0.00	0.05	0.31	0.26	1.00	0.13	0.22	-0.03
мэргэжлээрээ ажиллах	-0.11	-0.41	0.81	-0.84	-0.27	0.01	0.03	0.13	1.00	0.02	-0.86
Ажил хайх бэрхшээл	-0.08	0.36	0.11	-0.14	0.33	0.05	0.02	0.22	0.02	1.00	-0.14
ангилал	-0.10	0.10	-0.77	0.92	0.10	0.17	0.14	-0.03	-0.86	-0.14	1.00

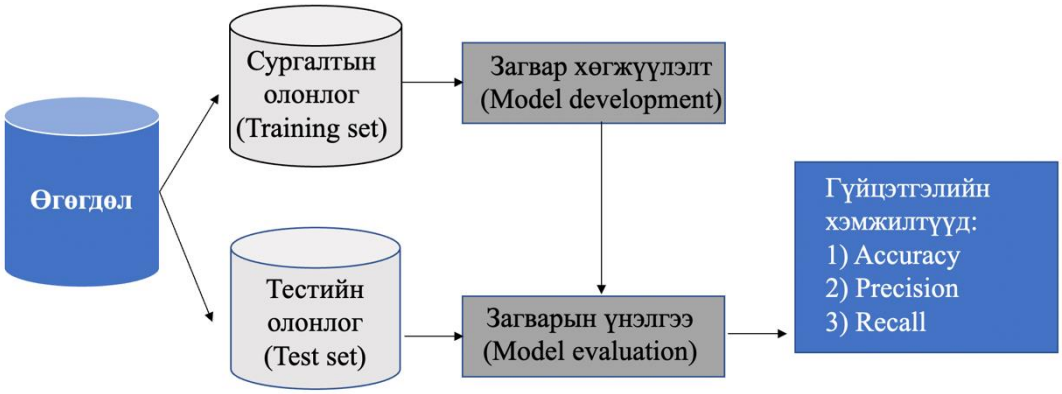
Эх сурвалж: Судлаачийн тооцоолол

4.4 Судалгааны арга зүйн хэрэгжүүлэлт

Ихэнх өгөгдөл олборлолтын чиглэлээр хийгдсэн судалгааны ажилд WEKA software программыг ашиглаж үнэлдэг нь судлагдсын байдлын явцад тодорхойлогдсон билээ. Тус программ нь өгөгдөл олборлолт болон машин сургалтын алгоритмуудыг хэрэгжүүлэхэд зориулагдсан ба ангилал болон бүлэглэлийн загварыг үүсгэх боломжтой, үнэ төлбөргүй гэдгээрээ давуу талтай байв. Төгсөгчдийн хөдөлмөр эрхлэлтийг урьдчилан таамаглах ангиллын загварыг бий болгохдоо суралцах болон үнэлгээний гэсэн хоёр үе шатаар ангиллын загварыг бий болгосон. Суралцах үе шатанд ангилагч алгоритм нь өгөгдсөн өгөгдлийн олонлог дээр загвараа сургадаг байна. Үнэлгээний үе шатанд ангилагчийн гүйцэтгэлийг шалгах ба гүйцэтгэлийн үнэн зөв байдал (accuracy), алдаа (error), нарийвчлал (precision), санах ой (recall) зэрэг янз бүрийн параметрүүд дээр үндэслэн үнэлсэн болно. К-Хамгийн ойрын хөрш, Нэйви Бэйес, Шийдвэрийн модны алгоритмын ангиллын загвар бий болгох процессыг Зураг IV.2-д дүрслэн харуулж байна. Сургалт болон тестийн олонлогийг үнэлэхдээ k дахин хөндлөн баталгаажуулалтын (k-fold Cross-Validation) аргыг ашигласан ба энэ нь өгөгдлийн олонлогоос дахин түүвэр үүсгэж

өгөгдөл олборлолт болон машин сургалтын аргуудыг үнэлэхэд ашигладаг механизм билээ. k параметр нь өгөгдлийн олонлогийг түүвэр болгон хуваах бүлгүүдийн тоог илэрхийлдэг ба энэхүү шинжилгээгээнд 10-дахин хөндлөн баталгаажуулалтын аргыг ашигласан болно. Энэ нь өгөгдлийн олонлог 10 бүлэг түүвэрт хуваагдсаны дараа k-1 бүлэг буюу 9 бүлэг нь сургалтын олонлог болж үлдсэн бүлэг нь тестийн олонлог болон загварыг үнэлэх явдал юм (An Introduction to Statistical Learning, 2013).

Зураг IV.2 Ангиллын загвар үүсэх процесс



Эх сурвалж: Avinash Naviani, 2018

4.4.1 Нэйви Бэйес алгоритмын хэрэгжүүлэлт

Нэйви Бэйес алгоритм нь нөхцөлт магадлалт ангилагч болхынхоо хувьд төгсөгчдийн хөдөлмөр эрхлэлтийн байдлыг төгсөгчийн ерөнхий мэдээлэл, хөдөлмөр эрхлэлтийн байдал, ажил хайх бэрхшээл, голч зэрэг нийт 12 чанарын болон тоон хувьсагчид дээр тулгуурлан салбар сургууль тус бүрээр ажил эрхлэлтийн байдлын ангиллын загвар үүсгэсэн болно. Эмпирик судалгааны арга зүйн хэсэгт авч үзсэний дагуу Нэйви Бэйес алгоритмыг дараах загвараар хэрэгжүүлсэн билээ.

Хүснэгт IV.7 Үнэлгээнд ашигласан Нэйви Бэйес алгоритмын загвар

Загвар	$\boldsymbol{P(x\|C_i)P(C_i) = P(A_1\|C_i)P(A_2\|C_i)P \dots P(A_{12}\|C_i) = P(C_i)\prod_{j=1}^{12}(A_j\|C_i)}$
Өгөгдлийн олонлог	$x = \{A_1, A_2, \dots, A_{12}\}$
Ангиллын олонлог	$C = \{C_1, C_2, \dots, C_9\}$
Нөхцөлт магадлал	$P(A_1\|C_1), \quad P(A_2\|C_1), \dots, \quad P(A_{12}\|C_1);$ $P(A_1\|C_2), \quad P(A_2\|C_2), \dots, \quad P(A_{12}\|C_2);$ $\dots \quad \dots \quad \dots$ $P(A_{12}\|C_9), \quad P(A_{12}\|C_9), \dots, \quad P(A_{12}\|C_9)$

Эх сурвалж: Судлаачийн тооцоолол

Дээрх загварын хувьд 12 шинж чанар бүхий хувьсагч тус бүрд нөхцөлт магадлал оноох зарчмаар ангилал үүсгэсэн болно. Жишээлбэл, Нягтлан бодогч мэргэжилтэй голч нь 3.4, эрэгтэй, хотод амьдардаг, Бизнесийн сургуульд сурдаг оюутны ажилтай болон ажилгүй байх магадлалаар нөхцөл тавин ангилал үүснэ. Дараах хүснэгтэд энэхүү алгоритмын нөхцөлт магадлалын үр дүнг харуулсан бөгөөд ажилд орох хамгийн магадлал өндөртэй нь C_5 буюу ОУХНУС байна. Харин ажилд орохгүй байх магадлал хамгийн өндөр нь C_4 буюу БС ангилал байна. Сүүлийн хоёр ангиллын хувьд нийт түүврээс бага хувь эзэлж байгаа тул өндөр магадлал гарсан гэж үзэж байна.

Хүснэгт IV.8 Нөхцөлт магадлалын үр дүн

Ангилал		**P**	
		Ажилтай	Ажилгүй
C_1	БУС	0.81	0.19
C_2	НУС	0.88	0.12
C_3	ХУС	0.89	0.11
C_4	БС	0.77	0.23
C_5	ОУХНУС	0.96	0.04
C_6	ХЗС	0.89	0.11
C_7	ХШУИС	0.86	0.14
C_8	ЗС	0.97	0.03
C_9	ЭС	0.97	0.03

Эх сурвалж: Судлаачийн тооцоолол

Ангилагчийн гүйцэтгэлийг үнэлэхэд ашигладаг хэд хэдэн хэмжилтүүд байдаг ба нарийвчлалаас гадна язгуур дундаж квадрат алдаа, ROC зэргийг ашиглана. Хүснэгт IV.8-д энэхүү алгоритмын үүсгэсэн ангиллын загварын нарийвчлал болон Каппа статистик, Дундаж үнэмлэхүй алдаа гэх мэт бусад статистик үнэлгээг харуулж байна.

Хүснэгт IV.9 Нэйви Бэйес ангиллын статистик мэдээлэл

Зөв ангилсан тохиолдол (Correctly Classified Instances)	91.76%
Буруу ангилсан тохиолдол (Incorrectly Classified Instances)	8.24%
Каппа статистик (Kappa statistic)	0.9032
Дундаж үнэмлэхүй алдаа (Mean absolute error)	0.0262
Дундаж квадрат алдааны язгуур (Root mean squared error)	0.1113
Харьцангуй үнэмлэхүй алдаа (Relative absolute error)	13.92%
Харьцангуй квадрат алдааны язгуур (Root relative squared error)	36.29%
Нийт түүврийн хэмжээ	1031

Эх сурвалж: Судлаачийн тооцоолол

Нийт 1031 түүврийг олонлогтой дээрх үнэлгээнээс харвал тус ангиллын загварын зөв ангилсан тохиолдол болон нарийвчлал (accuracy) нь 91.8% буюу 946 тохиолдолийг зөв ангилагдсан буйг илтгэж байна. Зөв ангилсан тохиолдол нь бодит эерэг буюу True

Positive, бодит сөрөг буюу True Negative утгуудын нийлбэр бөгөөд нарийвчлал нь уг утгыг нийт түүврийн тоонд хуваасантай тэнцүү байна. Энэхүү загварын үр дүнд 8.24% буюу 85 тохиолдол нь буруу ангилагдсан боловч 91.8%-ийг зөв ангилсан тул уг ангиллын загварыг үр дүнтэй ангилагч гэж үзнэ. Каппа статистик нь хувьсагчдын үнэлгээний найдвартай байдлыг хэмждэг ба 90.32% буй нь загварын найдвартай байдал сайн, бусад статистик алдааны хэмжүүрүүд нь бага байх тусам загварын нарийвчлал өндөр буйг илэрхийлэх бөгөөд Хүснэгт IV.9-өөс харахад дундаж үнэмлэхүй алдаа буюу MAE утга 0.0262 байна. Мөн түүнчлэн, RMSE буюу дундаж квадрат алдааны язгуур нь 11.13% буюу загварын нарийвчлал сайн гэж үзнэ.

Өмнөх хэсэгт дурдсанаар загварыг хэд хэдэн параметрээр үнэлсэн ба Хүснэгт IV.9-д загварын үнэлгээний мэдээлэл, эдгээр үнэлгээний жинлэсэн дундаж утга харагдаж байна. Уг үр дүнгээс харвал бодит эерэг хувь нь үр дүн эерэг байх үед ажиглалтын эерэг үр дүнг загвар нь урьдчилан таамаглах магадлал бөгөөд 1 утгад дөхөх тусам загвар сайн буйг илтгэнэ. Харин хуурмаг эерэг хувь буюу False Positive нь үр дүн нь сөрөг байх үед ажиглалтын сөрөг үр дүнг загвар нь урьдчилан таамаглах магадлалыг хэлэх бөгөөд 0 утгад дөхөх нь загварын таамаглалыг дэмжих билээ. ХУС болон ХШУИС сургуулийн ангиллын эерэг бодит утга хамгийн их буй нь ангиллын загвар тус сургуулиудын оюутнуудыг амжилттай тодорхойлж байна гэсэн үг юм. Мөн түүнчлэн ангиллын загварын гүйцэтгэлийг харьцуулахдаа нарийвчлал (Precision), санах ой (Recall), F-Хэмжүүр (F-Measure) зэргийг авч үзсэн болно. Дээрх хүснэгтэд, тус ангиллын загвар нь C_1 ангиллыг хамгийн сайн буюу 93% нь зөв ангилагдсан болохыг илэрхийлж байна. Recall буюу эргэн дуудах хувийн утга нь нийт ангиллын дунджаар 92% нь зөв ангилагдсан хувийн утгыг өгч байна.

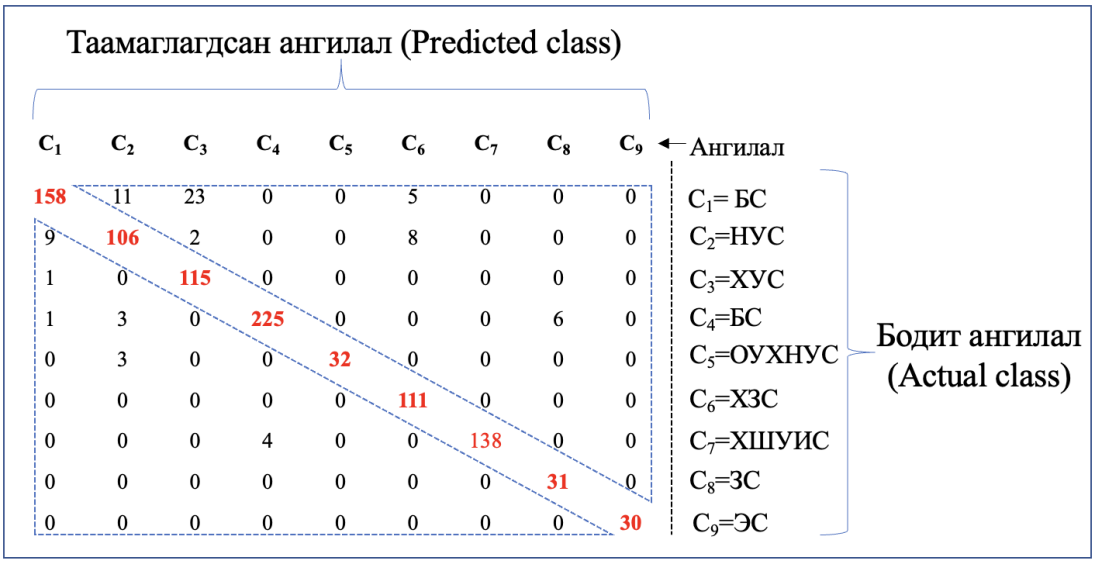
Хүснэгт IV.10 Нэйви Бэйес загварын нарийвчилсан үнэлгээ

	TP rate	FP rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Ангилал
	0.802	0.013	0.935	0.802	0.863	0.838	0.987	0.96	C_1
	0.848	0.019	0.862	0.848	0.855	0.835	0.991	0.932	C_2
	0.991	0.027	0.821	0.991	0.898	0.889	0.998	0.989	C_3
	0.926	0.005	0.983	0.926	0.953	0.94	0.999	0.997	C_4
	0.889	0	1	0.889	0.941	0.941	0.999	0.988	C_5
	1	0.014	0.895	1	0.945	0.939	1	0.997	C_6
	0.972	0	1	0.972	0.986	0.984	1	1	C_7
	1	0.009	0.796	1	0.87	0.873	1	1	C_8
	1	0.006	0.838	1	0.912	0.913	1	1	C_9
Жинлэсэн дундаж	0.918	0.011	0.924	0.918	0.917	0.905	0.996	0.981	

Эх сурвалж: Судлаачийн тооцоолол

Ангиллын алгоритмын гүйцэтгэлийг нэгтгэн дүгнэхэд хамгийн өргөн хэрэглэгддэг арга бол төөрөгдлийн матриц бөгөөд энэхүү матриц нь ангиллын асуудлыг шийдэхэд ашигладаг түгээмэл хэмжүүрүүдийн нэг билээ. Дараах зурагт Нэйви Бэйес алгоритмын төөрөгдлийн матрицыг дүрслэн харуулсан бөгөөд матрицын диагональ хэсэг нь True Positive буюу зөв ангилагдсан боломжуудыг илэрхийлж байна. Тухайлбал, нийт түүврээс 946 нь зөв ангилагдсаныг матрицын диагональ хэсэг харуулж байна.

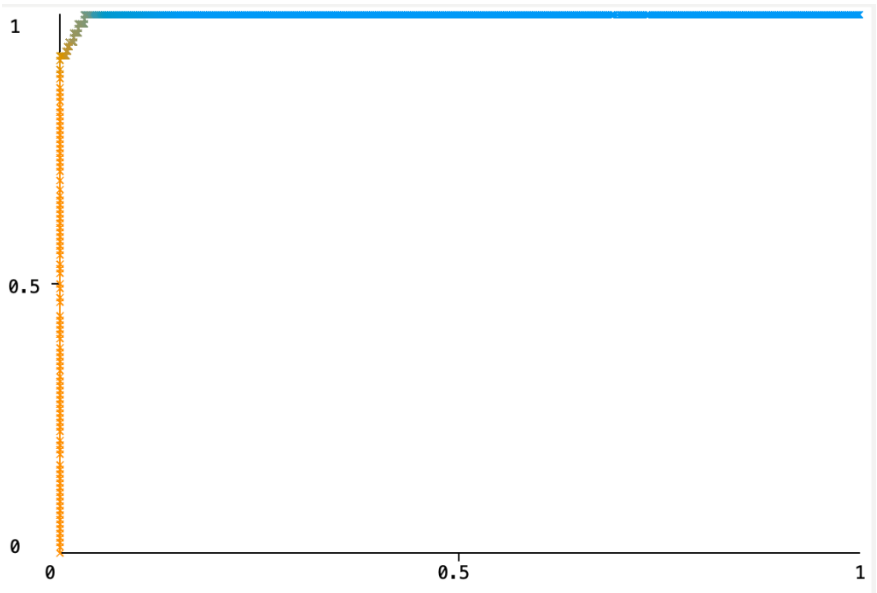
Зураг IV.3 Төөрөгдлийн матриц (Confusion Matrix)



Эх сурвалж: Судлаачийн тооцоолол

Гурвалжин дүрсээр илэрхийлэгдсэн хэсгүүд нь хуурмаг сөрөг (False Negative) болон хуурмаг эерэг (False Positive) утга гэж үзнэ. Энэ нь буруу ангилагдсан түүврийн нийлбэр ба диагоналиас дээш болон доош буй босоо утгууд нь хуурмаг сөрөг утга, харин диагоналиас хоёр тийш хөндлөн тоонууд нь хуурмаг эерэг үр дүн байна. Нийт 1031 түүврээс 946 нь зөв загварчлагдаж, 85 түүвэр буруу ангилагдсан ба ангилал тус бүр дээр хэд нь зөв, хэд нь буруу ангилагдсаныг дээрх матрицаас харах боломжтой.

Зураг IV.4 Reciever Operating Characteristic (ROC) муруй



Эх сурвалж: Судлаачийн тооцоолол

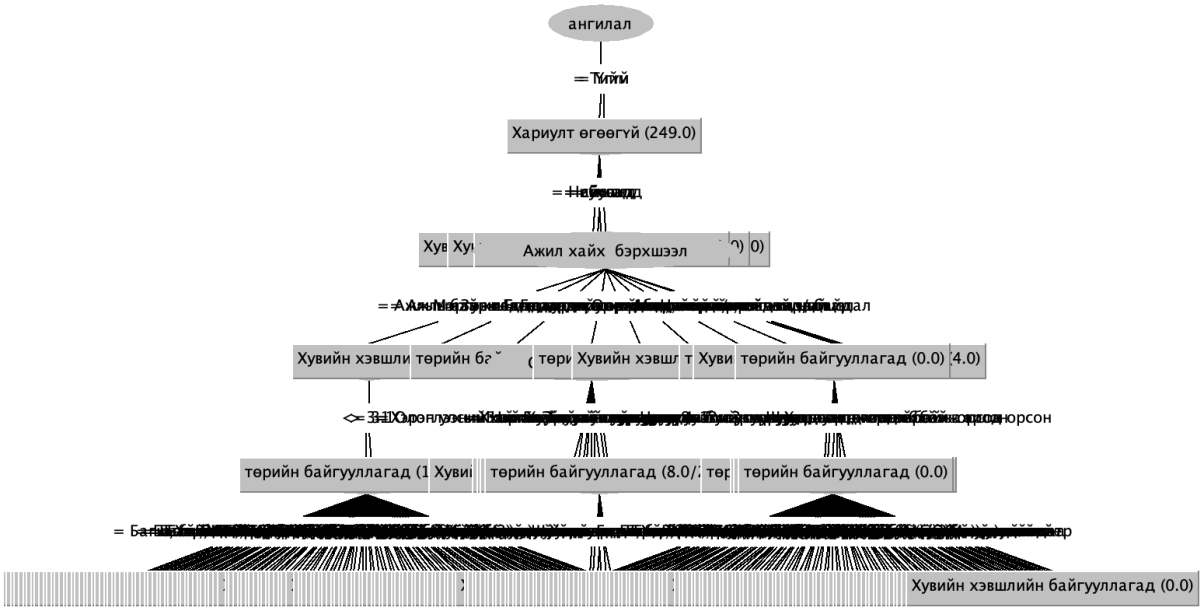
Зураг IV.4-т Нэйви Бэйес алгоритмын үүсгэсэн ангиллыг үнэлэх ROC муруй буюу загварыг үнэлэгч муруйг дүрслэн харуулж байна. Reciever Operating Characteristic (ROC) муруй нь бодит эерэг хувь болон хуурмаг эерэг хувийн утгуудаар тодорхойлогддог ба ROC area=1 байх нь төгс таамаглалыг, харин тус утга 0.5-аас бага буюу тэнцүү байх нь санамсаргүй таамаглагдсаныг илэрхийлдэг байна. Тухайлбал, ROC муруй нь зүүн дээд буланд ойртох тусам ангилал илүү үр дүнтэйг илтгэх ба Нэйви Бэйес алгоритмын хувьд

үүссэн ангиллын ROC муруй =0.996 байна. Хавсралт 6-д ангилал тус бүрд үүсгэсэн ROC муруйг дүрсэлсэн ба загварын үнэлгээний хэсгээс харвал C_1 ангилал хамгийн сул таамаглагдсан буюу 0.987 байна, C_6, C_7, C_8, C_9 зэрэг ангилал нь 1 буюу хамгийн өндөр таамаглагдсан гэж үзнэ. Ангилал тус бүр сайн загварчлагдсан тул бүх салбар сургуулийн оюутнуудад тэдний гүйцэтгэлийг сайжруулах, чадваруудыг нь нэмэх зэргээр зөвөлгөө өгөх боломж үүсч байна. Тиймээс бүхий л салбар сургуулийн оюутнуудын ажил эрхлэлтийн байдлыг урдчилан таамаглах ангилах боломжтой юм.

4.4.2 Шийдвэрийн мод алгоритмын хэрэгжүүлэлт

Шийдвэрийн мод нь шийдвэрийн зангилаа, боломжийн зангилаа болон төгсгөлийн зангилаа гэсэн гурван төрлийн зангилаанаас бүрдсэн шаталсан модны бүтэцтэй билээ. Энэхүү эмпирик шинжилгээгээр нийт 117 навчны зангилаатай модны хэмжээ буюу зангилаа нь 121 бүхий шийдвэрийн мод Зураг IV.5-д дүрсэлсэн байдлаар үүссэн. Шийдвэрийн мод нь мэргэжил, хүйс, хөдөлмөр эрхлэлтийн байдал, голч зэрэг үндсэн 12 хувьсагч бүхий асуултаас бүрдсэн ба төгсгөлийн зангилааг салбар сургууль тус бүрээр дүрсэлсэн байна. Шийдвэрийн зангилаа бүр нь ангилалаар тэмдэглэгдсэн ба хаалтанд байгаа эхний тоо нь тус зангилаанд хувиарлагдсан тохиолдлын тоо, хоёр дох тоо нь алдаатай зангилаанд оноогдсон тохиолдлын тоо байна. Жишээлбэл, Эрдэнэт сургуулийн хувьд 31 тохиолдлоос, 0 тохиолдол буруу ангилагдсан буйг илтгэнэ. Шийдвэрийн мод алгоритмын дэлгэрэнгүй дүрслэлийг Хавсралт 4-д оруулсан болно.

Зураг IV.5 Шийдвэрийн мод



Эх сурвалж: Судлаачийн тооцоолол

Өгөгдөл олборлолтын ангиллын загварыг үнэлэх параметрүүд нь ижил байх ба энэхүү хэсэгт Шийдвэрийн мод алгоритмын үүсгэсэн ангиллын загварын нарийвчлал болон Каппа статистик, Дундаж үнэмлэхүй алдаа гэх мэт бусад статистик үнэлгээг харуулж байна. Шийдвэрийн модны ангиллын загвар маань 99.612% нарийвчлалтай буюу 1027 тохиолол зөв ангилагджээ.

Хүснэгт IV.11 Шийдвэрийн мод ангиллын статистик мэдээлэл

Зөв ангилсан тохиолдол (Correctly Classified Instances)	99.612%
Буруу ангилсан тохиолдол (Incorrectly Classified Instances)	0.388%
Каппа статистик (Kappa statistic)	0.9954
Дундаж үнэмлэхүй алдаа (Mean absolute error)	0.0008
Дундаж квадрат алдааны язгуур (Root mean squared error)	0.0223
Харьцангуй үнэмлэхүй алдаа (Relative absolute error)	44.96%
Харьцангуй квадрат алдааны язгуур (Root relative squared error)	7.28%
Нийт түүврийн хэмжээ	1031

Эх сурвалж: Судлаачийн тооцоолол

Энэхүү загварын үр дүнд 0.38% буюу 4 тохиолдол нь буруу ангилагдсан тул уг ангиллын загварыг маш сайн ангилагч гэж үзнэ. Тиймээс ангиллын загвар дээр үндэслэн салбар сургууль бүрийн оюутнуудын хөдөлмөр эрхлэлт, суралцах хугацааны гүйцэтгэлийг тэдэнд зөвлөх болон бусад замаар сайжруулж болохоор харагдаж байна. Каппа статистик нь хувьсагчдын үнэлгээний найдвартай байдлыг хэмждэг ба 99.54% буй нь загварын найдвартай байдал сайн байна. Харин статистик алдааны хэмжүүрүүд нь бага байх тусам загварын найдвартай байдал өндөр буйг илэрхийлнэ. Хүснэгт IV.11-өөс харахад дундаж үнэмлэхүй алдаа буюу MAE утга 0.0008 байна. RMSE буюу дундаж квадрат алдааны язгуур нь 2.23% буюу 10%-аас бага тул загварын нарийвчлал маш сайн гэж үзнэ.

Хүснэгт IV.12 Шийдвэрийн мод алгоритмын загварын нарийвчилсан үнэлгээ

	TP rate	FP rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Ангилал
	0.995	0.004	0.985	0.995	0.990	0.988	0.987	0.96	C_1
	0.984	0	1	0.984	0.992	0.991	0.991	1	C_2
	0.991	0.001	0.991	0.991	0.991	0.990	0.990	0.991	C_3
	1	0	1	1	1	0.990	1	1	C_4
	1	0	1	1	1	1	1	1	C_5
	1	0	1	1	1	1	1	1	C_6
	1	0	1	1	1	1	1	1	C_7
	1	0	1	1	1	1	1	1	C_8
	1	0	1	1	1	1	1	1	C_9
Жинлэсэн дундаж	0.996	0.001	0.996	0.996	0.996	0.995	0.996	0.998	

Эх сурвалж: Судлаачийн тооцоолол

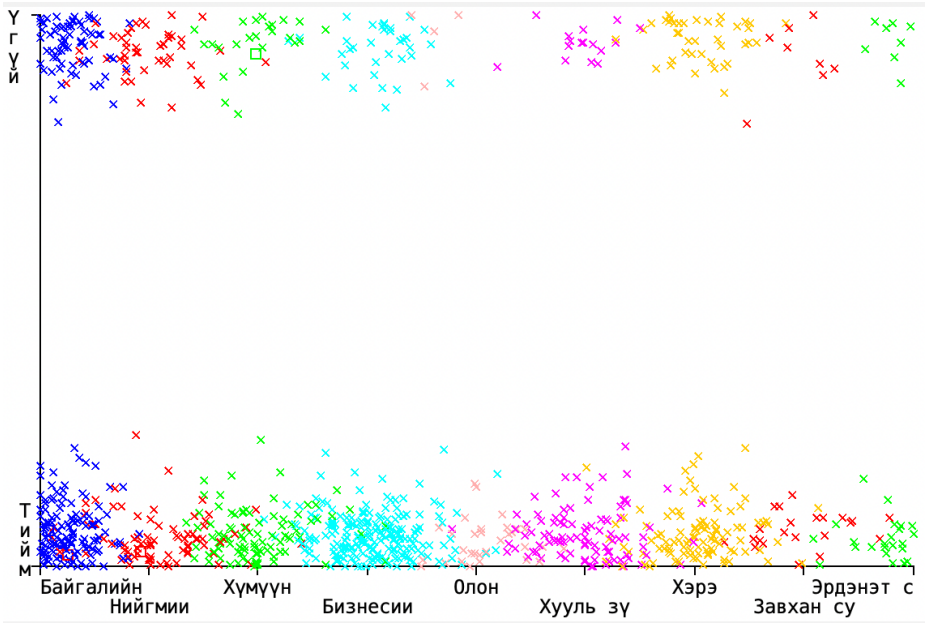
Хүснэгт IV.12-д Шийдвэрийн мод алгоритмын нарийвчилсан үнэлгээг харуулж байна. Бодит эерэг хувь нь 0.99 ба үр дүн эерэг байх үед ажиглалтын эерэг үр дүнг загвар нь урьдчилан таамаглах магадлал бөгөөд 1 утгад дөхсөн тул загвар сайн буйг илтгэнэ.

Харин хуурмаг эерэг хувь буюу False Positive нь үр дүн нь сөрөг байх үед ажиглалтын сөрөг үр дүнг загвар нь урьдчилан таамаглах магадлал бөгөөд 0.001 буй нь 0 утгад дөхсөн тул загварын таамаглалыг дэмжих билээ. Мөн өмнөх ангиллын загвартай адил

гүйцэтгэлийг харьцуулахдаа нарийвчлал (Precision), санах ой (Recall), F-Хэмжүүр (F-Measure) зэргийг авч үзсэн болно. Шийдвэрийн модны ангиллын загвар нь ангиллыг 99.6% нь зөв ангилагдсан болохыг илэрхийлж байна. Recall буюу эргэн дуудах хувийн утга нь нийт ангиллын дунджаар 99.6% нь зөв ангилагдсан гэх хувийн утгыг өгч байна.

Дараах зурагт салбар сургууль тус бүр дээр ажилтай болон ажилгүй ангиллын үнэлгээг харуулж байна. Ингэхдээ, хэвтээ тэнхлэгт салбар сургууль, босоо тэнхлэгт хөдөлмөр эрхлэлтийн байдал бөгөөд дөрвөлжин дүрсээр буруу ангилагдсан утга илэрхийлэгдэж байгаа бол зөв ангилагдсан утга нь х тэмдгээр дүрсэлсэн болно. Нийт төгсөгчдийн дийлэнх хувь нь ажилтай ба зөв таамаглагдсан байгаа тул ангилал бүрийн сурагчдад зөвлөгөө өгөх, гүйцэтгэлийг сайжруулах боломжтой гэж харж байна.

Зураг IV.6 Ангиллын загварын алдааны график



Эх сурвалж: Судлаачийн тооцоолол

Өгөгдөл олборлолтын эцсийн зорилго нь тодорхой бус байдлыг багасгах явдал ба үүнийг Шийвэрийн мод алгоритмын хувьд Энтроп H(S) болон Мэдээллийн хожоо IG(S,A) зэрэг хэмжүүрүүдээр хэмждэг билээ. Энэхүү алгоритмын тодорхой бус байдлын хэмжүүрийг дараах байдлаар тодорхойлсон болно (Энтропи болон Мэдээллийн хожооны дэлгэрэнгүй бодолтийг Хавсралт 3-д оруулсан). Судлаачийн тооцооллын хувьд энтроп утга 0.76 байгаа бөгөөд эх олонлогийн санамсаргүй байдал өндөр гэж үзнэ. Учир нь шийдвэрийн мод алгоритм нь энтроп 0-1 хооронд утга авах бөгөөд энтроп хамгийн бага байхыг илүүд үздэг билээ. Харин мэдээллийн хожоо нь 0.03 байсан энтропийн зөрүү маш бага байгаа нь харагдаж байна.

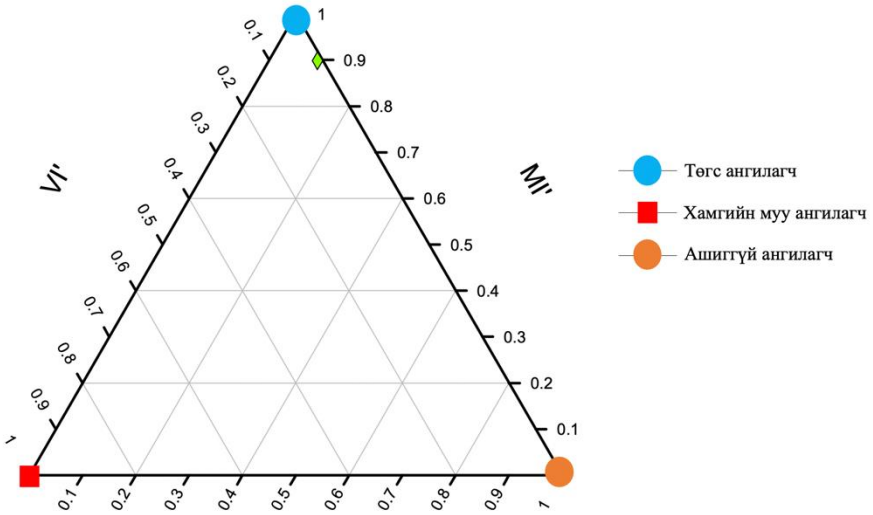
Хүснэгт IV.13 Энтроп ба Мэдээллийн хожоо

Энтроп	$H(S) = \sum_{x \in X} -p(x) log_2 p(x)$
Н(Хөдөлмөрийн байдал)	0.80
Н(Хөдөлмөрийн байдал|БУС)	0.883
Н(Хөдөлмөрийн байдал|НУС)	0.914
Н(Хөдөлмөрийн байдал|ХУС)	0.931
Н(Хөдөлмөрийн байдал|БС)	0.579
Н(Хөдөлмөрийн байдал|ОУХНУС)	0.503
Н(Хөдөлмөрийн байдал|ХЗС)	0.594
Н(Хөдөлмөрийн байдал|ХШУИС)	0.858
Н(Хөдөлмөрийн байдал|ЗС)	0.849
Н(Хөдөлмөрийн байдал|ЭС)	0.787
Жинлэсэн дундаж Н(S)	0.769
Мэдээллийн хожоо	$IG(S,A) = H(S) - \sum_{t \in T} p(t)H(t) = H(S) - H(S\|A)$
Н(Хөдөлмөрийн байдал)-Н(Хөдөлмөрийн байдал|Салбар сургууль)	0.030

Эх сурвалж: Судлаачийн тооцоолол

Өгөгдөл олборлолтын алгоритм нь энтропийн утгыг гурвалжин графикаар хэмждэг бөгөөд Зураг IV.7-д энтропийн гурвалжинг дүрслэн харууллаа. Өөрөр хэлбэл, Энтроп болон мэдээллийн хожооны утгыг бататгахын тулд дараах гурвалжингаар хэмжих бөгөөд энтропи гурвалжин нь ангиллын загварын сайн эсвэл муу ангилагч хэмээн үнэлдэг байна.

Зураг IV.7 Энтроп гурвалжин



Эх сурвалж: Судлаачийн тооцоолол

Дээрх гурвалжингийн оройнууд нь ангиллын загварын үнэлгээг өгдөг хэсэг ба энэхүү эмпирик шинжилгээний хувьд ангиллын загвар нь төгс ангилагч гэсэн бүлэгт багтаж

байна. Тиймээс шийдвэрийн мод алгоритм нь төгсөгчдийн хөдөлмөр эрхлэлтийн байдлыг загварчилхад сайн тохирно гэж үзэж болохоор байна.

4.4.3 К-Хамгийн ойрын хөрш аргын хэрэгжүүлэлт

К-Хамгийн ойрын хөрш алгоритмаар ангиллын загвар үүсгэхдээ сургалтын олонлог болон тестийн олонлог хоорондын зайг Манхаттэн зайны функц дээр үндэслэн тооцоолж эцсийн ангиллын үр дүнг тодорхойлсон билээ. Хүснэгт IV.14-д К-Хамгийн ойрын хөрш алгоритмын үүсгэсэн ангиллын загварын үнэлгээнүүдийг ба к утга бүрд харьцуулан харуулж байна. Хамгийн сайн k утгыг сонгох хэд хэдэн шалгууруудыг харгалзаж үзсэн болно. Үүнд: k=1 байх нь хамгийн ойрын хөрштэй ижил ангид хувиарлагдах ба хазайлт бага, мөн ангилалд хамаарагдаагүй байхын тулд k утга сондгой байх хэрэгтэй билээ. Мөн түүнчлэн k=1 болон 5, 11 гэж сонгосон ба тестийн олонлогийн ангиллыг тодорхойлохдоо сургалтын олонлогт хамгийн ойрхон эхний хөршүүдээр ангиллыг үүсгэсэн гэж үзнэ.

Хүснэгт IV.14 К-Хамгийн ойрын хөрш ангиллын статистик мэдээлэл

Манхаттэн зай	**K=1**	**K=5**	**K=11**
Зөв ангилсан тохиолдол (Correctly Classified Instances)	98.64%	90.49%	85.06%
Буруу ангилсан тохиолдол (Incorrectly Classified Instances)	1.35%	9.5%	14.93%
Каппа статистик (Kappa statistic)	0.984	0.887	0.829
Дундаж үнэмлэхүй алдаа (Mean absolute error)	0.004	0.0254	0.0424
Дундаж квадрат алдааны язгуур (Root mean squared error)	0.0454	0.1156	0.1452
Харьцангуй үнэмлэхүй алдаа (Relative absolute error)	2.12%	13.49%	22.5%
Харьцангуй квадрат алдааны язгуур (Root relative squared error)	14.7%	37.6%	47.31%
Зөв таамагласан түүврийн хэмжээ	1017	933	877

Эх сурвалж: Судлаачийн тооцоолол

К-Хамгийн ойрын хөрш алгоритмын загварын үр дүнд k=1 үед 98.64% буюу 1017 тохиолдол нь зөв, 1.35% буюу 14 тохиолдол буруу ангилагдсан байна. Тиймээс уг загварын гүйцэтгэлийг сайн гэж үзнэ. k утга 5 болон 11 үед зөв ангилагдсан тоо харьцангуй бага байна. k=1 үед алгоритм хамгийн сайн ажиллаж байгаа тул үнэлгээг нарийвчлан үнэлэхдээ 1 хөрштэй үеийн үр дүнг үзүүлэх болно. Учир нь тухайн сургуулийн оюутнуудын хөдөлмөр эрхлэлтийн байдлыг салбар сургууль тус бүрээр амжилттай ангилж байгаа юм. Энэхүү үр дүнд тулгуурлан ирээдүйн хөдөлмөр эрхлэлтийн байдлыг урдчилан таамаглах боломжтой болох бөгөөд салбар сургууль тус бүрээр үр дүнг сайжруулах боломж бүрдэж буй. Дээрх үнэлгээний хүснэгтээс, k=1 үед Каппа статистик буюу хувьсагчдын үнэлгээний найдвартай байдал нь 98.4% байна. Харин статистик алдааны параметрүүд нь бага бөгөөд үүнээс загварын найдвартай байдал өндөр буйг илэрхийлж буй билээ. Дундаж квадрат алдааны язгуур RMSE нь 4.5% болон Дундаж үнэмлэхүй алдаа MAE нь 0.4% тул загварын нарийвчлал маш сайн гэж үзнэ.

Хүснэгт IV.15 К- Хамгийн ойрын хөрш алгоритмын загварын нарийвчилсан үнэлгээ

	TP rate	FP rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Ангилал
	0.98	0.006	0.975	0.980	0.977	0.972	0.993	0.992	C_1
	0.96	0.002	0.984	0.960	0.972	0.968	0.996	0.992	C_2
	0.974	0.004	0.966	0.974	0.970	0.966	0.998	0.981	C_3
	1	0.001	0.996	1	0.998	0.997	0.999	0.990	C_4
	0.94	0.001	0.971	0.944	0.958	0.956	0.958	0.917	C_5
	1	0.001	0.991	1	0.996	0.995	1	1	C_6
	1	0	1	1	1	1	1	1	C_7
	1	0	1	1	1	1	1	1	C_8
	1	0	1	1	1	1	1	1	C_9
Жинлэсэн дундаж	0.986	0.002	0.986	0.918	0.986	0.984	0.996	0.990	

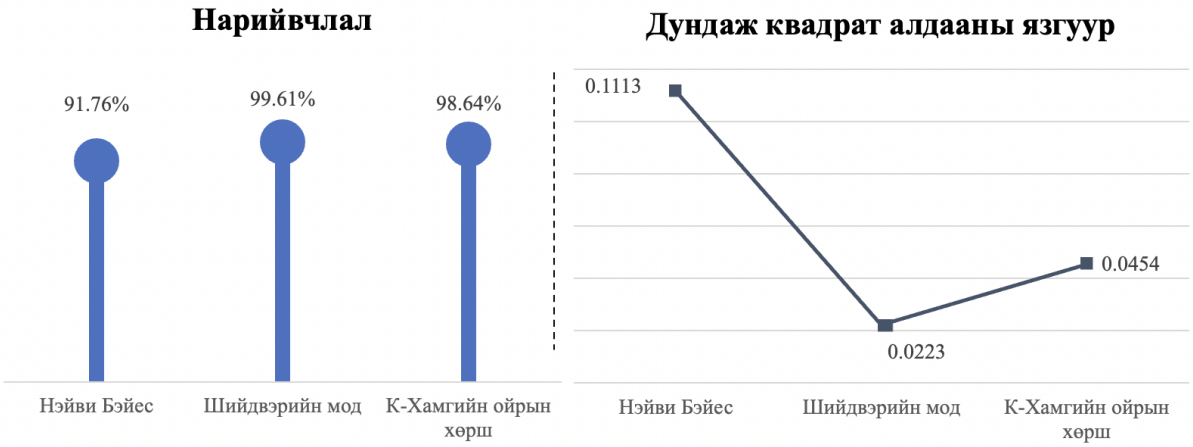
Эх сурвалж: Судлаачийн тооцоолол

Хүснэгт IV.14-д К-Хамгийн ойрын хөрш алгоритмын загварын нарийвчилсан үнэлгээг ангилал тус бүрээр харуулж байна. Бодит эерэг хувь нь 0.98 ба үр дүн эерэг байх үед ажиглалтын эерэг үр дүнг загвар нь урьдчилан таамаглах магадлал бөгөөд 1 утгад дөхсөн тул загвар сайн буйг илтгэнэ. Харин хуурмаг эерэг хувь буюу False Positive нь үр дүн нь сөрөг байх үед ажиглалтын сөрөг үр дүнг загвар нь урьдчилан таамаглах магадлал бөгөөд 0.002 буй нь 0 утгад дөхсөн тул загварын таамаглалыг дэмжинэ гэж үзнэ. Мөн өмнөх ангиллын загвартай адил гүйцэтгэлийг харьцуулахдаа нарийвчлал (Precision), санах ой (Recall), F-Хэмжүүр (F-Measure) зэргийг авч үзсэн болно. Шийдвэрийн модны ангиллын загвар нь ангиллыг 98.6% нь зөв ангилагдсан болохыг илэрхийлж байна. Recall буюу эргэн дуудах хувийн утга нь нийт ангиллын дунджаар 91.8% нь зөв ангилагдсан гэх хувийн утгыг өгч байна.

4.5 Үр дүнгийн харьцуулалт

Энэхүү хэсэгт өгөгдөл олборлолтын ангиллын загваруудын үр дүнг харьцуулах бөгөөд ингэхдээ нарийвчлал буюу зөв таамаглагдсан утга болон дундаж квадрат алдааны язгууpаар харьцуулах болно. Ангиллын алгоритмыг хооронд нь харьцуулах нь тухайн чиглэлийн судалгааны ажилд хамгийн тохиромжтой ангиллын алгоритмыг тодорхойлох зорилготой ба өөрөөр хэлбэл D хэмжээтэй өгөгдлийн олонлогт аль алгоритм нь илүү нарийвчлалтай ангилал үүсгэх боломжтойг харуулах явдал юм. Уг үр дүнг бий болгохын тулд хөндлөн баталгаажуулалтын аргыг ашиглан баталгаажуулдаг (Tiago,2021), өмнөх хэсэгт дурдсанаар тус эмпирик ажилд 10 дахин хөндлөн баталгаажуулалтын аргаар судалгааны үндсэн арга зүйг хэрэгжүүлсэн билээ. Зураг IV.8-д К-Хамгийн ойрын хөрш, Шийдвэрийн мод, Нэйви Бэйес алгоритмуудын нарийвчлал болон дундаж квадрат алдааны язгуур (RMSE)-ийг дүрслэн харуулж байна.

Зураг IV.8 Ангиллын алгоритмуудын үр дүнгийн харьцуулалт



Эх сурвалж: Судлаачийн тооцоолол

Уг судалгааны ажлаар төгсөгчийн ерөнхий мэдээлэл, хөдөлмөр эрхлэлтийн байдал, ажил хайх бэрхшээл, голч зэрэг нийт 12 чанарын болон тоон хувьсагчид дээр тулгуурлан салбар сургууль тус бүрээр ажил эрхлэлтийн байдлын ангиллын загвар үүсгэсэн билээ. Мөн төгсөгчдийн ажил эрхлэлтэд нөлөөлж буй шинж чанаруудыг тодорхойлох нь чухал асуудал байсан ба МУИС-ийг 2018-2019 онд төгссөн оюутнуудын бодит мэдээлэлд тулгуурлан төгсөгч ажилтай байсан эсэх, ажилгүй хэвээр байгааг тус сургуулийн салбар сургууль тус бүрээр ангилан, дээрх алгоритмуудыг ашиглан ангиллын загвар үүсгэсэн болно. Үр дүнгээс харахад: Нарийвчлалын хувьд 100% байвал төгс ангилагдсан гэж үзэх бөгөөд Нэйви Бэйес 91.76%, К-Хамгийн ойрын хөрш 98.64%, Шийдвэрийн мод алгоритмын ангиллын загвар 99.61% буюу хамгийн өндөр нарийвчлалтай байна. Тиймээс Шийдвэрийн модны алгоритм нь төгсөгчдийн хөдөлмөр эрхлэлтийн байдлыг урьдчилан таамаглахад илүү тохиромжтой гэж үзнэ. Учир нь шийдвэрийн модны алгоритм нь өндөр нарийвчлалтайгаар (discriminative model) ангилал үүсгэх чадвартай буюу илүү уян хатан, хэрэгжүүлэхэд хялбар тул илүү давуу талтай, харин бусад алгоритм нь үүсгэгч загвар (generative model) ажээ (Danny.V, 2018). Зургийн баруун талд эдгээр алгоритмуудын дундаж квадрат алдааны язгуурыг (RMSE) харуулж байна. Тус утга нь

бага байвал илүү сайн таамаглал гарах билээ. Тухайлбал, RMSE буюу дундаж квадрат алдааны язгуур нь загварын нарийвчлалыг үнэлэхдээ: RMSE < 10% үед маш сайн, RMSE 10% - 20% хооронд сайн, RMSE 20% - 30% үед боломжийн, RMSE > 30% үед муу байна. Энэхүү эмпирик шинжилгээний үр дүнд, Нэйви Бэйес 11%, Шийдвэрийн мод 2%, К-Хамгийн ойрын хөрш 4.5%-ийн алдаатай ангиллын загвар үүсгэжээ. Уг үнэлгээнд үндэслэн харвал Шийдвэрийн мод алгоритм нь мөн адил бусад аргуудаас илүү сайн таамаглал гаргасан байгаа юм. Тиймээс төгсөгчдийн хөдөлмөр эрхлэлтийн байдлыг урдчилан таамаглах судалгааны ажилд хамгийн тохиромжтой өгөгдөл олборлолтын алгоритм нь Шийдвэрийн мод буюу Decision Tree алгоритм байна.

ДҮГНЭЛТ, САНАЛ

Дэлхий даяар 2021 оны байдлаар 235 сая орчим оюутан их, дээд сургуульд суралцаж байна (UNESCO, 2021). Харин Монгол Улсад 2021 оны байдлаар нийт төгсөгчдийн тоо 58343 хүн буйгаас 79% ажилтай, 1.5% ажилгүй, 20% нь ажиллах хүчнээс гадуур байгаа тул дийлэнх нь ажил хайх итгэлээ алдсан гэж үзнэ. Уг статистик нь жил бүр их, дээд сургууль, боловсролын байгууллага төгсөгчдийн тоо нэмэгдэж буйг илтгэх бөгөөд төгсөгчдийг ажлын байраар хангах, хөдөлмөрийн зах зээлийг тэнцвэржүүлэх нь үндэсний хэмжээний асуудал болж байгааг харуулж байна. Тиймээс өнөөгийн хөдөлмөрийн зах зээлийн өөрчлөлтийн үйл явцтай уялдуулан төгсөгчдийн хөдөлмөр эрхлэлтийн төлөв байдлыг гаргах, ажилгүйдлийн шалтгаан, нөхцөлийг тодорхойлох, их сургуулиас бэлтгэгдэж буй мэргэжилтний ур чадвар хөдөлмөрийн зах зээлд нийцэж буй эсэх, төгсөгчдийн хөдөлмөр эрхлэлтэд тулгамдаж буй хүндрэл бэрхшээл зэргийг тодорхойлох шаардлагатай юм.

Тус судалгааны ажлын хүрээнд түүврийн хэмжээ хангалттай эсэх зэргийг харгалзан үзэж МУИС-ийн Оюутан, төгсөгчийн хэлтсээс явуулдаг “Төгсөгчдийн хөдөлмөр эрхлэлтийн судалгаа”-ийн өгөгдөлийг ашигласан болно. Тус судалгаа нь 2018-2019 оны хичээлийн жилд төгссөн 1031 оюутныг хамруулсан бөгөөд шинжилгээг хийхдээ хүчин зүйлсийн хамаарлын шинжилгээ буюу корреляцын шинжилгээг хийж, өгөгдөл олборлолтын ангиллын алгоритмууд болох К-Хамгийн ойрын хөрш, Нэйви Бэйес, Шийдвэрийн модны аргуудыг ашиглан загвараа үнэлсэн. Тухайлбал, ажил эрхлэлтийн онцлог, орлогын ялгаатай байдал, ажилгүй байгаа шалтгаан, сурлагын голч, ажилгүй байгаа шалтгаан зэргийг хүйсийн онцлогтой харьцуулан, дүн шинжилгээ хийсэн болно. Судалгаанд хамрагдсан төгсөгчдийг хүйсийн хувьд авч үзэхэд 67% нь эмэгтэй, 33% нь эрэгтэй бөгөөд нийт эрэгтэйн 81.3% нь ажилтай бол нийт эмэгтэйн 72.3% нь ажил эрхэлж байна. Бүх салбар сургуулиудын голч дүнгийн дундаж нь 2.80 ба нийт төгсөгчдийн 63.9% нь эзэмшисэн мэргэжилээрээ ажиллаж байна. Судалгаанд хамрагдсан төгсөгчдийн дундаж цалин ₮1,065,175.72 бөгөөд ХНХСИ-с эрхлэн гаргадаг мөн үеийн Монгол Улсын нийт төгсөгчдийн дундаж орлогтой харьцангуй ялгаатай байв. Төгсөгчдийн сарын дундаж орлогыг эрэгтэй, эмэгтэй хүйсийн хувьд харьцуулж үзэхэд бага зэрэг ялгаа харагдаж байсан ба эмэгтэйчүүдийн хувьд дундаж орлого ₮968.5 мянга бол эрэгтэйчүүд сард дунджаар ₮1255.4 мянгын орлого сард олдог буюу ₮286.9 мянгаар илүү орлоготой байна.

Мөн корреляцын шинжилгээний үр дүнгээс үзэхэд голч оноо ба ажил эрхлэлт хооронд хамааралтай буюу 0.9 байсан ба төгсөгчийн голч нь ажил эрхлэхтэй өндөр хамааралтайг илтгэж байна. Үүнээс сурлагын голч дүн ямар нэг байдлаар ажил хөдөлмөр эрхлэхэд нөлөө үзүүлдэг гэж үзнэ. Харин голч оноо бага байх тусам ажил эрхлэлтийн хувь хэмжээ буурахад нөлөө үзүүлдэг гэсэн хамаарал ажиглагдсан. Мөн тухайн төгсөгчийн мэргэжил ямар газар ажиллахтай эерэг хамааралтай буюу 0.6 байв. Түүнчлэн, төгсөгчдийн 63.2% нь нийслэлд амьдарч байгаа ба амьдарч буй байршлаас шалтгаалан орлогын түвшин ялгаатайгаас гадна ихэнх ажлын байр нийслэлд байдагтай холбоотой байна.

Судалгааны үндсэн арга зүйн хэрэгжүүлэлт хэсгийн үр дүнд Нэйви Бэйес алгоритмын нарийвчлал 91.76%, К-Хамгийн ойрын хөрш алгоритмын нарийвчлал 98.64%,

Шийдвэрийн мод алгоритмын ангиллын загвар 99.61% буюу хамгийн өндөр нарийвчлалтай байна. Тиймээс Шийдвэрийн модны алгоритм нь төгсөгчдийн хөдөлмөр эрхлэлтийн байдлыг урьдчилан таамаглахад илүү тохиромжтой гэж үзнэ. Үүний шалтгааныг уг алгоритмын давуу талууд болох илүү уян хатан, хэрэгжүүлэхэд хялбар зэргээр тайлбарлаж болох ба өмнө нь сонгогдоогүй байгаа шинж чанаруудыг харгалзан дэд олонлог бүр дээр давтагддаг алгоритмын ажиллах зарчимтай хамааралтай билээ.

Хэдийгээр, төгсөгчдөөс хангалттай мэдээлэл цуглуулж, дүн шинжилгээ хийдэг, цаашдын үйл ажиллагаанд ашиглаж буй их, дээд сургууль манай улсын хэмжээнд цөөн, мөн оюутны хөдөлмөр эрхлэлт, ахиц дэвшилд дүн шинжилгээ хийх, хянах тогтолцоо манай улсад байхгүй боловч хөдөлмөр эрхлэлтийн талаар баримтлах бодлогын 1.3.6-д энэ талаар тусгасанаас гадна 2021 онд БСШУЯ нь ХНХСИ-тэй хамтран “Төгсөгчдийн хөдөлмөр эрхлэлтийн судалгаа”-ны платформыг үүсгэсэн нь том ахиц дэвшил байв. Тиймээс цаашид энэ төрлийн судалгаа хийхэд тоон мэдээллээр хангагдах нөхцөл бүрдэж буй ба өгөгдөл олборлолтын арга ашиглан оновчтой шийдвэр гаргах, өгөгдлийг олборлон хэрэгтэй мэдээлэл үүсгэх боломж бүрдэж буйг илтгэж байна. Учир нь их, дээд сургуулиуд төгсөгчдөө хөдөлмөрийн зах зээлд нэвтрэхэд шаардлагатай чадваруудыг сайжруулж, ажлын байрны тодорхой эрэлт хэрэгцээг хангах, түүнд нийцэхүйц төгсөгч бэлтгэх нь чухал асуудал юм.

НОМ ЗҮЙ

Монгол хэл дээр хийгдсэн судалгаа:

Баттүшиг, П. Н. (2017). *Өгөгдөл олборлолт, түгээмэл хэрэглээ.* Улаанбаатар.

Сэлэнгэ. М, Э. (2018). *Оюутны хөдөлмөр эрхлэлт.* МИС-МТИС.

Чимгээ. Д (2020). *Мэдээллийн технологийн хөгжлийн ойрын ирээдүйн төлөв: BIG DATA. Улаанбаатар.*

Энхтуул.Б. (2018). *Үнэт цаасны багц сонголт хийхэд өгөгдөл олборлолтын аргуудыг ашиглах нь.* Улаанбаатар: Хэрэглээний стандарт сэтгүүл.

Энхтуул.Б. (2020). *Үнэт цаасны багц сонголт хийхэд өгөгдөл олборлолтын аргуудыг ашиглах нь.* Хэрэглээний стандарт сэтгүүл.

Гадаад хэл дээр хийгдсэн судалгаа:

Abdul, A. B. (2016). *Supervised and Unsupervised Learning in Data Mining for Employment Prediction of Fresh Graduate Students.* Perak Malaysia: Sultan Idris Education University.

Aziz, M. T. (2016). *Graduates employment classification using data mining approach.* AIP Conference Proceedings.

A.Paidi. (2012). *Data Mining : Future Trends and Applications.* Computer Science.

Bangsuk Jantawan. (2013). *The Application of Data Mining to Build Classification Model for Predicting Graduate Employment.* Thailand: International Journal of Computer Science and Information Security.

Bulao, J. (2022 оны 1 4). *25+ Impressive Big data statistics for 2021*. Techjury.net: https://techjury.net/blog/big-data-statistics/

Dorina Kabakchieva. (2012). *Student Performance Prediction by Using Data Mining Classification Algorithms.* International Journal of Computer Science and Management Research.

Jour, Yang Aimin, Wei Zhang. (2020). *Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA.* Frontiers in Bioengineering and Biotechnology.

J.Han, J. M. (2011). *Data mining: concepts and techniques.* Elsevier.

J.R.Quinlan. (1986). *ID3 algorithm.* wikipedia: https://en.wikipedia.org/wiki/ID3_algorithm

Kalpesh Adhatrao, A. G. (2013). *PREDICTING STUDENTS' PERFORMANCE USING ID3 AND C4.5 CLASSIFICATION ALGORITHM S. Navi Mumbai, Maharashtra, India: Department of Computer Engineering*

Lin, W. X. (2008). *College student employment data platform based on FPGA and machine learning.* Elsevier.

Li, Y. (2021). *Analysis of student employment prediction based on decision tree C4.5 algorithm* . Scientific Journal of Intelligent Systems Research.

Leticia Beana-Ruiz, Laura Garach. (2014). *Using Data Mining Techniques to Road Safety Improvement in Spanish Roads.* Procedia - Social and Behavioral Sciences.

Othman, Z. (2017). *Classification Techniques for Predicting Graduate Employability.* International Journal on Advanced Science Engineering information technology.

Micheal P.Todaro, S. C. (2020). *Economomic development. United Kingdom*: Pearson.

Pauzi, W. N. (2019). *Supervised and Unsupervised Data Mining Techniques on Employability of Public Higher Learning Institute Graduates in Malaysia*. ournal of Physics : Conference Series 2084:1, 012004.

Paul Zikopoulos, C. E. (2011*). Understanding Big Data: Analytics for Enterprice class Hadoop and streaming data. McGraw-Hill Obsorne Media*.

Shahiri, A. M. (2020). *The Third Information Systems International Conference A Review on Predicting Student's Performance using Data Mining Techniques.* Procedia computer science.

Subramanian, D. (2012). *A Simple Introduction to K-Nearest Neighbors Algorithm*. https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-

Wei Peng, J. C.*An Implementation of ID3 --- Decision Tree Learning Algorithm*. Sydney Australia: University of New South Wales, School of Computer Science

Yingying Wang, Y. L. (2017). *Improvement of ID3 Algorithm Based on Simplified Information Entropy and Coordination Degree*. MDPI.

Zulfany Erlisa Rasjid, R. S. (2017). *Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques.*

Zhao Daping , Fang Yong Zhang, Chaoliang Wang, Zongrun. (2019). *Portfolio Selection Based on Bayesian Theory.* Mathematical Problems in Engineering.

ХАВСРАЛТ

Хавсралт 1.

Судлагдсан байдлын оюуны зураглал

Эх сурвалж: Судлаачийн тооцоолол

Хавсралт 2. Шийдвэрийн мод алгоритмын тодорхой бус байдлын тооцоололын хүснэгт

Тэмдэглэгээ		E Ажилтай	U Ажилгүй	N Түүвэр	
E1	Байгалийн ухааны сургууль	134	58	192	N1
E2	Нийгмийн ухааны сургууль	59	29	88	N2
E3	Хүмүүнлэгийн ухааны сургууль	107	57	164	N3
E4	Бизнесийн сургууль	212	34	246	N4
E5	Олон улсын харилцаа, нийтийн удирдлагын сургууль	32	4	36	N5
E6	Хууль зүйн сургууль	95	16	111	N6
E7	Хэрэглээний шинжлэх ухаан, инженерчлэлийн сургууль	94	37	131	N7
E8	Завхан сургууль	21	8	29	N8
E9	Эрдэнэт сургууль	26	8	34	N9
		780	251	1031	N

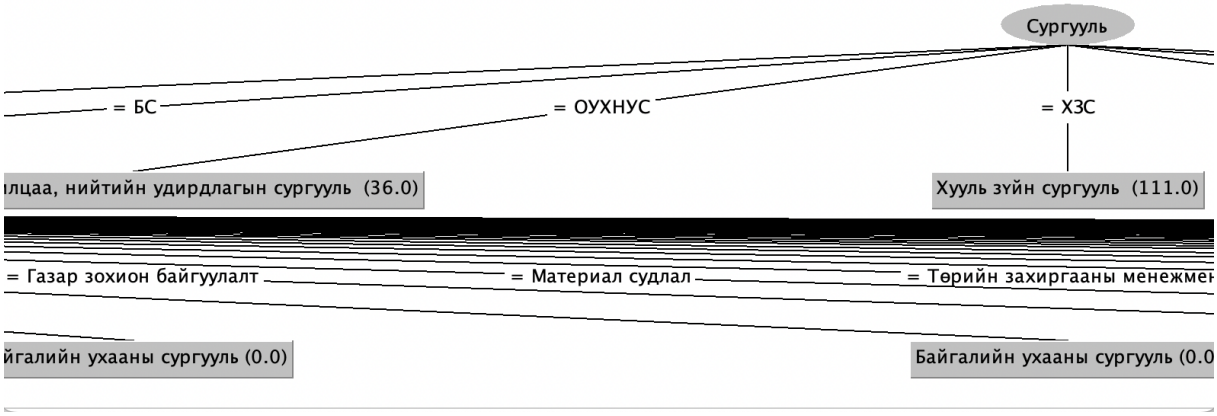
Эх сурвалж: Судлаачийн тооцоолол

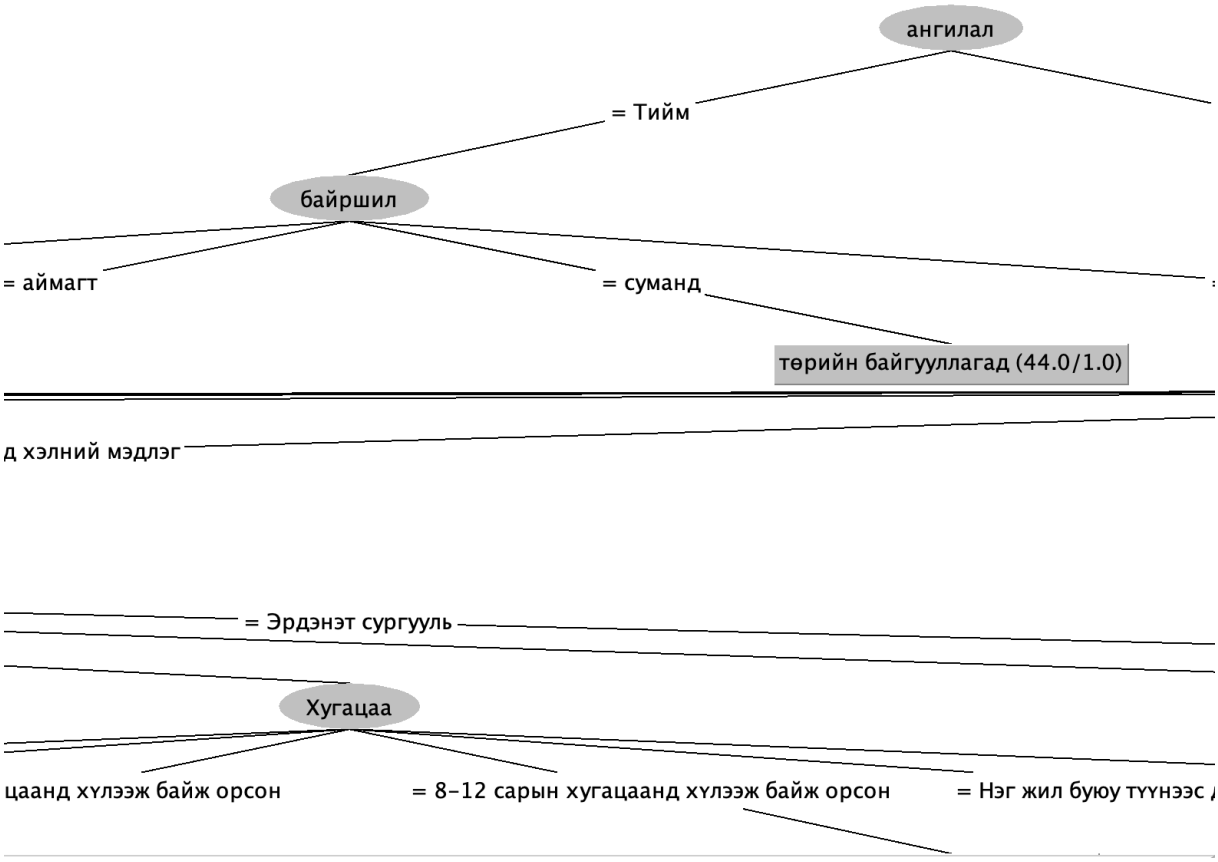
Хавсралт 3. Энтроп болон Мэдээллийн хожоо тооцоолол

Энтроп	$H(S) = \sum_{x \in X} -p(x) log_2 p(x)$	
H(Хөдөлмөрийн байдал)	=-E/N*Log2(-E/N)-U/N*Log2(-U/N)	0.80
H(Хөдөлмөрийн байдал|БУС)	=-E1/N1* Log2(-E1/ N1)-U1/N1*Log2(-U1/N1)	0.883
H(Хөдөлмөрийн байдал|НУС)	=-E2/N2* Log2(-E2/ N2)-U2/N2*Log2(-U2/N2)	0.914
H(Хөдөлмөрийн байдал|ХУС)	=-E3/N3* Log2(-E3/ N3)-U3/N3*Log2(-U3/N3)	0.931
H(Хөдөлмөрийн байдал|БС)	=-E4/N4* Log2(-E4/ N4)-U4/N4*Log2(-U4/N4)	0.579
H(Хөдөлмөрийн байдал|ОУХНУС)	=-E5/N5* Log2(-E5/ N5)-U5/N5*Log2(-U5/N5)	0.503
H(Хөдөлмөрийн байдал|ХЗС)	=-E6/N6* Log2(-E6/ N6)-U6/N6*Log2(-U6/N6)	0.594
H(Хөдөлмөрийн байдал|ХШУИС)	=-E7/N7* Log2(-E7/ N7)-U7/N7*Log2(-U7/N7)	0.858
H(Хөдөлмөрийн байдал|ЗС)	=-E8/N8* Log2(-E8/ N8)-U8/N8*Log2(-U8/N8)	0.849
H(Хөдөлмөрийн байдал|ЭС)	=-E9/N9* Log2(-E9/ N9)-U9/N9*Log2(-U9/N9)	0.787
Жинлэсэн дундаж	H(Хөдөлмөрийн байдал|Салбар сургууль)	0.769
Мэдээллийн хожоо	$IG(S,A) = H(S) - \sum_{t \in T} p(t)H(t) = H(S) - H(S\|A)$	
IG(S,A)	H(Хөдөлмөрийн байдал)- H(Хөдөлмөрийн байдал|Салбар сургууль)	0.030

Эх сурвалж: Судлаачийн тооцоолол

Хавсралт 4. Шийдвэрийн модны томруулсан хэсэг





Эх сурвалж:Судлаачийн тооцоолол

Хавсралт 5 Хувьсагчдын дэлгэрэнгүй тайлбар

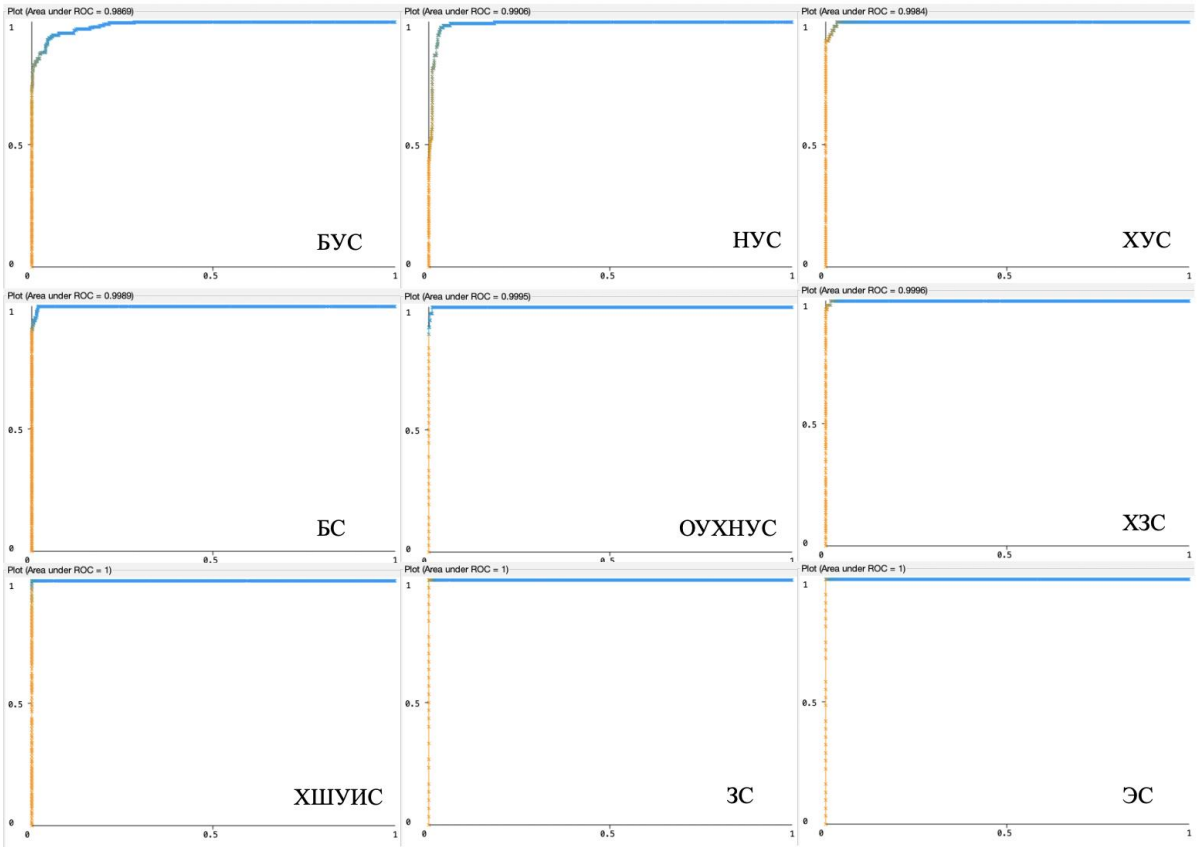
Хувьсагч	Тэмдэглэгээ	Утга	Нэгж	Тайлбар
Бүрэлдэхүүн сургууль	A_1	Чанарын	[1:7]	1- Шинжлэх ухааны сургууль 2- Хэрэглээний шинжлэх ухааны сургууль 3- Бизнесийн сургууль 4- Хууль зүйн сургууль 5- Завхан сургууль 6- Эрдэнэт сургууль
Салбар сургууль	A_2	Чанарын	[1:9]	1- Байгалийн ухааны сургууль 2- Нийгмийн ухааны сургууль 3- Хүмүүнлэгийн ухааны сургууль 4- Хэрэглээний шинжлэх ухаан, инженерчлэлийн сургууль 5- Олон улсын харилцаа нийтийн удирдлагын сургууль 6- Бизнесийн сургууль 7- Хууль зүйн сургууль 8- Завхан сургууль 9- Эрдэнэт сургууль

Хөдөлмөрийн байдал	A_3	Чанарын	[1:2]	1- Ажилтай 2- Ажилгүй
Мэргэжил	A_4	Чанарын	[1:109]	1- Багш байгалийн ухааны боловсрол 2- Нийгмийн ажил 3- Гадаад хэлний орчуулга 4- Компьютерын сүлжээ 5- Улс төр судлал 6- Хэрэглээний математик 7- Хүн ам зүй 8- Менежмент, худалдааны ... 109- Эдийн засаг статистик
Ажил	A_5	Чанарын	[1:8]	1- Гадаадад ажиллаж амьдардаг 2- Олон улсын байгууллагад 3- Төрийн байгууллагад 4- Төрийн бус байгууллагад 5- Төсвийн байгууллагад 6- Хувиараа ажил эрхлэдэг 7- Хувийн хэвшлийн байгууллагад 8- Хариулт өгөөгүй
Ажилд орсон хугацаа	A_6	Чанарын	[1:6]	1- Шууд ажилд орсон 2- 3 сарын дотор ажилд орсон 3- 4-7 сарын хугацаанд хүлээж байж орсон 4- 8-12 сарын хугацаанд хүлээж байж орсон 5- Нэг жил буюу түүнээс дээш хүлээж байж ажилд орсон 6- Хариулт өгөөгүй
Мэргэжилээрээ ажиллаж буй эсэх	A_7	Чанарын	[1:3]	1- Мэргэжлээрээ ажиллаж байгаа 2- Мэргэжлээрээ ажиллаагүй өөр ажил эрхэлж байгаа 3- Хариулт өгөөгүй
Ажил хайх бэрхшээл	A_8	Чанарын	[1:11]	1- Ажлын туршлага дутах 2- Мэргэжилд тохирох ажил олдохгүй 3- Ажлын орчин, ажлын цаг 4- Гадаад хэлний мэдлэг ...

				11-Өөртөө итгэлгүй байдал
Ажил хийхгүй байгаа шалтгаан	A_9	Чанарын	[1:14]	1- Ар гэрийн шалтгаанаар 2- Үргэлжлүүлэн суралцаж байгаа 3- Гадаад орон руу явсан ... 14- Хариулт өгөөгүй
Голч	A_{10}	Тоон	1.0-4.0	Төгсөх үеийн голч
Хүйс	A_{11}	Чанарын	[1:2]	3- Эмэгтэй 4- Эрэгтэй
Сарын дундаж орлого	A_{12}	Тоон	[1:1031]	Тухайн төгсөгчийн сарын дундаж орлого

Эх сурвалж:Судлаачийн тооцоолол

Хавсралт 6 ROC муруй



Эх сурвалж:Судлаачийн тооцоолол