دانشگاه شهید بهشتی

Shahid Beheshti University

موضوع : یافتن بهترین مدل پیشگویانه به منظور پیشگویی رده درآمدی خانوارهای شهری استان همدان با استفاده از داده‌های طرح هزینه و درآمد خانوار در سال ۱۳۹۸ است که توسط مرکز آمار ایران تهیه شده است.

گردآورنده : مجتبی کنعانی سرچشمه

استاد راهنما : آقای دکتر محمدرضا فقیهی حبیب آبادی

فهرست

۱. مقدمه

طرح آمارگیری از هزینه و درآمد خانوارهای شهری کشور از سال ۱۳۴۷ توسط مرکز آمار ایران آغاز شده است. این آمارگیری از سال ۱۳۵۳ علاوه بر هزینه، درآمد خانوارهای شهری را نیز شامل شده است و تاکنون غیر از سال‌های ۱۳۵۵، ۱۳۵۷، ۱۳۶۰ همه ساله در کشور اجرا شده است. هر ساله با بررسی‌های لازم و تطبیق با استانداردها و توصیه‌های بین المللی تغییرات مورد نیاز در راستای تکامل تهیه و اجرای طرح توسط کارشناسان مربوطه اعمال می‌شود.

➢ هدف

این پروژه هدف یافتن بهترین مدل پیشگویانه به منظور پیشگویی رده درآمدی خانوارهای شهری استان همدان با استفاده از داده‌های طرح هزینه و درآمد خانوار در سال ۱۳۹۸ است که توسط مرکز آمار ایران تهیه شده است.

۲. معرفی داده ها

اطلاعاتی که ما در این جا داریم توسط کارشناسان و به وسیله پرسشنامه از خانوار ها جمع آوری شده است.

پرسشنامه طرح هزینه و درآمد خانوارهای شهری شامل بخش‌های زیر است:

- خصوصیات اجتماعی اعضای خانوار
- مشخصات محل سکونت و تسهیلات و لوازم عمده زندگی
- هزینه‌های خوراکی و غیرخوراکی خانوار
- درآمدهای خانوار

جدول توضیحات هر یک از متغیر های مورد استفاده در این بخش در پیوست موجود است.

در بخش پاکسازی نیز اطلاعات تکمیلی مربوط به معرفی داده ها موجود است.

۳. پاکسازی داده ها

در ابتدا ماتریس داده های ما شامل ۷۸۸ ثبت و ۶۸ ستون است.

در این بخش قصد داریم عملیات پاکسازی کلی روی داده ها انجام دهیم ، این نوع از پاکسازی برای استفاده از داده ها در تمام انواع مدل های رده بندی لازم است.

در هر مدل رده بندی نیز بنا بر ویژگی های آن مدل تغییرات خاصی نیز روی داده ها انجام خواهیم داد که آن تغییرات را در بخش های مربوط به هر مدل بررسی میکنیم.

❖ آدرس خانوار – Address

این مقدار در پرسشنامه به شکل رو به رو از میباشد.

شماره ردیف خانوار در خوشه		ماه مراجعه		شماره خوشه				کد شهرستان		کد استان		شهری ۱ روستایی۲

در این جا تمام آدرس های ما با عدد ۱ شروع میشوند پس تمام این اطلاعات مربوط به مناطق شهری میباشد و لازم نیست ما این اطلاعات را ذخیره کنیم.
همچنین کد استان در تمام داده ها برابر با ۱۳ میباشد که نشان دهنده استان همدان میباشد و این بخش از آدرس ها را نیز نیاز نداریم.
ادامه آدرس نشان دهنده کد شهرستان، شماره خوشه، ماه مراجعه و شماره ردیف خانوار در خوشه میباشد که از بین این ۴ بخش، تنها بخش مهم برای ما بخش کد شهرستان است پس تنها این دو رقم را نگه میداریم و بقیه ارقام را حذف میکنیم.
سپس با توجه به فایل تقسیمات کشوری ، هر یک از این اعداد دو رقمی را به نام شهرستان مربوطه در استان همدان تبدیل میکنیم که تبدیلات ما و تعداد ثبت های مربوط به هر شهرستان به شکل زیر خواهد بود

نام شهرستان	کد شهرستان
tooyserkan	۰۱
malayer	۰۲
nahaavand	۰۳
hamedan	۰۴
kaboodarAhang	۰۵
asadAbaad	۰۶
bahaar	۰۷
razan	۰۸
faamenin	۰۹

asadAbaad	70
bahaar	76
faamenin	36
hamedan	217
kaboodarAhang	47
malayer	132
nahaavand	92
razan	47
tooyserkan	71

❖ کد استان – C.Ostan

تمام ثبت ها در این داده ها مربوط به استان همدان است و مقدار این ستون در تمام ثبت ها برابر با ۱۳ است و اطلاعات خاصی در اختیار ما قرار نمیدهد پس آن را حذف میکنیم.

❖ تعداد اعضای خانوار – Tedad.a

داده های این ستون پردازش خاصی نیاز ندارد.

شاخص های آماری مختلف مربوط به این ستون را در زیر مشاهده میکنید.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.000   2.000   3.000   3.151   4.000   8.000
```

❖ جنسیت سرپرست خانوار – Gender

مقادیر این ستون به صورت ۰ و ۱ است که ۰ نشان دهنده مذکر بودن سرپرست خانوار و ۱ نشان دهنده مؤنث بودن سرپرست خانوار است که این دو گروه را به Male و Female تغییر میدهیم.

تعداد ثبت های مربوط به هر یک از جنسیت ها را زیر رو مشاهده میکنید.

male	651
female	137

❖ سن سرپرست خانوار – Age

داده های این ستون پردازش خاصی نیاز ندازد.

شاخص های آماری مختلف مربوط به این ستون را در زیر مشاهده میکنید.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
21.00   40.00   50.00   52.16   62.00   95.00
```

❖ میزان سواد سرپرست خانوار – Savad

مقادیر این ستون به صورت ۰ و ۱ است که ۰ نشان دهنده با سواد بودن سرپرست خانوار و ۱ نشان دهنده بی سواد بودن سرپرست خانوار است که این دو گروه را به Yes و No تغییر میدهیم.

تعداد ثبت های مربوط به هر یک از وضعیت های سواد را زیر رو مشاهده میکنید.

No	147
Yes	641

❖ سرپرست خانوار تحصیل می‌کند یا خیر؟ – InEdu

مقادیر این ستون به صورت ۰ و ۱ است که ۰ نشان دهنده این است که سرپرست خانوار در حال تحصیل نیست و ۱ نشان دهنده این است که سرپرست خانوار در حال تحصیل است که این دو گروه را به Yes و No تغییر میدهیم.

همچنین اطلاعات ۱۴۷ ثبت در این متغیر بدون مقدار است که تمام این ثبت ها دقیقا مربوط به افرادی است که سرپرست خانوار بی سواد است و به احتمال بالا اکثر این افراد در حال تحصیل نیز نیستند پس مقادیر مربوط به این ثبت ها را نیز در گروه No قرار میدهیم.

تعداد ثبت ها در هر یک از این گروه ها به شکل زیر است.

No	777
Yes	11

❖ مدرک تحصیلی سرپرست خانوار – Madrak

مقادیر این ستون میتواند شامل اعداد ۱ تا ۹ باشد که هر کدام نشان دهنده سطوح مختلف سواد هستند که با توجه با جدول زیر مقادیر آن ها را با عبارت های متناسب جایگزین کردیم.

همچنین اطلاعات ۱۴۷ ثبت در این متغیر بدون مقدار است که تمام این ثبت ها دقیقا مربوط به افرادی است که سرپرست خانوار بی سواد است پس مقادیر مربوط به این ثبت ها را نیز در یک گروه جدا با عنوان illiterate قرار میدهیم.

تعداد ثبت ها در هر یک از این گروه ها به شکل زیر است.

پس از تبدیل	سطح سواد	کد
elementary	ابتدایی/ سوادآموزی	۱
mid_۱	راهنمایی/ متوسطه۱	۲
mid_۲	متوسطه / متوسطه۲	۳
Diploma	دیپلم و پیش دانشگاهی	۴
Associate_Degree	فوق دیپلم / کاردانی	۵
Bachelor	لیسانس / کارشناسی	۶
Masters	کارشناسی ارشد و دکترای حرفه ای	۷
PHD	دکترای تخصصی	۸
other	سایر و غیر رسمی	۹
illiterate	بی سواد	-

Associate_Degree	41
Bachelor	79
Diploma	134
elementary	191
illiterate	147
Masters	22
mid_1	141
mid_2	33

❖ وضعیت فعالیت سرپرست خانوار – Faaliat

مقادیر این ستون میتواند شامل اعداد ۱ تا ۶ باشد که هر کدام نشان دهنده انواع مختلف فعالیت هستند که با توجه با جدول زیر مقادیر آن ها را با عبارت های متناسب جایگزین کردیم.

تعداد ثبت ها در هر یک از این گروه ها به شکل زیر است.

کد	نوع فعالیت	پس از تبدیل
۱	شاغل	Employed
۲	بیکار (جویای کار)	job_seeker
۳	دارای درآمد بدون کار	income_without_work
۴	محصل	student
۵	خانه دار	housewife
۶	سایر	other

Employed	498
housewife	22
income_without_work	242
job_seeker	11
other	14
student	1

❖ تعداد اعضای شاغل در خانوار – T.shaghel

شاخص های آماری مختلف مربوط به این ستون را در زیر مشاهده میکنید.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
1.000   1.000   1.000   1.274   1.000   4.000     196
```

همچنین اطلاعات ۱۹۶ ثبت در این متغیر بدون مقدار است و با توجه به شاخص های آماری تصمیم گرفتم معیار میانه را برای جانهی این ثبت ها در نظر بگیرم. پس از جانهی شاخص های آماری مختلف این ستون به شکل زیر تغییر میکند.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.000   1.000   1.000   1.206   1.000   4.000
```

همان طور که مشاهده میکنید تغییر زیادی در شاخص های آماری داده ها به وجود نیامد
قطعا این کار ما باعث اضافه شدن مقدار اندکی خطا به اطلاعاتمان است اما اگر این کار را انجام ندهیم نمیتوانیم از این ثبت ها در مدل ها رده بندی استفاده کنیم زیرا در این مدل ها تنها ثبت هایی کاربرد دارند که اطلاعات تمام ستون هایشان به صورت کامل نوشته شده باشد پس قطعا این مقدار اندک خطا را به از دست دادن ۱۹۶ ثبت از داده هایمان ترجیح میدهیم.

❖ نحوه تصرف منزل مسکونی - n.t.m

مقادیر این ستون میتواند شامل اعداد ۱ تا ۷ باشد که هر کدام نشان دهنده انواع مختلف نحوه تصرف محل سکونت هستند که با توجه با جدول زیر مقادیر آن ها را با عبارت های متناسب جایگزین کردیم.

تعداد ثبت ها در هر یک از این گروه ها به شکل زیر است.

پس از تبدیل	نوع فعالیت	کد
landlord	ملکی عرصه و اعیان	۱
house_owner	ملکی اعیان	۲
Rent	اجاری	۳
Mortgage	رهن	۴
for_service	در برابر خدمت	۵
free	رایگان	۶
other	سایر	۷

for_service	5
free	83
house_owner	1
landlord	547
Mortgage	17
other	1
Rent	134

❖ تعداد اتاق در اختیار - T.O

داده های این ستون پردازش خاصی نیاز ندارد.

شاخص های آماری مختلف مربوط به این ستون را در زیر مشاهده میکنید.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.000   3.000   4.000   3.712   4.000   9.000
```

❖ سطح زیر بنای محل سکونت - S.Z

داده های این ستون پردازش خاصی نیاز ندارد.

شاخص های آماری مختلف مربوط به این ستون را در زیر مشاهده میکنید.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
12.00   73.00   90.00   98.06  120.00  500.00
```

❖ نوع اسکلت بنای محل سکونت – N.S

اطلاعات این ستون و ستون بعد را با یکدیگر تلفیق میکنیم و ستون بعد را حذف میکنیم.

مقادیر این ستون شامل اعداد ا تا ۳ میباشد که عنوان گروه ۱ "فلزی" است و عنوان گروه ۲ "بتون آرمه" است و گروه ۳ نیز به سایر موارد تعلق دارد و در صورتی که گروه ۳ در این ستون انتخاب شود باید اطلاعات دقیق تر آن در ستون بعدی مشخص شود، پس ما این دو ستون را تلفیق میکنیم و با توجه به جدول زیر مقادیر آن را با عبارت متناسب جایگزین میکنیم.

تعداد ثبت ها در هر یک از این گروه ها به شکل زیر است.

کد	ستون	نوع مصالح استفاده شده	پس از تبدیل
۱	N.S	فلزی	Metal
۲	N.S	بتون آرمه	Reinforced_concrete
۱	Masleh	آجر و آهن یا سنگ و آهن	Brick&M_Rock&M
۲	Masleh	آجر و چوب یا سنگ و چوب	Brick&W_Rock&W
۳	Masleh	بلوک سیمانی (با هر نوع سقف)	Cement_block
۴	Masleh	تمام آجر یا سنگ و آجر	Brick_Rock
۵	Masleh	تمام چوب	Wood
۶	Masleh	خشت و چوب	Adobe&W
۷	Masleh	خشت و گل	Adobe&Mud
۸	Masleh	سایر	other

Adobe&W	3
Brick&M_Rock&M	498
Brick&W_Rock&W	25
Brick_Rock	1
Cement_block	1
Metal	203
Reinforced_concrete	56
Wood	1

❖ مصالح عمده بنای محل سکونت - Masleh

اطلاعات این ستون را با ستون قبل تلفیق کردیم و این ستون را حذف میکنیم.

❖ دارایی ها

- اتومبیل شخصی – oto
- موتورسیکلت - motor
- دوچرخه – do
- رادیو – radio
- ضبط – zabt
- تلویزیون سیاه و سفید - TV.S
- تلویزیون رنگی - TV.r
- انواع ویدئو، VCD و DVD – DVD
- انواع یارانه و تبلت – Pc
- تلفن همراه – mobile
- فریزر – freeizer
- یخچال – yakhchal
- یخچال فریزر - yakhchal.f
- اجاق گاز – gaz
- جارو برقی - jaro.b
- ماشین لباسشویی - m.lebas
- چرخ خیاطی - charkh.kh
- پنکه – panke
- کولر آبی متحرک - cooler.a
- کولر گازی متحرک - cooler.g
- ماشین ظرفشویی - m.zarf
- مایکروویو و انواع فرهای هالوژن دار – microfer
- آب لوله‌کشی - ab.l
- برق – bargh
- گاز لوله‌کشی - gaz.l
- تلفن ثابت – tel
- دسترسی به اینترنت – internet
- حمام – hamam
- آشپزخانه – ashpazkhane
- کولر آبی ثابت - cooler.a.s
- برودت مرکزی - broodat.m
- حرارت مرکزی - hararat.m
- پکیج – package
- کولر گازی ثابت - cooler.g.s
- کولر گازی ثابت - cooler.g.s
- شبکه عمومی فاضلاب – fazelab

این ستون ها نشان دهنده دارایی های افراد هستند که به مقادیر این ستون ها به سه صورت کلی ۰ و ۱ و مقداردهی نشده میباشد که این ستون های مقدار دهی نشده در اکثر این ستون ها نشان دهنده این است که فرد آن دارایی را ندارد اما در برخی ستون ها نیز نشان دهنده یک دارایی بسیار ابتدایی است که از فرط ابتدایی بودن کسی به خود زحمت نداده که داشتن یا نداشتن آن را در پرسشنامه مشخص کند مانند آب لوله کشی.

۵ ستون وجود دارد که در آن تمام مقادیر به یک شکل است و این ستون ها اطلاعات خاصی به ما اضافه نمیکنند پس آن ها را حذف میکنیم.

این ستون ها عبارتند از:

- کولر گازی متحرک - cooler.g
- آب لوله‌کشی - ab.l
- برق – bargh
- گاز لوله‌کشی - gaz.l
- برودت مرکزی - broodat.m

در بین بقیه این ستون ها تمام مقادیر مقدار دهی نشده نشان دهنده عدم دارایی و تمام مقادیری که با ۱ مقدار دهی شده اند نشان دهنده ی دارا بودن است که مقدار دهی های این ستون ها را به حالت Yes و No تغییر میدهیم.

تعداد ثبت ها در هر یک از این ستون ها در هر یک از گروه ها به شکل زیر است.

oto	No	460
oto	Yes	328
motor	No	634
motor	Yes	154
do	No	607
do	Yes	181
radio	No	756
radio	Yes	32
zabt	No	773
zabt	Yes	15
TV.S	No	786
TV.S	Yes	2
TV.r	No	10
TV.r	Yes	778
DVD	No	652
DVD	Yes	136
Pc	No	565
Pc	Yes	223
mobile	No	35
mobile	Yes	753

freeizer	No	558
freeizer	Yes	230
yakhchal	No	426
yakhchal	Yes	362
yakhchal.f	No	347
yakhchal.f	Yes	441
gaz	No	3
gaz	Yes	785
jaro.b	No	95
jaro.b	Yes	693
m.lebas	No	177
m.lebas	Yes	611
charkh.kh	No	488
charkh.kh	Yes	300
panke	No	567
panke	Yes	221
cooler.a	No	675
cooler.a	Yes	113
m.zarf	No	717
m.zarf	Yes	71

microfer	No	711
microfer	Yes	77
tel	No	4
tel	Yes	784
internet	No	250
internet	Yes	538
hamam	No	643
hamam	Yes	145
ashpazkhane	No	2
ashpazkhane	Yes	786
cooler.a.s	No	339
cooler.a.s	Yes	449
hararat.m	No	778
hararat.m	Yes	10
package	No	716
package	Yes	72
cooler.g.s	No	760
cooler.g.s	Yes	28
fazelab	No	344
fazelab	Yes	444

❖ سوخت

- نوع سوخت برای پخت و پز - sookht.p
- نوع سوخت برای ایجاد گرما – sookht.g
- نوع سوخت برای تهیه آب گرم – sookht.ab

مقادیر این ستون ها میتواند شامل اعداد ۱ تا ۳۰ باشد که هر کدام نشان دهنده انواع مختلف سوخت هستند که با توجه با جدول زیر مقادیر آن ها را با عبارت های متناسب جایگزین کردیم.

کد	نوع سوخت	پس از تبدیل
۱-۱۱-۲۱	نفت سفید	Kerosene
۲-۱۲-۲۲	گازوئیل	Gasoline
۳-۱۳-۲۳	گاز مایع	liquid_gas
۴-۱۴-۲۴	گاز طبیعی (شبکه عمومی)	natural_gas
۵-۱۵-۲۵	برق	Electricity
۶-۱۶-۲۶	هیزم و زغال	Timber
۷-۱۷-۲۷	سوخت حیوانی	Biofuel
۸-۱۸-۲۸	زغال سنگ	coal
۹-۱۹-۲۹	سایر سوخت ها	other
۱۰-۲۰-۳۰	هیچکدام	none

تعداد ثبت ها در ستون sookht.p در هر یک از این گروه ها به شکل رو به رو است.

liquid_gas	3
natural_gas	784
other	1

تعداد ثبت ها در ستون sookht.g در هر یک از این گروه ها به شکل رو به رو است.

Electricity	1
liquid_gas	2
natural_gas	785

تعداد ثبت ها در ستون sookht.ab در هر یک از این گروه ها به شکل رو به رو است.

Electricity	1
liquid_gas	3
natural_gas	784

❖ هزینه ها

- هزینه‌های خوراکی و دخانیات خانوار در یکماه گذشته - H_Khorakivadokhani
- هزینه‌های نوشیدنی خانوار در یکماه گذشته - H_Noshidani
- هزینه‌های پوشاک خانوار در یکماه گذشته - H_Pushak
- هزینه‌های مسکن- آب، سوخت، روشنایی و... - H_Maskan
- هزینه‌های لوازم خانگی خاوار در یکماه گذشته - H_mobleman
- هزینه‌های بهداشتی خانوار در یکماه گذشته - H_Behdasht
- هزینه‌های حمل و نقل خانوار در یکماه گذشته - H_Hamlonaghl
- هزینه ارتباطات خانوار در یکماه گذشته - H_Ertebatat
- هزینه های تفریحات خانوار در ماه گذشته - H_Tafrihat
- هزینه‌های غذای آماده هتل و رستوران‌های خانوار در یکماه گذشته - H_Ghazayeamade
- هزینه کالاها یا خدمات متفرقه خانواردر یکماه گذشته - H_kalavakhadamat

این ستون ها شامل اطلاعاتی است که ممکن است مقدار دهی نشده باشند، در تمامی ستون ها به جز ستون H_Khorakivadokhani فرض ما بر این است که بی مقدار بودن نشان دهنده صفر بودن مقدار این هزینه در خانوار است.

اما در ستون H_Khorakivadokhani نمیتوانیم اطلاعات مقدار دهی نشده را برابر صفر در نظر بگیریم چون ابتدایی ترین هزینه هر خانواده هزینه خوراکی است و صفر بودن اطلاعات این ستون بی معنی است.

۴ ثبت در این ستون مقدار دهی نشده اند و ما این مقادیر را با میانه هزینه های خوراکی و دخانی سایر خانوار ها جانهی میکنیم.

شاخص های آماری در هر یک از این ستون ها به شکل زیر است.

نام ستون	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
H_Khorakivadokhani	۱۴۰۹۰۰	۴۴۶۰۲۷۶	۶۹۶۲۷۰۰	۸۵۶۵۴۹۱	۱۰۶۳۶۰۰۰	۸۶۲۶۷۳۶۰
H_Noshidani	۰	۰	۰	۴۲۵۷۶	۲۸۵۰۰۰	۱۰۸۵۰۰۰۰
H_Pushak	۰	۰	۱۵۰۰۰۰	۱۴۴۰۱۵۴	۱۶۵۰۰۰۰	۴۱۰۰۰۰۰۰
H_Maskan	۵۰۰۰۰۰	۵۵۵۵۰۰۰	۷۹۵۲۵۰۰	۹۷۹۰۴۴۰	۱۱۴۶۲۷۵۰	۶۳۴۵۵۰۰۰
H_mobleman	۰	۱۲۰۰۰۰	۴۲۰۰۰۰	۶۷۷۵۹۸	۸۱۵۰۰۰	۱۱۲۱۰۰۰۰
H_Behdasht	۰	۶۳۷۵۰	۶۳۹۰۰۰	۲۳۹۹۶۸۴	۲۱۷۲۵۰۰	۱۰۲۰۰۰۰۰۰
H_Hamlonaghl	۰	۲۵۰۰۰۰	۵۸۰۰۰۰	۱۶۱۱۲۵۰	۲۰۶۲۵۰۰	۲۶۸۹۰۰۰۰
H_Ertebatat	۰	۲۳۰۰۰۰	۴۵۰۰۰۰	۵۷۲۸۳۸	۷۵۰۰۰۰	۸۵۰۰۰۰۰
H_Tafrihat	۰	۰	۰	۲۸۳۳۰۷	۲۴۰۰۰۰	۶۸۷۰۰۰۰
H_Ghazayeamade	۰	۰	۰	۵۴۳۸۴۹	۳۰۰۰۰۰	۱۲۲۲۵۰۰۰۰
H_kalavakhadamat	۰	۱۶۰۰۰۰	۳۹۷۵۰۰	۶۵۷۳۳۱	۸۰۰۰۰۰	۱۲۴۵۰۰۰۰

❖ درآمد ها

- D_Mozd - درآمد مزد خانوار در ۱۲ ماه گذشته
- D_Azad - درآمد آزاد خانوار در ۱۲ ماه گذشته
- D_Motefaraghe - درآمدهای متفرقه خانوار در ۱۲ ماه گذشته
- D_Yarane - مبلغ دریافتی یارانه نقدی در ۱۲ ماه گذشته

در صورت بی مقدار بودن هر کدام از این بخش ها آن را صفر در نظر میگیریم.

همچنین با بررسی مقادیر درآمد تمام خانوار ها متوجه شدم مقدار درآمد ۱۳ خانوار در یک یا چند مورد از این ۴ ستون مقداری منفی است و از آن جا که در هیچ بخشی از این پرسشنامه تعریفی برای اعداد منفی در نظر گرفته نشده است فرض ما بر این است که این اعداد اشتباه هستند و داده هایی که این ویژگی را دارند از داده های اصلی حذف میکنیم.

شاخص های آماری در هر یک از این ستون ها به شکل زیر است.

نام ستون	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
D_Mozd	۰	۰	۰	۱۱۵۰۷۷۰۳۹	۱۹۱۰۰۰۰۰۰	۱۱۸۰۰۰۰۰۰۰
D_Azad	۰	۰	۰	۱۰۸۵۵۰۴۳۷	۱۴۴۰۰۰۰۰۰	۳۱۶۰۰۰۰۰۰۰
D_Motefaraghe	۰	۱۵۰۰۰۰۰	۱۲۹۰۰۰۰۰	۱۰۷۴۰۳۴۸۱	۱۳۸۰۰۰۰۰۰	۴۲۸۰۰۰۰۰۰۰
D_Yarane	۰	۱۰۹۲۰۰۰۰	۱۶۳۸۰۰۰۰	۱۶۵۶۰۲۳۹	۲۱۸۴۰۰۰۰	۳۸۲۲۰۰۰۰

سپس با جمع کردن مقادیر این ۴ ستون برای هر خانوار مقداری جدید میسازیم که نشان دهنده درآمد کلی خانوار در ۱۲ ماه گذشته است و نام این ستون را income میگذاریم و این ۴ ستون را از داده هایمان حذف میکنیم.

شاخص های آماری برای این ستون به شکل زیر است.

```
   Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
4095000  172510000  272280000  347591196  409020000 5426380000
```

سپس ابتدا مقدار درآمدی را پیدا میکنیم که از درآمد ۰.۷ خانواده ها بزرگتر باشد و از درآمد ۰.۳ خانواده ها کوچکتر باشد. این مقدار برابر است با :

70%: 379698000

سپس یک ستون جدید با عنوان Y میسازیم و مقدار این ستون رو برای خانوار هایی که درآمدشان از سرحد درآمدی بالاتر باشد High و برای سایر خانوار ها برابر با Low میگذاریم.

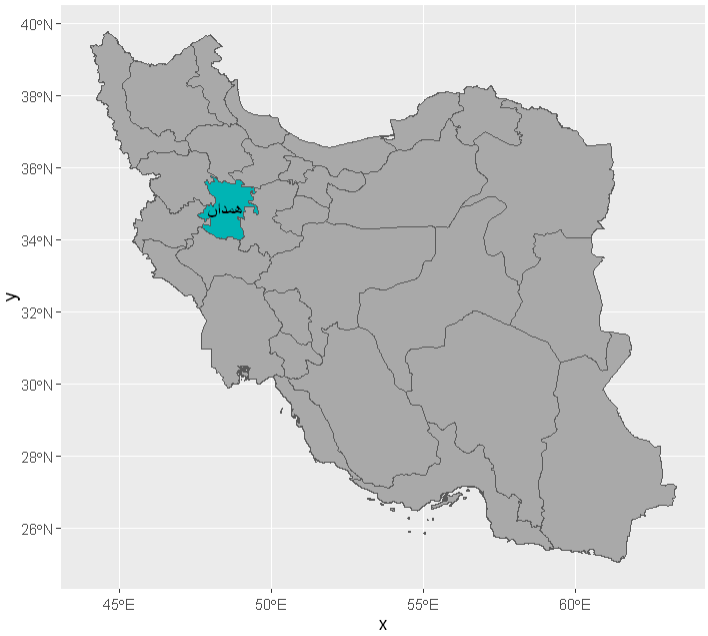
تعداد ثبت ها در این ستون به شکل رو به رو است.

Low	542
High	233

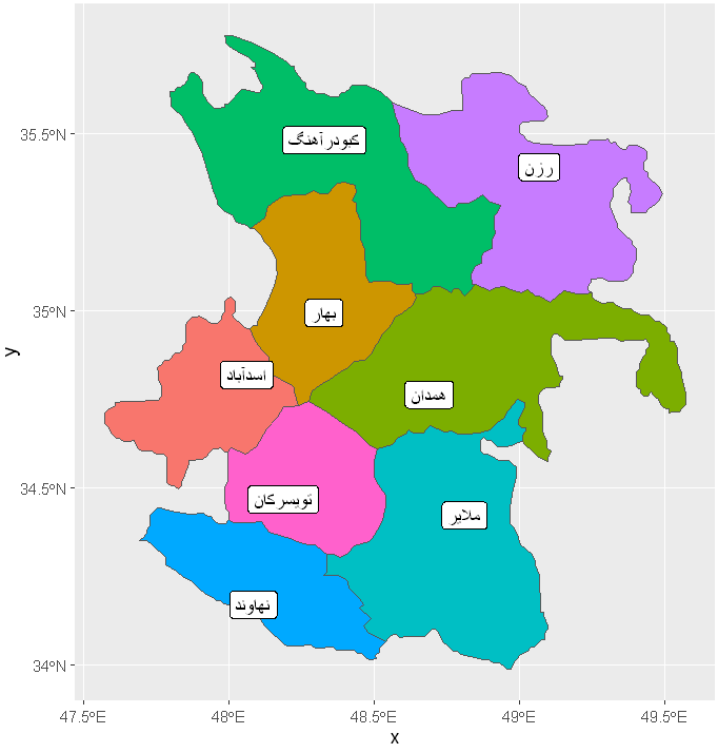
در انتها ماتریس داده های ما شامل ۷۷۵ ثبت و ۵۹ ستون است.

٤. تصویری سازی

❖ موقعیت مکانی استان همدان در ایران



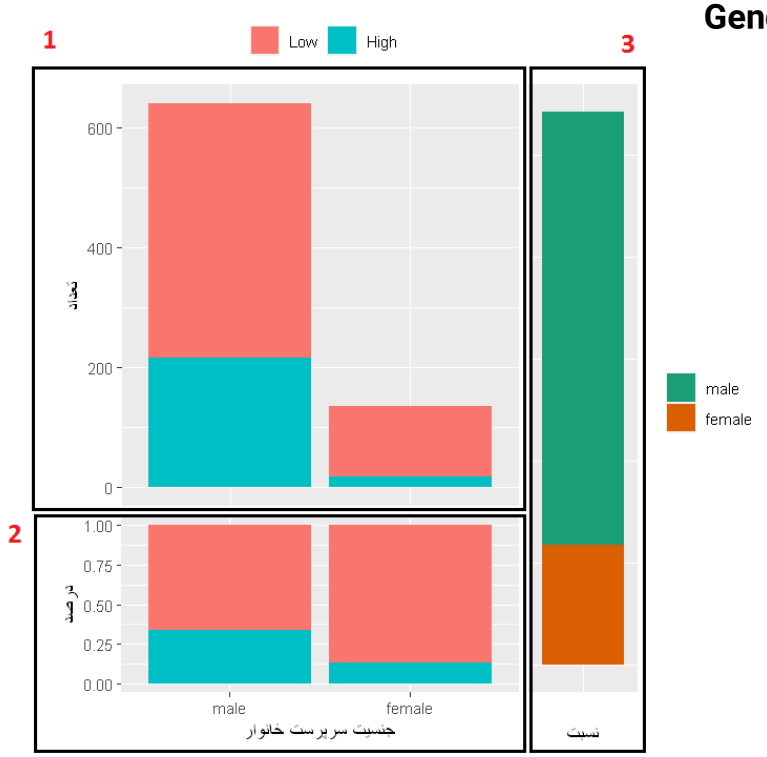
❖ موقعیت مکانی شهرستان های استان همدان



❖ نمودار های تک متغیر نسبت به متغیر هدف

برای این نوع از نمودار ها یک قالب کلی طراحی کردم که در ابتدای کار برای یکی از ستون ها این قالب کلی را توضیح میدهم ، توضیحات مربوط به سایر نمودار ها با این قالب نیز به همین شکل میباشد و برای پرهیز از تکرار گفته ها این توضیحات را در هر بخش تکرار نمیکنیم.

➤ جنسیت سرپرست خانوار - Gender



این قالب شامل ۳ بخش اصلی است.

بخش ۱ : این بخش نشان دهنده تعداد ثبت ها در هر یک از گروه های ستون مورد نظر به تفکیک پردرآمد یا کم درآمد بودن آن خانوار است.

بخش ۲ : این بخش نشان دهنده درصد گروه پردرآمد و کم درآمد در هر یک از رده های ستون مورد نظر است که رده ها در این نمودار به ترتیب نزول درصد خانوار پردرآمد مرتب شده اند.

بخش ۳ : این بخش نشان دهنده درصد هر یک از رده ها در ستون مورد نظر است.

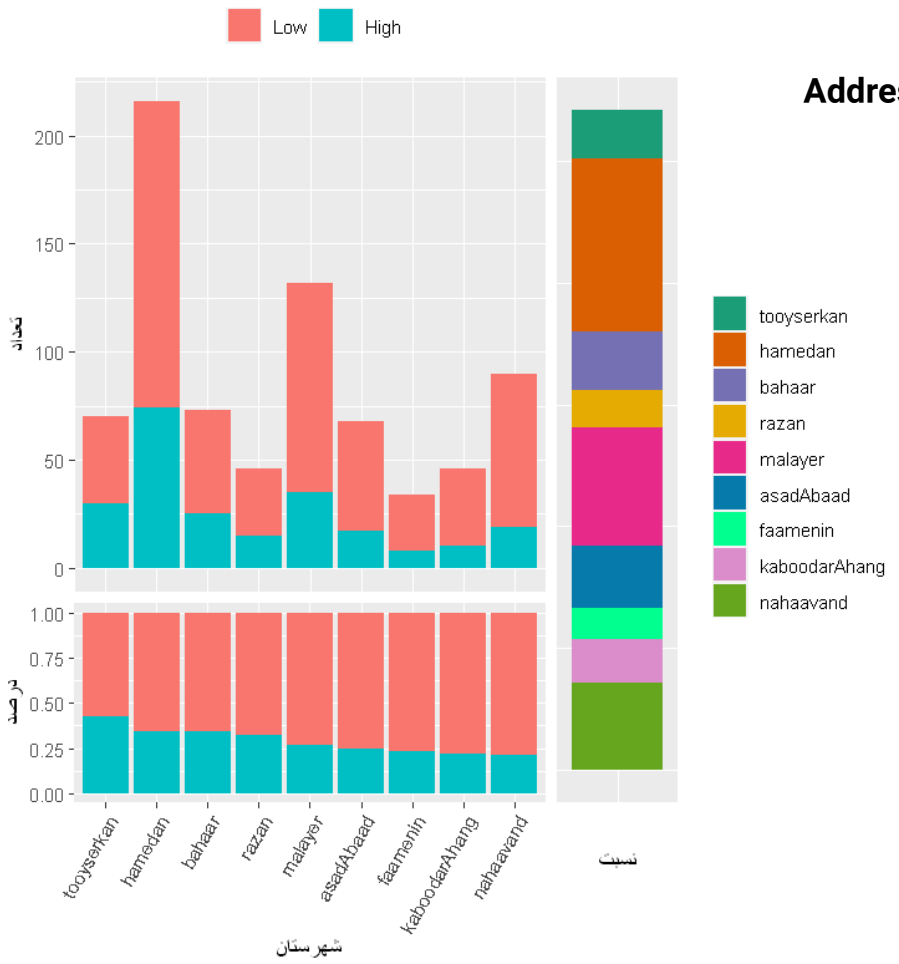
راهنما ۱ : راهنمای بالای نمودار ها راهنمای بخش های ۱ و ۲ میباشد.

پینوشت : این دو رنگ در تمام بخش های این گزارش برای دو گروه کم در آمد و پردرآمد در نظر گرفته شده است.

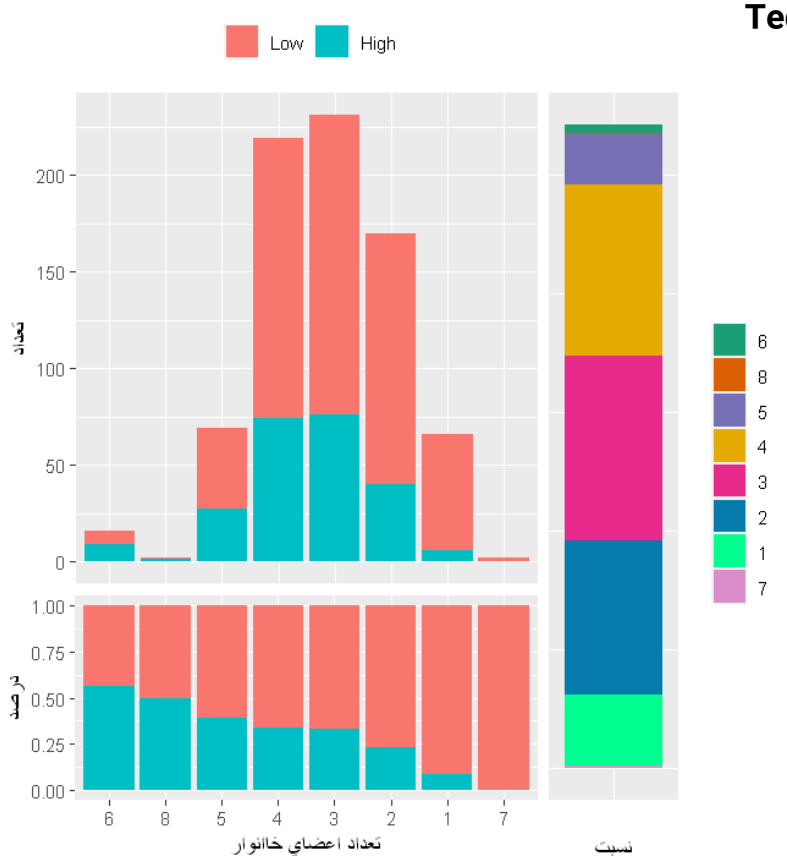
راهنما ۲ : راهنمای سمت راست نمودار ها راهنمای بخش ۳ میباشد.

برای مثال این نمودار نشان میدهد که سرپرست خانوار های مذکر بخش بزرگتری از سرپرستان خانوار را در بر میگیرند و همچنین درصد بیشتری از خانوار هایی که سرپرستشان مذکر است در گروه پر در آمد قرار میگیرند.

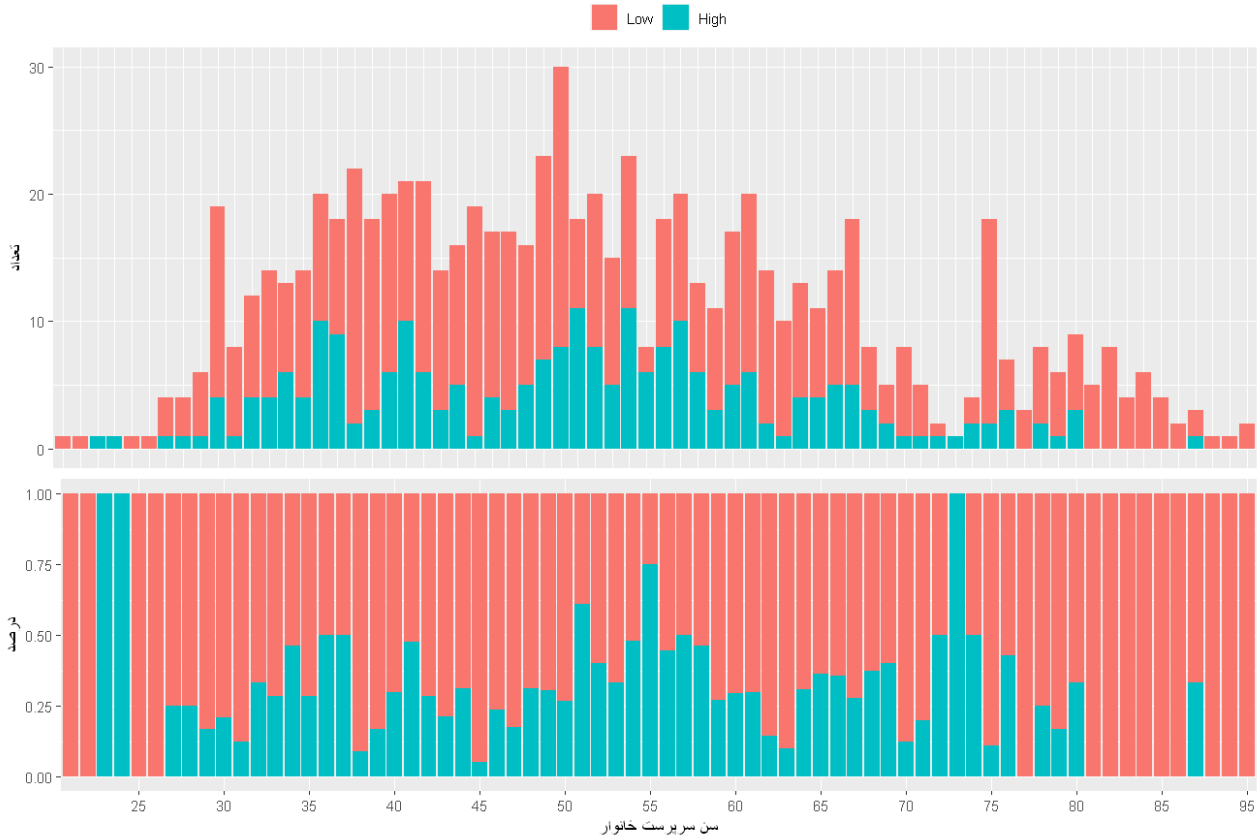
Address – آدرس خانوار



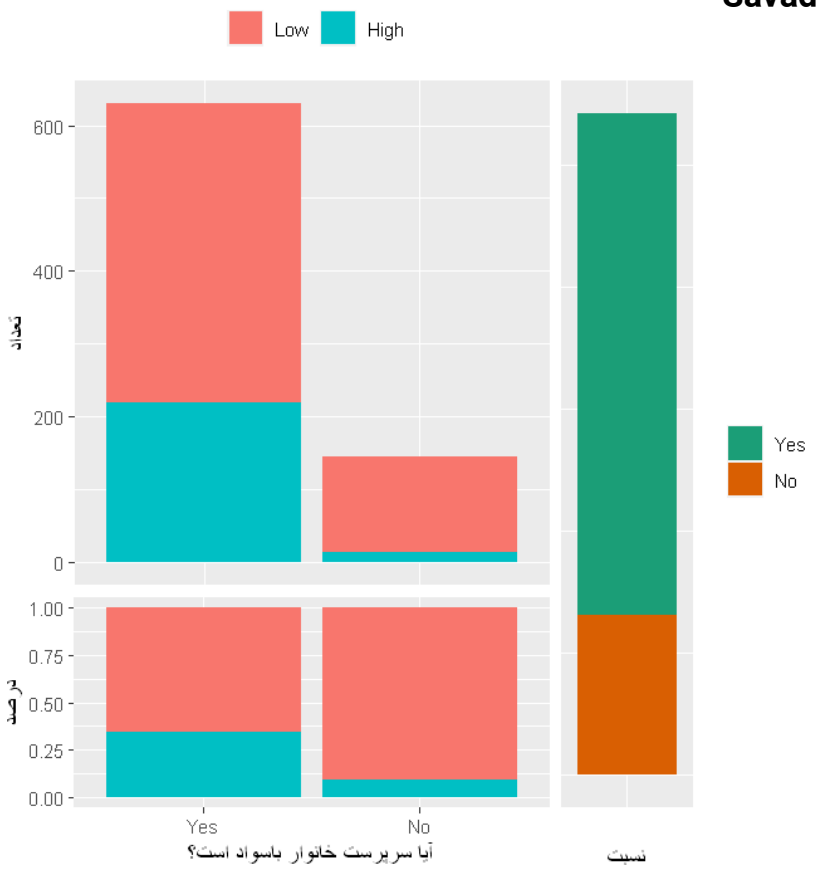
Tedad.a – تعداد اعضای خانوار



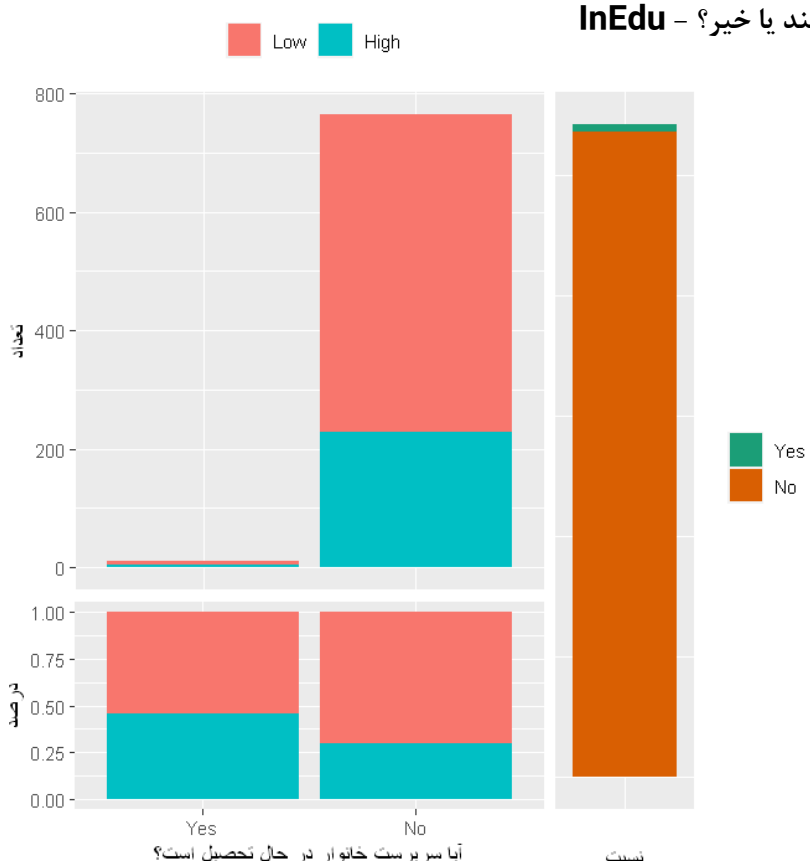
سن سرپرست خانوار – Age



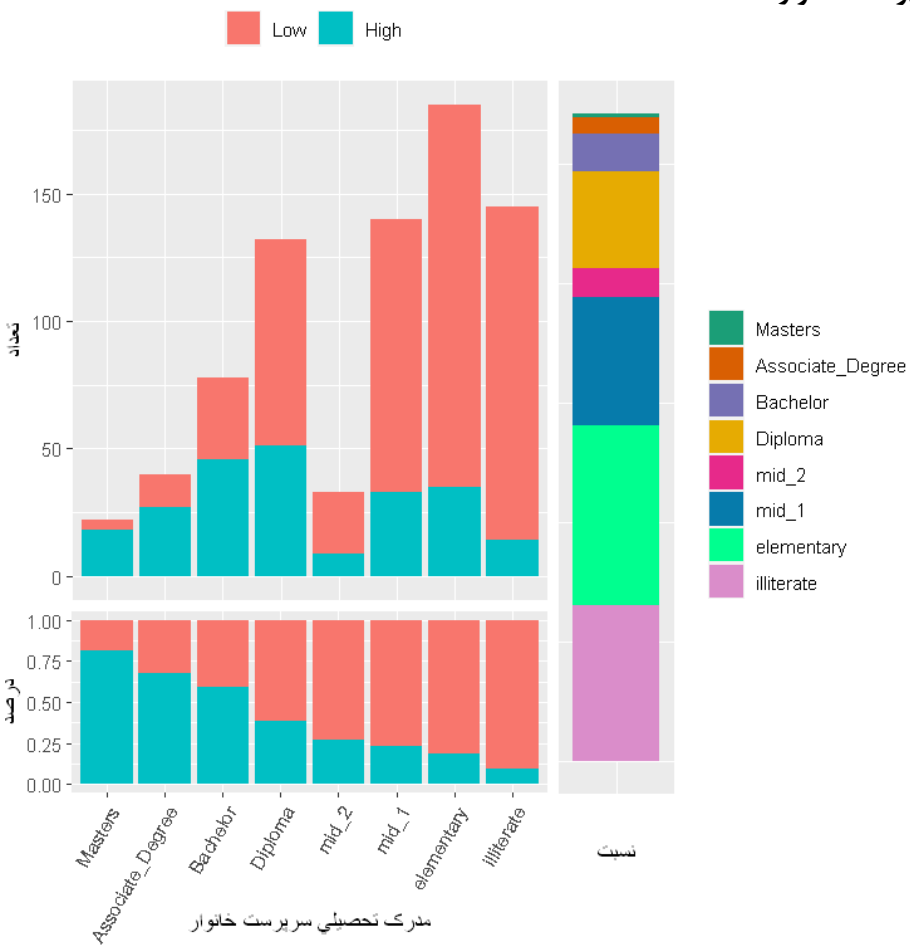
میزان سواد سرپرست خانوار – Savad



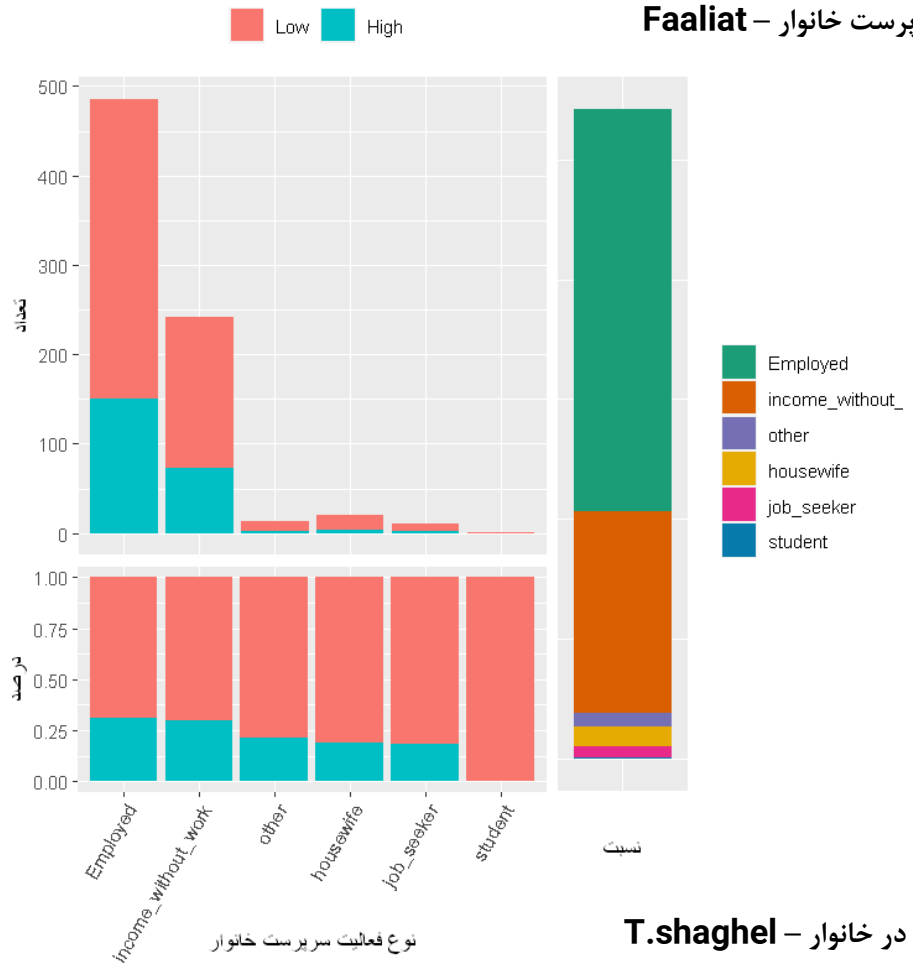
- سرپرست خانوار تحصیل می‌کند یا خیر؟ – **InEdu**



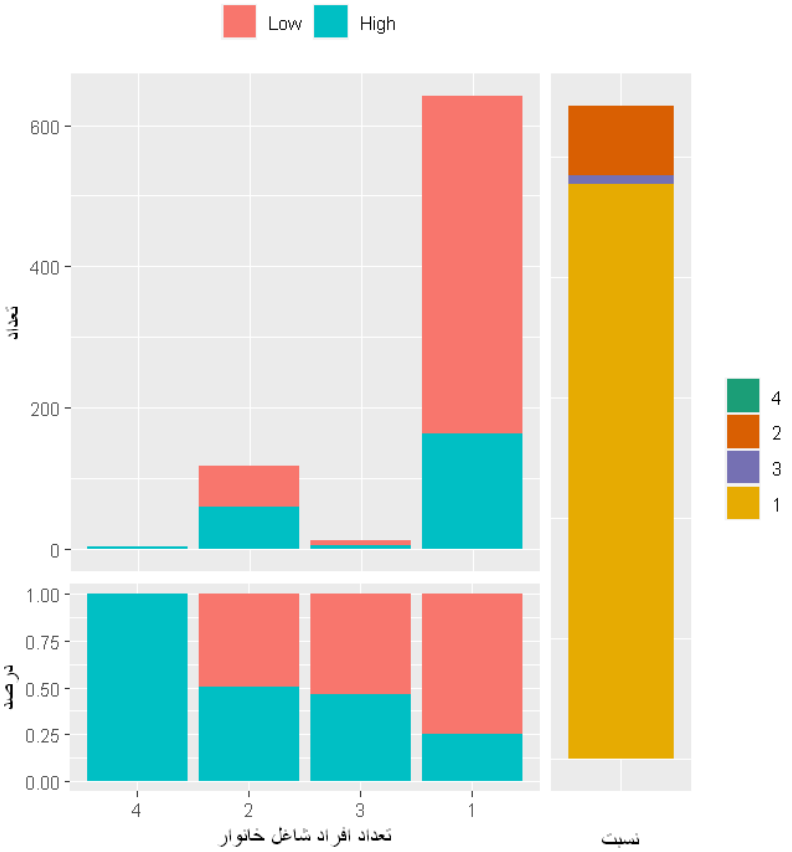
- مدرک تحصیلی سرپرست خانوار – **Madrak**



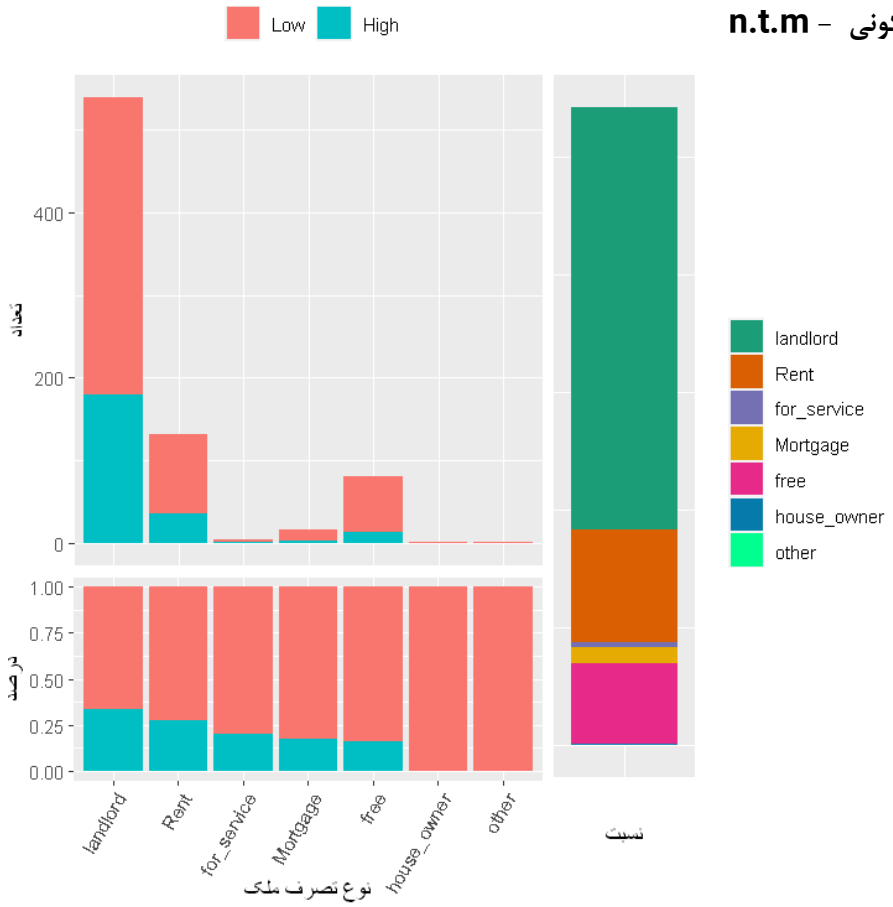
وضعیت فعالیت سرپرست خانوار – Faaliat



تعداد اعضای شاغل در خانوار – T.shaghel

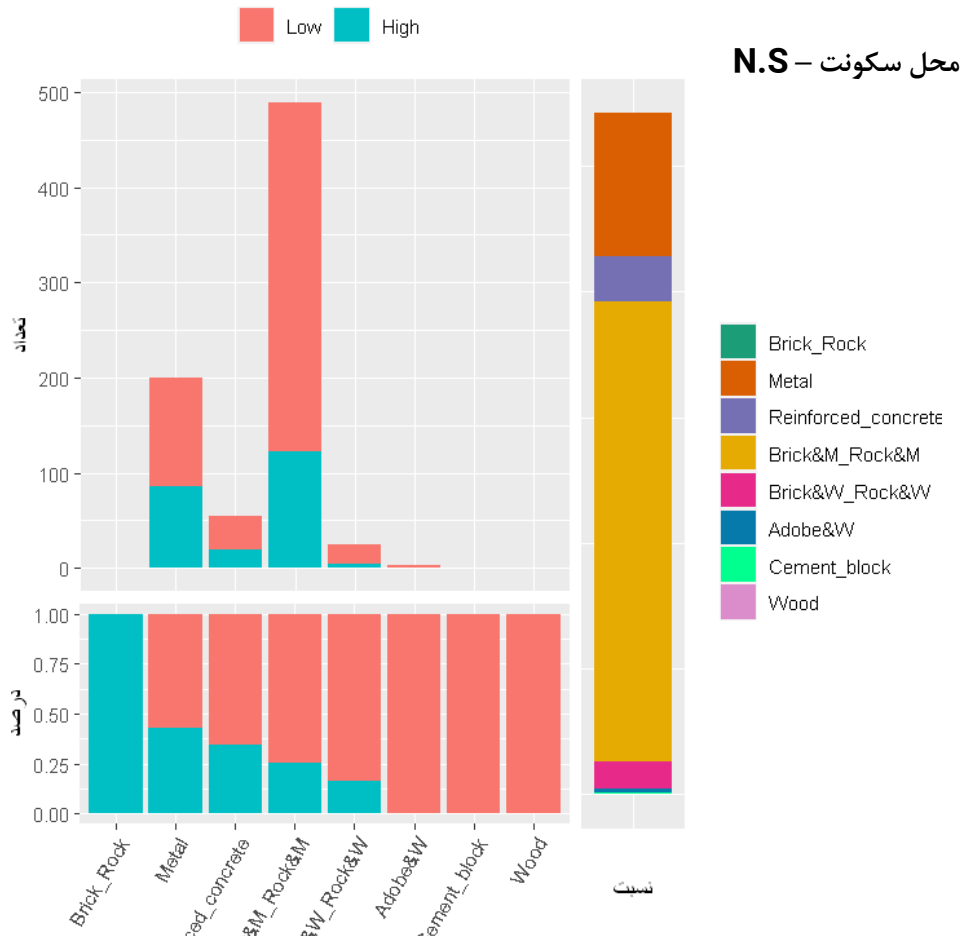


- **نحوه تصرف منزل مسکونی – n.t.m**

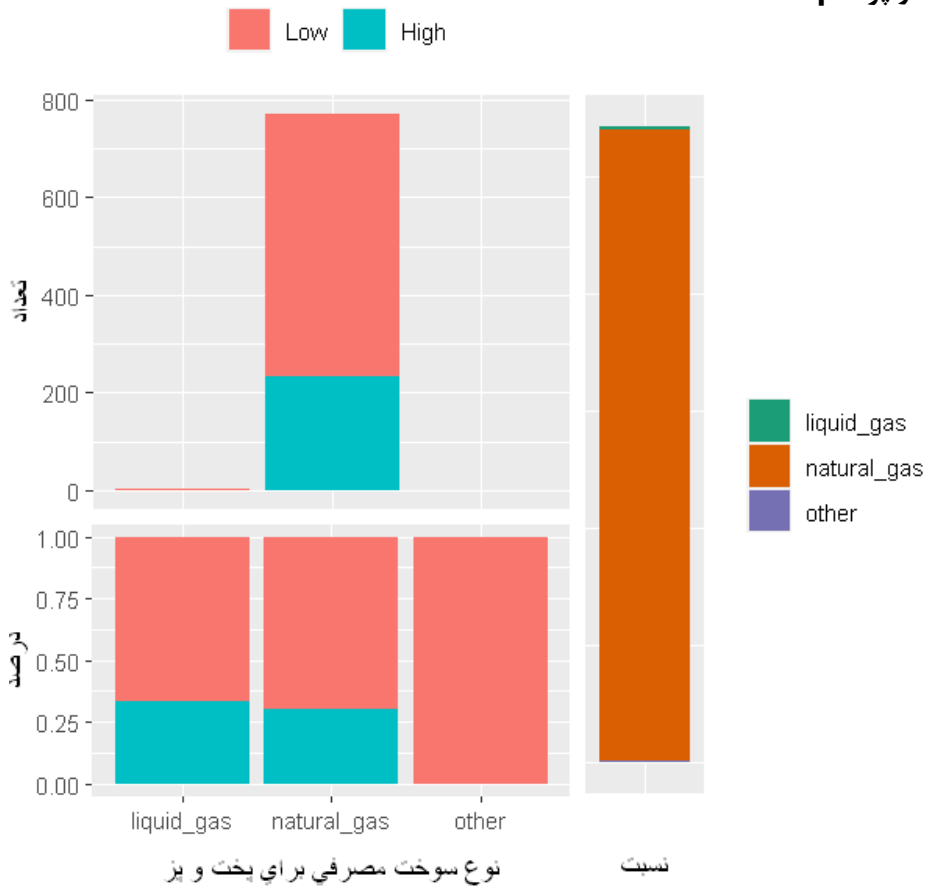


- **تعداد اتاق در اختیار – T.O**

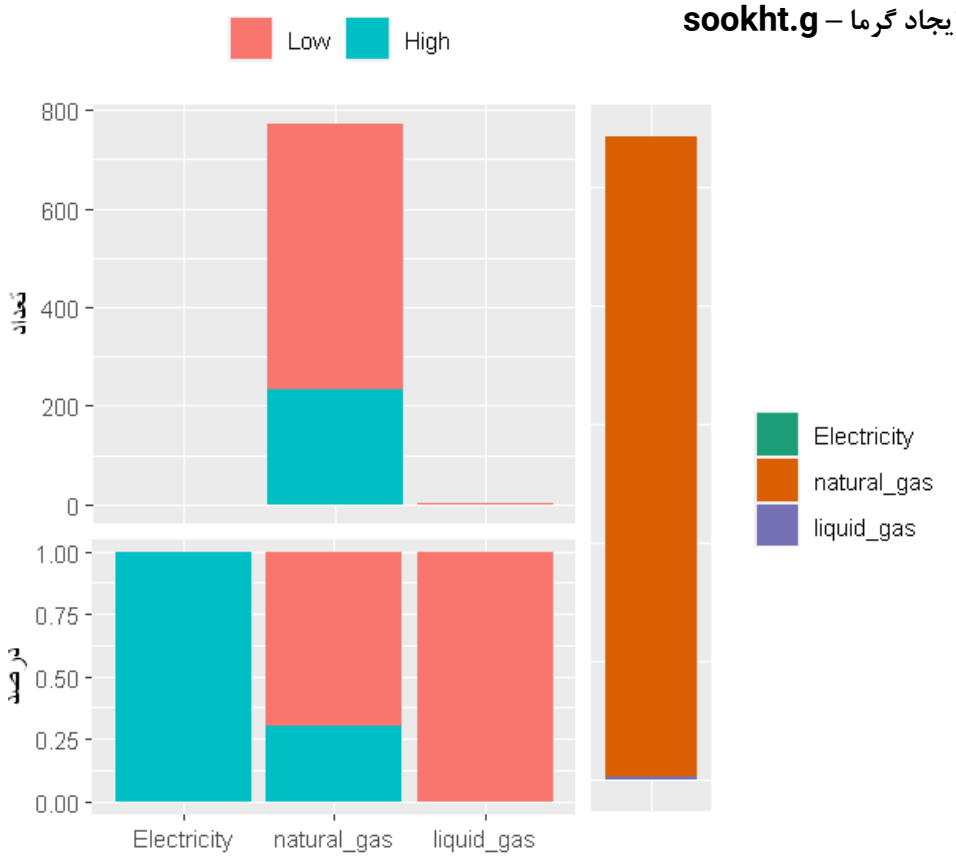
- نوع اسکلت بنای محل سکونت – **N.S**



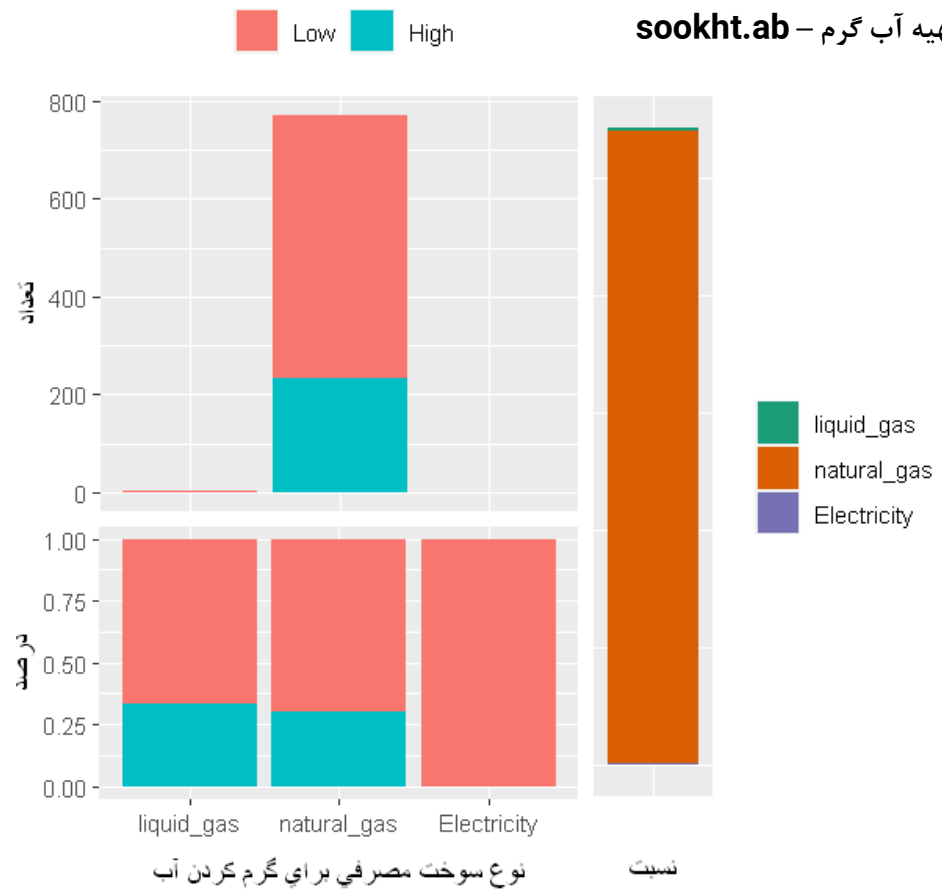
- نوع سوخت برای پخت و پز – **sookht.p**



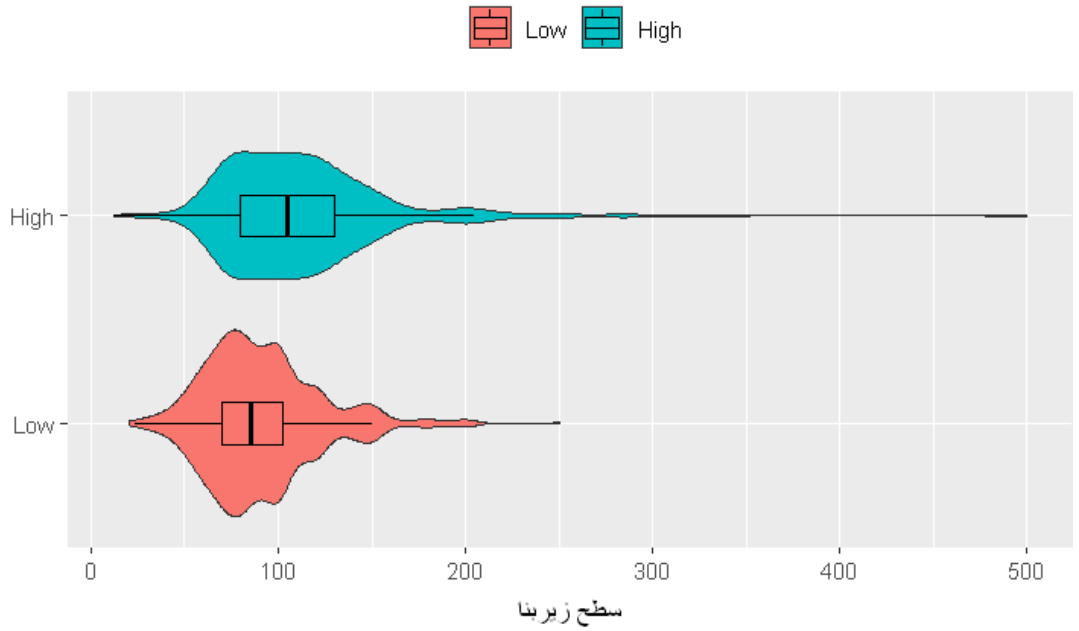
➤ نوع سوخت برای ایجاد گرما – **sookht.g**



➤ نوع سوخت برای تهیه آب گرم – **sookht.ab**



❖ سطح زیر بنای محل سکونت – S.Z

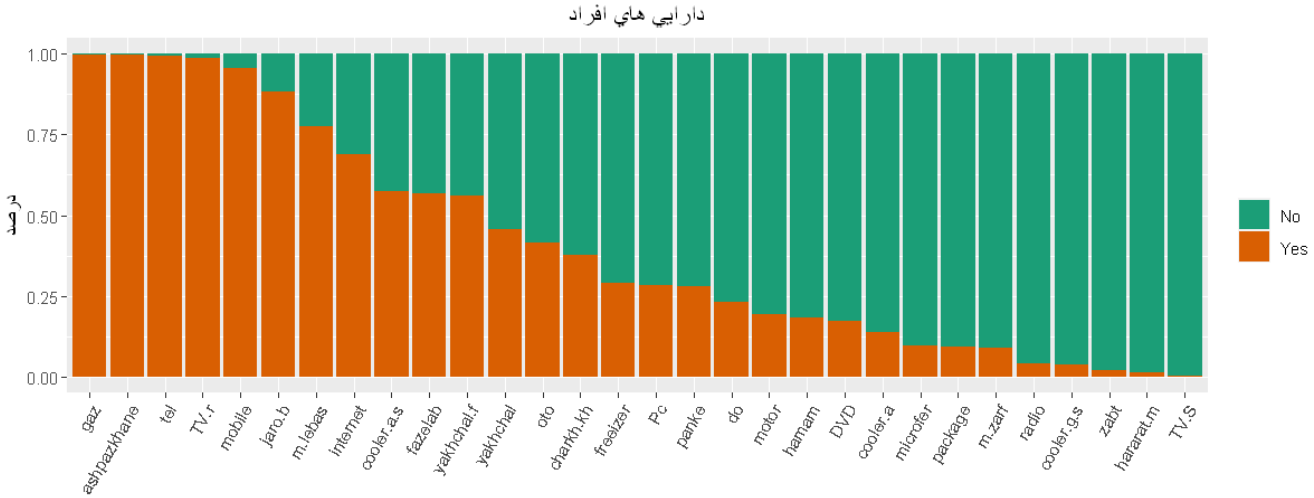


نمودار ویالن + نمودار جعبه ای

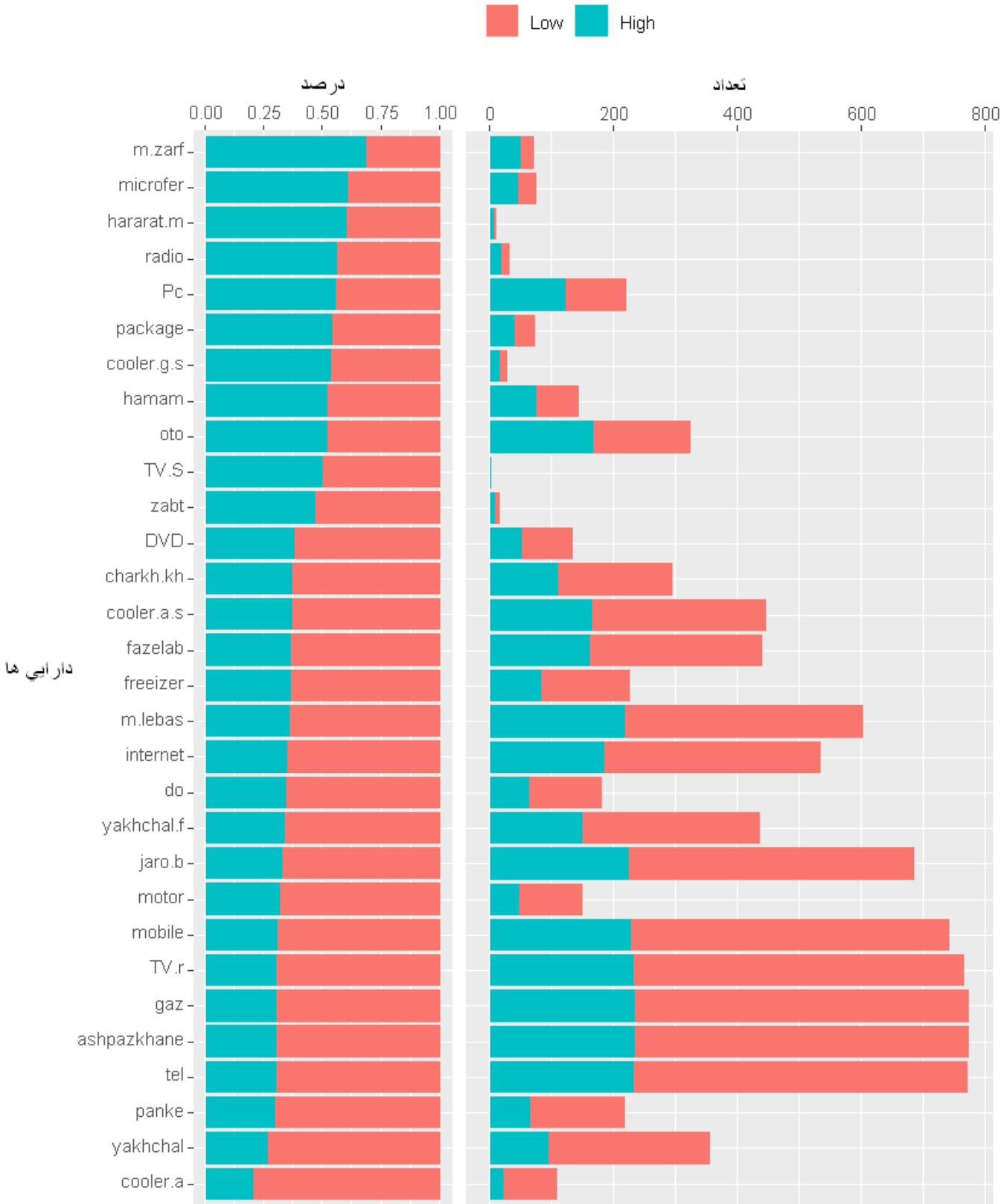
این نمودار پراکندگی سطح زیر بنای ثبت ها را به تفکیک گروه در آمدی خانوار ها نشان میدهد و همان طور که مشاهده میکنید سطح زیر بنای محل سکونت خانوار های پردرآمد نسبتا بالاتر از سطح زیر بنای محل سکونت خانواده های کم درآمد است.

❖ دارایی ها

برای بررسی نسبت دارایی ها به گروه در آمدی خانوار ها ابتدا لازم است ببینیم چه درصدی از خانوار ها یک دارایی را دارا هستند که نمودار زیر نشان دهنده این موضوع است.



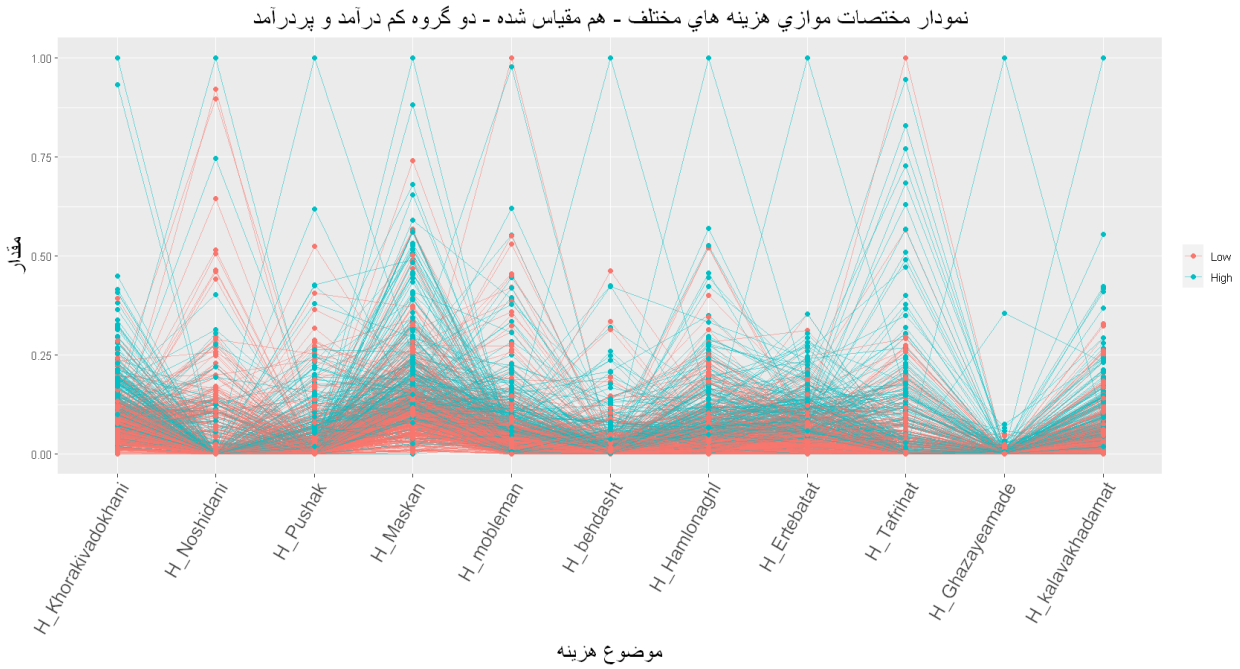
حال میتوانیم بررسی کنیم که چند درصد از افرادی که یک دارایی خاص را دارا هستند جزو گروه پردرآمد هستند و چند درصد از این افراد جزو گروه کم درآمد هستند که نمودار زیر این موضوع را نشان میدهد.



❖ هزینه ها

قصد داریم با استفاده از ۳ نمودار مختلف درک بهتری راجع به ستون های مربوط به هزینه خانوار به دست بیاوریم.

➤ نمودار مختصات موازی



در نمودار مختصات موازی ، خطوط افقی (شکسته) هر یک نشان دهنده یکی از ثبت ها است و هر یک از ستون های هزینه نیز به شکل یک خط فرضی عمودی در نظر گرفته میشود که پایین ترین نقطه این خط عمودی به کمترین میزان هزینه شده در این مورد اشاره دارد و بالاترین نقطه آن به بیشترین مقدار هزینه شده در این مورد اشاره دارد و در این نمودار هر یک از ثبت ها با توجه به مقدار هزینه ای که در هر یک از این موارد انجام داده اند نقاط مختلفی را روی این خطوط عمودی انتخاب میکنند و این نقاط را به یکدیگر متصل میکنند.

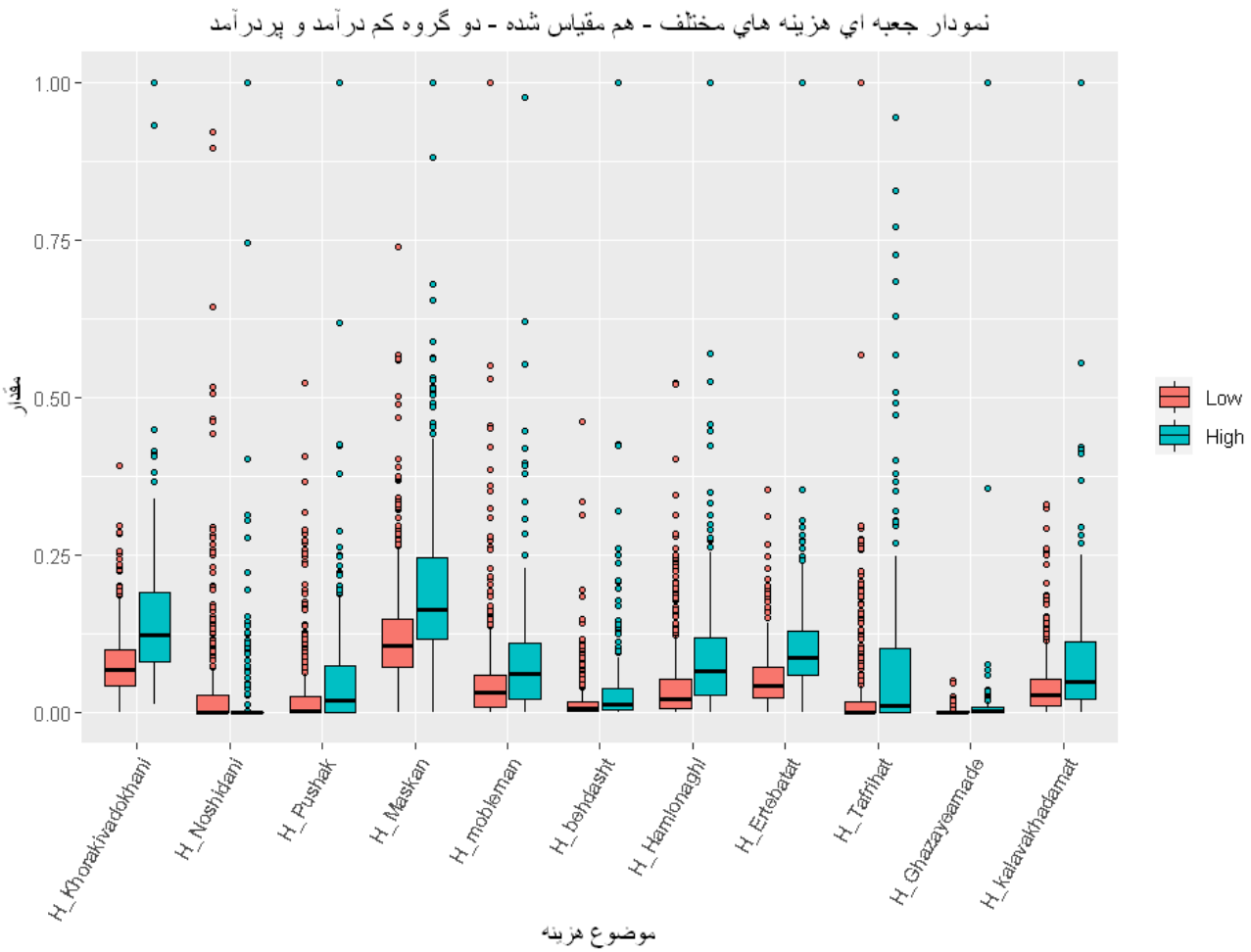
به این ترتیب ما میتوانیم پراکندگی ثبت ها رو هر یک از این متغیر ها را در کنار هم بررسی کنیم.

برای این که این ستون ها با یکدیگر قابل مقایسه باشد ابتدا مقادیر هر یک از این ستون ها را به بازه **[۰,۱]** انتقال داده ایم.

همان طور که مشاهده میکنید تراکم رنگ قرمز در این نمودار بیشتر سطوح پایین هزینه است و تراکم رنگ آبی بیشتر روی سطوح بالای هزینه است که نشان دهنده این است که خانوار های کم درآمد هزینه کمتری نیز در بخش های مختلف انجام میدهند.

همچنین به این نکته نیز توجه کنید که تنها ۰.۳ ثبت ها مربوط به خانوار های پردرآمد است و اگر تعداد ثبت های بیشتری به این گروه تعلق داشت تراکم رنگ آبی در بخش بالای نمودار بیشتر مشخص میشد.

➤ نمودار جعبه ای

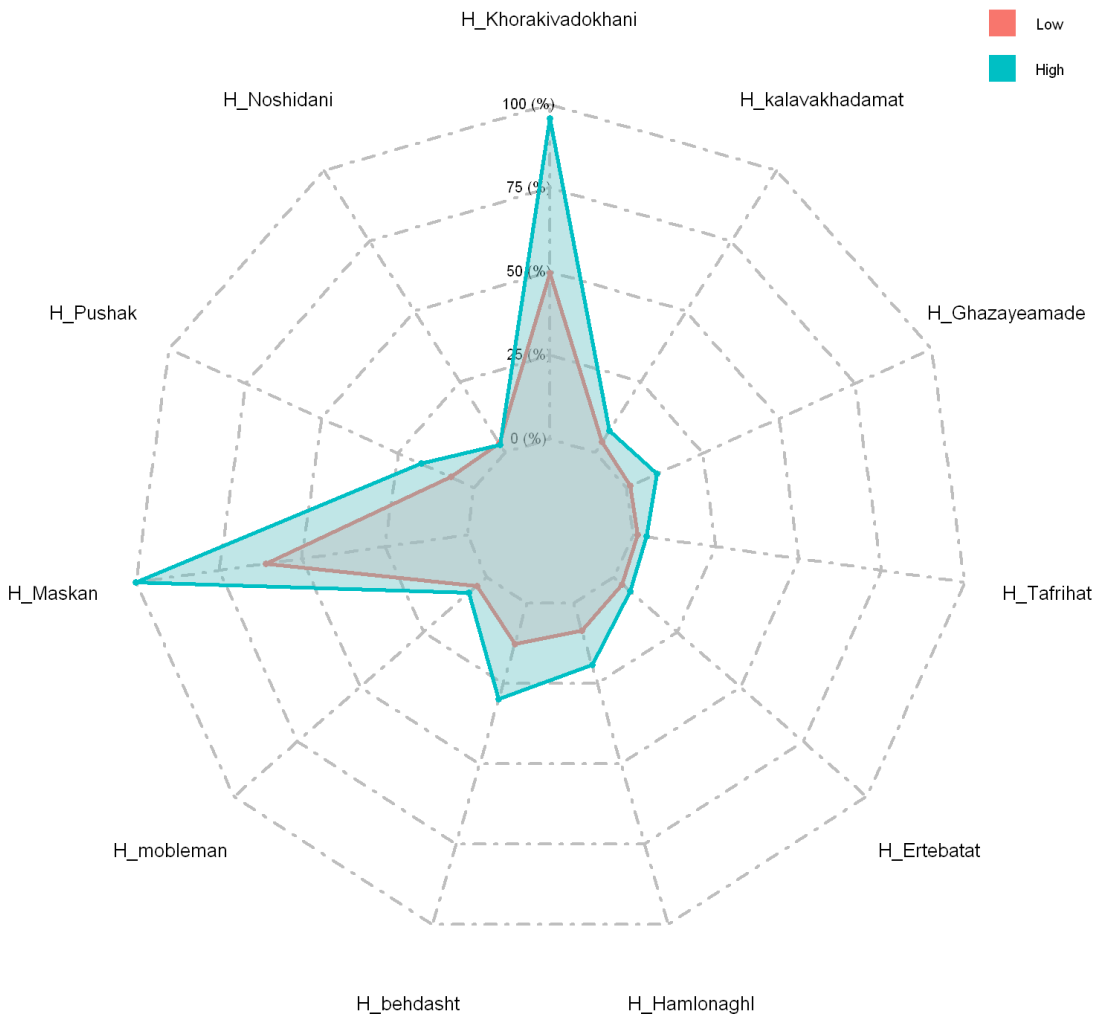


این نمودار نیز هزینه های خانوار ها در بخش های مختلف را به تفکیک گروه درآمدی نشان میدهد

برای این که این ستون ها با یکدیگر قابل مقایسه باشد ابتدا مقادیر هر یک از این ستون ها را به بازه **[۰,۱]** انتقال داده ایم.

همان طور که مشاهده میکنید معمولا هزینه های خانوار های پر درآمد در بخش های مختلف بیشتر است.

نمودار عنکبوتی



این نمودار نیز میانگین هزینه های خانوار های کم درآمد و پردرآمد را نشان میدهد.

در این نمودار مقادیر هم مقیاس نشده اند و میتوان میانگین هزینه های بخش های مختلف را نیز با یکدیگر مقایسه کرد.

اصلی ترین هزینه خانوار ها مربوط به هزینه مسکن و هزینه خوراکی و دخانی است و پس از آن هزینه بهداشت و پوشاک و حمل و نقل بیشترین هزینه را دارد.

❖ نمودار شبکه ای

این نوع از نمودار ها مانند یک گراف وزن دار است.

هر یک از راس های این گراف گروه های مختلف داده ها را نشان میدهد.

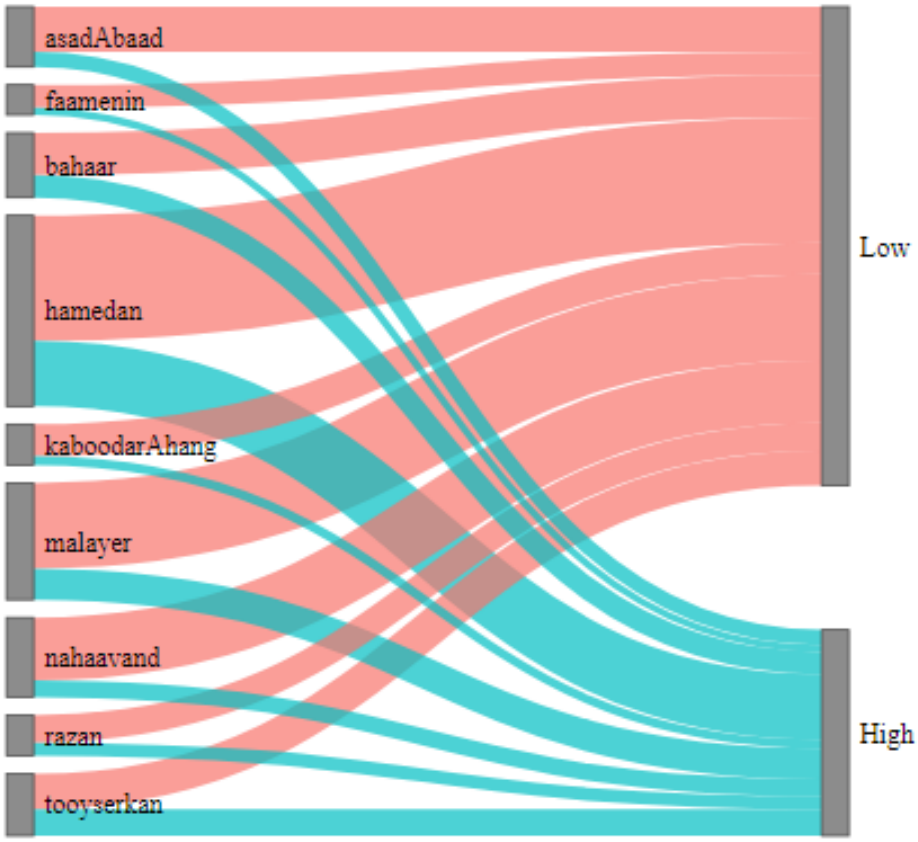
هر یک از یال های این گراف تعداد ثبت هایی را نشان میدهد که ویژگی راس ابتدایی و انتهایی آن یال را دارا هستند.

در این جا ما قصد داریم اطلاعات چند ستون مختلف را نسبت به ستون گروه درآمدی رسم کنیم.

پس یک گروه از راس ها را به ستون گروه درآمدی نسبت میدهیم که شامل دو راس High و Low است.

گروه دیگر را نیز به ستون های دیگر در نظر میگیریم.

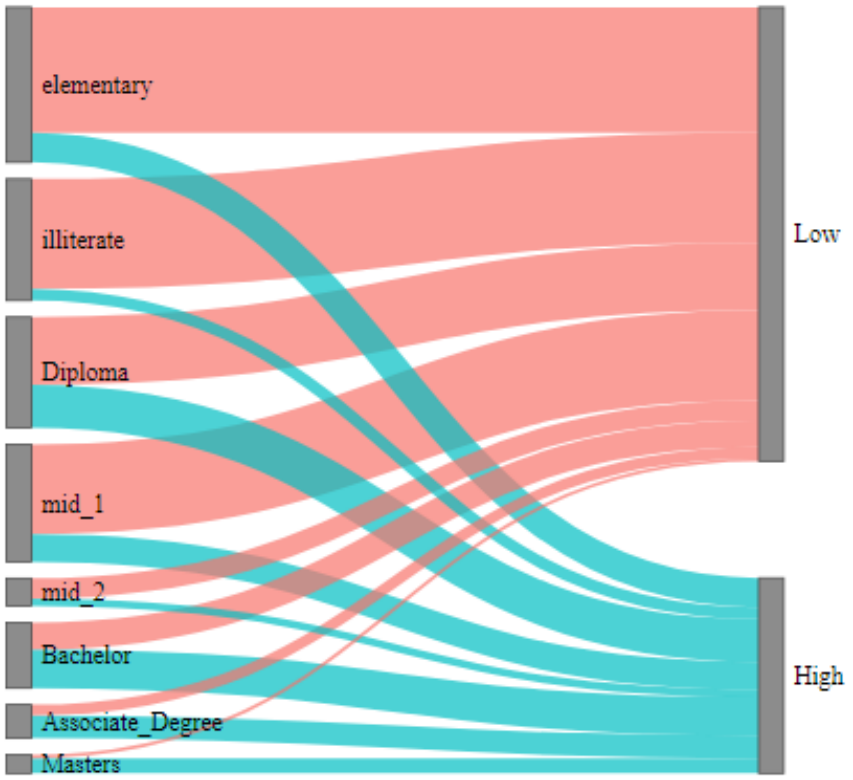
➤ آدرس خانوار – Address



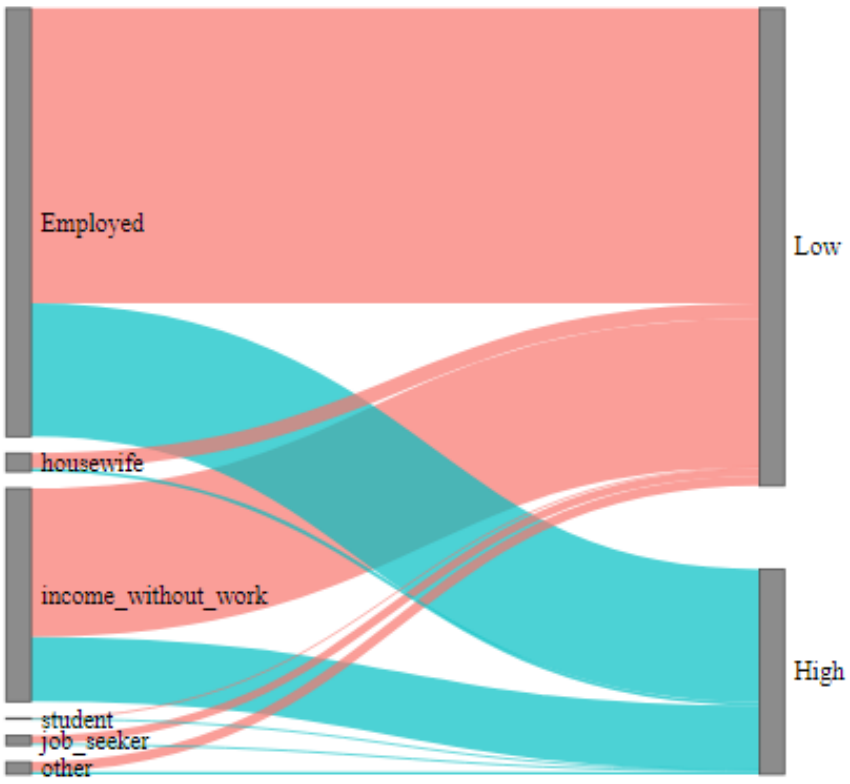
به وسیله این نمودار میتوانیم نسبت اندازه گروه های مختلف را به یکدیگر بهتر متوجه شویم

همچنین میتوانیم متوجه شویم که بخش اصلی گروه های پردرآمد و کم درآمد مربوط به کدام رده (در این مثال شهرستان) میباشد.

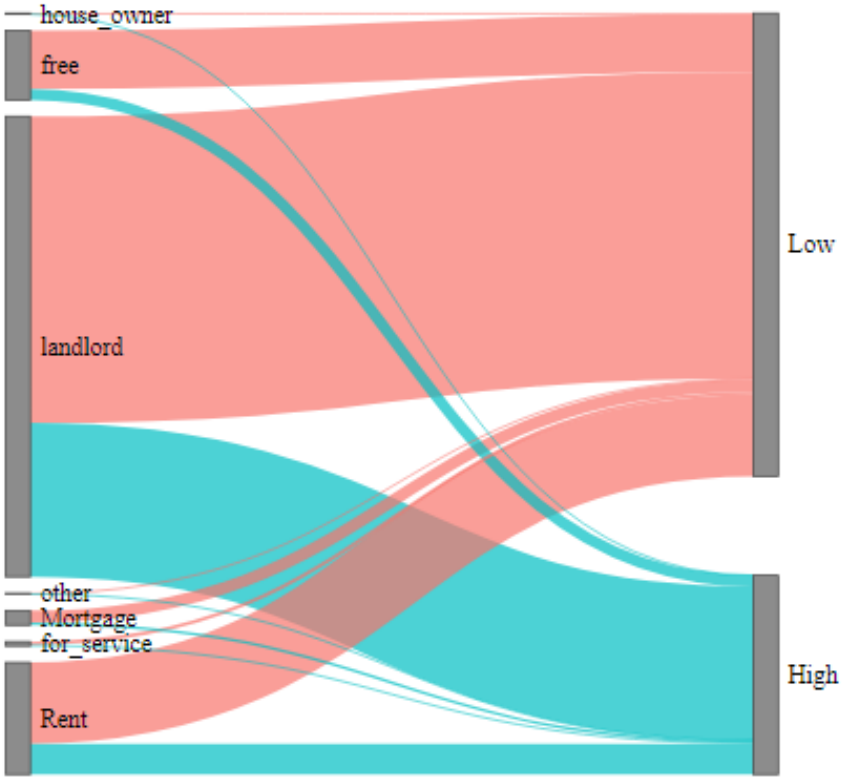
➤ مدرک تحصیلی سرپرست خانوار – **Madrak**



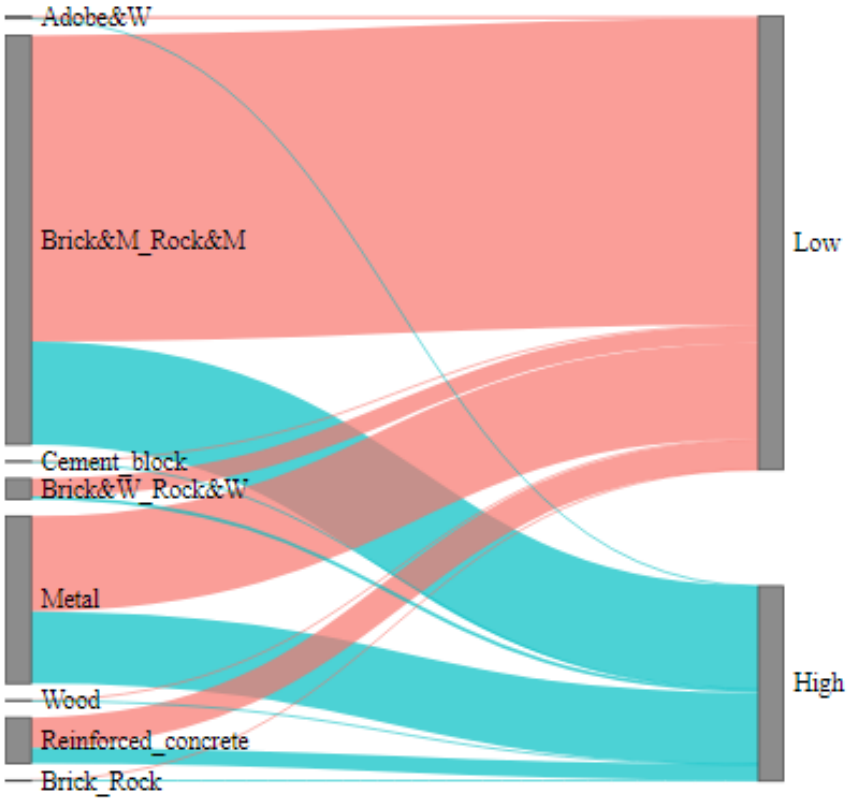
➤ وضعیت فعالیت سرپرست خانوار – **Faaliat**



➤ نحوه تصرف منزل مسکونی - **n.t.m**



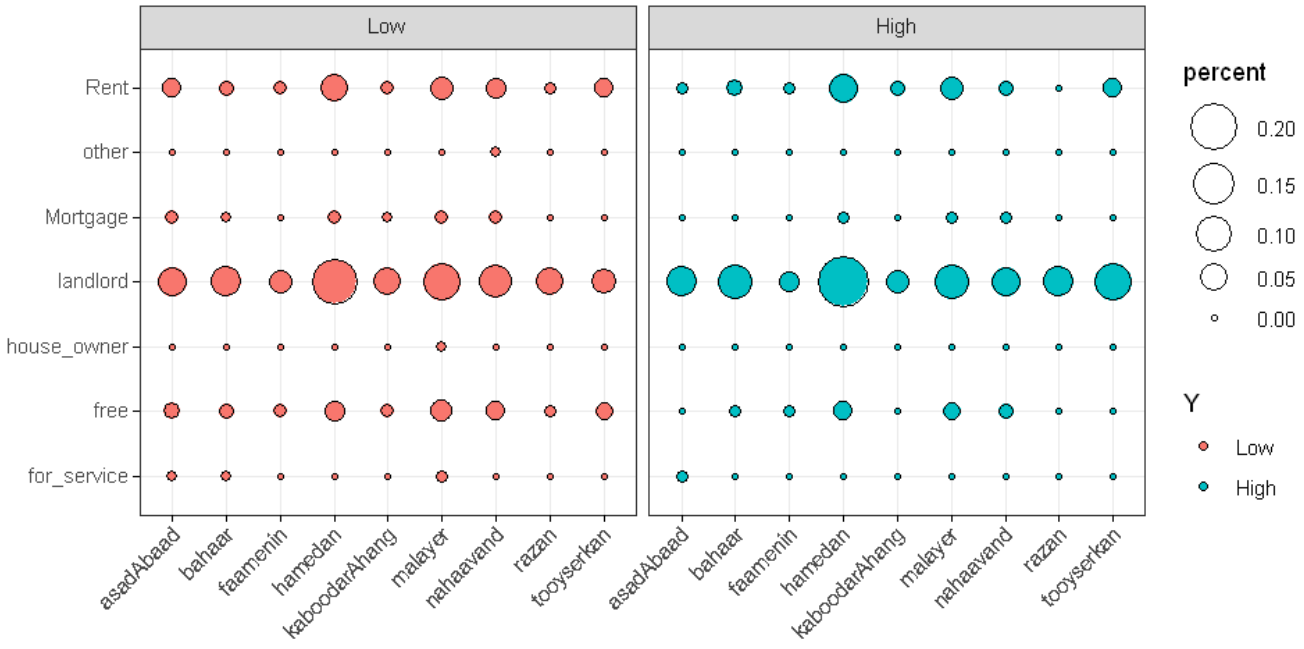
➤ نوع اسکلت بنای محل سکونت – **N.S**



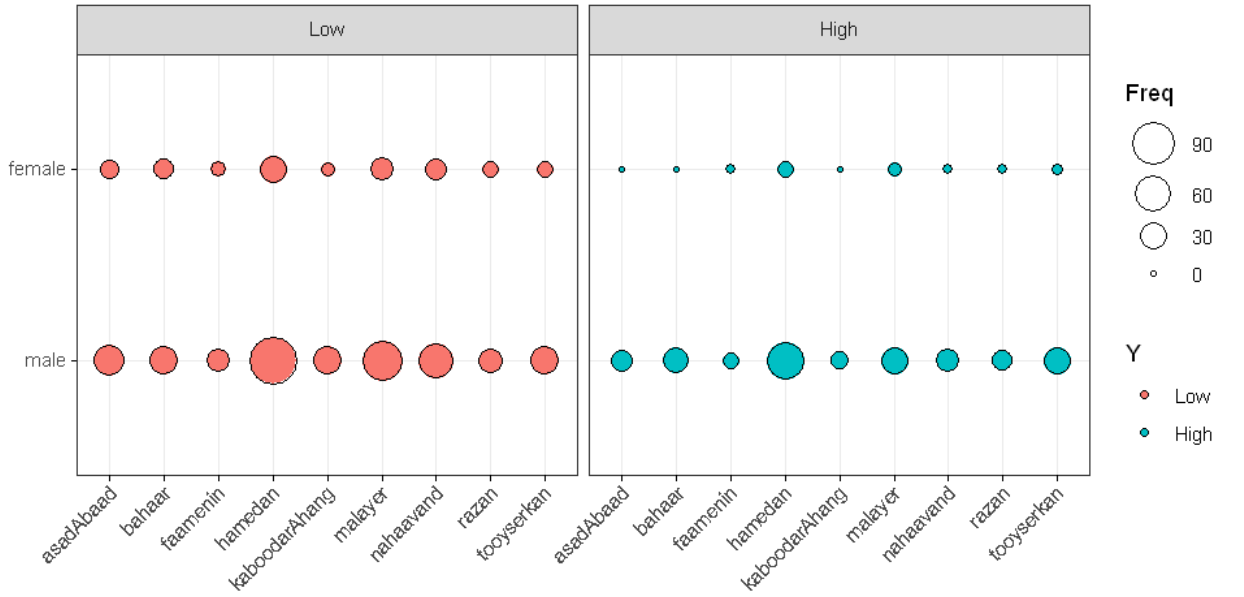
❖ نمودار های چند متغیر نسبت به متغیر هدف

برای این بخش از نمودار بالنی استفاده کردیم که اندازه هر دایره در این نمودار نشان دهنده تعداد ثبت های مربوط به آن گروه است.

Address & n.t.m ➤

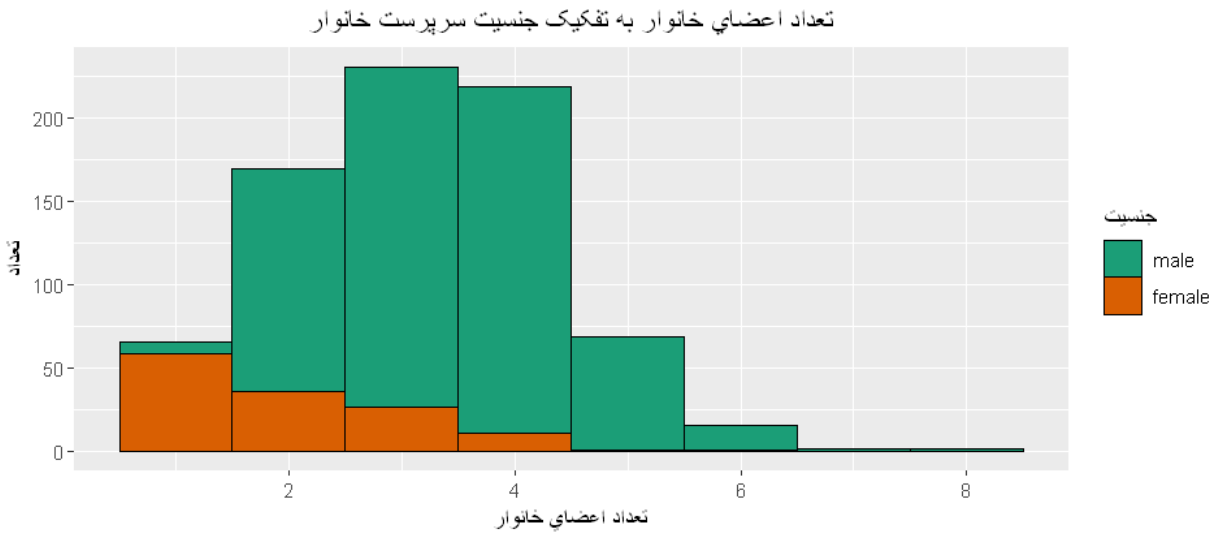


Address & Gender ➤



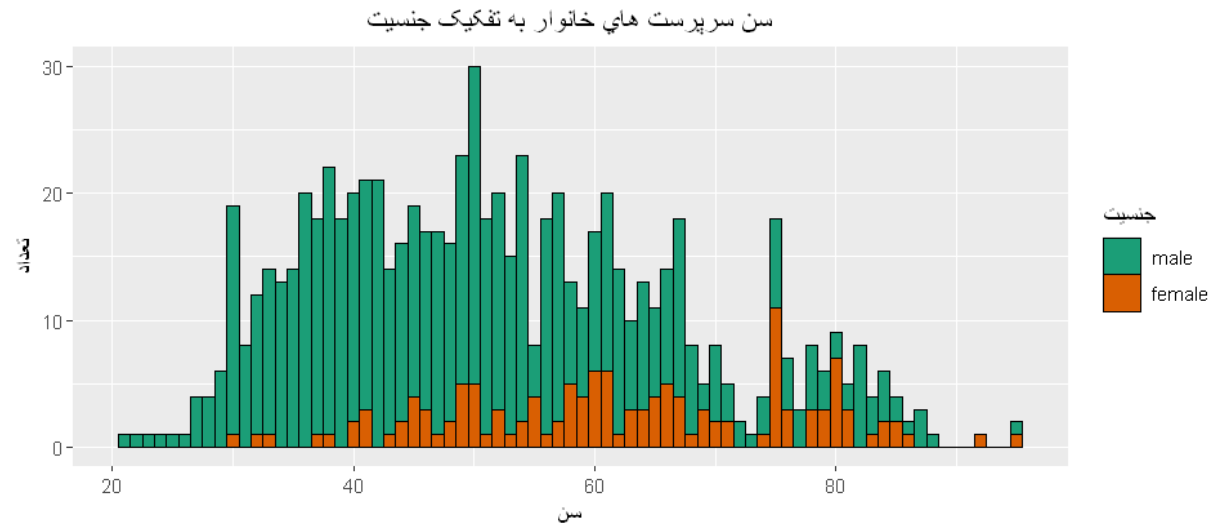
❖ نمودار متغیر های دیگر نسبت به هم

➤ Tedad.a & Gender



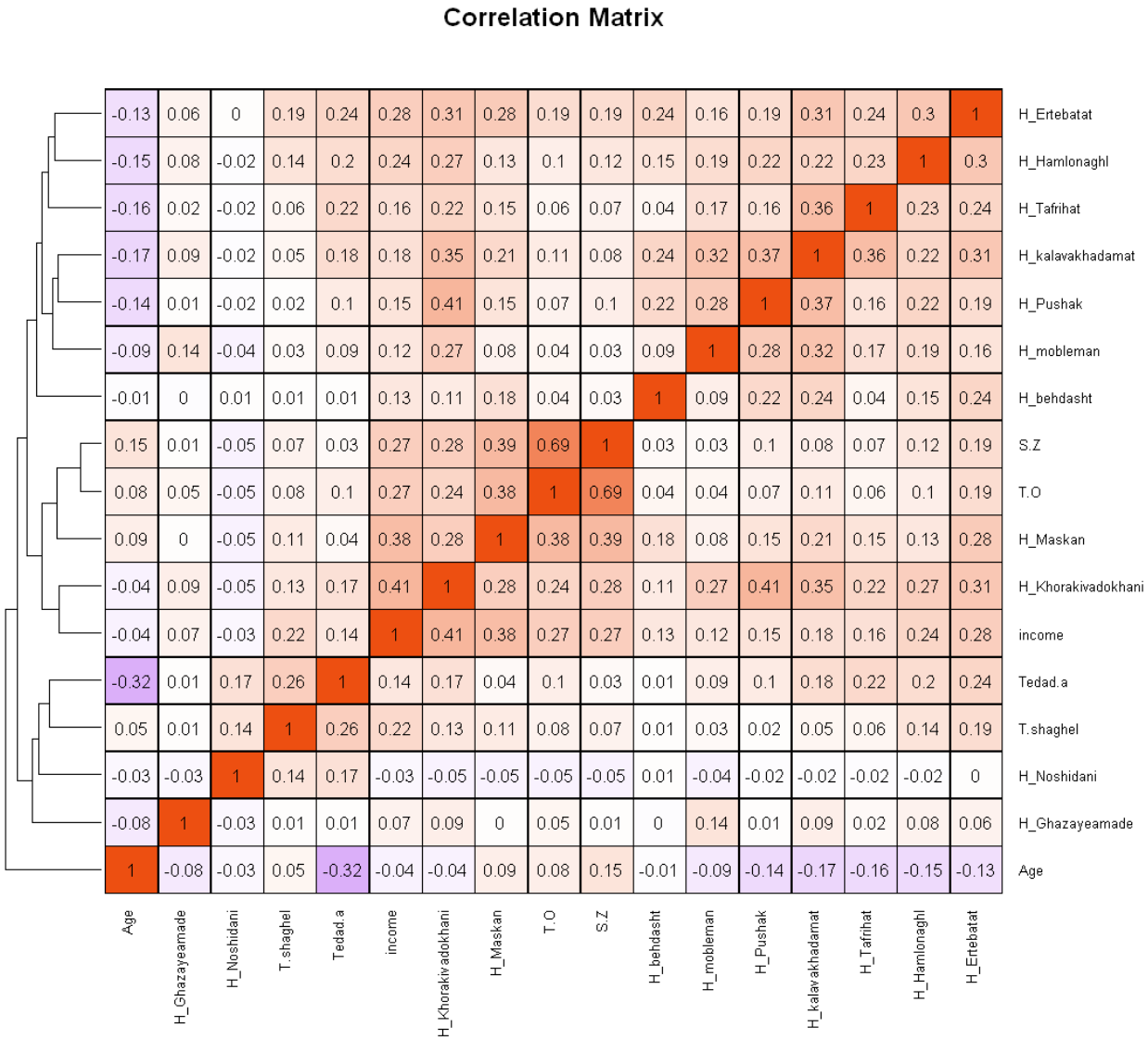
در خانوار های با جمعیت کم ، سرپرستان عمدتا بانوان هستند و در خانوار های با جمعیت زیادتر ، سرپرستان عمدتا آقایان هستند

➤ نمودار میله ای Age & Gender



میتواند نشان دهنده این موضوع باشد که عمدتا آقایان زودتر از بانوان فوت میکنند و این موضوع باعث میشود با افزایش سن سرپرستان خانوار ، تعداد سرپرستان خانوار بانو افزایش پیدا کند.

❖ نمودار حرارتی متغیر های عددی و درآ مد کل



اعداد روی این نمودار نشان دهنده همبستگی بین ستون ها است و همبستگی های مثبت تر با رنگ نارنجی و همبستگی های منفی تر با رنگ بنفش نشان داده شده است.

مقادیر هزینه شده در بخش های مختلف توسط خانوار ها نیز عموما با یکدیگر همبستگی مثبت دارند که یعنی با افزایش توانایی خانوار برای هزینه کردن در یک بخش تواناییش برای هزینه کردن در سایر بخش ها نیز عموما افزایش میابد.

با افزایش سن سرپرست خانوار تعداد اعضای خانواده کاهش میباند و همچنین قدرت هزینه کردن خانوار نیز کاهش میابد.

با افزایش سطح زیر بنا تعداد اتاق های ملک افزایش میابد.

هزینه خوراکی و هزینه مسکن و هزینه ارتباطات بیشترین تاثیر مثبت را روی درآمد خانوار دارند

۵. اجرای مدل های مختلف

در این بخش قصد داریم تا با استفاده از ۴ نوع مدل ، رده بندی را در این داده ها انجام دهیم.

این ۴ نوع مدل عبارتند از :

- K – نزدیک ترین همسایه
- درخت رده بندی
- رگرسیون لجستیک
- شبکه عصبی مصنوعی

برای هر یک از این ۴ نوع مدل چندین مدل مختلف با خصوصیات مختلف را برازش میدهیم و در نهایت با بررسی عملکرد این مدل ها روی داده های آزمایشی بهترین مدل های هر بخش را انتخاب میکنیم و در نهایت بین بهترین مدل های هر بخش یکی را به عنوان بهرین مدل رده بندی انتخاب میکنیم.

برای بررسی عملکرد مدل های رده بندی از جدولی با عنوان جدول در هم ریختگی استفاده میکنیم که حالت کلی آن به شکل زیر است.

هر یک از عناصر جدول به شرح ذیل می‌باشد:

TP : بیانگر تعداد ثبت هایی است که دسته واقعی آنها مثبت بوده و الگوریتم دسته بندی نیز دسته آنها را بدرستی مثبت تشخیص داده است.

TN : بیانگر تعداد ثبت هایی است که دسته واقعی آنها منفی بوده و الگوریتم دسته بندی نیز دسته آنها را بدرستی منفی تشخیص داده است.

FP : بیانگر تعداد ثبت هایی است که دسته واقعی آنها منفی بوده و الگوریتم دسته بندی دسته آنها را به اشتباه مثبت تشخیص داده است.

FN : بیانگر تعداد ثبت هایی است که دسته واقعی آنها مثبت بوده و الگوریتم دسته بندی دسته آنها را به اشتباه منفی تشخیص داده است.

در این تعاریف مثبت و منفی نشان دهنده دو رده داده است که در مثال ما به High و Low تغییر میکند.

شاخص های متفاوت دیگری نیز میتوان بسته به نیاز از این جدول استخراج کرد که در زیر برخی از مهم ترین این شاخص ها را بررسی میکنیم.

مهمترین معیار برای تعین کارایی کلی یک الگوریتم دسته بندی دقت یا نرخ دسته بندی (Classification Accuracy) است که این معیار دقت کل یک دسته بند را محاسبه می‌کند. در واقع این معیار مشهورترین و عمومی‌ترین معیار محاسبه کارایی الگوریتم‌های دسته بندی است که نشان می‌دهد، دسته بند طراحی شده چند درصد از کل مجموعه رکوردهای آزمایشی را بدرستی دسته بندی کرده است.

دقت دسته بندی با استفاده از رابطه زیر بدست می‌آید که بیان می‌کند دو مقدار TP و TN مهمترین مقادیری هستند که در یک مسئله دودسته ای باید بیشینه شوند.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

معیار خطای دسته بندی (Error Rate) دقیقاً برعکس معیار دقت دسته بندی است که با استفاده از رابطه ۱ - ACC بدست می‌آید. کمترین مقدار آن برابر صفر است زمانی که بهترین کارایی را داریم و بطور مشابه بیشترین مقدار آن برابر یک است زمانی که کمترین کارائی را داریم.

یکی دیگر از معیار های مهم کارایی که مناسب زمان هایی است که هدف ما به دست آوردن بهترین دقت ممکن روی یک رده خاص است معیار حساست یا (Sensitivity) است که به آن نرخ پاسخ‌های مثبت درست (True Positive Rate) نیز میگویند که این شاخص نشان دهنده این است که مدل ما چند درصد از مواردی که واقعا درگروه مثبت هستند را به درستی در گروه مثبت قرار داده است و از رابطه زیر بدست می آید. زمانی که از این شاخص برای انتخاب بهترین مدل استفاده میکنیم هدف ما دستیابی به مدلی است که در رده بندی گروه مثبت بهترین عملکرد را دارا باشد.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

در مقابل گاهی ممکن است هدف اصلی ما دستیابی به بهترین عملکرد در رده بندی گروه منفی باشد ، در این مواقع از شاخص دیگری استفاده که شباهت زیادی به شاخص حساسیت دارد و نام آن شاخص مشخصه سازی یا (Specificity) است که به آن نرخ پاسخ های منفی درست (True Negative Rate) نیز میگویند و از رابطه زیر بدست می آید.

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

در این جا رده بندی درست خانوار های پردرآمد برای ما بسیار حائز اهمیت است پس یکی از این دو شاخص برای ما مهم هستند.

K – نزدیک ترین همسایه (KNN)

در مدل knn ما با استفاده از پیدا کردن نزدیک ترین ثبت ها به ثبت مورد نظرمان که قصد رده بندی آن را داریم و بررسی رده های این ثبت ها رده ای را برمیگزینیم که بیشترین تکرار را بین همسایه ها داشته باشد.

برای محاسبه این فاصله از فاصله اقلیدسی بین نقاط استفاده میکنیم که فرمول آن به شکل زیر است.

$$d(\mathbf{p},\mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

که در این جا اعداد ۱ تا n مشخص کننده ویژگی های مختلف ثبت های p و q هستند.

پیش پردازش

در بخش پیش پردازش باید دو عمل اصلی انجام دهیم

- **تبدیل ستون های رسته ای به ستون های ظاهری**

مدل knn برای پیدا کردن فاصله ثبت ها از فاصله اقلیدسی بین ثبت ها استفاده میکند و برای محاسبه این فاصله نیاز است که مقادیر ستون های مختلف داده ها به شکل عددی باشند به همین دلیل نمیتوان از متغیر های رسته ای برای این کار استفاده کرد پس باید راهی پیدا کنیم تا بتوان متغیر های رسته ای را به صورت عددی نمایش دهیم.
اگر متغیر رسته ای مورد نظر حالت ترتیبی داشته باشد ، مثلا متغیری که مقادیر مختلف آن "بد" "متوسط" "خوب" باشند، در این جا میتوان متغیر ها را به اعدادی ترتیبی تبدیل کرد مثلا در این جا به ۱- ، ۰ ، ۱ میتوان تبدیل کرد. اما اگر این متغیر ها ترتیبی نداشته باشند باید آن ها را به متغیر ظاهری تبدیل کنیم که به آن نمایش one hot encoding نیز میگویند که به ازای هر رسته در هر متغیر رسته ای یک متغیر جدید ساخته میشود که اگر آن داده عضو آن رسته باشد مقدارش برابر ۱ و در غیر اینصورت برابر ۰ خواهد بود و در نهایت ستون اول ساخته شده در این فرایند نیز حذف میشود زیرا سایر ستون ها تمام اطلاعات را نمایش میدهند و این ستون اطلاعات اضافه ای ندارد.

- **نرمال کردن ستون های عددی**

در پیدا کردن فاصله بین ثبت ها اگر مقیاس اعداد یکی از ستون ها از سایر ستون ها خیلی بزرگتر باشد باعث میشود که تاثیر سایر ستون ها در محاسبه فاصله بین ثبت ها بسیار کم شود و این موضوع باعث از دست رفتن اطلاعات و ضعیف شدن مدل ما میشود به همین منظور ابتدا باید داده ها را هم مقیاس کنیم.
برای این کار میانگین هر ستون را از مقادیر آن ستون کم میکنیم و سپس مقادیر به دست آمده را به واریانس آن ستون تقسیم میکنیم.

➤ افراز داده ها

در این جا ۰.۸ ثبت ها را به صورت تصادفی انتخاب میکنیم و برای آموزش مدل در نظر میگیریم که به آن داده آموزشی میگوییم که خود این ثبت ها را به دو بخش آموزشی و آزمایشی تقسیم میکنیم که در بخش بعد مختصرا دلیل آن را توضیح میدهیم.

۰.۲ باقی مانده از ثبت ها را نیز برای بررسی عملکرد مدل کنار میگذاریم و برای آموزش مدل هیچ استفاده ای از این ثبت ها نمیکنیم که به این داده ها داده های آزمایشی میگوییم.

➤ اعتبار سنجی متقابل (cross validation)

اغلب در مدل‌سازی احتیاج به برآورد پارامترهای مدل داریم. در صورتی که تعداد پارامترها زیاد باشد، پیچیدگی مدل زیاد شده و ممکن است محاسبات به سادگی قابل انجام نباشند. اعتبارسنجی متقابل یکی از راه‌هایی است که می‌توان تعداد پارامترها (متغیرهای) مدل را بصورت بهینه تعیین کرد.

برای مدل KNN پارامتری که میخواهیم مقدار بهینه آن را پیدا کنیم عدد K است که مشخص کننده تعداد همسایه هایی است که برای رده بندی هر ثبت جدید باید بررسی شود.

روش اعتبار سنجی متقابل به این شکل است که هر بار یک بخش از داده های آموزشی را برای ارزیابی مدل به دست آمده کنار میگذارد که به آن داده اتبارسنجی میگوییم ، و پس از برازش مدل روی داده های آموزشی عملکرد مدل را روی داده های اعتبارسنجی بررسی میکنیم و این کار را در چند تکرار انجام میدهیم و هر بار یک بخش از داده های آموزشی را به منظور ارزیابی عملکرد مدل بدست آمده کنار میگذاریم و در نهایت میانگین عملکرد مدل در این تکرار ها تا حد خوبی نشان دهنده عملکرد کلی مدل روی داده های جدید خواهد بود.

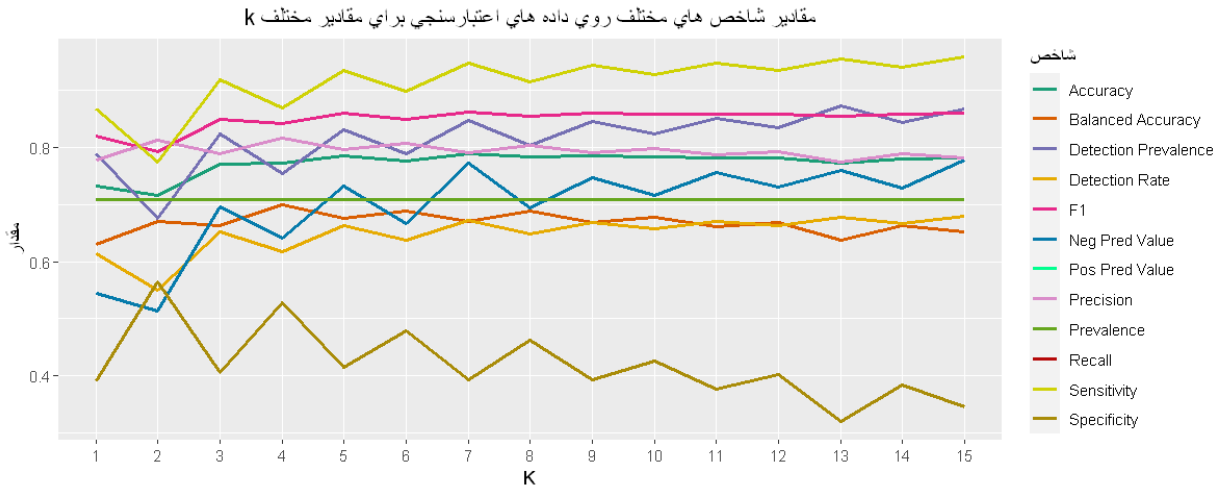
بنابراین پس از انجام اعتبار سنجی متقابل بهترین مقدار پارامتر مورد نظرمان را برای استفاده در مدل اصلی انتخاب میکنیم و مدل اصلی ساخته شده را روی داده های آزمایشی بررسی میکنیم که هیچ تاثیری در آموزش مدل نداشته است و عملکرد مدل روی این ثبت ها نشان دهنده عملکرد نهایی مدل است.

در نهایت پس از پیدا کردن بهترین مقدار برای پارامتر مورد نظر میتوانیم از تمام داده های آموزشی و اعتبارسنجی برای آموزش مدل اصلی استفاده کنیم و به این ترتیب قدرت مدل را به بیشترین حد آن برسانیم.

➤ مدل

در این جا اعتبار سنجی متقابل را برای K های ۱ تا ۱۵ به ۵ تکرار انجام میدهیم و میانگین شاخص های مختلف عملکرد مدل را در این ۵ تکرار ذخیره میکنیم.

خروجی به دست آمده به شکل زیر است.



در این نمودار چندین شاخص دیگر نیز وجود دارد که به توضیح آن ها نمیپردازیم.

در این جا رده High ، کلاس منفی ما میباشد به همین دلیل شاخص مهم مد نظر برای ما در این بخش شاخص Specificity است و بالاترین مقدار این شاخص در K=۲ رخ داده است پس از این پارامتر برای برازش مدل نهایی استفاده میکنیم.

خروجی مدل روی داده های آموزشی به شکل زیر است.

خروجی سایر مدل ها را نیز در چنین قالبی نمایش میدهیم.

یک بار در این جا بخش های مهم این خروجی را توضیح میدهیم.

بخش ۱ : ماتریس در هم ریختگی مدل

بخش ۲ : دقت مدل

بخش ۳ : حساسیت و مشخصه سازی مدل

بخش ۴ : مشخص کننده این که کدام رده را باید رده مثبت در نظر بگیریم و کدام رده را رده منفی در نظر بگیریم.

```
          Reference
Prediction Low High
      Low  383    0
      High  56  181
```
1

```
               Accuracy : 0.9097
                 95% CI : (0.8843, 0.9311)
    No Information Rate : 0.7081
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7997

 Mcnemar's Test P-Value : 1.987e-13

            Sensitivity : 0.8724
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.7637
             Prevalence : 0.7081
         Detection Rate : 0.6177
   Detection Prevalence : 0.6177
      Balanced Accuracy : 0.9362

       'Positive' Class : Low
```
2 3 4

همان طور که انتظار میرود مدل روی داده های آموزشی عملکرد بسیار خوبی دارد اما برای ما ملاک عملکرد مدل روی داده هایی است که تا به حال ندیده یعنی داده های آزمایشی.

عملکرد مدل روی داده های آزمایشی به شکل زیر است.

```
          Reference
Prediction Low High
      Low   83   18
      High  20   34

               Accuracy : 0.7548
                 95% CI : (0.6794, 0.8203)
    No Information Rate : 0.6645
    P-Value [Acc > NIR] : 0.009507

                  Kappa : 0.4553

 Mcnemar's Test P-Value : 0.871131

            Sensitivity : 0.8058
            Specificity : 0.6538
         Pos Pred Value : 0.8218
         Neg Pred Value : 0.6296
             Prevalence : 0.6645
         Detection Rate : 0.5355
   Detection Prevalence : 0.6516
      Balanced Accuracy : 0.7298

       'Positive' Class : Low
```

در این جا همان طور که مشاهده میکنید مدل توانسته ۶۵ درصد خانوار های پردرآمد را درست رده بندی کند و در کل نیز ۷۴ درصد داده ها را به درستی رده بندی کند و عملکرد قابل قبولی داشته است.

برای بهتر کردن این مدل میتوان به جای استفاده از میانگین ساده از میانگین وزن دار استفاده کرد و تاثیر فاصله ثبت ها را نیز در رده بندی در نظر گرفت.

➢ درخت تصمیم

درخت تصمیم دقیقا مانند یک درخت است با این تفاوت که از ریشه به سمت پایین (برگ) رشد کرده است. در الگوریتم درخت تصمیم نمونه‌ها را دسته‌بندی می‌کنیم که در واقع دسته‌ها در انتهای گره‌های برگ قرار دارد و تمام راس های میانی مانند یک شرط عمل میکند که در صورت داشتن یا نداشتن آن شرط به سمت چپ یا راست آن راس هدایت میشویم. درخت تصمیم در مسائلی کاربرد دارد که بتوان آنها را به صورتی مطرح نمود که پاسخ واحدی به صورت نام یک دسته یا کلاس ارائه دهند.

همچنین یکی از نکات مهم در مورد مدل درخت این است که تفسیر آن بسیار ساده است و هر فرد غیر متخصصی نیز میتواند با نگاه انداختن به درخت متوجه فرایند تصمیم گیری آن شود و این یکی از نقاط قوت اصلی درخت نسبت به بسیاری از روش های داده کاوی است.

➢ پیش پردازش

در این بخش تنها باید یک پردازش انجام دهیم

- **تبدیل ستون های رسته ای به ستون های ظاهری**

  توضیحات این فرایند در مدل KNN داده شده است.

➢ افراز داده ها

۰.۸ ثبت ها را به صورت تصادفی برای آموزش مدل انتخاب میکنیم که به آن داده های آموزشی میگوییم.

۰.۲ باقی مانده ثبت ها را برای آزمایش مدل کنار میگذاریم که به آن داده های آزمایشی میگوییم.

➢ مدل ها

درخت رده بندی شامل انواع مختلقی از مدل ها میشود که در اینجا آن ها را بررسی میکنیم.

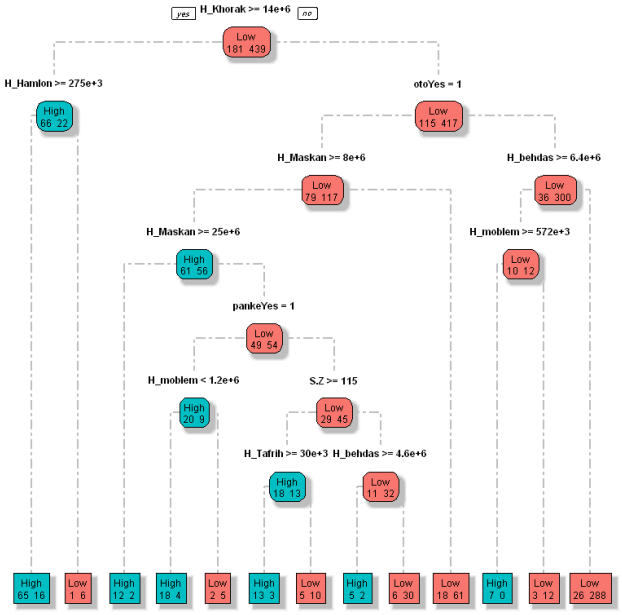
❖ **مدل ۱ : درخت رده بندی بدون تنظیم کردن پارامتر ها**

آزمایشی

```
          Reference
Prediction High Low
      High   29   9
      Low    23  94

               Accuracy : 0.7935
                 95% CI : (0.7212, 0.8543)
    No Information Rate : 0.6645
    P-Value [Acc > NIR] : 0.0002887

                  Kappa : 0.5039

 Mcnemar's Test P-Value : 0.0215563

            Sensitivity : 0.5577
            Specificity : 0.9126
         Pos Pred Value : 0.7632
         Neg Pred Value : 0.8034
             Prevalence : 0.3355
         Detection Rate : 0.1871
   Detection Prevalence : 0.2452
      Balanced Accuracy : 0.7352

       'Positive' Class : High
```

آموزشی

```
          Reference
Prediction High Low
      High  120  27
      Low    61 412

               Accuracy : 0.8581
                 95% CI : (0.8281, 0.8846)
    No Information Rate : 0.7081
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.6366

 Mcnemar's Test P-Value : 0.0004351

            Sensitivity : 0.6630
            Specificity : 0.9385
         Pos Pred Value : 0.8163
         Neg Pred Value : 0.8710
             Prevalence : 0.2919
         Detection Rate : 0.1935
   Detection Prevalence : 0.2371
      Balanced Accuracy : 0.8007

       'Positive' Class : High
```



عوامل مهم : هزینه خوراکی و دخانی - هزینه حمل ونقل – اتومبیل - هزینه مسکن – هزینه بهداشت – هزینه مبلمان پنکه – سطح زیربنا – هزینه تفریح

🡸 همان طور که مشاهده میکنید در این جا کلاس مثبت رده High را نشان میدهد پس در اینجا شاخص مهم برای ما شاخص حساسیت است.

همان طور که مشاهده میکنید این مدل در داده های آموزشی عملکرد نسبتا خوبی داشته و ۶۶ درصد ثبت های مربوط به خانوار های پر درآمد را درست رده بندی کرده اما در داده های آزمایشی تنها ۵۵ درصد از داده های مربوط به گروه پردرآمد را درست رده بندی کرده است و با این که دقت کلی مدل تقریبا ۸۰ درصد است اما عملکرد کلی مدل برای ما قابل قبول نیست.

❖ **مدل ۲ : عمیق ترین درخت رده بندی**

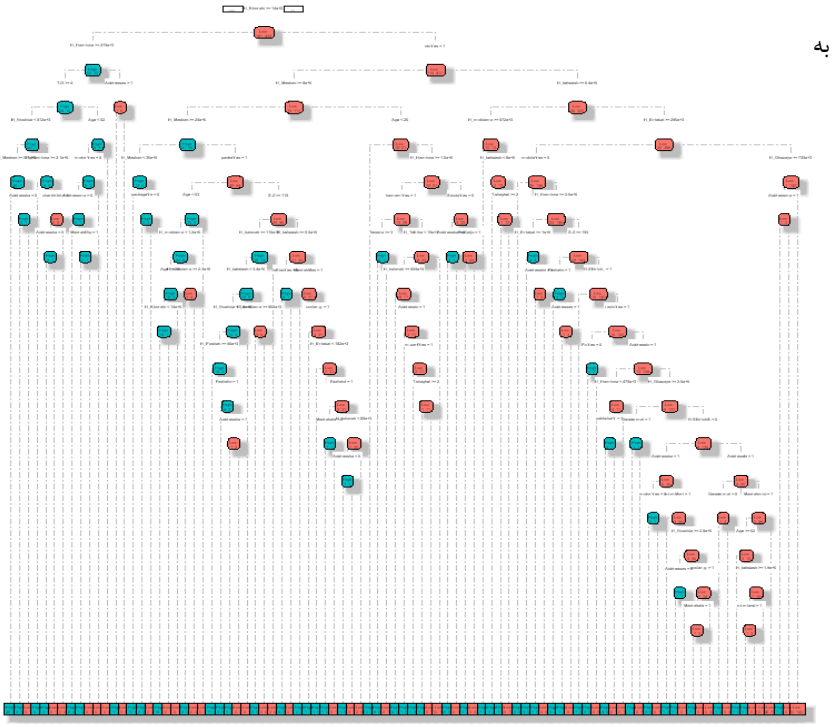
آزمایشی

```
          Reference
Prediction High Low
      High   27  25
      Low    25  78

               Accuracy : 0.6774
                 95% CI : (0.5977, 0.7502)
    No Information Rate : 0.6645
    P-Value [Acc > NIR] : 0.4027

                  Kappa : 0.2765

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.5192
            Specificity : 0.7573
         Pos Pred Value : 0.5192
         Neg Pred Value : 0.7573
             Prevalence : 0.3355
         Detection Rate : 0.1742
   Detection Prevalence : 0.3355
      Balanced Accuracy : 0.6383

       'Positive' Class : High
```

آموزشی

```
          Reference
Prediction High Low
      High  181   0
      Low     0 439

               Accuracy : 1
                 95% CI : (0.9941, 1)
    No Information Rate : 0.7081
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.2919
         Detection Rate : 0.2919
   Detection Prevalence : 0.2919
      Balanced Accuracy : 1.0000

       'Positive' Class : High
```

در مدل عمیق ترین درخت ، شاخه های درخت تا جایی پیشروی میکنند که تمام ثبت ها در رده درست قرار بگیرند و به همین دقت مدل روی داده های آموزشی ۱۰۰ درصد است ، اما این امر باعث بیش برازش مدل میشود که به این معنی است که مدل توانایی داشتن عملکرد خوب روی داده های جدید را از دست میدهد و اطلاعات جزئی بیش از حدی را فرا میگیرد که باعث میشود عملکرد مدل روی داده های آزمایشی ضعیف باشد.



همان طور که مشاهده میکنید این مدل تنها موفق به رده بندی درست ۵۱ درصد از خانوار های پردرآمد شده است که عملکرد ضعیفی است.

مدل های درخت عمیق راس های زیادی دارند که باعث میشود نمایش آن ها سخت شود.

متغیر های مهم در این نمودار خیلی مشخص نیست چون تعداد زیادی از متغیر ها تاثیرگذار هستند

❖ **مدل ۳ : عمیق ترین درخت رده بندی (هرس شده با کمترین xerror)**

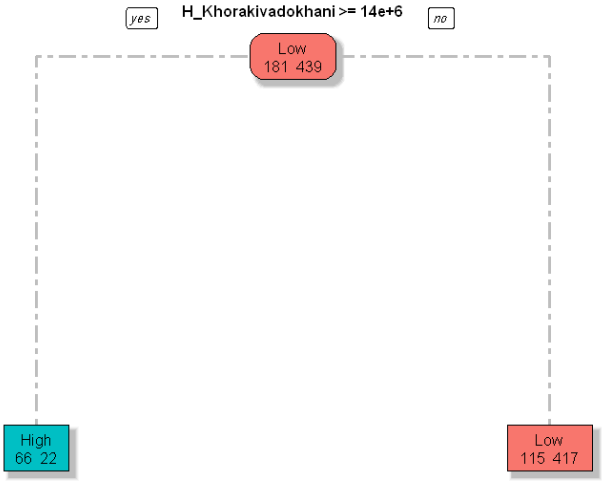
آزمایشی

```
          Reference
Prediction High Low
      High   20   6
      Low    32  97

               Accuracy : 0.7548
                 95% CI : (0.6794, 0.8203)
    No Information Rate : 0.6645
    P-Value [Acc > NIR] : 0.009507

                  Kappa : 0.3725

 Mcnemar's Test P-Value : 0.00005002

            Sensitivity : 0.3846
            Specificity : 0.9417
         Pos Pred Value : 0.7692
         Neg Pred Value : 0.7519
             Prevalence : 0.3355
         Detection Rate : 0.1290
   Detection Prevalence : 0.1677
      Balanced Accuracy : 0.6632

       'Positive' Class : High
```

آموزشی

```
          Reference
Prediction High Low
      High   66  22
      Low   115 417

               Accuracy : 0.779
                 95% CI : (0.7443, 0.8111)
    No Information Rate : 0.7081
    P-Value [Acc > NIR] : 0.000040592536126692

                  Kappa : 0.3705

 Mcnemar's Test P-Value : 0.000000000000003839

            Sensitivity : 0.3646
            Specificity : 0.9499
         Pos Pred Value : 0.7500
         Neg Pred Value : 0.7838
             Prevalence : 0.2919
         Detection Rate : 0.1065
   Detection Prevalence : 0.1419
      Balanced Accuracy : 0.6573

       'Positive' Class : High
```

عوامل مهم : هزینه خوراکی و دخانی



	Y					cover
2	0.25	when	H_Khorakivadokhani	>=	14454014	14%
3	0.78	when	H_Khorakivadokhani	<	14454014	86%

➔ از نکات مثبت این مدل این است که مدل بسیار ساده ای است و تنها با توجه به اطلاعات ۱ ستون خانوار ها را رده بندی میکند اما عملکرد این مدل بسیار ضعیف است و کم برازش شده است زیرا عملکرد این مدل حتی روی داده های آموزشی نیز ضعیف است این مدل در داده های آموزشی تنها ۳۶ درصد ثبت های پردرآمد را درست رده بندی کرده و دقت کلی ۷۷ درصد را داشته است و روی داده های آزمایشی نیز ۳۸ درصد ثبت های پردآمد را درست رده بندی کرده و دقت کلی ۷۵ درصد داشته و در کل مدل قابل قبولی نیست.

❖ **مدل ۴ : عمیق ترین درخت رده بندی (بهترین هرس)**

آزمایشی

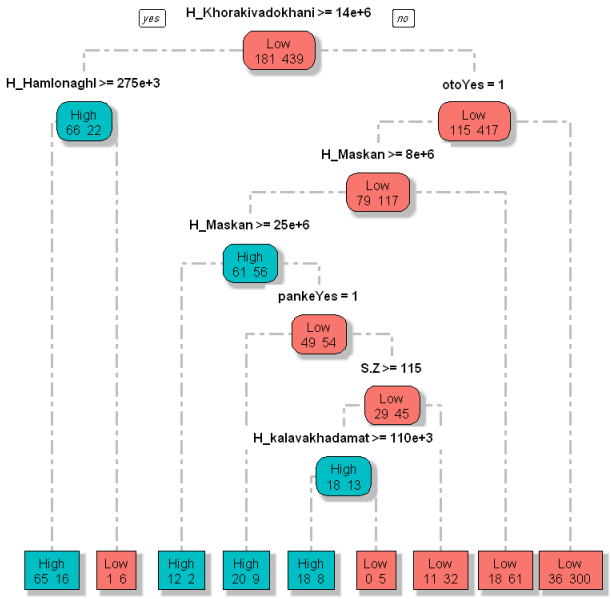
```
          Reference
Prediction High Low
      High   29  13
      Low    23  90

               Accuracy : 0.7677
                 95% CI : (0.6932, 0.8317)
    No Information Rate : 0.6645
    P-Value [Acc > NIR] : 0.003398

                  Kappa : 0.453

 Mcnemar's Test P-Value : 0.133614

            Sensitivity : 0.5577
            Specificity : 0.8738
         Pos Pred Value : 0.6905
         Neg Pred Value : 0.7965
             Prevalence : 0.3355
         Detection Rate : 0.1871
   Detection Prevalence : 0.2710
      Balanced Accuracy : 0.7157

       'Positive' Class : High
```

آموزشی

```
          Reference
Prediction High Low
      High  115  35
      Low    66 404

               Accuracy : 0.8371
                 95% CI : (0.8056, 0.8653)
    No Information Rate : 0.7081
    P-Value [Acc > NIR] : 0.00000000000005835

                  Kappa : 0.5851

 Mcnemar's Test P-Value : 0.002835

            Sensitivity : 0.6354
            Specificity : 0.9203
         Pos Pred Value : 0.7667
         Neg Pred Value : 0.8596
             Prevalence : 0.2919
         Detection Rate : 0.1855
   Detection Prevalence : 0.2419
      Balanced Accuracy : 0.7778

       'Positive' Class : High
```

این مدل از بهترین هرس عمیق ترین درخت بدست آمده است که در این روش علاوه بر xerror ، خطای محاسبه xerror نیز برای هرس کردن لحاظ میشود.

عوامل مهم : هزینه خوراکی و دخانی – هزینه حمل و نقل – اتومبیل ۰ هزینه مسکن – پنکه – سطح زیر بنا – هزینه کالا و خدمات



➔ عملکرد مدل روی داده های آموزشی معمولی است و ۶۳ درصد این ثبت های پردرآمد را درست رده بندی کرده است و در کل دقت ۸۳ درصد داشته است

اما عملکردش روی داده های آزمایشی مناسب نیست و تنها ۵۵ درصد ثبت های پردرآمد را درست رده بندی کرده است و در کل دقت ۷۶ درصد داشته است.

❖ مدل ۵ : جنگل تصادفی

جنگل تصادفی یک الگوریتم یادگیری نظارت شده محسوب می‌شود. همانطور که از نام آن مشهود است، این الگوریتم جنگلی را به طور تصادفی می‌سازد. «جنگل» ساخته شده، در واقع گروهی از درخت‌های تصمیم است.

جنگل تصادفی، تصادفی بودن افزوده‌ای را ضمن رشد درختان به مدل اضافه می‌کند. این الگوریتم، به جای جست‌وجو به دنبال مهم‌ترین ویژگی‌ها هنگام تقسیم کردن یک گره ، به دنبال بهترین ویژگی‌ها در میان مجموعه تصادفی از ویژگی‌ها می‌گردد. این امر منجر به تنوع زیاد و در نهایت مدل بهتر می‌شود. بنابراین، در جنگل تصادفی، تنها یک زیر مجموعه از ویژگی‌ها توسط الگوریتم برای تقسیم یک گره در نظر گرفته می‌شود.

در این جا برای ساخت این مدل از ۵۰۰ درخت تصمیم استفاده کردیم که نتایج آن به شرح زیر است.

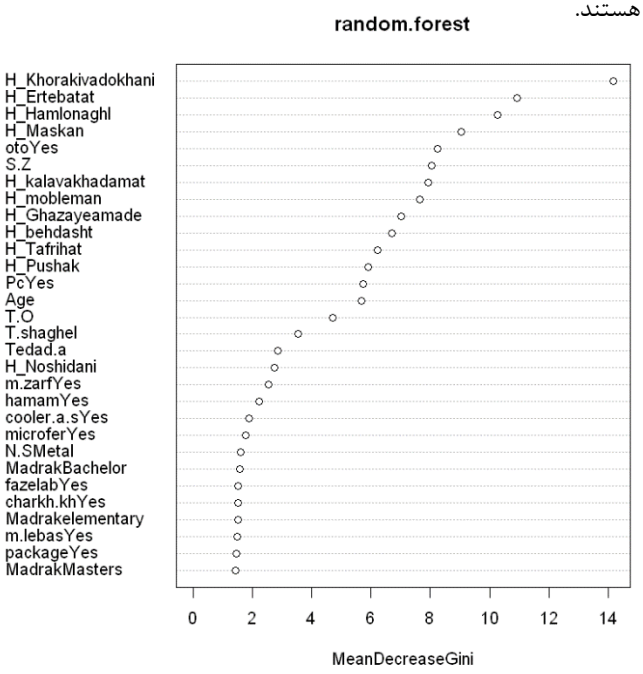
آموزشی

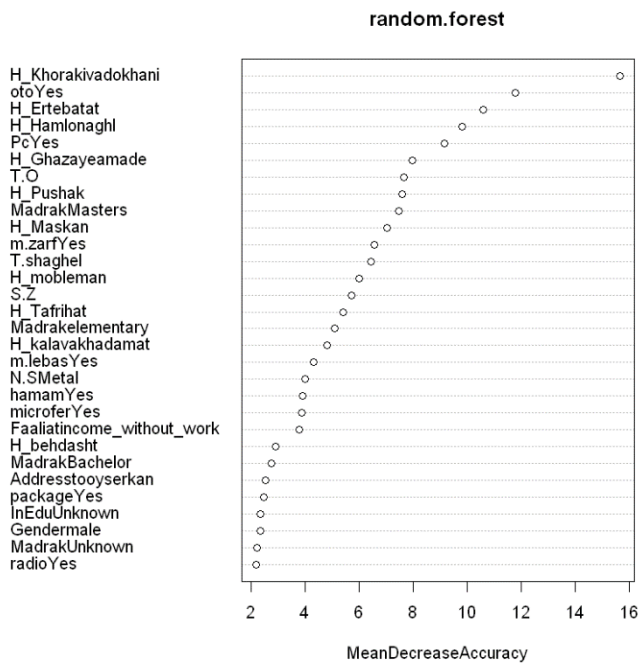
```
          Reference
'rediction High Low
      High  164   0
      Low    17 439

               Accuracy : 0.9726
                 95% CI : (0.9565, 0.9839)
    No Information Rate : 0.7081
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.9318

 Mcnemar's Test P-Value : 0.0001042

            Sensitivity : 0.9061
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9627
             Prevalence : 0.2919
         Detection Rate : 0.2645
   Detection Prevalence : 0.2645
      Balanced Accuracy : 0.9530

       'Positive' Class : High
```

آزمایشی

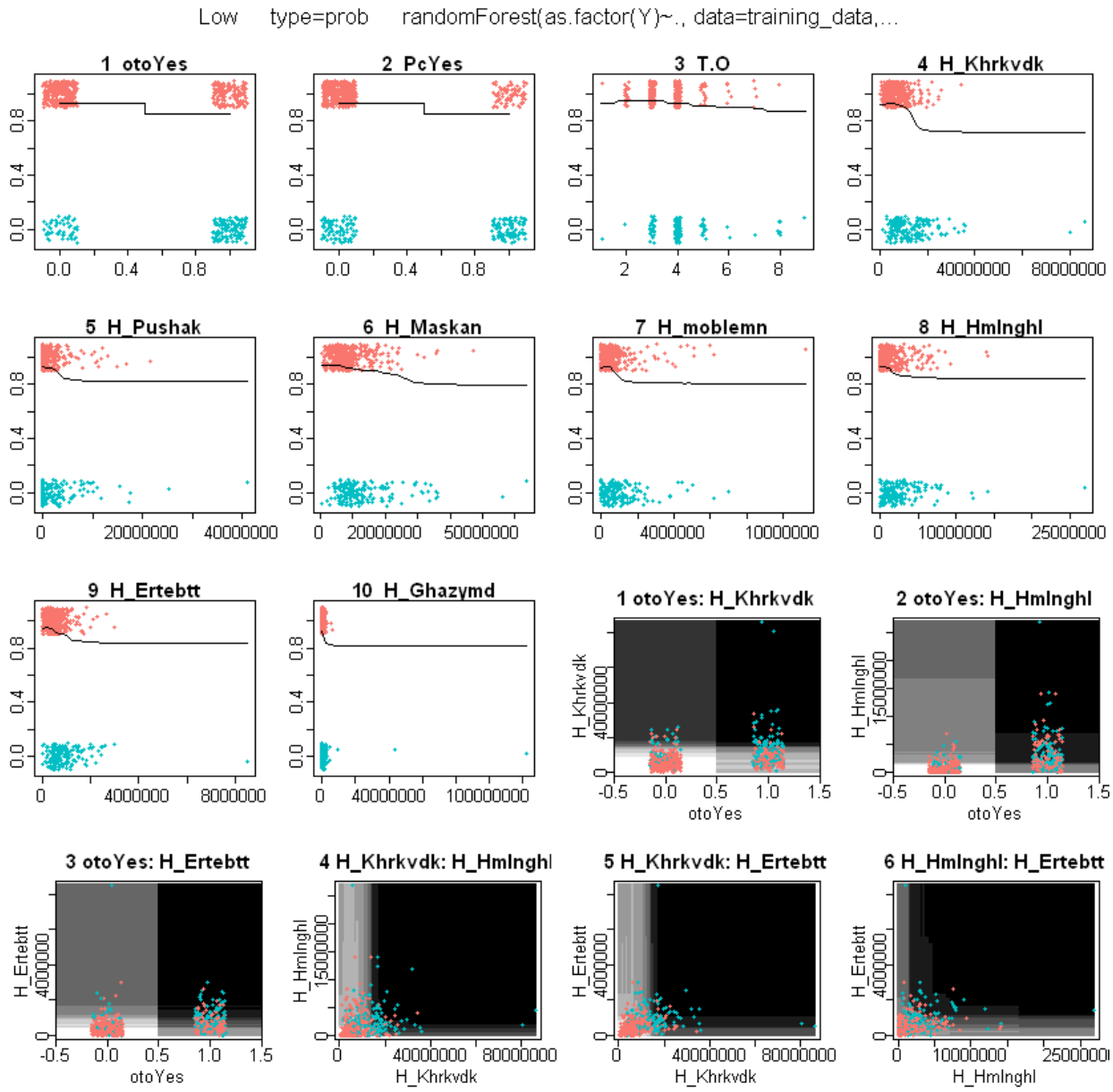
```
          Reference
Prediction High Low
      High   28   3
      Low    24 100

               Accuracy : 0.8258
                 95% CI : (0.7568, 0.882)
    No Information Rate : 0.6645
    P-Value [Acc > NIR] : 0.000005686

                  Kappa : 0.5659

 Mcnemar's Test P-Value : 0.0001186

            Sensitivity : 0.5385
            Specificity : 0.9709
         Pos Pred Value : 0.9032
         Neg Pred Value : 0.8065
             Prevalence : 0.3355
         Detection Rate : 0.1806
   Detection Prevalence : 0.2000
      Balanced Accuracy : 0.7547

       'Positive' Class : High
```

عوامل مهم : مواردی که در دو نمودار زیر در بخش های بالاتر قرار دارند عوامل مهمی برای افزایش دقت و کاهش خطای جینی این مدل هستند.





- عملکرد مدل روی داده های آموزشی بسیار خوب بوده است و دقت کلی ۹۷ درصد و حساسیت ۹۰ درصد را به دست آورده است. عملکرد مدل روی داده های آزمایشی متوسط بوده است ، دقت کلی ۸۲ درصد را به دست آورده است اما فقط توانسته ۵۳ درصد ثبت های پردرآمد را درست رده بندی کند اما ۹۷ درصد ثبت های مربوط به خانوار های کم درآمد را درست رده بندی کرده است.



این نمودار ها نیز احتمال رده بندی هر یک از ثبت ها در گروه پردرآمد یا کم درآمد با توجه به ستون های مختلف ماتریس داده ها نمایش میدهد.

در ۱۰ نمودار اول ستون عمودی نشان دهنده احتمال رده بندی در گروه پردرآمد یا کم درآمد است و در ۶ نمودار بعدی نیز رنگ سفید نشان دهنده احتمال رده بندی در گروه کم درآمد و رنگ سیاه نشان دهنده احتما رده بندی در گروه پر درآمد است.

❖ **مدل ۶ : درخت تقویت شده**

در این جا یک دنباله از درخت ها برازش میشود که هر درخت بر ثبت های بدرده بندی شده در درخت قبلی تمرکز میکند.
در این جا یک دنباله ۱۰۰ تایی از درخت ها را برازش میدهیم.

آزمایشی

```
          Reference
Prediction High Low
      High   28  10
      Low    24  93

               Accuracy : 0.7806
                 95% CI : (0.7072, 0.8431)
    No Information Rate : 0.6645
    P-Value [Acc > NIR] : 0.001063

                  Kappa : 0.4729

 Mcnemar's Test P-Value : 0.025782

            Sensitivity : 0.5385
            Specificity : 0.9029
         Pos Pred Value : 0.7368
         Neg Pred Value : 0.7949
             Prevalence : 0.3355
         Detection Rate : 0.1806
   Detection Prevalence : 0.2452
      Balanced Accuracy : 0.7207

       'Positive' Class : High
```

آموزشی

```
          Reference
Prediction High Low
      High  181   0
      Low     0 439

               Accuracy : 1
                 95% CI : (0.9941, 1)
    No Information Rate : 0.7081
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.2919
         Detection Rate : 0.2919
   Detection Prevalence : 0.2919
      Balanced Accuracy : 1.0000

       'Positive' Class : High
```

عوامل مهم : در جدول زیر عوامل مهم به ترتیب اهمیتشان مشخص شده اند

	importance
H_Khorakivadokhani	9.1351958
H_Ertebatat	7.1602587
H_mobleman	6.3998906
H_Maskan	5.8468992
H_Hamlonaghl	5.8293001
H_behdasht	5.8286385
S.Z	5.5200080
H_kalavakhadamat	5.5069571
Age	5.4764040
H_Pushak	5.4502262
H_Ghazayeamade	3.5297656
H_Tafrihat	2.6133224
H_Noshidani	2.3318902
T.O	1.8142312
Tedad.a	1.7246699

- همان طور که مشاهده میکنید در این مدل نیز مانند بسیاری دیگر از مدل ها ستون ها هزینه جزو مهم ترین ستون ها برای رده بندی مدل بوده اند.
- در این مدل نیز بیش برازش روی داده های آموزشی رخ داده است که باعث شده درصد دقت روی این داده ها برابر با ۱۰۰ باشد و این موضوع باعث میشود که عملکرد مدل روی داده های آزمایشی چندان مطلوب نباشد.

دقت کلی مدل روی داده های آزمایشی ۷۸ درصد است و مدل تنها موفق به رده بندی درست ۵۳ درصد از ثبت های خانوار های پردرآمد شده است اما توانسته ۹۰ درصد ثبت های خانوار های کم درآمد را به درستی رده بندی کند.

➤ مقایسه عملکرد مدل ها

از آن جا که در تمام مدل های بعدی کلاس مثبت کلاس گروه کم درآمد است و کلاس منفی کلاس گروه پردرآمد است در این چند مدل نیز همین فرض را در نظر میگیریم تا در نهایت بتوانیم راحت تر این مدل ها را با یکدیگر مقایسه کنیم.

به همین دلیل مقادیر شاخص های حساسیت و مشخصه سازی که در این دو جدول وجود دارد نسبت به نتایجی که در بالا مشاهده نمودید جا به جا است.

	آموزشی (Train)			آزمایشی (Test)		
	Accuracy	Sensetivity	Specificity	Accuracy	Sensetivity	Specificity
مدل ۱	85	93	66	79	91	55
مدل ۲	100	100	100	67	75	51
مدل ۳	77	94	36	75	94	38
مدل ۴	83	92	63	76	87	55
مدل ۵	97	100	90	82	97	53
مدل ۶	100	100	100	78	90	53

برای انتخاب بهترین مدل در هر بخش مدلی را برمیگزینیم که بالاترین شاخص مشخصه سازی را در ثبت های آزمایشی داشته باشد یعنی بهترین عملکرد را در رده بندی ثبت های گروه پردرآمد داشته باشد و اگر چند مدل در این شاخص با یکدیگر برابر بودند با توجه به شاخص دقت کل بهترین مدل را برمیگزینیم.

در اینجا هر دو مدل ۱ و ۴ توانسته اند مشخصه سازی ۵۵ درصد داشته باشند اما چون دقت کلی مدل ۱ بهتر است این مدل را به عنوان بهترین مدل این بخش برمیگزینیم.

➢ رگرسیون لجستیک

همان طور که می‌دانیم در رگرسیون خطی، متغیر وابسته یک متغیر کمی در سطح فاصله‌ای یا نسبی است و پیش بینی کننده ها از نوع متغیرهای پیوسته، گسسته یا ترکیبی از این دو هستند. اما هنگامی که متغیر وابسته کمی نباشد، یعنی به صورت دو یا چند رده ای باشد، از رگرسیون لجستیک استفاده می‌کنیم که امکان پیش‌بینی عضویت گروهی را فراهم می‌کند.

➢ پیش پردازش

در این بخش تنها باید یک پردازش انجام دهیم

- **تبدیل ستون های رسته ای به ستون های ظاهری**

  توضیحات این فرایند در مدل KNN داده شده است.

➢ افراز داده ها

۰.۸ ثبت ها را به صورت تصادفی برای آموزش مدل انتخاب میکنیم که به آن داده های آموزشی میگوییم.

۰.۲ باقی مانده ثبت ها را برای آزمایش مدل کنار میگذاریم که به آن داده های آزمایشی میگوییم.

➢ مدل ها

درخت رده بندی شامل انواع مختلفی از مدل ها میشود که در اینجا آن ها را بررسی میکنیم.

❖ **مدل ۱: رگرسیون لجستیک با تمام متغیر های پیشگو**

در ساخت این مدل از تمام ستون های ماتریس داده ها به عنوان متغیر پیشگو استفاده کرده ایم. خروجی مدل روی ثبت های آموزشی و آزمایشی به شکل زیر است.

آزمایشی

```
          Reference
Prediction Low High
      Low   90   20
      High  13   32

               Accuracy : 0.7871
                 95% CI : (0.7142, 0.8487)
    No Information Rate : 0.6645
    P-Value [Acc > NIR] : 0.0005641

                  Kappa : 0.506

 Mcnemar's Test P-Value : 0.2962699

            Sensitivity : 0.8738
            Specificity : 0.6154
         Pos Pred Value : 0.8182
         Neg Pred Value : 0.7111
             Prevalence : 0.6645
         Detection Rate : 0.5806
   Detection Prevalence : 0.7097
      Balanced Accuracy : 0.7446

       'Positive' Class : Low
```

آموزشی

```
          Reference
Prediction Low High
      Low  407   54
      High  32  127

               Accuracy : 0.8613
                 95% CI : (0.8316, 0.8875)
    No Information Rate : 0.7081
    P-Value [Acc > NIR] : < 0.0000000000000002

                  Kappa : 0.6521

 Mcnemar's Test P-Value : 0.02354

            Sensitivity : 0.9271
            Specificity : 0.7017
         Pos Pred Value : 0.8829
         Neg Pred Value : 0.7987
             Prevalence : 0.7081
         Detection Rate : 0.6565
   Detection Prevalence : 0.7435
      Balanced Accuracy : 0.8144

       'Positive' Class : Low
```

عوامل مهم : شهرستان اسدآباد – شهرستان همدان – هزینه خوراکی – تعداد شاغل – اتومبیل – شهرستان ملایر – هزینه تفریحات

- در این جا کلاس مثبت مربوط به رده Low است پس در اینجا رده بندی کلاس منفی برای ما اهمیت اصلی را دارد که برای بررسی این موضوع باید عملکرد مدل را بر اساس شاخص مشخصه سازی بررسی کنیم که نشان دهنده درصد ثبت هایی که به درستی در رده خانوار های پردرآمد قرار گرفته نسبت به کل ثبت های مربوط به خانوار های پردرآمد است.
- دقت کلی مدل روی داده های آموزشی برابر با ۸۶ درصد است و مدل توانسته ۷۰ درصد خانوار های پردرآمد را درست رده بندی کند.
دقت کلی مدل روی داده های آموزشی برابر با ۷۸ درصد است و مدل تنها موفق شده ۶۱ درصد خانوار های پردرآمد را درست رده بندی کند.
- در این مدل به دلیل زیاد بودن تعداد متغیر های ممورد استفاده در مدل این اطلاعات را در ین جا نمایش نداده ایم ، توضیحات بیشتر درمورد نحوه بررسی خلاصه مدل در مدل بعدی بررسی میشود.

- **مدل ۲ : رگرسیون لجستیک قدم به قدم پیشرو**

در این نوع از مدل رگرسیون از یک مدل بدون متغیر های پیشگو شروع میکنیم و متغیر ها را یکی یکی به مدل اضافه میکنیم و عملکرد مدل را بررسی میکنیم تا مدلی بسازیم که به تعداد کمتری از متغیر ها برای رده بندی نیاز داشته باشد و عملکرد قابل قبولی نیز داشته باشد ، همچنین این نوع برازش مدل از بیش برازش تا حد خوبی جلوگیری میکند.

خروجی مدل روی ثبت های آموزشی و آزمایشی به شکل زیر است.

آموزشی

```
          Reference
Prediction Low High
      Low  403   66
      High  36  115

               Accuracy : 0.8355
                 95% CI : (0.8039, 0.8638)
    No Information Rate : 0.7081
    P-Value [Acc > NIR] : 0.0000000000001237

                  Kappa : 0.5817

 Mcnemar's Test P-Value : 0.004086

            Sensitivity : 0.9180
            Specificity : 0.6354
         Pos Pred Value : 0.8593
         Neg Pred Value : 0.7616
             Prevalence : 0.7081
         Detection Rate : 0.6500
   Detection Prevalence : 0.7565
      Balanced Accuracy : 0.7767

       'Positive' Class : Low
```

آزمایشی

```
          Reference
Prediction Low High
      Low   89   23
      High  14   29

               Accuracy : 0.7613
                 95% CI : (0.6863, 0.826)
    No Information Rate : 0.6645
    P-Value [Acc > NIR] : 0.005777

                  Kappa : 0.4407

 Mcnemar's Test P-Value : 0.188445

            Sensitivity : 0.8641
            Specificity : 0.5577
         Pos Pred Value : 0.7946
         Neg Pred Value : 0.6744
             Prevalence : 0.6645
         Detection Rate : 0.5742
   Detection Prevalence : 0.7226
      Balanced Accuracy : 0.7109

       'Positive' Class : Low
```

بررسی خلاصه مدل : خلاصه مدل های رگرسیون در قالبی به صورت تصویر زیر ارائه میشود که در این جا بخش های مهم این خلاصه را به اختصار توضیح میدهیم.

Coefficients:

1	2			3
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.55913758933	1.59860409323	-3.477	0.000506 ***
H_Khorakivadokhani	0.00000012890	0.00000002487	5.182	0.000000219 ***
otoYes	0.92363201052	0.27139957651	3.403	0.000666 ***
H_Ertebatat	0.00000091861	0.00000030856	2.977	0.002911 **
T.O	0.36529402528	0.18722295435	1.951	0.051043 .
Addresstooyserkan	1.45093967241	0.39615292890	3.663	0.000250 ***
MadrakBachelor	0.95761747703	0.39021526698	2.454	0.014125 *
T.shaghel	0.87316541772	0.24921861619	3.504	0.000459 ***
H_Tafrihat	0.00000066069	0.00000025509	2.590	0.009597 **
N.SBrick_Rock	20.54937445419	882.74372835474	0.023	0.981428
m.lebasYes	0.74887823482	0.40878839268	1.832	0.066959 .
Madrakmid_1	-0.78486216728	0.35518791592	-2.210	0.027125 *
H_Ghazayeamade	0.00000021373	0.00000019390	1.102	0.270337
PcYes	0.50371713582	0.27804070551	1.812	0.070038 .
AddressasadAbaad	-0.79406173127	0.48854346964	-1.625	0.104085
S.Z	0.00790300310	0.00458178520	1.725	0.084550 .
SavadYes	0.73053295301	0.44513217660	1.641	0.100764
sookht.gnatural_gas	-2.64039976602	1.46586506177	-1.801	0.071662 .
MadrakMasters	1.18705374275	0.85743407114	1.384	0.166228

بخش ۱ : نام متغیر های مورد استفاده در مدل

بخش ۲ : ضریب هر یک از متغیر ها در مدل نهایی

بخش ۳ : اهمیت هر یک از متغیر ها در مدل نهایی

- مثبت و منفی بودن ضرایب در این مدل ها نشان دهنده این است که با افزایش مقدار این متغیر در یک ثبت مدل تمایل بیشتری به رده بندی این ثبت در گروه مثبت دارد یا گروه منفی.
- نمیتوانیم از اندازه ضرایب نتیجه گیری خاصی بکنیم زیرا ما در اینجا متغیر ها را هم مقیاس نکرده ایم به همین دلیل این کار درست نیست. برای مثال هزینه خوراکی در اینجا ضریب بسیار کوچکی دارد و سنگی یا آجری بودن اسکلت ساختمان ضریب خیلی بزرگتری دارد اما همان طور که در مدل های قبلی نیز مشاهده کردیم مهم ترین ستون در این داده های هزینه خوراکی و دخانی است و کوچک بودن ضریب این ستون در این جا به دلیل این است که مقادیر این ستون اعدادی در مقیاس میلیون هستند چون هزینه خانوار را نشان میدهند اما در مقابل مقادیر ستون جنس اسکلت ساختمان اعداد ۰ یا ۱ هستند چون این متغیر متغیر ظاهری است ، پس نمیتوانیم به وسیله اندازه ضرایب اهمیت آن ها را با یکدیگر مقایسه کنیم.
- در اینجا ستون های گاز طبیعی و مدرک متوسطه و شهرستان عباس آباد در کم درآمد رده بندی شدن خانوار ها تاثیر گذار هستند.
- باقی ستون ها نیز در پردرآمد رده بندی شدن خانوار ها تاثیر گذار هستند.
- مثلا ضریب ستون هزینه خوراکی و دخانی مثبت است پس هر چه قدر هزینه مصرف شده در این حوزه در یک ثبت بیشتر باشید احتمال اینکه این ثبت در گروه پردرآمد رده بندی شود نیز بیشتر است.
- بخش سوم نیز همان طور که گفته شد اهمیت هر یک از این ستون ها را نشان میدهد که مقدار این اهمیت از p-value محاسبه میشود.
- هر چه p-value کمتر باشد و ستاره ها بیشتر باشد اهمیت آن ستون در مدل بیشتر است.
- **عوامل مهم :** هزینه خوراکی و دخانی – اتومبیل – تویسرکان – تعداد شاغل
- دقت کلی مدل روی داده های آموزشی برابر با ۸۳ درصد است و مدل توانسته ۶۳ درصد خانوار های پردرآمد را درست رده بندی کند. دقت کلی مدل روی داده های آموزشی برابر با ۷۶ درصد است و مدل تنها موفق شده ۵۵ درصد خانوار های پردرآمد را درست رده بندی کند.

- **مدل ۳ : رگرسیون لجستیک قدم به قدم پسرو**

در این نوع از رگرسیون ابتدا یک مدل با استفاده از تمام پیشگو ها میسازیم و سپس در هر مرحله یکی از پیشگو های اضافی را حذف میکنیم.

آموزشی

```
          Reference
Prediction Low High
      Low  406   64
      High  33  117

               Accuracy : 0.8435
                 95% CI : (0.8125, 0.8713)
    No Information Rate : 0.7081
    P-Value [Acc > NIR] : 0.000000000000002571

                  Kappa : 0.6015

 Mcnemar's Test P-Value : 0.002319

            Sensitivity : 0.9248
            Specificity : 0.6464
         Pos Pred Value : 0.8638
         Neg Pred Value : 0.7800
             Prevalence : 0.7081
         Detection Rate : 0.6548
   Detection Prevalence : 0.7581
      Balanced Accuracy : 0.7856

       'Positive' Class : Low
```

آزمایشی

```
          Reference
Prediction Low High
      Low   88   24
      High  15   28

               Accuracy : 0.7484
                 95% CI : (0.6725, 0.8146)
    No Information Rate : 0.6645
    P-Value [Acc > NIR] : 0.01515

                  Kappa : 0.4104

 Mcnemar's Test P-Value : 0.20018

            Sensitivity : 0.8544
            Specificity : 0.5385
         Pos Pred Value : 0.7857
         Neg Pred Value : 0.6512
             Prevalence : 0.6645
         Detection Rate : 0.5677
   Detection Prevalence : 0.7226
      Balanced Accuracy : 0.6964

       'Positive' Class : Low
```

عوامل مهم : اسدآباد – همدان - ملایر

تعداد اعضای شاغل – هزینه خوراکی و دخانی

هزینه ارتباطات – اتومبیل

عوامل در راستای رده بندی در رده کم درآمد :

اسدآباد – بهار – فامنین – همدان – کبودرآهنگ

ملایر – نهاوند – رضن – متوسطه ۱

عوامل در راستای رده بندی در رده پردرآمد :

سایر عوامل

```
Coefficients:
                          Estimate      Std. Error z value       Pr(>|z|)
(Intercept)          -8.11300426263   1.23112037379  -6.590 0.000000000044 ***
AddressasadAbaad     -2.27565391586   0.60990105152  -3.731       0.000191 ***
Addressbahaar        -1.28908054050   0.56763594927  -2.271       0.023149 *
Addressfaamenin      -1.39691024685   0.74345643634  -1.879       0.060253 .
Addresshamedan       -1.62761024221   0.44961093688  -3.620       0.000295 ***
AddresskaboodarAhang -1.87470627538   0.76761773327  -2.442       0.014596 *
Addressmalayer       -1.37192241033   0.47843346149  -2.868       0.004137 **
Addressnahaavand     -1.24314519667   0.52684509294  -2.360       0.018295 *
Addressrazan         -1.05787112871   0.61820804400  -1.711       0.087046 .
SavadYes              0.62242419230   0.44883090400   1.387       0.165513
MadrakBachelor        0.83589648290   0.39455966780   2.119       0.034128 *
Madrakmid_1          -0.90658574081   0.36381346906  -2.492       0.012706 *
N.SBrick_Rock        20.47370237969 882.74373310374   0.023       0.981496
N.SMetal              0.51840862215   0.31219490283   1.661       0.096808 .
otoYes                0.89576869834   0.27469272535   3.261       0.001110 **
PcYes                 0.54596070634   0.28147284560   1.940       0.052421 .
yakhchalYes           1.09242690343   0.73422305424   1.488       0.136786
yakhchal.fYes         1.41266403406   0.74037668633   1.908       0.056387 .
m.lebasYes            0.73158437657   0.42017320275   1.741       0.081657 .
T.shaghel             0.96127328940   0.25987534341   3.699       0.000216 ***
T.O                   0.32256582239   0.19609224727   1.645       0.099976 .
S.Z                   0.00943610509   0.00486093314   1.941       0.052232 .
H_Khorakivadokhani    0.00000013039   0.00000002534   5.146 0.000000265646 ***
H_Ertebatat           0.00000092803   0.00000031364   2.959       0.003087 **
H_Tafrihat            0.00000059789   0.00000025855   2.312       0.020750 *
H_Ghazayeamade        0.00000020645   0.00000019589   1.054       0.291922
```

- اندازه ضرایبی که مقیاس مقادیرشان شبیه هم هست را میتوان با یکدیگر مقایسه کرد ، مثلا تمام ستون های ظاهری را میتوان با هم مقایسه کرد چون مقادیر ثبت ها در این ستون ها برابر ۰ یا ۱ است.
- مثلا در این جا ثبت های عضو شهرستان اسدآباد به احتمال بیشتری در گروه کم درآمد قرار میگیرند تا ثبت های عضو شهرستان همدان
- دقت کلی مدل روی داده های آموزشی برابر با ۸۳ درصد است و مدل توانسته ۶۳ درصد خانوار های پردرآمد را درست رده بندی کند.

دقت کلی مدل روی داده های آموزشی برابر با ۷۶ درصد است و مدل تنها موفق شده ۵۵ درصد خانوار های پردرآمد را درست رده بندی کند.

- **مدل ۴ : رگرسیون لجستیک قدم به قدم دو سویه**

رگرسیون قدم به قدم مانند رگرسیون پسرو میباشد و از یک مدل با تمام متغیر ها شروع میشود با این تفاوت که در هر مرحله پس از حذف کردن یکی از متغیر ها بررسی میکند که آیا اضافه کردن یکی دیگر از متغیر ها میتواند AIC را پایین بیاورد یا خیر اگر چنین متغیری پیدا کرد آن را به مدل اضافه میکند.

آزمایشی

```
          Reference
Prediction Low High
      Low   88   24
      High  15   28

               Accuracy : 0.7484
                 95% CI : (0.6725, 0.8146)
    No Information Rate : 0.6645
    P-Value [Acc > NIR] : 0.01515

                  Kappa : 0.4104

 Mcnemar's Test P-Value : 0.20018

            Sensitivity : 0.8544
            Specificity : 0.5385
         Pos Pred Value : 0.7857
         Neg Pred Value : 0.6512
             Prevalence : 0.6645
         Detection Rate : 0.5677
   Detection Prevalence : 0.7226
      Balanced Accuracy : 0.6964

       'Positive' Class : Low
```

آموزشی

```
          Reference
Prediction Low High
      Low  406   64
      High  33  117

               Accuracy : 0.8435
                 95% CI : (0.8125, 0.8713)
    No Information Rate : 0.7081
    P-Value [Acc > NIR] : 0.00000000000000002571

                  Kappa : 0.6015

 Mcnemar's Test P-Value : 0.002319

            Sensitivity : 0.9248
            Specificity : 0.6464
         Pos Pred Value : 0.8638
         Neg Pred Value : 0.7800
             Prevalence : 0.7081
         Detection Rate : 0.6548
   Detection Prevalence : 0.7581
      Balanced Accuracy : 0.7856

       'Positive' Class : Low
```

```
Coefficients:
                          Estimate       Std. Error z value       Pr(>|z|)
(Intercept)          -8.11300426263   1.23112037379  -6.590 0.000000000044 ***
AddressasadAbaad     -2.27565391586   0.60990105152  -3.731       0.000191 ***
Addressbahaar        -1.28908054050   0.56763594927  -2.271       0.023149 *
Addressfaamenin      -1.39691024685   0.74345643634  -1.879       0.060253 .
Addresshamedan       -1.62761024221   0.44961093688  -3.620       0.000295 ***
AddresskaboodarAhang -1.87470627538   0.76761773327  -2.442       0.014596 *
Addressmalayer       -1.37192241033   0.47843346149  -2.868       0.004137 **
Addressnahaavand     -1.24314519667   0.52684509294  -2.360       0.018295 *
Addressrazan         -1.05787112871   0.61820804400  -1.711       0.087046 .
SavadYes              0.62242419230   0.44883090400   1.387       0.165513
MadrakBachelor        0.83589648290   0.39455966780   2.119       0.034128 *
Madrakmid_1          -0.90658574081   0.36381346906  -2.492       0.012706 *
N.SBrick_Rock        20.47370237969 882.74373310374   0.023       0.981496
N.SMetal              0.51840862215   0.31219490283   1.661       0.096808 .
otoYes                0.89576869834   0.27469272535   3.261       0.001110 **
PcYes                 0.54596070634   0.28147284560   1.940       0.052421 .
yakhchalYes           1.09242690343   0.73422305424   1.488       0.136786
yakhchal.fYes         1.41266403406   0.74037668633   1.908       0.056387 .
m.lebasYes            0.73158437657   0.42017320275   1.741       0.081657 .
T.shaghel             0.96127328940   0.25987534341   3.699       0.000216 ***
T.O                   0.32256582239   0.19609224727   1.645       0.099976 .
S.Z                   0.00943610509   0.00486093314   1.941       0.052232 .
H_Khorakivadokhani    0.00000013039   0.00000002534   5.146 0.000000265646 ***
H_Ertebatat           0.00000092803   0.00000031364   2.959       0.003087 **
H_Tafrihat            0.00000059789   0.00000025855   2.312       0.020750 *
H_Ghazayeamade        0.00000020645   0.00000019589   1.054       0.291922
```

عوامل مهم : اسدآباد – همدان – تعداد شاغل

خوراکی و دخانی – هزینه ارتباطات – ملایر

اتومبیل

عوامل در راستای رده بندی در رده کم درآمد :

اسدآباد – بهار – فامنین – همدان – کبودرآهنگ

ملایر – نهاوند – رضن – متوسطه ۱

عوامل در راستای رده بندی در رده پردرآمد :

سایر عوامل

🡸 مدل به دست آمده در این روش دقیقا معادل مدل به دست آمده از روش قبل شده است

➔ مدل جستجوی فرسا برای زمانی مناسب است که تعداد متغیر های پیشگو کم باشد به همین دلیل این مدل در این جا بررسی نشده است.

➤ مقایسه عملکرد مدل ها

	آموزشی (Train)			آزمایشی (Test)		
	Accuracy	Sensetivity	Specificity	Accuracy	Sensetivity	Specificity
مدل ۱	86	92	70	78	87	61
مدل ۲	83	91	63	76	86	55
مدل ۳	84	92	64	74	85	53
مدل ۴	84	92	64	74	85	53

در این بخش بهترین مدل ۱ توانسته است بهترین عملکرد را برای رده بندی درست ثبت های مربوط به گروه پردرآمد در داده های آزمایشی را داشته باشد به همین دلیل این مدل را به عنوان مدل برتر این بخش انتخاب میکنیم.

شبکه عصبی مصنوعی

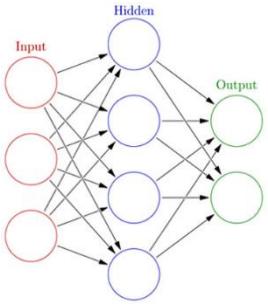
شبکه های عصبی مصنوعی (ANN) یا Artificial Neural Networks و به عبارت دیگر سیستم های اتصالگر، سیستم های محاسبه کننده ای هستند که از شبکه های عصبی زیستی الهام گرفته شده اند.

این سسیستم ها از بخش های کوچکتری به نام نورون تشکیل شده اند که هر یک از این نورون ها توانایی پردازش دارند و میتوانند اطلاعات را به یکدیگر انتقال دهند.

این سیستم ها، با بررسی مثال ها میتوانند الگو های مختلف در این مثال ها را فرا بگیرند و پس از فراگیری این اطلاعات توسط مدل ها میتوانیم از آن ها برای رده بندی ثبت ها استفاده کنیم.

این سیستم ها از تعدادی لایه نورونی و یال های وزن دار تشکیل شده اند که وزن یال های شدت اتصال بین نورون ها را نشان میدهد.

شبکه های عصبی از ۳ نوع لایه ساخته شده اند.



لایه ورودی : تعداد نورون های موجود در این لایه برابر با تعداد متغیر های ورودی میباشد.

لایه خروجی : تعداد نورون های موجود در لایه برابر با تعداد رده های متغیر برآمد است.

لای های پنهان : در این جا میتوان از یک یا چند لایه نورونی استفاده کرد که در هر لایه یک یا چند نورون وجود دارد که بنا به قدرت محاسباتی که نیاز داریم میتوانیم تعداد لایه ها و نورون ها را در بخش تنظیم کنیم.

اتصالات بین لایه ها به صورت کامل است.

در این نوع مدل ها چندین پارامتر مختلف برای تنظیم مدل وجود دارد که برای دستیابی به یک مدل خوب مقادیر آن ها باید تنظیم شود.

یکی از نکات منفی راجع به شبکه های عصبی این است که برعکس مدل هایی مانند درخت که عملکرد آن ها بسیار واضح است و عملکرد بسیار ناواضحی دارند و به اصطلاح مانند یک جعبه سیاه عمل میکنند و نمیتوان متوجه شد کدام عوامل در انتخاب های شبکه بیشترین تاثیر را داشته اند.

اما در مقابل این شبکه ها قدرت بسیار بالایی دارند که از آن ها برای حل مسائل و مشکلات بسیار سخت و بزرگ نیز میتوان استفاده کرد که در بسیاری از روش های دیگر این امکان وجود ندارد.

برخی از این پارامتر ها عبارتند از :

- تعداد لایه های پنهان و نورون های موجود در هر لایه
- تعداد نورون های لایه ورودی به تعداد متغیر های مورد استفاده در هر مدل است
- تعداد نورون های لایه خروجی به تعداد رده های متغیر برآمد یعنی ۲ است.

- تابع فعالسازی : در این جا از تابع logistic استفاده میکنیم.
- از بهینه ساز resilient backpropagation چون سرعت آن از backpropagation کلاسیک بهتر است و میتوان learning rate را بر صورت پویا و با توجه به مسئله تغییر داد و در شرایطی مختلف learning rate را کم و زیاد میکند.
- آستانه Threshold : نشان دهنده سرحد خطایی است که اگر عملکرد مدل ما تا آن حد پیشرفت کند مدل را قابل قبول ارزیابی میکنیم. هر چقدر این مقدار کم تر باشد به این معنی است که مدل باید سعی کند ثبت های آموزشی را به بهترین شکل ممکن رده بندی کند که در این صورت عملکرد مدل روی داده های آموزشی بسیار خوب خواهد بود اما این موضوع باعث میشود که این مدل روی داده های آزمایشی عملکرد قابل قبولی نداشته باشد و بیش برازش رخ دهد ، در مقابل اگر مقدار این آستانه خیلی بالا باشد باعث میشود که مدل به خوبی الگو های موجود در ثبت های آموزشی را فرا نگیرد و عملکرد ضعیفی روی داده های آزمایشی نیز داشته باشد و کم برازش رخ دهد.
- ضریب یادگیری : این ضریب سرعت اصلاح وزن های شبکه را مشخص میکند که در ابتدای کار برابر با ۰.۱ است که در میانه یادگیری ، در صورتی که مدل در حال گرفتار شدن در یک اکسترمم موضعی کوچک باشد این مقدار ۲۰ درصد افزایش پیدا میکند و در صورتی که مدل در حال رسیدن به اکسترمم موضعی خوب یا اکسترمم کلی باشد این مقدار ۲۰ درصد کاهش پیدا میکند تا بتواند راحت تر به هدف برسد.

پیش پردازش

در این بخش تنها باید یک پردازش انجام دهیم

- **تبدیل ستون های رسته ای به ستون های ظاهری**

توضیحات این فرایند در مدل KNN داده شده است.

افراز داده ها

۰.۸ ثبت ها را به صورت تصادفی برای آموزش مدل انتخاب میکنیم که به آن داده های آموزشی میگوییم.
۰.۲ باقی مانده ثبت ها را برای آزمایش مدل کنار میگذاریم که به آن داده های آزمایشی میگوییم.

مدل ها

در صورت استفاده از متغیر های زیاد در این نوع از مدل ها به شبکه های پیچیده ای نیاز داریم که برازش آن ها زمان زیادی طول میکشد. به همین دلیل در این بخش تنها از متغیر هایی استفاده میکنیم که اهمیت آن ها در مدل های قبلی مشخص شده است.
با استفاده از مجموعه متغیر های مشخص شده توسط مدل های قبلی ، چندین مدل مختلف شبکه عصبی آموزش دادم که در هر بخش ویژگی های بهترین مدل تولید شده را بررسی میکنیم.

- **مدل ۱ : شبکه عصبی با استفاده از متغیر های مهم در درخت تصمیم با بهترین هرس**

تعداد لایه ها : ۱ تعداد نورون های لایه ها : ۲ آستانه : ۰.۰۰۰۱

متغیر های استفاده شده :

اتومبیل – پنکه – سطح زیربنا – هزینه خوراکی و دخانی – هزینه مسکن – هزینه حمل و نقل – هزینه کالا و خدمات

آزمایشی

```
          Reference
Prediction Low High
      Low   92   20
      High  11   32

               Accuracy : 0.8
                 95% CI : (0.7283, 0.8599)
    No Information Rate : 0.6645
    P-Value [Acc > NIR] : 0.0001424

                  Kappa : 0.5314

 Mcnemar's Test P-Value : 0.1507628

            Sensitivity : 0.8932
            Specificity : 0.6154
         Pos Pred Value : 0.8214
         Neg Pred Value : 0.7442
             Prevalence : 0.6645
         Detection Rate : 0.5935
   Detection Prevalence : 0.7226
      Balanced Accuracy : 0.7543

       'Positive' Class : Low
```

آموزشی

```
          Reference
Prediction Low High
      Low  397   73
      High  42  108

               Accuracy : 0.8145
                 95% CI : (0.7816, 0.8444)
    No Information Rate : 0.7081
    P-Value [Acc > NIR] : 0.0000000007788

                  Kappa : 0.5276

 Mcnemar's Test P-Value : 0.00515

            Sensitivity : 0.9043
            Specificity : 0.5967
         Pos Pred Value : 0.8447
         Neg Pred Value : 0.7200
             Prevalence : 0.7081
         Detection Rate : 0.6403
   Detection Prevalence : 0.7581
      Balanced Accuracy : 0.7505

       'Positive' Class : Low
```

- کلاس مثبت مربوط به رده Low است پس در این جا شاخص مهم برای ما مشخصه سازی Specificity است.
- دقت کلی مدل روی داده های آموزشی ۸۱ درصد است.
- در داده های آموزشی مدل توانسته ۹۰ درصد ثبت های مربوط به خانوار های کم درآمد را به درستی رده بندی کند.
- در داده های آزمایشی مدل توانسته ۸۹ درصد ثبت های مربوط به خانوار های کم درآمد را به درستی رده بندی کند.
- دقت کلی مدل روی داده های آزمایشی ۸۰ درصد است.
- در داده های آموزشی مدل توانسته ۵۹ درصد ثبت های مربوط به خانوار های پر درآمد را به درستی رده بندی کند.
- در داده های آزمایشی مدل توانسته ۶۱ درصد ثبت های مربوط به خانوار های پر درآمد را به درستی رده بندی کند.

- **مدل ۲ : شبکه عصبی با استفاده از متغیر های مهم در جنگل تصادفی**

تعداد لایه ها : ۱ **تعداد نورون های لایه ها : ۳** **آستانه : ۰.۰۰۱**

متغیر های استفاده شده : اتومبیل – سطح زیربنا – هزینه های خوراکی و دخانی - مسکن - حمل و نقل - ارتباطات - کالا و خدمات

آزمایشی

```
          Reference
Prediction Low High
      Low   98   22
      High   5   30

               Accuracy : 0.8258
                 95% CI : (0.7568, 0.882)
    No Information Rate : 0.6645
    P-Value [Acc > NIR] : 0.000005686

                  Kappa : 0.5749

 Mcnemar's Test P-Value : 0.002076

            Sensitivity : 0.9515
            Specificity : 0.5769
         Pos Pred Value : 0.8167
         Neg Pred Value : 0.8571
             Prevalence : 0.6645
         Detection Rate : 0.6323
   Detection Prevalence : 0.7742
      Balanced Accuracy : 0.7642

       'Positive' Class : Low
```

آموزشی

```
          Reference
Prediction Low High
      Low  401   83
      High  38   98

               Accuracy : 0.8048
                 95% CI : (0.7714, 0.8353)
    No Information Rate : 0.7081
    P-Value [Acc > NIR] : 0.00000002408

                  Kappa : 0.4907

 Mcnemar's Test P-Value : 0.00006334248

            Sensitivity : 0.9134
            Specificity : 0.5414
         Pos Pred Value : 0.8285
         Neg Pred Value : 0.7206
             Prevalence : 0.7081
         Detection Rate : 0.6468
   Detection Prevalence : 0.7806
      Balanced Accuracy : 0.7274

       'Positive' Class : Low
```

- کلاس مثبت مربوط به رده Low است پس در این جا شاخص مهم برای ما مشخصه سازی Specificity است.
- دقت کلی مدل روی داده های آموزشی ۸۰ درصد است.
- در داده های آموزشی مدل توانسته ۹۱ درصد ثبت های مربوط به خانوار های کم درآمد را به درستی رده بندی کند.
- در داده های آزمایشی مدل توانسته ۹۵ درصد ثبت های مربوط به خانوار های کم درآمد را به درستی رده بندی کند.
- دقت کلی مدل روی داده های آزمایشی ۸۲ درصد است.
- در داده های آموزشی مدل توانسته ۵۴ درصد ثبت های مربوط به خانوار های پر درآمد را به درستی رده بندی کند.
- در داده های آزمایشی مدل توانسته ۵۷ درصد ثبت های مربوط به خانوار های پر درآمد را به درستی رده بندی کند.

- **مدل ۳ : شبکه عصبی با استفاده از متغیر های مهم در رگرسیون لوجستیک**

تعداد لایه ها : ۱ تعداد نورون های لایه ها : ۲ آستانه : ۰.۰۰۱

متغیر های استفاده شده : نهاوند – سواد – مدرک ابتدایی – اتومبیل – تعداد شاغل – سطح زیر بنا – هزینه خوراکی و خوردنی

آزمایشی

```
          Reference
Prediction Low High
      Low   97   24
      High   6   28

               Accuracy : 0.8065
                 95% CI : (0.7354, 0.8654)
    No Information Rate : 0.6645
    P-Value [Acc > NIR] : 0.00006759

                  Kappa : 0.5252

 Mcnemar's Test P-Value : 0.001911

            Sensitivity : 0.9417
            Specificity : 0.5385
         Pos Pred Value : 0.8017
         Neg Pred Value : 0.8235
             Prevalence : 0.6645
         Detection Rate : 0.6258
   Detection Prevalence : 0.7806
      Balanced Accuracy : 0.7401

       'Positive' Class : Low
```

آموزشی

```
          Reference
Prediction Low High
      Low  403   75
      High  36  106

               Accuracy : 0.821
                 95% CI : (0.7885, 0.8504)
    No Information Rate : 0.7081
    P-Value [Acc > NIR] : 0.00000000006422

                  Kappa : 0.5377

 Mcnemar's Test P-Value : 0.00031

            Sensitivity : 0.9180
            Specificity : 0.5856
         Pos Pred Value : 0.8431
         Neg Pred Value : 0.7465
             Prevalence : 0.7081
         Detection Rate : 0.6500
   Detection Prevalence : 0.7710
      Balanced Accuracy : 0.7518

       'Positive' Class : Low
```

- کلاس مثبت مربوط به رده Low است پس در این جا شاخص مهم برای ما مشخصه سازی Specificity است.
- دقت کلی مدل روی داده های آموزشی ۸۲ درصد است.
- در داده های آموزشی مدل توانسته ۹۱ درصد ثبت های مربوط به خانوار های کم درآمد را به درستی رده بندی کند.
- در داده های آزمایشی مدل توانسته ۹۴ درصد ثبت های مربوط به خانوار های کم درآمد را به درستی رده بندی کند.
- دقت کلی مدل روی داده های آزمایشی ۸۰ درصد است.
- در داده های آموزشی مدل توانسته ۵۸ درصد ثبت های مربوط به خانوار های پر درآمد را به درستی رده بندی کند.
- در داده های آزمایشی مدل توانسته ۵۳ درصد ثبت های مربوط به خانوار های پر درآمد را به درستی رده بندی کند.

- **مدل ۴ : شبکه عصبی با استفاده از متغیر های ساخته شده توسط PCA**

در روش Principal Component Analysis یا آنالیز مولفه اصلی ما به جای کار با ستون های اصلی داده ها با تعدادی ستون جدید کار میکنیم که هر کدام از این ستون ها ترکیبی از ستون های اصلی داده ها را شامل میشوند. به این ترتیب میتوانیم با استفاده از تعداد ستون کمتری اطلاعات مهم موجود در ماتریس داده ها را نمایش دهیم.

در این جا توانستیم تنها با استفاده از ۶ متغیر جدید ۹۸ درصد واریانس داده ها را نمایش دهیم.

از این ۶ متغیر جدید در ساخت مدل زیر استفاده میکنیم.

تعداد لایه ها : ۱ **تعداد نورون های لایه ها : ۷** **آستانه : ۰.۰۰۰۱**

آزمایشی

```
          Reference
Prediction Low High
      Low   94    9
      High   6   46

               Accuracy : 0.9032
                 95% CI : (0.8454, 0.9448)
    No Information Rate : 0.6452
    P-Value [Acc > NIR] : 0.000000000000135

                  Kappa : 0.786

 Mcnemar's Test P-Value : 0.6056

            Sensitivity : 0.9400
            Specificity : 0.8364
         Pos Pred Value : 0.9126
         Neg Pred Value : 0.8846
             Prevalence : 0.6452
         Detection Rate : 0.6065
   Detection Prevalence : 0.6645
      Balanced Accuracy : 0.8882

       'Positive' Class : Low
```

آموزشی

```
          Reference
Prediction Low High
      Low  319   47
      High  24  230

               Accuracy : 0.8855
                 95% CI : (0.8578, 0.9095)
    No Information Rate : 0.5532
    P-Value [Acc > NIR] : < 0.0000000000000002

                  Kappa : 0.7665

 Mcnemar's Test P-Value : 0.00903

            Sensitivity : 0.9300
            Specificity : 0.8303
         Pos Pred Value : 0.8716
         Neg Pred Value : 0.9055
             Prevalence : 0.5532
         Detection Rate : 0.5145
   Detection Prevalence : 0.5903
      Balanced Accuracy : 0.8802

       'Positive' Class : Low
```

- همان طور که مشاهده میکنید با کاهش تعداد متغیر ها و جمع آوری بخش عظیمی از واریانس داده ها توانستیم اطلاعات بیشتری را به مدل منتقل کنیم و به همین دلیل نتایج مدل نیز بهبود عمده ای پیدا کرد.
- کلاس مثبت مربوط به رده Low است پس در این جا شاخص مهم برای ما مشخصه سازی Specificity است.
- دقت کلی مدل روی داده های آموزشی ۸۸ درصد است.
- در داده های آموزشی مدل توانسته ۹۳ درصد ثبت های مربوط به خانوار های کم درآمد را به درستی رده بندی کند.
- در داده های آزمایشی مدل توانسته ۹۴ درصد ثبت های مربوط به خانوار های کم درآمد را به درستی رده بندی کند.
- دقت کلی مدل روی داده های آزمایشی ۹۰ درصد است.
- در داده های آموزشی مدل توانسته ۸۳ درصد ثبت های مربوط به خانوار های پر درآمد را به درستی رده بندی کند.
- در داده های آزمایشی مدل توانسته ۸۳ درصد ثبت های مربوط به خانوار های پر درآمد را به درستی رده بندی کند.

➤ مقایسه عملکرد مدل ها

	آموزشی (Train)			آزمایشی (Test)		
	Accuracy	Sensetivity	Specificity	Accuracy	Sensetivity	Specificity
مدل ۱	81	90	59	80	89	61
مدل ۲	80	91	54	82	95	57
مدل ۳	82	91	58	80	94	53
مدل ۴	88	93	83	90	94	83

در این بخش بهترین مدل ۴ توانسته است بهترین عملکرد را برای رده بندی درست ثبت های مربوط به گروه پردرآمد در داده های آزمایشی را داشته باشد به همین دلیل این مدل را به عنوان مدل برتر این بخش انتخاب میکنیم.

٦. نتیجه گیری

	آموزشی (Train)			آزمایشی (Test)		
	Accuracy	Sensetivity	Specificity	Accuracy	Sensetivity	Specificity
۲ نزدیک ترین همسایه	90	87	100	75	80	65
درخت بدون پارامتر	85	93	66	79	91	55
لجستیک با تمام متغیر ها	86	92	70	78	87	61
شبکه عصبی + PCA	88	93	83	90	94	83

بهترین مدل در این جا مدل شبکه عصبی به کمک PCA است که توانسته است عملکرد بهتری نسبت به سایر مدل ها روی ثبت های ما داشته باشد.

در مدل PCA توانستیم با جمع آوری ۹۸ درصد اطلاعات موجود در ثبت ها در تنها ۶ ستون و با استفاده از شبکه عصبی مصنوعی با ۷ نورون در لایه پنهان به خوبی الگو های موجود در ثبت ها را پیدا کنیم و به دقت بالایی هم در رده بندی ثبت های با رده پردرآمد و هم ثبت های با رده کم درآمد برسیم.

یکی از مهم ترین و تاثیر گذار ترین اطلاعات موجود در این ثبت ها اطلاعات هزینه خانوار ها علی الخصوص هزینه خوراکی و دخانی خانوار است که پردرآمد یا کم درآمد بودن خانوار را مشخص میکند.

همان طور که انتظار میرود خانوار هایی که هزینه بیشتری در بخش های مختلف انجام میدهند عموما خانوار های پردرآمد هستند.

در صورت بیشتر بودن تعداد ثبت ها قطعا میتوانستیم مدل رده بندی بهتری برای این داده ها طراحی کنیم.

۷. پیوست

- **معرفی متغیر ها**

ردیف	تعریف متغیر	نام متغیر
۱	آدرس خانوار	**Address**
۲	کد استان	**C.Ostan**
۳	جنسیت سرپرست خانوار	**Gender**
٤	سن سرپرست خانوار	**Age**
٥	میزان سواد سرپرست خانوار	**Savad**
٦	سرپرست خانوار تحصیل می‌کند یا خیر؟	**InEdu**
۷	مدرک تحصیلی سرپرست خانوار	**Madrak**
۸	وضعیت فعالیت سرپرست خانوار	**Faaliat**
۹	تعداد اعضای خانوار	**Tedad.a**
۱۰	نحوه تصرف منزل مسکونی	**n.t.m**
۱۱	تعداد اتاق در اختیار	**T.O**
۱۲	سطح زیر بنای محل سکونت	**S.Z**
۱۳	نوع اسکلت بنای محل سکونت	**N.S**
١٤	مصالح عمده بنای محل سکونت	**Masleh**
١٥	اتومبیل شخصی	**oto**
١٦	موتورسیکلت	**motor**
۱۷	دوچرخه	**do**
۱۸	رادیو	**radio**
۱۹	ضبط	**zabt**
۲۰	تلویزیون سیاه و سفید	**TV.S**
۲۱	تلویزیون رنگی	**TV.r**
۲۲	انواع ویدئو، VCD و DVD	**DVD**
۲۳	انواع یارانه و تبلت	**Pc**
٢٤	تلفن همراه	**mobile**
٢٥	فریزر	**freeizer**
٢٦	یخچال	**yakhchal**
۲۷	یخچال فریزر	**yakhchal.f**

۲۸	اجاق گاز	**gaz**
۲۹	جارو برقی	**jaro.b**
۳۰	ماشین لباسشویی	**m.lebas**
۳۱	چرخ خیاطی	**charkh.kh**
۳۲	پنکه	**panke**
۳۳	کولر آبی متحرک	**cooler.a**
۳٤	کولر گازی متحرک	**cooler.g**
۳۵	ماشین ظرفشویی	**m.zarf**
۳٦	مایکروویو و انواع فرهای هالوژن دار	**microfer**
۳۷	آب لوله‌کشی	**ab.l**
۳۸	برق	**bargh**
۳۹	گاز لوله‌کشی	**gaz.l**
٤۰	تلفن ثابت	**tel**
٤۱	دسترسی به اینترنت	**internet**
٤۲	حمام	**hamam**
٤۳	آشپزخانه	**ashpazkhane**
٤٤	کولر آبی ثابت	**cooler.a.s**
٤۵	برودت مرکزی	**broodat.m**
٤٦	حرارت مرکزی	**hararat.m**
٤۷	پکیج	**package**
٤۸	کولر گازی ثابت	**cooler.g.s**
٤۹	شبکه عمومی فاضلاب	**fazelab**

ردیف	**نوع سوخت عمده مصرفی خانوار**	
	تعریف متغیر	**نام متغیر**
۵۰	نوع سوخت برای پخت و پز	**sookht.p**
۵۱	نوع سوخت برای ایجاد گرما	**sookht.g**
۵۲	نوع سوخت برای تهیه آب گرم	**sookht.ab**

ردیف	هزینه‌های خانوار	
	تعریف متغیر	نام متغیر
۵۳	هزینه‌های بهداشتی خانوار در یکماه گذشته	**H_Behdasht**
۵٤	هزینه ارتباطات خانوار در یکماه گذشته	**H_Ertebatat**
۵۵	هزینه‌های غذای آماده هتل و رستوران‌های خانوار در یکماه گذشته	**H_Ghazayeamade**
۵٦	هزینه‌های حمل و نقل خانوار در یکماه گذشته	**H_Hamlonaghl**
۵۷	هزینه کالاها یا خدمات متفرقه خانواردر یکماه گذشته	**H_kalavakhadamat**
۵۸	هزینه‌های خوراکی و دخانیات خانوار در یکماه گذشته	**H_Khorakivadokhani**
۵۹	هزینه‌های لوازم خانگی خاوار در یکماه گذشته	**H_mobleman**
٦۰	هزینه‌های مسکن- آب، سوخت، روشنایی و...	**H_Maskan**
٦۱	هزینه‌های نوشیدنی خانوار در یکماه گذشته	**H_Noshidani**
٦۲	هزینه های تفریحات خانوار در ماه گذشته	**H_Tafrihat**
٦۳	هزینه‌های پوشاک خانوار در یکماه گذشته	**H_Pushak**

ردیف	درآمدهای خانوار	
	تعریف متغیر	نام متغیر
٦٤	مبلغ دریافتی یارانه نقدی در ۱۲ ماه گذشته	**D_Yarane**
٦۵	درآمد آزاد خانوار در ۱۲ ماه گذشته	**D_Azad**
٦٦	درآمدهای متفرقه خانوار در ۱۲ ماه گذشته	**D_Motefaraghe**
٦۷	درآمد مزد خانوار در یک سال گذشته	**D_Mozd**

- **موقعیت استان همدان در نقشه ایران**

```
sf <- st_read(dsn="path/to/adm", layer="IRN_adm1")
shape <- readOGR(dsn="path/to/adm", layer="IRN_adm1")

pt_centroid<-sf %>% sf::st_centroid()
pts<-st_coordinates(pt_centroid)
p<-cbind(pts,pt_centroid)

labels = as.list(rep('',31))
labels[10] = list('همدان')

colors = as.list(rep('darkgrey',31))
colors[10] = list('#00b4b4')

ggplot(sf) +
    geom_sf(data = sf, aes(fill = colors)) +
    theme(legend.position="none")+
    geom_sf_text(aes(label = labels),size=4)
```

- **موقعیت شهرستان های استان همدان**

```
sf <- st_read(dsn="path.to.adm", layer="IRN_adm2")
sf = sf[sf$NAME_1 == 'Hamadan',]
shape <- readOGR(dsn="path.to.adm", layer="IRN_adm2")
shape = shape[shape@data$ID_1 == 10,]

coord = sf %>% sf::st_centroid() %>% st_coordinates(pt_centroid)
labels = a = c('تویسرکان','رزن','نهاوند','ملایر','کبودرآهنگ','همدان','بهار','اسدآباد')

ggplot(sf) +
  geom_sf(aes(fill=NAME_2)) +
  ggrepel::geom_label_repel(
    data = sf,
    aes(label = labels, geometry = geometry),
    stat = "sf_coordinates",
  )
```

- **قالب کلی استفاده شده در نمودار های یک متغیر نسبت به متغیر هدف**

```
calculate_percent = function(indexes) {
    temp               = as.data.frame(table(data[indexes]))
    total              = temp %>% group_by(temp[1]) %>% dplyr::summarise(tot=sum(Freq))
    temp$tot           = total[c(temp[[1]]),]$tot
    temp$percent       = temp$Freq / temp$tot
    temp[[1]]          = factor(temp[[1]],levels = c(as.character(temp[[1]][order(temp[1:(dim(temp)[1]/2),]$percent)])))
    colnames(temp)[1:2] = c('Var1','Var2')
    return (temp)
}
```

```
df = calculate_percent(c(1,59))
plot1 = ggplot(df, aes(fill=Var2, y=Freq, x=Var1))+
        geom_bar(position="stack", stat="identity") +
        labs(title="",x='', y = "تعداد",fill='') +
        theme(legend.position="top",
              axis.title.x = element_blank(),
              axis.text.x = element_blank(),
              axis.ticks.x = element_blank(),
              plot.margin = unit(c(0,0,1,0), "cm"))

plot2 = ggplot(df, aes(fill=Var2, y=Freq, x=Var1))+
    geom_bar(position="fill", stat="identity") +
    labs(title="",x='شهرستان', y = "درصد",fill='') +
    theme(legend.position="none",
          plot.title = element_text(hjust = 0.5),
          axis.text.x = element_text(angle = 60, hjust = 1),
          plot.margin = unit(c(-1.55,0,0,-0.09), "cm"))

plot3 = ggplot(df[1:(dim(df)[1]/2),], aes(fill=Var1, y=Freq, x=Var2))+
        geom_bar(position="stack", stat="identity") +
        labs(title="",x='نسبت', y = "",fill='') +
        theme(legend.position="right",
              axis.title.x = element_text(vjust=-7.6),
              axis.text.x = element_blank(),
              axis.ticks.x = element_blank(),
              axis.title.y = element_blank(),
              axis.text.y = element_blank(),
              axis.ticks.y = element_blank(),
            plot.margin = unit(c(1.38,0,2.09,.1), "cm")) +
        scale_fill_manual(values = colors)


grid.arrange(plot1,plot2,plot3,
             ncol = 2, nrow = 2,
             layout_matrix = rbind(c(1,3), c(2,3)),
             widths = c(3, 2), heights = c(3, 1.3))
```

1 = Address

59 = Y

- **نمودار ویالن + نمودار جعبه ای**

```
ggplot(data,aes(x    = S.Z,
                y    = Y,
                fill= Y)) +

    geom_violin() +

    geom_boxplot(width         = 0.2,
                 color         ="black",
                 alpha         = 0.2,
                 outlier.size  = -1) +

    labs(title ='',
         x     ='سطح زیربنا',
         y     = '',
         fill  ='') +

    theme(legend.position = "top",
          axis.title.x    = element_text(vjust=-0.5),
          plot.margin     = unit(c(-0.2,0,1,0), "cm"))
```

- **نمودار درصد افرادی که یک دارایی را دارند یا ندارند.**

```
ggplot(df, aes(fill = variable,
               y    = as.numeric(as.character(value)),
               x    = type)) +

    geom_bar(position ="fill",
             stat     ="identity") +

    labs(title = 'دارایی های افراد',
         x     = '',
         y     = 'درصد',
         fill  ='') +

    theme(plot.title  = element_text(hjust = 0.5),
          axis.text.x = element_text(angle = 60,
                                     hjust = 1)) +

    scale_fill_manual(values = colors)
```

- **نمودار درصد قرار گرفتن در رده پردرآمد یا کم درآمد در صورت داشتن یک دارایی**

```
property_y = melt(cbind(as.data.frame(ifelse(data[14:43] =='No',0,1)),
                        data[59]))
property_y = as.data.frame(table(property_y))
property_y = property_y[property_y$value==1,][c(1,2,4)]
total      = property_y %>%
             group_by(variable) %>%
             dplyr::summarise(tot = sum(Freq))
property_y$total    = rep(total[total$variable==as.character(property_y$variable[seq(1,dim(property_y)[1],2)]),]$tot,
                          each=2)
property_y$percent  = property_y$Freq / property_y$total
property_y          = property_y[c(order(property_y$percent[seq(1,dim(property_y)[1],2)])*2,
                                   order(property_y$percent[seq(1,dim(property_y)[1],2)])*2-1),]
property_y$variable = factor(property_y$variable,
                             levels=as.character(property_y$variable[1:(dim(property_y)[1]/2)]))
```

```
plot1 = ggplot(property_y, aes(fill =Y,
                               y    =Freq,
                               x    =variable))+
        geom_bar(position ="stack",
                 stat     ="identity") +
        labs(title ="",
             x     ='',
             y     = "تعداد",
             fill  ='') +
        theme(legend.position="none",
              axis.title.x   = element_blank(),
              axis.text.x    = element_blank(),
              axis.ticks.x   = element_blank(),
              axis.text.y    = element_text(angle = 90,
                                            hjust = 0.5),
              plot.margin    = unit(c(-0.5,2.1,1.5,1.085), "cm"))

plot2 = ggplot(property_y, aes(fill = Y,
                               y    = Freq,
                               x    = variable))+
    geom_bar(position = "fill",
             stat     = "identity") +
    labs(title ="",
         x     ='دارایی ها',
         y     = "درصد",
         fill  ='') +
    theme(legend.position ="none",
          plot.title      = element_text(hjust = 0.5),
          axis.text.x     = element_text(angle = 90, hjust = 1,vjust=.2),
          axis.title.x    = element_text(angle = 90,hjust=0.5,vjust=0.5),
          axis.text.y     = element_text(angle = 90, hjust = 0.5),
          plot.margin     = unit(c(-2,2.1,4.5,1.07), "cm"))


ggarrange(plot1 + rremove('x.text'),plot2,ncol=1,nrow=2,common.legend = T,legend='top')
```

- **نمودار مختصات موازی**

```
ggparcoord(data,
           columns     = 47:57,
           groupColumn = 'Y',
           alphaLines  = .5,
           scale       ="uniminmax",
           showPoints  = T) +

    scale_color_manual(values=c("#F8766D","#00BFC4")) +

    labs(title ="نمودار مختصات موازی هزینه های مختلف - هم مقیاس شده - دو گروه کم درآمد و پردرآمد",
         x     ="موضوع هزینه",
         y     = "مقدار",
         color ='')+

    theme(axis.text.x  = element_text(angle = 60, hjust = 1,size=15),
          axis.title.x = element_text(angle = 0, hjust = 0.5,size=18),
          axis.title.y = element_text(angle = 90, hjust = 0.5,size=18),
          plot.title   = element_text(angle = 0, hjust = 0.5,size=20))
```

- **نمودار جعبه ای مقایسه هزینه های دو رده خانوار**

```
scaled_h = as.data.frame(lapply(data[47:57],normalize,na.rm = T))
scaled_h = as.data.frame(cbind(scaled_h,Y= data$Y))
scaled_h = melt(scaled_h)

ggplot(scaled_h,aes(x=variable , y=value , fill=Y)) +
    geom_boxplot(aes(colour = skeleton_type) ,outlier.colour="black", outlier.shape=21,outlier.size=1,colour="black") +
    labs(title ="نمودار جعبه ای هزینه های مختلف - هم مقیاس شده - دو گروه کم درآمد و پردرآمد",
         x     ="موضوع هزینه",
         y     = "مقدار",
         fill  ='') +
    theme(axis.text.x = element_text(angle = 60, hjust = 1),
          plot.title  = element_text(angle = 0, hjust = 0.5))
```

- **نمودار عنکبوتی**

```
df = rbind(c(rep(0,10)),
           as.data.frame(lapply(data[data$Y=='Low',][47:57],mean,na.rm=T)),
           as.data.frame(lapply(data[data$Y=='High',][47:57],mean,na.rm=T)))
max = rep(max(df),10)
df  = rbind(max,df)
rownames(df) = c('Max','Min','Low','High')

colors_border=c( "#F8766D", "#00BFC4")
colors_in=c( rgb(247/255, 183/255, 183/255,0.4),rgb(98/255, 193/255, 196/255,0.4) )

radarchart( df , axistype=1 , maxmin=T,
    pcol=colors_border , pfcol=colors_in , plwd=4 , plty=1,
    cglcol="grey", cglty=4, axislabcol="black", cglwd=3.5,
    vlcex=1.3,
    )
legend(x=1,
       y=1.3,
       legend = rownames(df[-c(1,2),]),
       bty = "n",
       pch=15 ,
       col=colors_border ,
       text.col = "Black",
       cex=1,
       pt.cex=4)
```

- **نمودار شبکه ای**

```
links = as.data.frame(table(data$Address,data$Y))
nodes <- data.frame(
  name = c(as.character(links$Var1),as.character(links$Var2)) %>% unique()
)

links$IDsource <- match(links$Var1, nodes$name)-1
links$IDtarget <- match(links$Var2, nodes$name)-1

nodes$group <- as.factor(c("node"))
my_color <- 'd3.scaleOrdinal() .domain(["L", "H","node"]) .range(["#F8766D","#00BFC4","grey"])'

p <- sankeyNetwork(Links = links, Nodes = nodes, Source = "IDsource", Target = "IDtarget",
                   Value = "Freq", NodeID = "name",
                   colourScale=my_color, LinkGroup="Var2",NodeGroup='group',fontSize=15,sinksRight=F)

p
```

- **نمودار بالنی**

```
ggballoonplot(df, x = "Address", y = "n.t.m", size = "percent",fill='Y',facet.by = 'Y',
              ggtheme = theme_bw())
```

- **نمودار میله ای متغیر ها نسبت به یکدیگر**

```
ggplot(data, aes(Tedad.a)) +

    geom_histogram(aes(fill=Gender),
                   binwidth = 1,
                   col='black',
                   size=.1) +

    labs(title="تعداد اعضای خانوار به تفکیک جنسیت سرپرست خانوار",
         x="تعداد اعضای خانوار",
         y = "تعداد",
         fill='جنسیت')+

    theme(plot.title = element_text(hjust = 0.5)) +

    scale_fill_manual(values = colors)
```

- **نمودار حرارتی همبستگی متغیر ها**

```
to_cor = data[c(2,4,9,11,12,47:58)]
cormat = cor(to_cor)

lmat <- rbind( c(5,3,4), c(2,1,4) )
lhei <- c(.1, .9)
lwid <- c(.5, 5, 0.1)

heatmap.2(
  data         = cormat,
  lmat         = lmat,
  lhei         = lhei,
  lwid         = lwid,
  keysize      = 1,
  notecex      = 1.2,
  key.title    = '',
  key          = F,
  colsep       = 0:nrow(cormat),
  rowsep       = 0:nrow(cormat),
  sepcolor     = 'black',
  sepwidth     = c(0.01,0.01),
  cellnote     = round(cormat,2),
  main         = "Correlation Matrix",
  notecol      = "black",
  density.info = "density",
  trace        = "none",
  margins      = c(14,8),
  col          = colorRampPalette(c('#9511ed', '#ffffff','#ed4f11'))(100),
  dendrogram   = "row")
```

- **نمودار بررسی عملکرد مدل KNN**

```
colors = c('#1B9E77','#D95F02', '#7570B3', '#E6AB02',
           '#E7298A',"#067bab","#00ff8f","#da8dca",
           '#66A61E',"#b50909","#cfd206",'#a98c09')

ggplot(data = to_plot_df) +
geom_line(mapping = aes(x     = k,
                        y     = value,
                        group = metric,
                        color = metric),
          size=1) +
labs(title ="مقادیر شاخص های مختلف روی داده های اعتبارسنجی برای مقادیر مختلف k",
     x     ='K',
     y     = "مقدار",
     color ='شاخص') +
theme(plot.title  = element_text(hjust = 0.5),
      axis.text.x = element_text(angle = 0,
                                 hjust = 0.5))+

scale_color_manual(values = colors)
```

- **نمودار درخت تصمیم**

```
rpart.plot(class.tree,
           type          = 1,
           split.font    = 2,
           extra         = 1,
           box.palette   = c("#00BFC4","#F8766D"),
           xcompact      = FALSE,
           ycompact      = TRUE,
           varlen        = -8,
           branch.lty    = 6,
           branch.col    = 8,
           branch.lwd    = 3,
           leaf.round    = 0,
           shadow.col    = 'gray',
           fallen.leaves = TRUE,
           cex           = 0.6)
```

- **نمودار جنگل تصادفی**

```
plotmo(random.forest,
       type   = "prob",
       type2  = "image",
       ngrid2 = 100,
       pt.col = ifelse(training_data$Y == "High", '#00BFC4' , '#F8766D'))
```

KNN ▪

```
k_results = list()
for (k in 1:15){
    fold_results = list()
    fold_acc     = list()
    for (z in 1:5){
        fold_number = z

        train_cv = data[train_folds[[fold_number]],]
        valid_cv = data[valid_folds[[fold_number]],]

        norm.values = preProcess(train_cv[c(2,4,9,11,12,47:57)],method=c('center','scale'))

        train.norm.numerical = as.data.frame(predict(norm.values,  train_cv[c(2,4,9,11,12,47:57)]))
        valid.norm.numerical = as.data.frame(predict(norm.values,  valid_cv[c(2,4,9,11,12,47:57)]))

        train.dummy.categorical = categorical_predictors[train_folds[[fold_number]],]
        valid.dummy.categorical = categorical_predictors[valid_folds[[fold_number]],]

        train.x = as.data.frame(c(train.norm.numerical,train.dummy.categorical))
        valid.x = as.data.frame(c(valid.norm.numerical,valid.dummy.categorical))

        train.y = factor(target[train_folds[[fold_number]],]$Y,levels=c('Low','High'))
        valid.y = factor(target[valid_folds[[fold_number]],]$Y,levels=c('Low','High'))

        CM =
        confusionMatrix(
            knn(train = train.x,
                test  = valid.x,
                cl    = train.y,
                k     = k),
            as.factor(valid.y))

        results = CM$byClass
        acc     = CM$overall[1]

        fold_results[[z]] = results
        fold_acc[[z]]     = acc
    }
    k_results[[k]] = c(rowMeans(as.data.frame(fold_results)),
                       rowMeans(as.data.frame(fold_acc)))
}
results        = as.data.frame(k_results)
names(results) = seq(1:length(names(results)))
results
```

```
knn(train = train_valid_x,
    test  = train_valid_x,
    cl    = train_valid_y,
    k     = 2)
```

Deepest Tree

```
deep.class.tree = rpart(formula  = Y~. ,
                        data     = training_data,
                        cp       = 0,
                        minsplit = 1,
                        method   = "class")
```

Pruned Tree Using Least X-Error

```
pruned.deep.class.tree = prune(deep.class.tree,
                               cp=deep.class.tree$cptable[which.min(deep.class.tree$cptable[,'xerror']),'CP'])
```

Random Forest

```
random.forest = randomForest(as.factor(Y) ~ . ,
                             data       = training_data,
                             ntree      = 500,
                             mtry       = 4,
                             nodesize   = 5,
                             importance = TRUE)
```

Boosted Tree

```
boosted.tree = boosting(Y ~ . ,
                        data = training_data)
```

Base Logistic

```
logit.reg = glm(Y ~ . ,
                data = training_data ,
                family = 'binomial')
```

Forward Logistic

```
logit.reg.forward = glm(Y ~ 1 ,
                        data   = training_data ,
                        family = 'binomial')

formula.all.variables = formula(glm(Y ~ . ,
                                     data   = training_data ,
                                     family = 'binomial'))

forward.reg = step(logit.reg.forward,
                   direction = 'forward',
                   scope     = formula.all.variables)
```

Backward Logistic

```
backward.reg = step(glm(Y ~ . ,
                        data   = training_data ,
                        family = 'binomial'),
                    direction  = 'backward')
```

Stepwise Logistic

```
stepwise.reg = step(glm(Y ~ . ,
                        data   = training_data ,
                        family = 'binomial') ,
                    direction  = 'both')
```

Neural Network ▪

```
nn = neuralnet(formula             = Y ~ .,
               data                = train.pca,
               hidden              = c(2),
               threshold           = 0.001,
               stepmax             = 1e+04 ,
               algorithm           = "rprop+",
               err.fct             = "sse",
               learningrate.factor = list(minus = .7 ,plus = 1.4),
               act.fct             = "logistic",
               linear.output       = FALSE,
               lifesign            = 'full',
               lifesign.step       = 100)
```

۸. منابع

https://www.r-graph-gallery.com

http://www.sthda.com/english

http://stackoverflow.com

https://blog.faradars.org

Data Mining For Business Analytics Book