به نام خدا

پروژه داده کاوی

استاد راهنما: دکتر فقیهی حبیب آبادی

عنوان پروژه: طرح هزینه و درامد خانوار سال ۱۳۹۹

فهرست

پیشگفتار

طرح آمارگیری هزینه و درامد خانوار از سال ۱۳۴۲ در مناطق روستایی و از سال ۱۳۴۷ در مناطق شهری به اجرا درآمده است.از سال ۱۳۵۳ علاوه بر هزینه‌های خانوار،اطلاعات درامد نیز گردآوری شد.هدف کلی طرح آمارگیری هزینه و درامد خانوار،براورد متوسط هزینه‌ها و درامد یک خانوار شهری و روستایی در سطح کشور و استان است.اهمیت این طرح در امکان بررسی روند مصرف کالاها و خدمات، مطالعه روابط متقابل ویژگی‌های اجتماعی اقتصادی خانوارها، ارزیابی آثار سیاست-های اقتصادی در زمینه تأمین عدالت اجتماعی و بررسی توزیع درامد و سرانجام نقش مهم آن در تأمین اطلاعات مورد نیاز حساب‌های ملی و منطقه‌ای، به خصوص در بررسی‌ها و برنامه‌ریزی‌های اقتصادی اجتماعی کشور است.

طرح آمارگیری هزینه‌های و درامد خانوار سال ۱۳۹۹ بر اساس آمارگیری از ۱۹۳۰۶ خانوار نمونه در نقاط شهری و ۱۸۲۵۱ خانوار نمونه در نقاط روستایی کشور اجرا شده است.

نتایج تفصیلی این طرح شامل ۲۳۴ جدول در قالب ۲ نشریه به تفکیک شهری و روستایی توسط مرکز آمار ایران هر ساله منتشر می شود.همچنین نتایج تفصیلی طرح در قالب ۵۱۲ جدول، شامل جزئی-ترین اطلاعات در سطح ۳۱ استان کشور و کل کشور برای استفاده از طریق واحد اطلاع‌رسانی مرکز آمار ایران در اختیار عموم علاقمندان، محققین، برنامه‌ریزان و پژوهشگران قرار می‌گیرد.

جامعه هدف این طرح شامل همه‌ی خانوارهای معمولی ساکن و گروهی در نقاط شهری یا روستایی است.این خانوارهای نمونه از ۳۸۷ شهرستان در مناطق شهری و از ۳۹۵ شهرستان در مناطق روستایی کل کشورانتخاب شده‌اند[۱]

فصل اول: معرفی مجموعه داده‌ها

مقدمه

به دلیل وسعت کشور ایران و وجود اقلیم‌های متنوع مطمئنا میزان هزینه و درامد خانوار در همه استان‌ها و خانوارهای مختلف به یک میزان نمی‌باشد. پس برای اجرای عدالت اجتماعی اقتصادی در بین اقشار مختلف جامعه همواره باید الگویی خاص برقرار شود تا خانوارهایی که مستحق دریافت یارانه کمک معیشتی هستند از خانوارهایی که

[۱]- بر اساس آخرین تقسیمات کشوری تا پایان سال ۱۳۹۵، کشور ایران از ۳۱ استان و ۴۲۹ شهرستان تشکیل شده است.

نیازی به یارانه مذکور ندارند، تفکیک شوند، تا علاوه بر کاهش هزینه-های اقتصادی کشور به بهبود برنامه‌ریزی اجتماعی اقتصادی منجر شود.

به گزارش ایسنا: "در ۱۹ اردیبهشت ماه سال ۱۴۰۱ وزیر امور اقتصادی و دارایی کشور از آغاز طرح واریز یارانه کمک معیشتی به میزان ۳۰۰ و ۴۰۰ هزار تومان برای تعداد اعضای خانوار خبر دادند."[۱]

این مبالغ به حساب ۷۷ میلیون نفر از جمعیت کشور که تا پیش از این یارانه ماهیانه دریافت می‌کردند، به شرح زیر واریز شد:

سه دهک پایین جامعه که از درامد کمی برخوردار بودند، هر یک ۴۰۰ هزار تومان، شش دهک میانی جامعه که از درامد متوسط برخوردار بودند، هر یک ۳۰۰ هزار تومان دریافت نمودند و دهک دهم جامعه یارانه مذکور را دریافت ننمودند.

پس از سیل عظیمی از اعتراضات منوط به عدم قرارگیری صحیح در دهک بندی صورت گرفته، تصمیم برآن شد که این حمایت مالی از دهک بالای جامعه نیز انجام گیرد.

هدف

حال هدف این پروژه برآن است تا مدلی ارائه نماید که این دهک بندی ها را به نحوی انجام دهد تا افراد قرار گرفته در دهک دهم تا جایی که ممکن است با خطای کمتری شناسایی گردند.

خلاصه هدف: شناسایی صحیح خانوارهای قرار گرفته در دهک دهم.

ضرورت هدف: اجرای صحیح دهک بندی و کاهش تعداد خانوارهای معترض در طرح یارانه کمک معیشتی و شناسایی خانوارهایی که نیازی به یارانه مذکور ندارند.

معرفی مجموعه داده‌ها

مجموعه داده‌های هزینه و درامد خانوار سال ۱۳۹۹ مرکز آمار ایران[۲] در قالب یک فایل (Access) و توضیحات مرتبط با این فایل به همراه پرسشنامه‌ای است که این داده‌ها از آن استخراج گردیده است، می-باشد.

در این مجموعه داده‌ها اسامی افراد خانوار وآدرس محل سکونت به دلیل رعایت حریم شخصی و حفظ حقوق شهروندان با شناسه هایی کدگذاری شده‌اند.تنها معیار برای افراز این خانوارها سه رقم اول این کدگزاری می‌باشد،که نشان دهنده استان محل سکونت خانوار است.(جدول شماره ۱)

ردیف	نام استان	نمونه شناسه کدگذاری شده
۱	مرکزی	۱۰۰۱۰۰۰۹۳۱۸
۲	گیلان	۱۰۱۰۶۰۱۸۴۲۶
۳	مازندران	۱۰۲۱۰۰۳۱۴۱۳
۴	آذربایجان‌شرقی	۱۰۳۱۲۰۴۴۸۱۷
۵	آذربایجان‌غربی	۱۰۴۰۱۰۴۸۸۲۹
۶	کرمانشاه	۱۰۵۰۵۰۶۴۱۱۸
۷	خوزستان	۱۰۶۰۳۰۷۳۹۲۳
۸	فارس	۱۰۷۲۴۰۹۴۹۲۶
۹	کرمان	۱۰۸۰۶۰۹۹۴۲۶
۱۰	خراسان رضوی	۱۰۹۱۶۱۱۱۵۱۴
۱۱	اصفهان	۱۱۰۱۲۱۳۰۸۲۹
۱۲	سیستان و بلوچستان	۱۱۱۰۶۱۴۳۴۱۵
۱۳	کردستان	۱۱۲۰۴۱۵۳۱۲۹
۱۴	همدان	۱۱۳۰۳۱۶۲۴۲۹
۱۵	چهارمحال‌وبختیاری	۱۱۴۰۲۱۷۵۴۱۷
۱۶	لرستان	۱۱۵۰۹۱۹۱۵۲۳
۱۷	ایلام	۱۱۶۰۱۱۹۲۵۲۳
۱۸	کهگیلویه و بویراحمد	۱۱۷۰۶۲۰۸۸۱۷
۱۹	بوشهر	۱۱۸۰۳۲۱۴۰۲۶
۲۰	زنجان	۱۱۹۰۴۲۲۳۸۲۱
۲۱	سمنان	۱۲۰۰۲۲۳۱۸۲۲
۲۲	یزد	۱۲۱۰۵۲۴۴۴۱۵
۲۳	هرمزگان	۱۲۲۰۵۲۵۸۹۲۹
۲۴	تهران	۱۲۳۰۱۲۷۷۹۱۹
۲۵	اردبیل	۱۲۴۰۱۲۹۶۷۲۶
۲۶	قم	۱۲۵۰۱۳۰۴۴۱۹
۲۷	قزوین	۱۲۶۰۳۳۱۹۷۱۷
۲۸	گلستان	۱۲۷۱۴۳۴۴۸۲۰
۲۹	خراسان‌شمالی	۱۲۸۰۴۳۵۳۲۱۳
۳۰	خراسان‌جنوبی	۱۲۹۰۱۳۵۸۱۲۱
۳۱	البرز	۱۳۰۰۱۳۷۶۱۲۳

جدول ۱: نمونه شناسه‌های موجود در دیتا خام

این کدگذاری استان‌ها و شهرستان‌ها در درگاه ملی آمار ایران صورت گرفته است.[۳]

---

۱ -برای جزئیات خبر میتوانید از طریق پیوند زیر اطلاعات بیشتری کسب نماید. (https://www.isna.ir/news/1401021912430/)

۲ - پیوند مجموعه داده‌ها: (https://drive.google.com/file/d/1QEcXCzVlWOjp1IRGqNnrkpryrsNBxppa/view?usp=sharing)

۳ - فایل تقسیمات کشوری سال ۹۹ را می‌توانید از طریق پیوند زیر مشاهده نمایید. (https://www.amar.org.ir/Portals/0/Geo/GEO99.xlsx)

همانطور که در قسمت مقدمه ذکر شد و از آنجایی که میزان هزینه و درامد خانوار در استان‌ها متفاوت است پس استان‌های کشور را در ۷ گروه‌بندی کلی به شرح زیر در نظر می‌گیریم. (جدول شماره ۲)

گروه	استان های قرار گرفته در گروه
۱	کرمان،خراسان‌رضوی،خراسان‌شمالی،خراسان‌جنوبی، سیستان و بلوچستان
۲	گیلان، مازندران، گلستان
۳	تهران، قزوین، البرز
۴	مرکزی، اصفهان، همدان، سمنان، قم، یزد
۵	آذربایجان‌شرقی، آذربایجان‌غربی، زنجان، اردبیل
۶	کرمانشاه، کردستان، چهارمحال‌وبختیاری، لرستان،ایلام،کهگیلویه و بویر-احمد
۷	خوزستان، فارس، بوشهر، هرمزگان

جدول ۲: گروه بندی استان‌های همسایه

مطابق گروه‌بندی صورت‌گرفته در کل کشور، استان‌های همسایه با یک دیگر در یک گروه قرار گرفته‌اند.

در این پروژه قصد داریم هزینه و درامد خانوار استان‌های گروه شماره ۶ را مورد بررسی قرار دهیم. پس نسبت به جداسازی این گروه اقدامات لازم را انجام می‌دهیم. همچنین با توجه به اینکه در هر ثبت داده خام، متغیرهایی اعم از ریز هزینه‌های خوراک، پوشاک، مسکن و ... وجود دارد پس تمام این ریز هزینه‌ها را تجمیع کرده و در متغیر مربوطه با یک عنوان وارد می‌نماییم.

در نتیجه مجموعه داده‌های گروه ۶ با ۳۱۰۷ ثبت و ۶۹ متغیر در دسترس قرار گرفته است.

معرفی متغیرهای مجموعه داده‌ها ۱

حال به معرفی متغیرهای مجموعه داده میپردازیم. (جدول شماره ۳)

ر	نام متغیر	معرفی و توضیحات
۱	X	اندیس شماره سطر
۲	Ostan	استان محل سکونت
۳	gorooh	گروه بندی صورت گرفته
۴	address	شناسه خانوار
۵	NoeKhanevar	نوع خانوار(۱: معمولی ساکن، ۲: گروهی)
۶	TedadAza	تعداد اعضا خانوار(عدد صحیح)
۷	Sarparstjensiat	جنسیت سرپرست(۱: مرد ، ۲: زن)
۸	SarparstSen	سن سرپرست(عدد صحیح)
۹	SarparstSavad	سواد سرپرست(۱: با سواد، ۲: بیسواد)
۱۰	SarparstTahsil	در حال تحصیل سرپرست(۱:بله، ۰: خیر)
۱۱	SarparstMadrak	مدرک‌تحصیلی‌سرپرست (۱:ابتدایی،۲:راهنمایی،۳:متوسطه،۴:دیپلم،۵:کاردانی،۶ :کارشناسی،۷:کارشناسی‌ارشد،۸:دکتراتخصصی، ۹:سایر و غیر رسمی)
۱۲	SarparstFaaliat	فعالیت‌سرپرست(۱:شاغل،۲:بیکار(جویای‌کار)،۳:دارای درامد بدون کار،۴:محصل، ۵:خانه دار،۶:سایر)
۱۳	Sarparstzanashoyi	زناشویی سرپرست(۱:دارای همسر، ۲: بی‌همسر بر اثر فوت همسر، ۳:طلاق، ۴:هرگز ازدواج نکرده)
۱۴	NahveTasarof	نحوه تصرف محل سکونت(۱:ملکی عرصه و اعیان، ۲:ملکی اعیان، ۳:اجاری، ۴:رهن، ۵: در برابر خدمت، ۶: رایگان)
۱۵	TedadOtagh	تعداد اتاق (عدد صحیح)
۱۶	SatheZirbana	مساحت زیر بنا (برحسب متر مربع ، عدد صحیح)
۱۷	NoeEskelet	نوع اسکلت(۱: فلزی، ۲:بتون ارمه ، ۳: سایر)
۱۸	mashin	خودرو (۱: دارد ، ۰: ندارد)
۱۹	motor	موتورسیکلت(۱: دارد ، ۰: ندارد)
۲۰	docharkhe	دوچرخه(۱: دارد ، ۰: ندارد)
۲۱	radio	رادیو(۱: دارد ، ۰: ندارد)
۲۲	radiozabt	ضبط(۱: دارد ، ۰: ندارد)
۲۳	tv	تلویزیون سیاه سفید(۱: دارد ، ۰: ندارد)
۲۴	tvrangi	تلویزیون رنگی(۱: دارد ، ۰: ندارد)
۲۵	video	ویدیو و dvd(۱: دارد ، ۰: ندارد)
۲۶	computer	رایانه و تبلت(۱: دارد ، ۰: ندارد)
۲۷	mobile	گوشی همراه(۱: دارد ، ۰: ندارد)
۲۸	freezer	فریز(۱: دارد ، ۰: ندارد)
۲۹	yakhchal	یخچال(۱: دارد ، ۰: ندارد)
۳۰	yakhchalfreezer	یخچال فریزر(۱: دارد ، ۰: ندارد)
۳۱	ojaghgaz	اجاق گاز(۱: دارد ، ۰: ندارد)
۳۲	jarobarghi	جاروبرقی(۱: دارد ، ۰: ندارد)
۳۳	lebasshoyi	لباسشویی(۱: دارد ، ۰: ندارد)
۳۴	khayati	چرخ خیاطی(۱: دارد ، ۰: ندارد)
۳۵	panke	پنکه(۱: دارد ، ۰: ندارد)
۳۶	coolerabimoteharek	کولر ابی متحرک(۱: دارد ، ۰: ندارد)
۳۷	coolergazimoteharek	کوار گازی متحرک(۱: دارد ، ۰: ندارد)
۳۸	zarfshoyi	ظرفشویی(۱: دارد ، ۰: ندارد)
۳۹	microfer	مایکروفر(۱: دارد ، ۰: ندارد)
۴۰	hichkodam	هیچکدام(۱: دارد ، ۰: ندارد)
۴۱	lolekeshiab	لوله کشی اب(۱: دارد ، ۰: ندارد)
۴۲	bargh	برق(۱: دارد ، ۰: ندارد)
۴۳	gazlolekeshi	لوله کشی گاز(۱: دارد ، ۰: ندارد)
۴۴	telephone	تلفن(۱: دارد ، ۰: ندارد)
۴۵	internet	اینترنت(۱: دارد ، ۰: ندارد)
۴۶	hamam	حمام(۱: دارد ، ۰: ندارد)
۴۷	ashpazkhane	اشپزخانه(۱: دارد ، ۰: ندارد)
۴۸	coolerabisabet	کولر ابی ثابت(۱: دارد ، ۰: ندارد)
۴۹	borodatmarkazi	سرمایش مرکزی(۱: دارد ، ۰: ندارد)
۵۰	hararatmarkazi	گرمایش مرکزی(۱: دارد ، ۰: ندارد)
۵۱	package	پکیج(۱: دارد ، ۰: ندارد)
۵۲	coolergazisabet	کولر گازی ثابت(۱: دارد ، ۰: ندارد)
۵۳	Fazelabshahri	فاضلاب شهری(۱: دارد ، ۰: ندارد)
۵۴	nsokhtpokhtpaz	سوخت پخت و پز(۳:گاز مایع،۴: گاز طبیعی شبکه)
۵۵	nsokhtgarma	سوخت‌گرمایش(۱۱:نفت‌سفید،۱۳:گازمایع،۱۴: گاز-طبیعی شبکه)
۵۶	nsokhtabgarm	سوخت‌اب‌گرم(۲۱:نفت‌سفید،۲۳:گازمایع،۲۴: گاز طبیعی شبکه)
۵۷	HazineKhoraki	هزینه خوراکی(برحسب ریال ، عددی)

۵۸	HazineNoshidani	هزینه نوشیدنی(برحسب ریال ، عددی)
۵۹	HazinePoshak	هزینه پوشاک(برحسب ریال ، عددی)
۶۰	HazineMaskan	هزینه مسکن(برحسب ریال ، عددی)
۶۱	HazineLavazemKhanegi	هزینه لوازم خانگی(برحسب ریال ، عددی)
۶۲	HazineDarmani	هزینه درمانی(برحسب ریال ، عددی)
۶۳	HazineHamlonaghl	هزینه حمل‌و نقل(برحسب ریال ، عددی)
۶۴	HazineErtebatat	هزینه ارتباطات(برحسب ریال ، عددی)
۶۵	HazineTafrihatFarhangi	هزینه تفریحات فرهنگی(برحسب ریال ، عددی)
۶۶	HazineGhazaAmade	هزینه غذا آماده(برحسب ریال ، عددی)
۶۷	HazineKalaMotefaregheh	هزینه کالا متفرقه(برحسب ریال ، عددی)
۶۸	HazineKalaBadavam	هزینه کالا بادوام (برحسب ریال ، عددی)
۶۹	Target	درامد(برحسب ریال ، عددی)

جدول ۳: معرفی متغیرهای مجموعه داده

همان طور که مشاهده میکنید اکثر متغیر های فوق نیازی به توضیحات خاصی ندارند. متغیرهای Hazine همگی مرتبط به ماه گذشته می‌باشدبه غیر از HazineKalaBadavam که مرتبط با هزینه ها در یکسال گذشته می باشد.

تنها متغیر Target که از مشورت و همفکری دانشجویان دانشگاه شهید بهشتی[1]طبق معادله زیر ایجاد گردیده است.

$$Target = (\text{SarmayeGozari} + \text{DaramadNKH12} + \text{AzadDaramadKh} + \text{MotefaregheHoghogh} + \text{Motefareghe Ejareh} + \text{MotefareghePasandaz} + \text{MotefaregheKomakhazine} + \text{MotefaregheForoshMasnoat} + \text{MotefaregheDaryafti})/12$$

حال به بررسی دقیق‌تر برخی از متغیرها می‌پردازیم.

- متغیر Ostan مرتبط با محل سکونت خانوار می‌باشد و تا قبل از گروه بندی کلی مورد استفاده گرفته است و با توجه اینکه در ادامه پروژه اطلاعات مفیدی در اختیار ما نمیگذارد و تمام اطلاعات لازم در متغیر آدرس موجود است، پس می بایست حذف گردد.
- متغیر gorooh نیز مرتبط با موضوع فوق است و با توجه به یکسان بودن آن برای هر سطر نیز در ادامه هیچ اطلاعاتی برای ما ندارد.
- متغیر address نیز با توجه به اینکه شناسه هر خانوار است و برای هر خانوار یکتاست پس همانند دو متغیر فوق در ادامه پروژه کمک شایانی نمیکند.
- متغیر SarparastTahsil و SarparastMadrak هر کدام دارای ۶۰۸ مقدار NA یا همان مقادیر ناموجود می‌باشند. این مقادیر ناموجود به متغیر SarparastSavad که باسواد بودن یا بی‌سواد بودن سرپرست خانوار را تعیین میکند، ارتباط دارد. این مقادیر به دلیل بی‌سواد بودن سرپرست خانوار می‌باشد. با توجه به تعداد زیاد این مقادیر و اهمیت این متغیر این مقادیر ناموجود را با مقدار "۰" جایگزین میکنیم.
- متغیر Hchkodam به دلیل یکسان بودن در تمامی ثبت‌ها و همچنین با توجه به سوال پرسشنامه اطلاعاتی در اختیار ما قرار نمیدهد. پس این متغیر نیز باید حذف گردد.
- متغیر bargh نیز همانند متغیر فوق در تمامی ثبت‌ها دارای مقدار ۱ می‌باشد.این نشان دهنده آن است که در این نمونه گیری هیچ خانوار بدون اشتراک برق وجود ندارد و از آنجایی که این نمونه گیری از تمامی بخش‌های استان‌های مذکور جمع‌آوری شده است، پس نتیجه میگیریم در تمام بخش‌های استان‌های مذکور برق شبکه خانگی وجود دارد. پس این متغیر نیز در ارائه مدل هیچ تاثیری ندارد.

نتیجه‌گیری فصل اول:

بنا به توضیحات فوق متغیرهای Ostan و gorooh و address و Hchkodam و bargh و NA حذف گردید و مقادیر با صفر جایگزین گردید، در شکل شماره ۱ خلاصه‌ای از هر متغیر را مشاهده می‌کنید.

همچنین برای درک بهتر مدل، از طریق دهک‌بندی، خانوارهای قرارگرفته در رده دهم را (با توجه به هدف پروژه) با یک و خانوارهای دیگر را با صفر نشان می‌دهیم.

همانطور که در خروجی زیر مشاهده میکنید دهک دهم، درامدی بین ۷۰۸۸۲۶۶۷ ریال و ۶۵۴۸۳۳۳۳۳ ریال دارند.

```
> quantile(final.df$Target, probs = seq(.1, .9, by = .1))
     10%      20%      30%      40%      50%      60%      70%      80%      90%
 8690000 15113333 19694000 23881333 28956667 34389667 41790333 51916667 70882667
> #calculate deciles of dataset
> quantile(final.df$Target, probs = seq(.9, by = .1))
      90%      100%
 70882667 654833333
```

و همچنین تعداد خانوار قرارگرفته در هر رده به شرح زیر است.

```
> as.data.frame(table(final.df$decile))
   Var1 Freq
1     1  311
2     2  311
3     3  311
4     4  311
5     5  311
6     6  311
7     7  311
8     8  310
9     9  310
10   10  310
```

پس مطابق گفته‌های بالا داریم.

```
> as.data.frame(table(final.df$decile))
  Var1 Freq
1    0 2797
2    1  310
```

۱ -دانشجویان دانشگاه شهید بهشتی، مقطع کارشناسی ارشد رشته ریاضی کاربردی- گرایش علوم داده، ورودی سال ۱۴۰۰

متغیر **decile** متغیر برامد است.

```
       X            NoeKhanevar       TedadAza
 Min.   :  685   Min.   :1.000   Min.   : 1.000
 1st Qu.: 5548   1st Qu.:1.000   1st Qu.: 3.000
 Median :10366   Median :1.000   Median : 4.000
 Mean   : 9359   Mean   :1.001   Mean   : 3.551
 3rd Qu.:12280   3rd Qu.:1.000   3rd Qu.: 4.000
 Max.   :17117   Max.   :2.000   Max.   :11.000
 SarparstJensiat  SarparstSen    SarparstSavad
 Min.   :1.000   Min.   :18.00   Min.   :1.000
 1st Qu.:1.000   1st Qu.:39.00   1st Qu.:1.000
 Median :1.000   Median :49.00   Median :1.000
 Mean   :1.128   Mean   :51.04   Mean   :1.196
 3rd Qu.:1.000   3rd Qu.:62.00   3rd Qu.:1.000
 Max.   :2.000   Max.   :99.00   Max.   :2.000
 SarparstTahsil  SarparstMadrak SarparstFaaliat
 Min.   :0.000   Min.   :0.0     Min.   :1.00
 1st Qu.:2.000   1st Qu.:1.0     1st Qu.:1.00
 Median :2.000   Median :2.0     Median :1.00
 Mean   :1.603   Mean   :2.7     Mean   :1.72
 3rd Qu.:2.000   3rd Qu.:4.0     3rd Qu.:3.00
 Max.   :2.000   Max.   :9.0     Max.   :6.00
 SarparstZanashoyi  NahveTasarof     TedadOtagh
 Min.   :1.000      Min.   :1.000    Min.   :1.00
 1st Qu.:1.000      1st Qu.:1.000    1st Qu.:3.00
 Median :1.000      Median :1.000    Median :4.00
 Mean   :1.195      Mean   :1.943    Mean   :3.76
 3rd Qu.:1.000      3rd Qu.:3.000    3rd Qu.:4.00
 Max.   :4.000      Max.   :6.000    Max.   :8.00
  SatheZirbana      NoeEskelet         mashin
 Min.   : 12.0   Min.   :1.000   Min.   :0.0000
 1st Qu.: 80.0   1st Qu.:2.000   1st Qu.:0.0000
 Median :100.0   Median :3.000   Median :0.0000
 Mean   :107.1   Mean   :2.556   Mean   :0.4493
 3rd Qu.:125.0   3rd Qu.:3.000   3rd Qu.:1.0000
 Max.   :350.0   Max.   :3.000   Max.   :1.0000
     motor          docharkhe          radio
 Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
 Median :0.0000   Median :0.0000   Median :0.00000
 Mean   :0.1007   Mean   :0.1435   Mean   :0.09495
 3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
   radiozabt            tv               tvrangi
 Min.   :0.00000   Min.   :0.000000   Min.   :0.0000
 1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:1.0000
 Median :0.00000   Median :0.000000   Median :1.0000
 Mean   :0.03669   Mean   :0.004506   Mean   :0.9781
 3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1.0000
 Max.   :1.00000   Max.   :1.000000   Max.   :1.0000
     video            computer          mobile
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000
 Median :0.0000   Median :0.0000   Median :1.0000
 Mean   :0.2111   Mean   :0.2414   Mean   :0.9652
 3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
    freezer          yakhchal      yakhchalfreezer
 Min.   :0.0000   Min.   :0.000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000
 Median :0.0000   Median :0.000   Median :1.0000
 Mean   :0.2066   Mean   :0.383   Mean   :0.6292
 3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.000   Max.   :1.0000
    ojaghgaz       jarobarghi       lebasshoyi
 Min.   :0.00   Min.   :0.0000   Min.   :0.0000
 1st Qu.:1.00   1st Qu.:1.0000   1st Qu.:1.0000
 Median :1.00   Median :1.0000   Median :1.0000
 Mean   :0.99   Mean   :0.8999   Mean   :0.7734
 3rd Qu.:1.00   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :1.00   Max.   :1.0000   Max.   :1.0000
    khayati            panke         coolerabimoteharek
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :0.2716   Mean   :0.2095   Mean   :0.1059
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
 coolergazimoteharek   zarfshoyi
 Min.   :0.00000      Min.   :0.00000
 1st Qu.:0.00000      1st Qu.:0.00000
 Median :0.00000      Median :0.00000
 Mean   :0.01867      Mean   :0.02961
 3rd Qu.:0.00000      3rd Qu.:0.00000
 Max.   :1.00000      Max.   :1.00000
```

```
    microfer         lolekeshiab      gazlolekeshi
 Min.   :0.00000   Min.   :0.000   Min.   :0.0000
 1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:1.0000
 Median :0.00000   Median :1.000   Median :1.0000
 Mean   :0.04731   Mean   :0.999   Mean   :0.9865
 3rd Qu.:0.00000   3rd Qu.:1.000   3rd Qu.:1.0000
 Max.   :1.00000   Max.   :1.000   Max.   :1.0000
   telephone         internet          hamam
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:1.0000
 Median :0.0000   Median :1.0000   Median :1.0000
 Mean   :0.3949   Mean   :0.7512   Mean   :0.9958
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
  ashpazkhane     coolerabisabet   borodatmarkazi
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000000
 1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000000
 Median :1.0000   Median :1.0000   Median :0.0000000
 Mean   :0.9907   Mean   :0.5401   Mean   :0.0009656
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000000
 hararatmarkazi       package       coolergazisabet
 Min.   :0.000000   Min.   :0.00000   Min.   :0.0000
 1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.0000
 Median :0.000000   Median :0.00000   Median :0.0000
 Mean   :0.001931   Mean   :0.03991   Mean   :0.1381
 3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.0000
 Max.   :1.000000   Max.   :1.00000   Max.   :1.0000
 fazelabshahri    nsokhtpokhtpaz   nsokhtgarma
 Min.   :0.0000   Min.   :3.000   Min.   :11.00
 1st Qu.:0.0000   1st Qu.:4.000   1st Qu.:14.00
 Median :1.0000   Median :4.000   Median :14.00
 Mean   :0.5452   Mean   :3.985   Mean   :13.96
 3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:14.00
 Max.   :1.0000   Max.   :4.000   Max.   :14.00
  nsokhtabgarm   HazineKhoraki       HazineNoshidani
 Min.   :21.00   Min.   :   600000   Min.   :      0
 1st Qu.:24.00   1st Qu.:  8426250   1st Qu.:      0
 Median :24.00   Median : 12199000   Median :      0
 Mean   :23.98   Mean   : 13623076   Mean   : 319900
 3rd Qu.:24.00   3rd Qu.: 16815000   3rd Qu.:      0
 Max.   :25.00   Max.   :112235000   Max.   :9000000
    HazinePoshak       HazineMaskan
 Min.   :       0   Min.   :   400000
 1st Qu.:       0   1st Qu.:  5900000
 Median :  380000   Median :  8220000
 Mean   : 1990246   Mean   :  9248438
 3rd Qu.: 2480000   3rd Qu.: 11500000
 Max.   :68150000   Max.   :120000000
 HazineLavazemKhanegi HazineDarmani
 Min.   :       0     Min.   :       0
 1st Qu.:  700000     1st Qu.:  200000
 Median : 1090000     Median :  520000
 Mean   : 1317564     Mean   : 2162923
 3rd Qu.: 1600000     3rd Qu.: 1600000
 Max.   :26800000     Max.   :500300000
 HazineHamlonaghl   HazineErtebatat
 Min.   :       0   Min.   :       0
 1st Qu.:  450000   1st Qu.:  400000
 Median : 1000000   Median :  670000
 Mean   : 1502440   Mean   :  803620
 3rd Qu.: 2000000   3rd Qu.: 1000000
 Max.   :47800000   Max.   :20100000
 HazineTafrihatFarhangi HazineGhazaAmade
 Min.   :         0     Min.   :        0
 1st Qu.:         0     1st Qu.:        0
 Median :         0     Median :        0
 Mean   :    202387     Mean   :   224337
 3rd Qu.:         0     3rd Qu.:   200000
 Max.   :180250000     Max.   :100000000
 HazineKalaMotefaregheh HazineKalaBadavam
 Min.   :       0       Min.   :0.000e+00
 1st Qu.:  450000       1st Qu.:1.126e+07
 Median :  850000       Median :5.183e+07
 Mean   : 1111492       Mean   :1.075e+08
 3rd Qu.: 1400000       3rd Qu.:1.350e+08
 Max.   :16790000       Max.   :2.187e+09
     Target
 Min.   :       0
 1st Qu.: 17553333
 Median : 28956667
 Mean   : 37834783
 3rd Qu.: 46404167
 Max.   :654833333
```

شکل ۱:خلاصه‌ای از وضعیت هر متغیرمجموعه داده‌ها

در نهایت این مجموعه داده را با نسبت ۱۰/۳۰/۶۰ به ترتیب مجموعه آموزشی، مجموعه اعتبارسنجی و مجموعه آزمون تقسیم می‌کنیم.

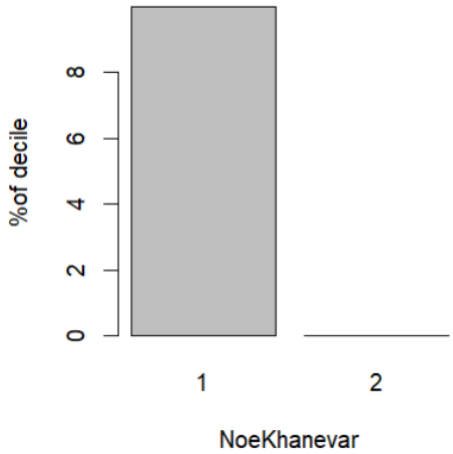
فصل دوم: تصویری سازی

در این فصل، مجموعه‌ای از نمودارها را که می‌توانند برای اکتشاف ماهیت چندمتغیره یک مجموعه داده، مورد استفاده قرار گیرند را تشریح می‌کنیم و بر استفاده از ارائه‌های گرافیکی با هدف اکتشاف داده‌ها، به ویژه در رابطه با تحلیل خودکار پیشگویانه، تمرکز می‌کنیم.[۱] جمله معروف(یک عکس *بیش از هزار کلمه می ارزد*)، به توانایی فشرده‌سازی اطلاعات شفاهی پراکنده در یک تصویر گرافیکی جمع و جور و سریعا فهمیدنی، باز می‌گردد. در مورد اعداد، تصویری‌سازی داده‌ها و خلاصه سازی عددی، ما را به یک ابزار قدرتمند برای اکتشاف داده‌ها و یک راه موثر برای ارائه نتایج مجهز می‌سازد.[۲]

بسته "**ggplot**" که هاردلی ویکمن تهیه کرده‌اند، به رایج‌ترین بسته گرافیکی اختصاصی در **R** برای تصویری‌سازی در گستره وسیعی از زمینه‌ها بدل شده است.[۳]

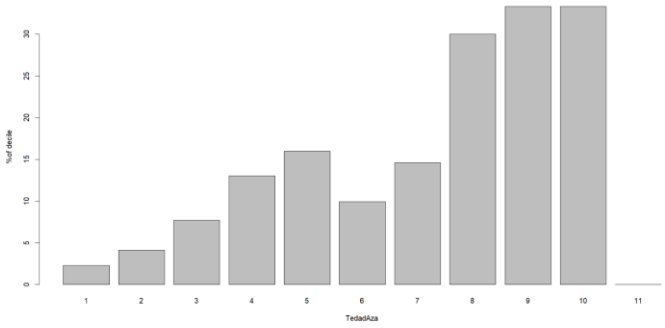
در ابتدا برای درک بهتر، متغیرها را با نمودارهایی نشان می‌دهیم. سه نمودار پایه که موثرترین نمودارها هستند، عبارت‌اند از نمودارهای میله‌ای، نمودارهای خطی و نمودارهای پراکنش.

حال نمودارهایی را براساس متغیر پیشگو و متغیرهای برامد نشان می-دهیم.



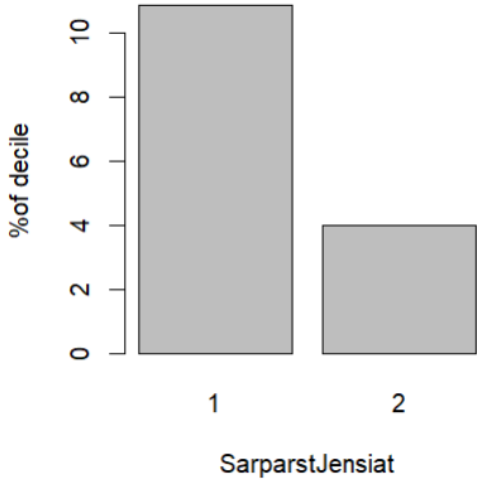
نمودار ۱: نمودار درصد خانوار دهک دهم نسبت به نوع خانوار

در ابتدا درصد خانوار دهک دهم را نسبت به نوع خانوار نشان می‌دهیم. نمودار ۱ نشان دهنده آن است که ۹۹,۹۸ درصد از خانوارهای نوع ۱ که همان معمولی ساکن است، در دهک دهم درامدی قرار دادند. پس یعنی می توان گفت که تقریبا تمام خانوارهای قرار گرفته دهک دهم نمونه، جز این دسته از خانوارها هستند.



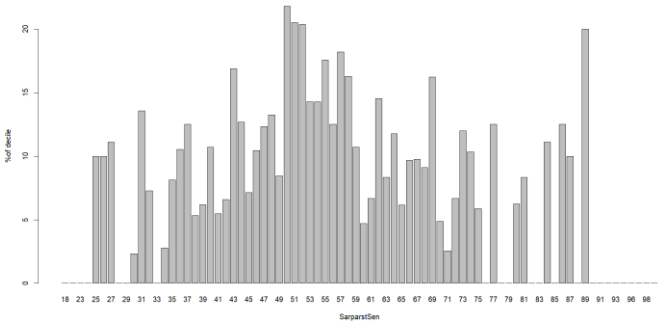
نمودار ۲: نمودار درصد خانوار دهک دهم نسبت به تعداد اعضا خانوار

در نمودار ۲، به راحتی میتوان مشاهده نمود که خانوارهای دارای ۸، ۹و ۱۰ عضو با درصد بیشتری در دهک دهم این نمونه را قرار گرفته‌اند.



نمودار ۳:درصد خانوارهای قرار گرفته در دهک دهم نسبت به جنسیت سرپرست.

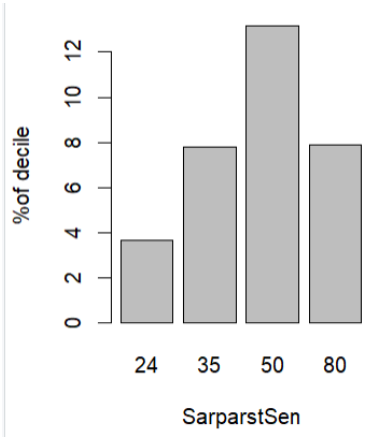
همانطور که در نمودار ۳ مشاهده می‌کنید عمده سرپرستان خانوار قرار گرفته در دهک دهم را مردان تشکیل می‌دهند.



نمودار ۴:درصد خانوارهای دهک دهم نسبت به سن سرپرستان خانوار

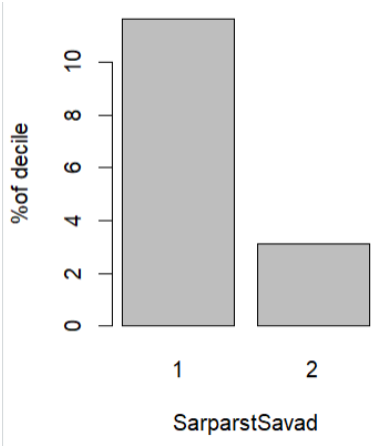
همانطور که در نمودار ۴ مشاهده می‌نمایید،عمده سرپرستان خانوار در رده سنی ۳۴ الی ۶۵ سال می‌باشند. پس میتوان این رده های سنی را در ۴ دسته کلی به صورت زیر تقسیم نمود.(نمودار ۵)

- ۱۸ الی ۳۰ سال به عنوان نوجوان(۲۴)
- ۳۱ الی ۴۰ سال به عنوان جوان (۳۵)
- ۴۱الی ۶۰ سال به عنوان میانسال (۵۰)
- ۶۰ الی ۱۰۰ سال به عنوان سالمند(۸۰)



نمودار ۵: نمودار درصدخانوار دهک دهم نسبت به رده بندی سنی

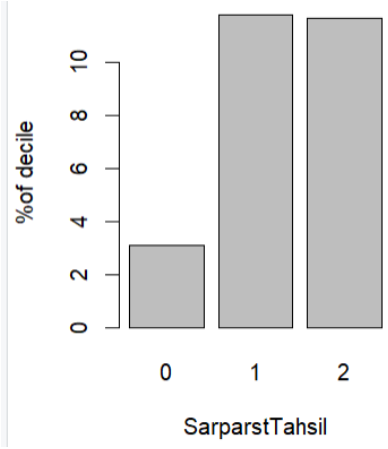
در نمودار ۶ ،مشهود است که درصد افراد باسواد قرارگرفته در دهک دهم به مراتب بالاتر از درصد افرادی بی‌سوادی است که در این دهک قرارگرفته اند. مطابق نمودار ۶ باسواد بودن سرپرستان در قرارگیری در دهک دهم متاثر است.



نمودار۶ : باسواد یا بی‌سواد بودن سرپرستان خانوار در دهک دهم

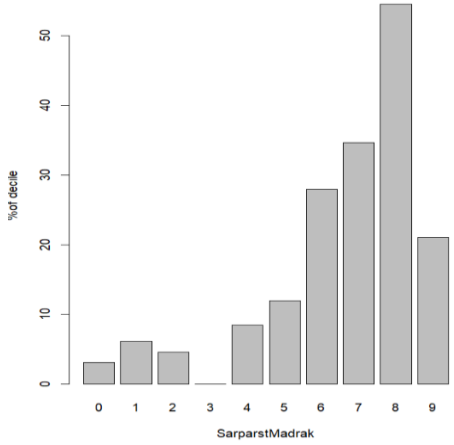
همچنین میتوان درصد سرپرستان در حال تحصیل و فارغ التحصیل را نسبت به دهک درامدی مشاهده نمود. لازم به ذکر است مقادیر صفر نشان دهنده افراد بی‌سواد است.[1]

از آنجایی که درصد سرپرست خانوارهای در حال تحصیل و فارغ-التحصیل قرار گرفته در دهک دهم، تفاوت معناداری ندارد پس می‌توان گفت که این متغیر تاثیری در ارائه مدل بهتر ندارد.



نمودار ۷: نسبت دهک درامدی دهم به وضعیت تحصیل سرپرست خانوار

در رابطه با سطح تحصیلات سرپرست خانوار می‌توانیم متغیرهای در این خصوص را ادغام کنیم. در واقع همانطور که ذکر شد می‌توانیم اطلاعات لازم را از متغیر مدرک تحصیلی نیز استخراج کنیم .نمودار ۸، نشان دهنده نوع مدرک تحصیلی سرپرستان خانوار می‌باشد.



نمودار ۸: تفکیک مدرک تحصیلی سرپرستان خانوار در دهک دهم نمونه

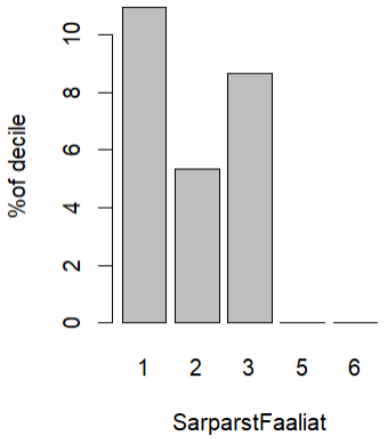
همانطور که در نمودار ۸ مشاهده میکنید، درصد بیشتری از سرپرستان خانوار قرار گرفته در دهک دهم درامدی، مطابق بخش تعریف متغیرها، دارای مدرک تحصیلی‌کارشناسی/لیسانس، کارشناسی‌ارشد و دکتری عمومی/ فوق لیسانس، دکترا تخصصی هستند.پس بنابراین می‌توان نتیجه گرفت که مدرک تحصیلی سرپرستان در درامد آن‌ها تاثیر به سزایی دارد. به عنوان مثال سرپرست خانواری که دارای مدرک تحصیلی دکترا تخصصی می‌باشد حتما در دهک دهم قرار می‌گیرد.

در بررسی متغیر **SarpartFaaliat** در میابیم که سرپرستان شاغل درصد بیشتری از دهک درامدی دهم را شامل می‌شوند. همچنین تعداد افراد دارای درامد بدون کار نیز تعداد قابل توجه‌ای دارند. البته لازم به

---

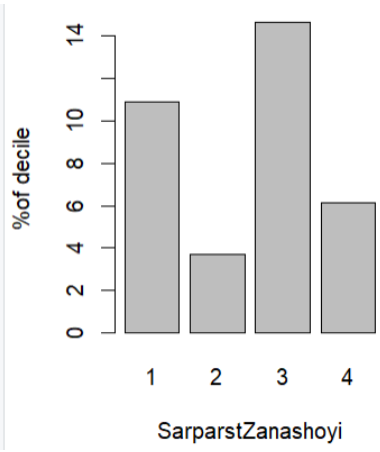
۱ - به بخش معرفی متغیرها در فصل اول مراجعه شود.

ذکر است با توجه به درصد افراد جویای کار قرارگرفته در این دهک می‌بایست در خصوص این افراد بررسی های دقیق تری صورت پذیرد.



نمودار ۹:نسبت خانوار قرارگرفته در دهک دهم به وضعیت شغلی سرپرستان خانوار

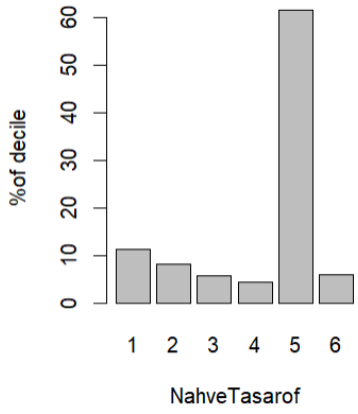
در بررسی متغیر **SarpartZanashoyi** در میابیم که وضعیت زناشویی افراد قرارگرفته در دهک دهم در وضعیت طلاق بسیار بیشتر از سایر وضعیت‌هاست، و این نشان دهنده آن است که به دلیل طلاق و مشکلات ناشی از آن به طرز محسوسی درامد خانوار افزایش یافته است.



نمودار ۱۰:نمودار نسبت خانواردهک دهم درامدی به وضعیت زناشویی‌سرپرست خانوار

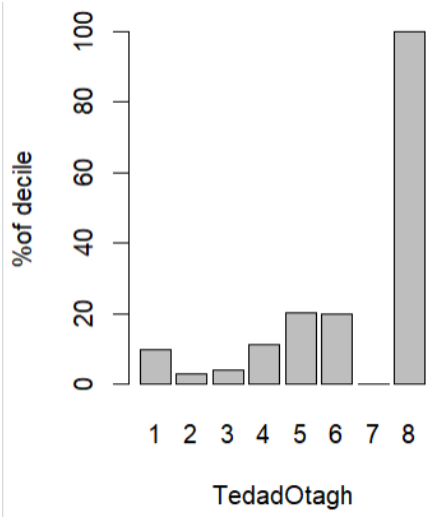
در ادامه به وضعیت محل سکونت خانوار و ارتباط آن با درامد افراد می‌پردازیم.

همانطور که در نمودار ۱۱ به وضوح دیده می‌شود، اکثر خانوارهایی که نحوه تصرف محل سکونتشان از نوع در برابر خدمت است. یعنی خانوار به ازای کارکرد در ارگان‌ها یا سازمان‌های دولتی و غیردولتی مشمول سکونت در خانه‌های به اصطلاح سازمانی گردیده است. درصد قرارگیری آن‌ها در دهک مذکور به مراتب بیشتر از سایر وضعیت‌ها است. به عبارتی ۶۰ درصد خانوار هایی که نحوه تصرف محل سکونتشان از نوع مذکور است در دهک دهم قرار گرفته اند.



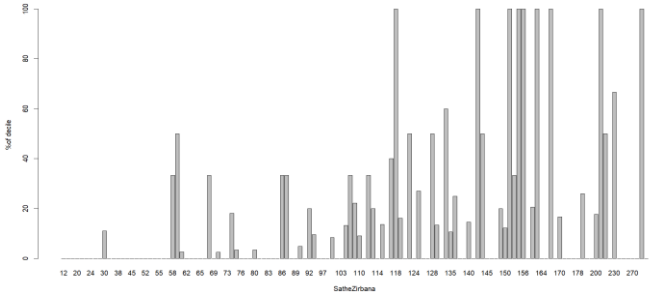
نمودار ۱۱: نمودار نسبت دهک دهم خانوار به وضعیت محل سکونت خانوار

از نمودار متغیر **TedadOtagh** در میابیم که اگر خانواری ۸ اتاق داشته-باشد، ۱۰۰ درصد در دهک دهم قرار میگیرد.البته لازم به ذکر است با توجه به تعداد اتاق سایر خانوارهای قرار گرفته در دهک دهم بدهی است که خانوارهای با ۵ اتاق و بالاتر با درصد بیشتری در دهک دهم قرار می‌گیرند.



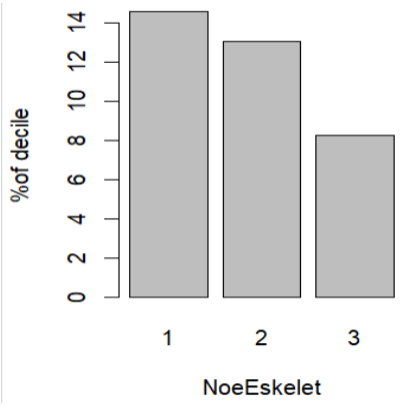
نمودار ۱۲: درصد قرارگیری خانوارها در دهک دهم نسبت به تعداد اتاق محل سکونت

در بررسی متغیر **SatheZirbana** درمی‌یابیم که همانند متغیر تعداد اتاق در اختیار خانوار، سطح زیر بنا نیز دقیقا چنین رفتاری دارد.یعنی در نمودار ۱۳ خانوارهایی که در خانه‌هایی با متراژ بالای ۱۱۰ متر مربع سکونت می‌کنند با درصد بیشتری در دهک درامدی‌دهم قرار گرفته‌اند.



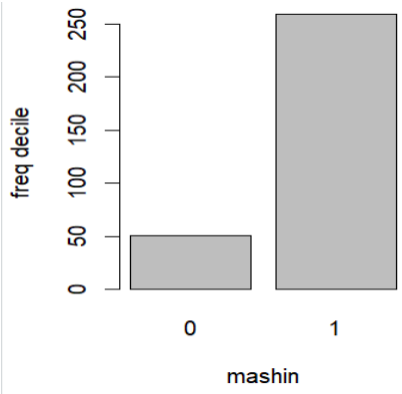
نمودار ۱۳:نمودار درصد خانوارهای قرارگرفته در دهک دهم نسبت به متراژ محل سکونت

در نمودار ۱۴، درصد بیشتر نوع اسکلت ساختمان خانوارهای قرارگرفته در دهک دهم، از نوع فلزی است.

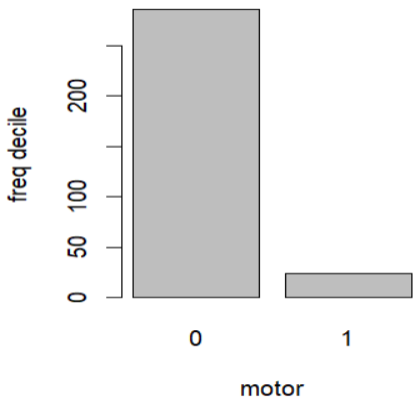


نمودار ۱۴درصد:نوع اسکلت خانوارهای قرارگرفته در دهک دهم درامدی

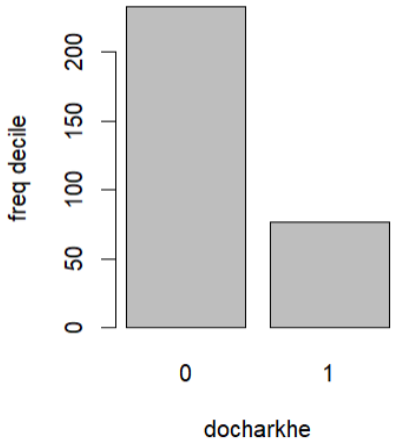
در نمودارهای زیر نیز تعداد خانوارهای دارای وسایل حمل و نقل قرارگرفته در دهک دهم را مشاهده می‌نمایید. لذا از توضیح این نمودار-ها به دلیل سادگی صرف نظر می‌کنیم.



نمودار ۱۵: تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن خودرو

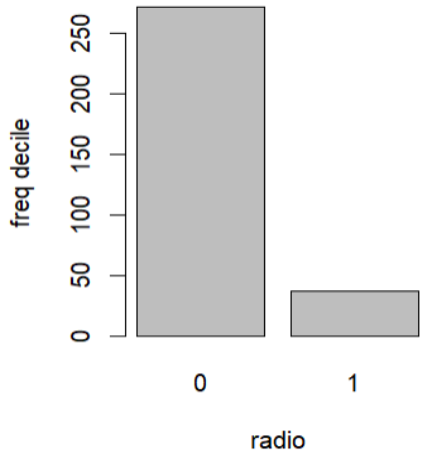


نمودار ۱۶: تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن موتورسیکلت

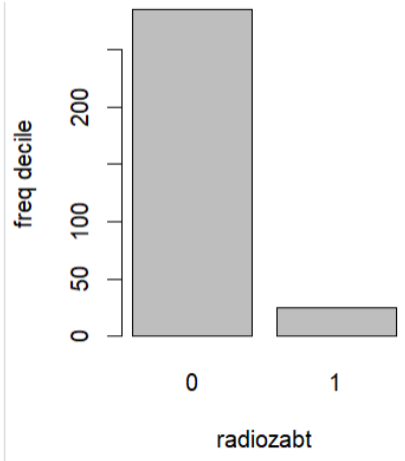


نمودار ۱۷: تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن دوچرخه

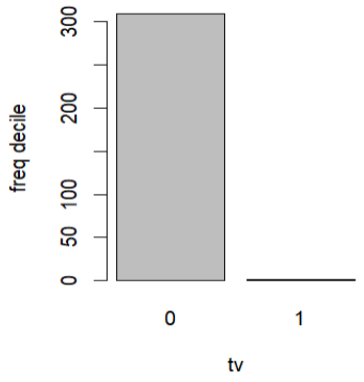
در نمودارهای زیر نیز دارابودن یا نبودن خانوارهای قرار گرفته در دهک دهم درامدی از لحاظ لوازم خانگی مورد نیاز زندگی بررسی می‌شود.



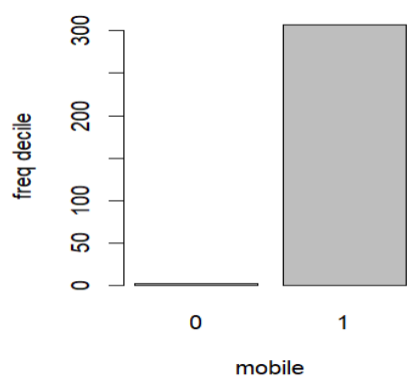
نمودار ۱۸: تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن رادیو



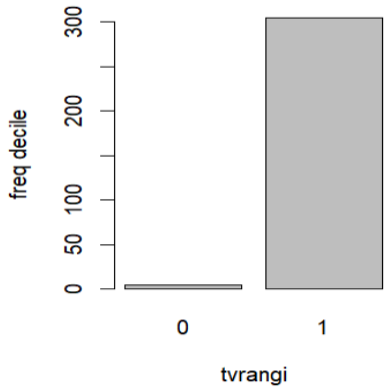
نمودار ۱۹: تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن رادیوضبط



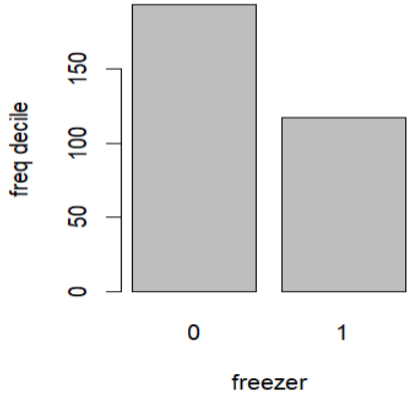
نمودار ۲۰: تعداد خانوارهای قرارگرفته در دهک دهم از لحاظ دارا بودن تلویزیون سیاه سفید



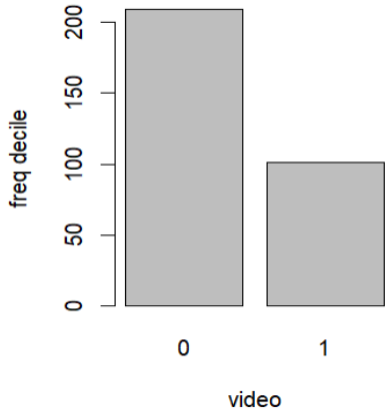
نمودار ۲۴: تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن گوشی همراه



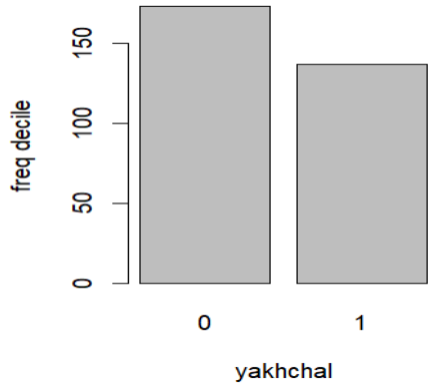
نمودار ۲۱: تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن تلویزیون رنگی



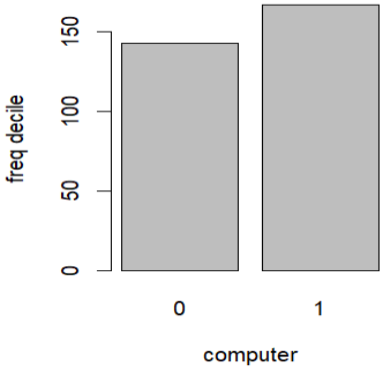
نمودار ۲۵: تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن فریزر



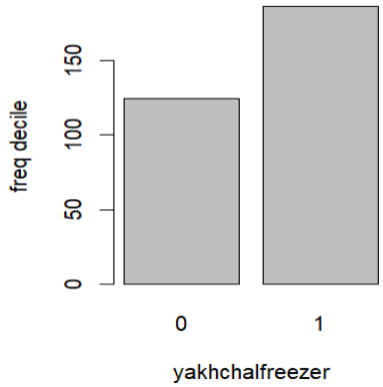
نمودار ۲۲: تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن دستگاه ویدیو



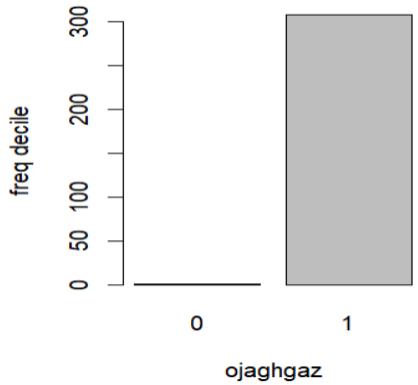
نمودار ۲۶: تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن یخچال



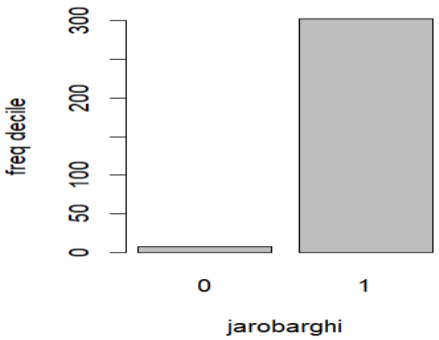
نمودار ۲۳: تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن کامپیوتر و تبلت



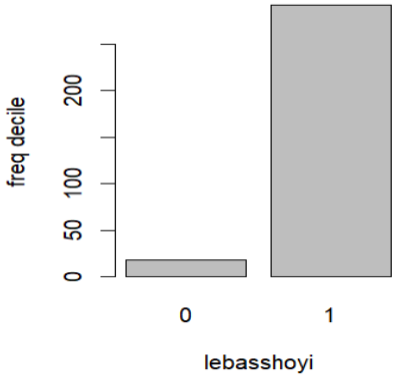
نمودار ۲۷: تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن یخچال فریزر



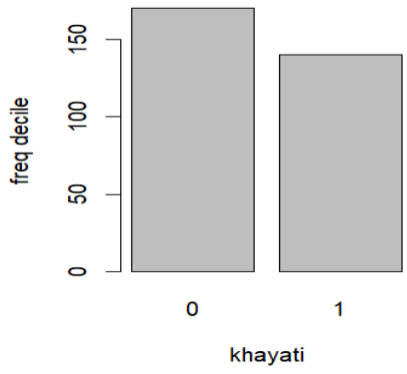
نمودار ۲۸:تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن اجاق گاز



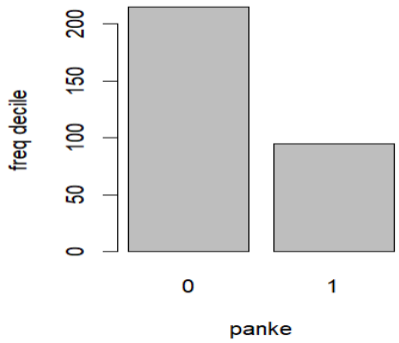
نمودار ۲۹:تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن جاروبرقی



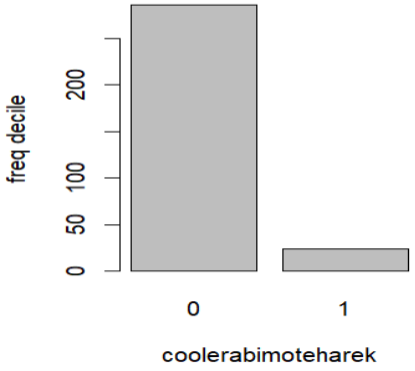
نمودار ۳۰:تعداد خانوارهای قرار گرفته در دهک دهم از لحاظ دارا بودن ماشین لباسشویی



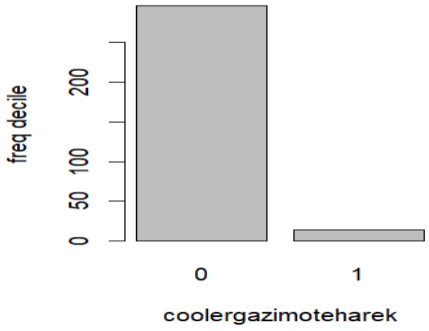
نمودار ۳۱:تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن چرخ خیاطی



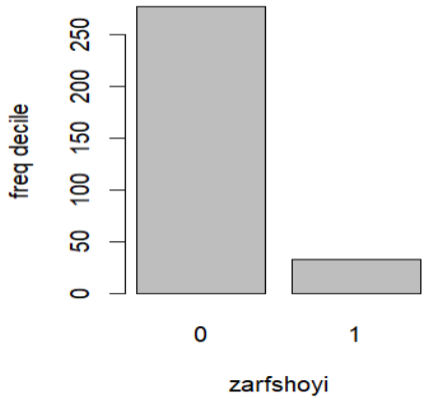
نمودار ۳۲:تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن پنکه



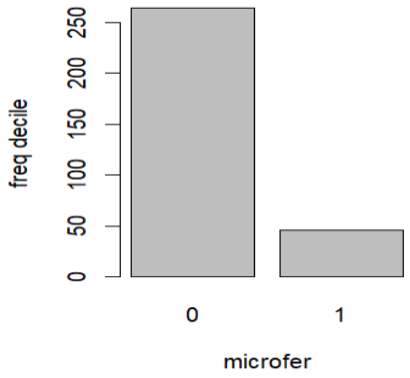
نمودار ۳۳:تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن کولر ابی متحرک



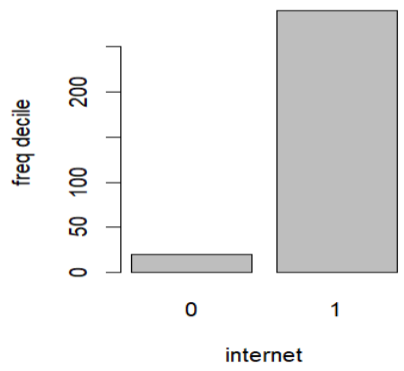
نمودار ۳۴:تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن کولرگازی متحرک



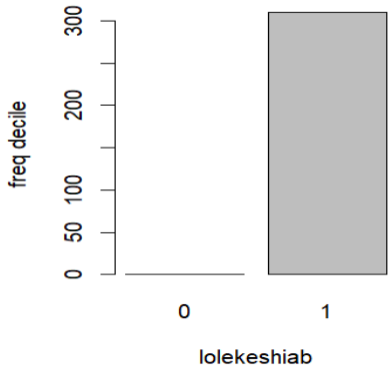
نمودار ۳۵:تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن ماشین ظرفشویی



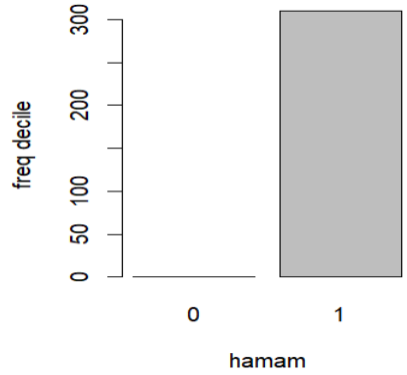
نمودار ۳۶:تعداد خانوار های قرار گرفته در دهک دهم از لحاظ دارا بودن مایکروفر



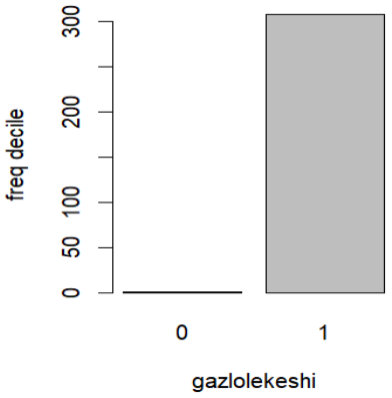
نمودار ۴۰:تعداد خانوارهای دارای دسترسی به شبکه اینترنت قرارگرفته در دهک دهم



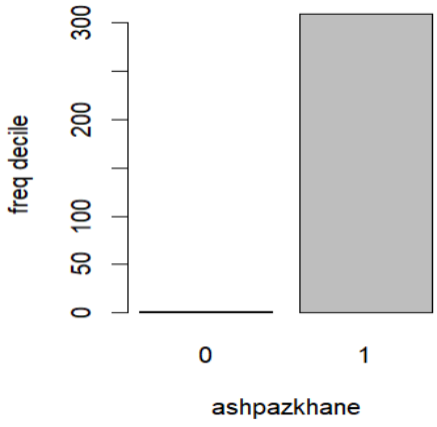
نمودار ۳۷:تعداد خانوارهای قرارگرفته در دهک دهم از لحاظ دارا بودن انشعاب اب لوله‌کشی



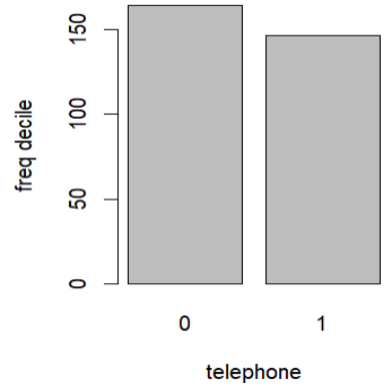
نمودار ۴۱:تعداد خانوار قرارگرفته در دهک دهم از لحاظ دارا بودن حمام



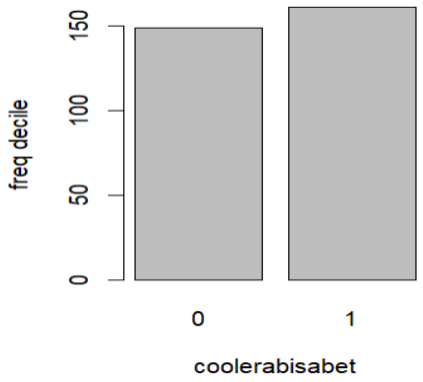
نمودار ۳۸:تعداد خانوارهای قرارگرفته در دهک دهم از لحاظ دارا بودن انشعاب گاز لوله‌کشی



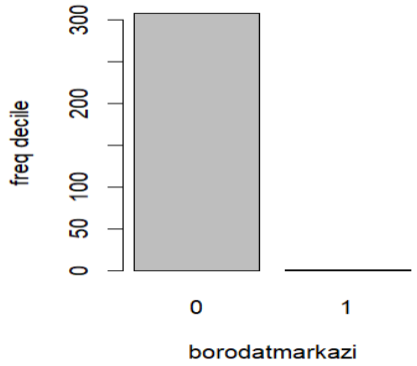
نمودار ۴۲:تعداد خانوار قرارگرفته در دهک دهم از لحاظ دارا بودن اشپزخانه



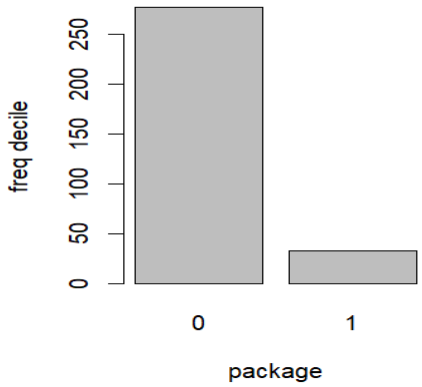
نمودار ۳۹:تعداد خانوارهای قرارگرفته در دهک دهم از لحاظ دارا بودن خط تلفن



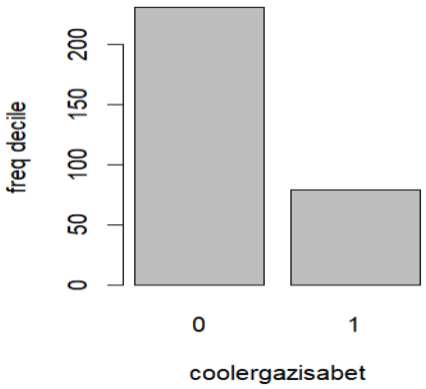
نمودار ۴۳:تعداد خانوار قرار گرفته در دهک دهم از لحاظ دارابودن کولر ابی ثابت



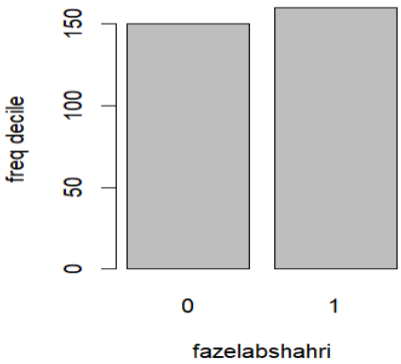
نمودار ۴۴:تعداد خانوار قرار گرفته در دهک دهم از لحاظ دارا بودن برودت مرکزی



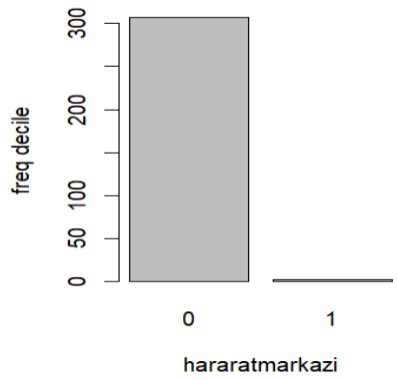
نمودار ۴۵:تعداد خانوار قرار گرفته در دهک دهم از لحاظ دارا بودن پکیج



نمودار ۴۶:تعداد خانوار قرارگرفته در دهک دهم از لحاط دارا بودن کولر گازی ثابت

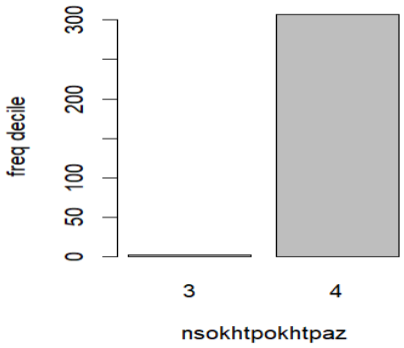


نمودار ۴۷: تعداد خانوار قرارگرفته در دهک دهم از لحاط دارا بودن انشعاب فاضلاب شهری

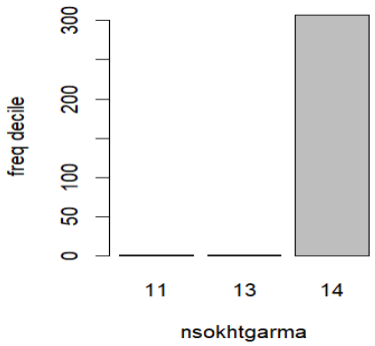


نمودار ۴۸:تعداد خانوار قرارگرفته در دهک دهم از لحاظ دارا بودن سیستم حرارت مرکزی

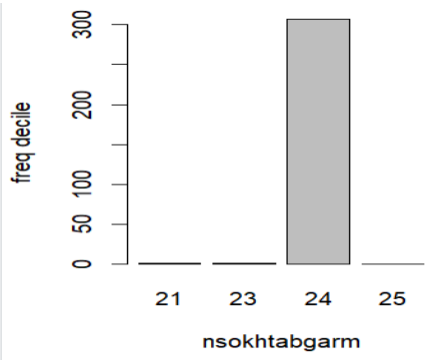
درادامه به بررسی نوع سوخت مورد نیاز برای پخت وپز و ایجاد گرما برای مصارف گوناگون می‌پردازیم.



نمودار ۴۹:نوع سوخت خانوار های قرار گرفته در دهک دهم برای پخت و پز



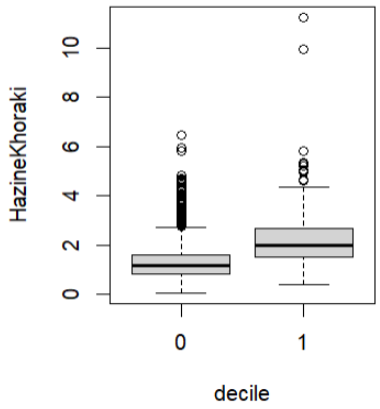
نمودار ۵۰:نوع سوخت خانوار قرار گرفته در دهک دهم برای گرمایش



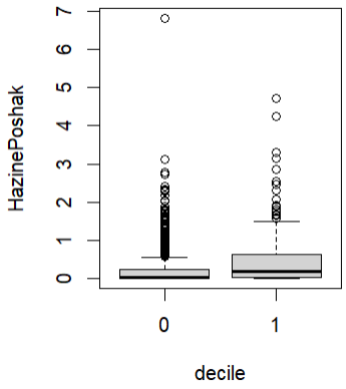
نمودار ۵۱:نوع سوخت مورد استفاده خانوارهای قرار گرفته در دهک دهم برای تهیه ابگرم

در ادامه به بررسی متغیرهای هزینه خانوار در زمینه‌های مختلف را بررسی می‌کنیم.

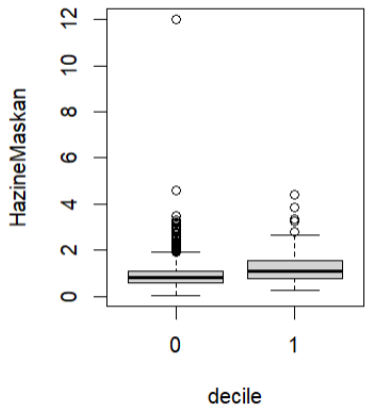
در نمودار ۵۲ ، با استفاده از نمودار پهلو به پهلو میزان هزینه‌ها در قسمت خوراکی خانوار را مشاهده می نمایید.



نمودار ۵۲:هزینه خوراکی خانوار قرار گرفته در دهک دهم



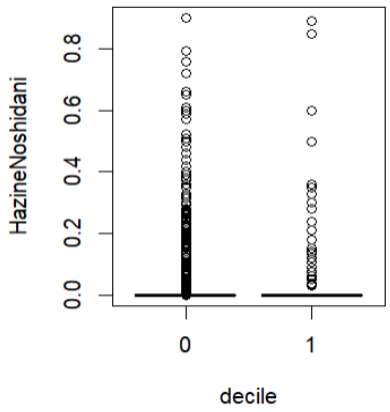
نمودار ۵۴: هزینه پوشاک در خانوارهای قرار گرفته در دهک دهم



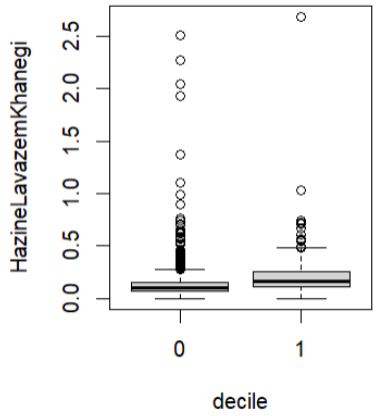
نمودار ۵۵:هزینه مربوط به خانه در خانوارهای قرار گرفته ر دهک دهم

همانطور که مشهود است در صدک ۵۰ خانوارهای قرار گرفته در دهک دهم میزان هزینه بیش از ۲ میلیون تومان است.

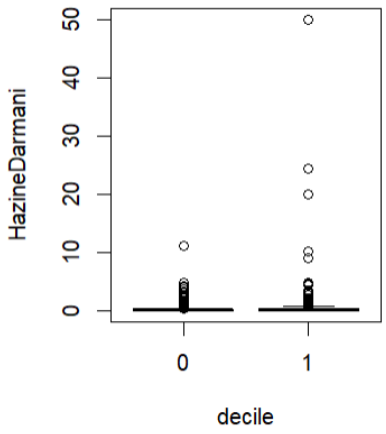
در نمودار ۵۳ نیز می توان به همین نحو نتیجه‌گیری کرد.



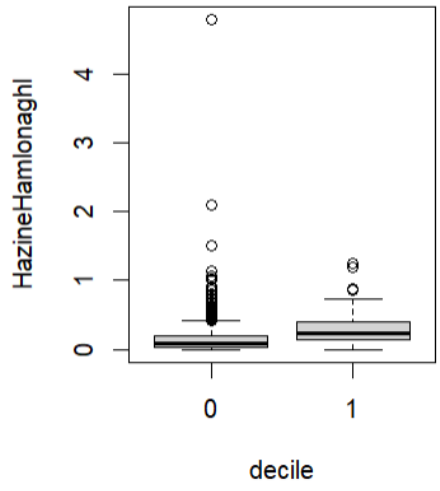
نمودار ۵۳: هزینه نوشیدنی خانوار قرار گرفته در دهک دهم



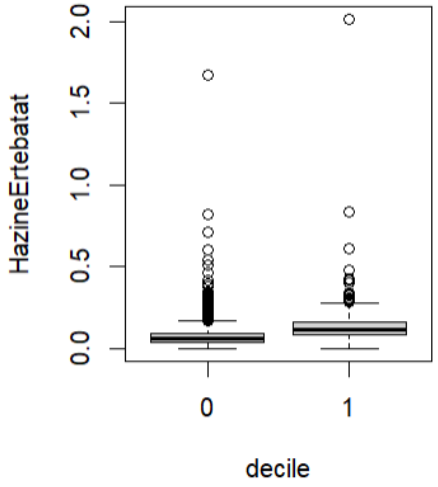
نمودار ۵۶: هزینه لوازم خانگی خانوار قرار گرفته در دهک دهم



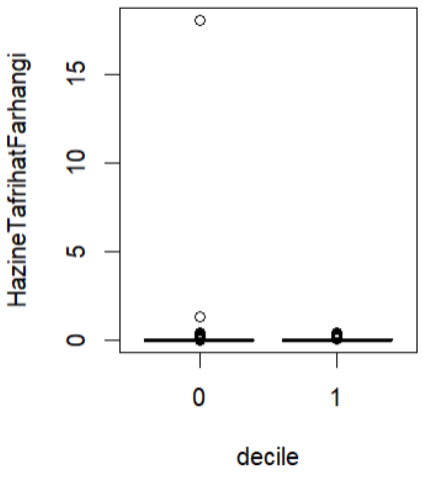
نمودار ۵۷: هزینه درمانی در خانوارهای قرار گرفته در دهک دهم



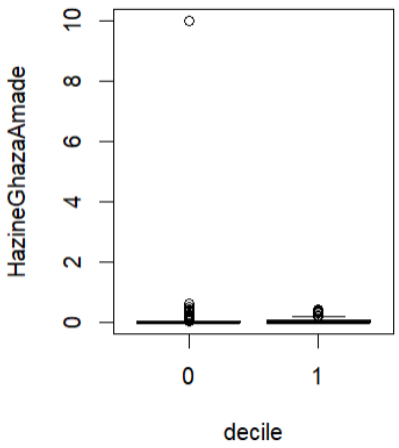
نمودار ۵۸:هزینه حمل و نقل خانوارهای قرار گرفته در دهک دهم



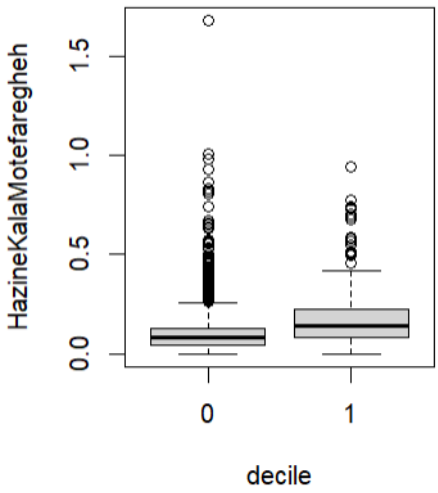
نمودار ۵۹: هزینه ارتباطات خانوارهای قرار گرفته در دهک دهم



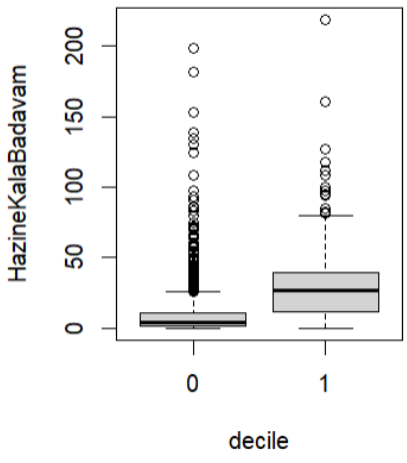
نمودار ۶۰:هزینه تفریحات فرهنگی خانوارهای قرار گرفته در دهک دهم



نمودار ۶۱:هزینه غذا اماده و رستوران خانوارهای قرار گرفته در دهک دهم



نمودار ۶۲:هزینه کالای متفرقه خانوارهای قرار گرفته در دهک دهم



نمودار ۶۳: هزینه کالای بادوام خانوارهای قرار گرفته در دهک دهم در یکسال

توضیحات نمودارها به صورت کلی

در نمودارهای ۱۵ ، ۱۶ و ۱۷ میزان استفاده و علاقه مندی دهک دهم درامدی نسبت به وسایل نقلیه را مشاهده می‌کنید.همانطور که مشهود است اکثر خانوار های قرار گرفته در دهک دهم دارای خودرو می‌باشند.

و از آنجایی که درصد خانوارهای دارای موتورسیکلت و دوچرخه در این دهک بسیار کم است، پس این دو متغیر را حذف میکنیم.

در نمودارهای ۱۸ ، ۱۹ و ۲۰ نیز همانطور که مشاهده می‌کنید، متغیر-های رادیو،رادیو ضبط و تلویزیون سیاه سفید نیز معیار مناسبی جهت تشخیص دهک دهم درامدی نمیباشد. به دلیل انکه اکثریت خانوارهای قرار گرفته در این دهک این وسایل را ندارند. پس این متغیر ها را نیز می‌بایست حذف کنیم.

در نمودارهای ۲۲ ، ۲۳ نیز با توجه به اینکه تفاوت معناداری بین داشتن یا نداشتن دستگاه ویدیو و DVD ، کامپیوتر و تبلت در این دهک نیست .پس بنابراین این متغیرها ها را نمیتوانیم حدف کنیم.

در نمودارهای ۲۱، ۲۴، ۲۸، ۲۹ ، ۳۰، ۳۷، ۳۸، ۴۰، ۴۱، ۴۲ با توجه به اینکه ۹۹/۹۵ درصد از خانوارهای قرار گرفته در دهک دهم از آن برخوردارند، پس این متغیرها میتواند ما را در تشخیص واقعی این دهک در امدی راهنمایی کند

در نمودارهای ۲۵ ، ۲۶، ۲۷ نیز تفاوتی انچنان بین داشتن و نداشتن وسیله مذکور وجود ندارد پس میباست حفظ شوند.

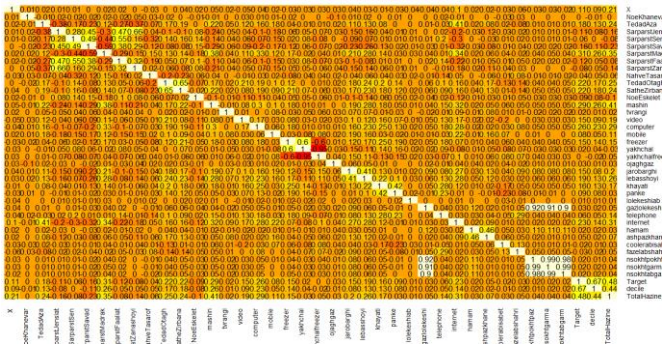
در نمودارهای ۳۳ ، ۳۴ ، ۳۵ ، ۳۶، ۴۴ ، ۴۵ ، ۴۶ ، ۴۸ همانطور که مشهود است، این امکانات و تجهیزات در خانوارهای دهک دهم نیز به ندرت دیده می‌شود، و از انجایی که هدف پروژه تشخیص دهک مذکور است، پس این متغیر ها هیچ کمکی به ارائه مدل بهتر نمیکند. پس می‌بایست حذف گردند.

در بررسی نوع سوخت خانوارها در میابیم که ۹۹/۹ درصد از خانوارهای قرار گرفته در این دهک درامدی از شبکه گاز طبیعی به عنوان منبع گرما در زندگی خود استفاده می کنند.. پس می‌توانیم ان متغیرها را به دو گروه کلی شبکه گاز طبیعی و سایر تقسیم بندی کنیم.

همچنین در نمودارهای ۵۳، ۵۴، ۵۵، ۵۶، ۵۷، ۵۸، ۵۹، ۶۰، ۶۱، ۶۲، ۶۳ که مرتبط با هزینه‌های خانوار در ماه گذشته می‌باشد. لذا می-توانیم،تمام هزینه‌های مربوطه را جمع کرده و در متغیر جدید طبق معادله ۲ جایگزین کنیم.

$$\text{TotalHazine} = \frac{1}{12}\text{HazineKalaBadavam} + \text{HazineKalaMotefaregheh} + \text{HazineGhazaAmade} + \text{HazineTafrihatFarhangi} + \text{HazineErtebatat} + \text{HazineHamlonaghl} + \text{HazineDarmani} + \text{HazineLavazemKhanegi} + \text{HazineMaskan} + \text{HazinePoshak} + \text{HazineNoshidani} + \text{HazineKhoraki}$$

حال نمودار حرارتی این متغیر ها را مشاهده میکنیم.

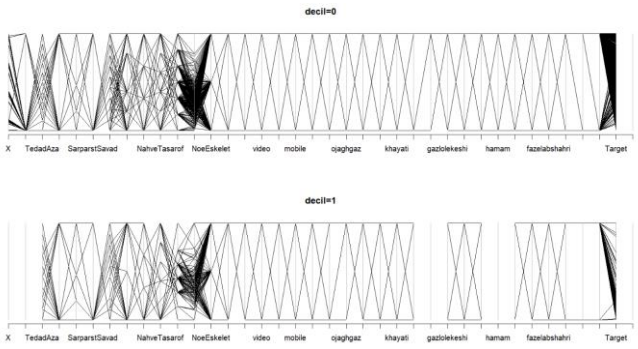


نمودار ۶۴: نمودار حرارتی همبستگی بین متغیرها مجموعه داده

همانطور که در نمودار ۶۴ مشهود است متغیرهای **freezer**،**yakhchal** و **yakhchalfreezer** همبستگی زیادی مشاهده می‌شود و همچنین همبستگی بسیار کمی بین **sarparastsen** و **sarparstfaaliat** مشاهده می‌شود.

پس می‌توانیم یکی از متغیرهای یخچال یا یخچال‌فریزر را حذف کنیم.

در نهایت ۳۱۰۷ ثبت و ۳۹ متغیر داریم.



نمودار ۶۵:نمودار محور های موازی متغیرها

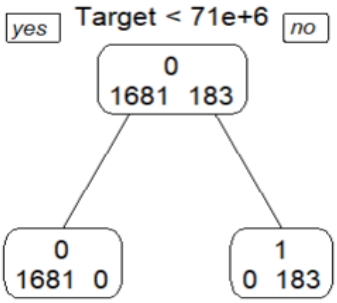
فصل سوم : درخت

درخت رده‌بندی

این فصل یک روش انعطاف پذیر داده رهنمون را تشریح می‌کند که می‌توان برای رده بندی و پیشگویی از آن استفاده کرد.در میان روش-های داده رهنمون،درختان، شفاف ترین و ساده ترین روش تفسیرند، که بر جداسازی ثبت‌ها به زیرگروه‌ها، از طریق ایجاد بخش‌هایی در پیشگویی‌ها مبتنی‌اند.

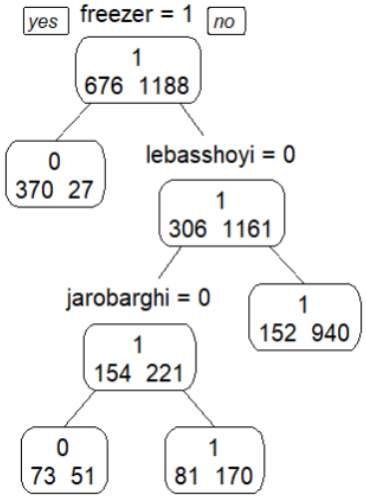
اگر یک فن رده بندی وجود داشته باشدکه بدون نیاز به تلاش چندان تحلیلگر،در طیف وسیعی از موقعیت‌ها خوب عمل کند و به راحتی توسط مصرف‌کننده تحلیل درک‌پذیر باشد، همانا مدعی قوی روانشناسی درخت است که بری‌من و همکاران توسعه‌اش داده‌اند.

در ابتدا با اجرا درخت به روی مجموعه آموزشی برای رده بندی دهک درامدی داریم:



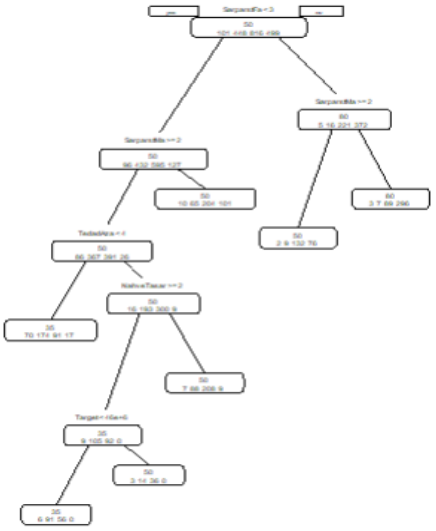
نمودار ۶۶:درخت رده‌بندی بر اساس دهک درامدی دهم

سپس درخت رده‌بندی زیر را به صورت دقیق تر داریم.



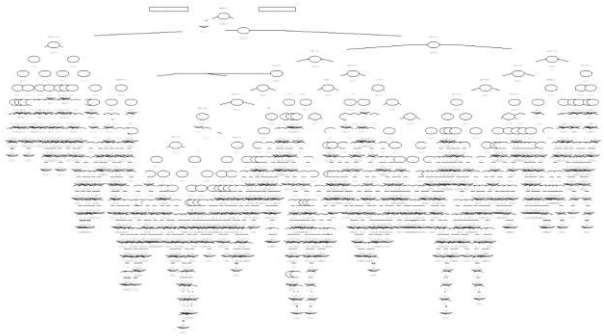
نمودار ۶۷:درخت رده‌بندی بر اساس متغیر yakhchalfreezer

می‌توانیم درخت رده‌بندی را با توجه به سن سرپرست به صورت زیر مشاهده کنیم.



نمودار ۶۸:درخت رده‌بندی بر اساس سن سرپرست

همچنین با اجرا درخت رده‌بندی عمیق به روی مجموعه آموزشی به شرح زیر داریم.



نمودار ۶۹:درخت رده‌بندی عمیق متغیر سن سرپرست

سپس با اجرا درخت به روی داده‌های اعتبارسنجی و آزمون از صحت عملکرد درخت رده‌بندی مطلع می‌شویم.

همانطور که در خروجی زیر مشاهده میکنید ارزیابی عملکرد به روی مجموعه آموزشی ۱۰۰ درصد است.

```
> confusionMatrix(default.ct.point.pred.train,train.data$decile)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1681    0
         1    0  183

               Accuracy : 1
                 95% CI : (0.998, 1)
    No Information Rate : 0.9018
    P-Value [Acc > NIR] : < 2.2e-16

                  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 : 0.9018
         Detection Rate : 0.9018
   Detection Prevalence : 0.9018
      Balanced Accuracy : 1.0000

       'Positive' Class : 0
```

خروجی ۱: ارزیابی عملکرد درخت با ماتریس درهم‌ریختگی به روی داده آموزشی

همچنین ارزیابی آن به روی داده اعتبار سنجی نیز ۱۰۰ درصد است.

```
> confusionMatrix(default.ct.point.pred.valid,valid.data$decile)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 833   0
         1   0  99

               Accuracy : 1
                 95% CI : (0.996, 1)
    No Information Rate : 0.8938
    P-Value [Acc > NIR] : < 2.2e-16

                  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 : 0.8938
         Detection Rate : 0.8938
   Detection Prevalence : 0.8938
      Balanced Accuracy : 1.0000

       'Positive' Class : 0
```

خروجی ۲: ارزیابی عملکرد درخت به روی داده اعتبارسنجی

همچنین دقیقا همین مقادیر برای درخت عمیق متغیر دهک درامدی ثبت شد.

در ارزیابی عملکرد درخت رده‌بندی متغیر سن سرپرست به روی مجموعه آموزشی به شرح زیر داریم.

```
> confusionMatrix(default.ct.point.pred.train,train.data$SarparstSen)
Confusion Matrix and Statistics

          Reference
Prediction  24  35  50  80
        24   0   0   0   0
        35  76 265 147  17
        50  22 176 580 186
        80   3   7  89 296

Overall Statistics

               Accuracy : 0.6121
                 95% CI : (0.5896, 0.6343)
    No Information Rate : 0.4378
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4049

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 24 Class: 35 Class: 50 Class: 80
Sensitivity            0.00000    0.5915    0.7108    0.5932
Specificity            1.00000    0.8305    0.6336    0.9275
Pos Pred Value             NaN    0.5248    0.6017    0.7494
Neg Pred Value         0.94582    0.8653    0.7378    0.8618
Prevalence             0.05418    0.2403    0.4378    0.2677
Detection Rate         0.00000    0.1422    0.3112    0.1588
Detection Prevalence   0.00000    0.2709    0.5172    0.2119
Balanced Accuracy      0.50000    0.7110    0.6722    0.7603
```

خروجی ۳:ارزیابی عملکرد درخت با ماتریس درهم‌ریختگی به روی مجموعه آموزشی

همانطور که مشاهده میکنید، رده‌بندی مدل به روی داده‌های آموزشی ۶۱ درصد صحت دارد ، از پیشگویی درخت به روی داده اعتبار سنجی خروجی زیر حاصل می‌شود.

```
> confusionMatrix(default.ct.point.pred.valid,valid.data$SarparstSen)
Confusion Matrix and Statistics

          Reference
Prediction  24  35  50  80
        24   0   0   0   0
        35  38 127  82   8
        50   7  88 275  91
        80   0   2  49 165

Overall Statistics

               Accuracy : 0.6084
                 95% CI : (0.5762, 0.6399)
    No Information Rate : 0.4356
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4022

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 24 Class: 35 Class: 50 Class: 80
Sensitivity            0.00000    0.5853    0.6773    0.6250
Specificity            1.00000    0.8210    0.6464    0.9237
Pos Pred Value             NaN    0.4980    0.5965    0.7639
Neg Pred Value         0.95172    0.8671    0.7219    0.8617
Prevalence             0.04828    0.2328    0.4356    0.2833
Detection Rate         0.00000    0.1363    0.2951    0.1770
Detection Prevalence   0.00000    0.2736    0.4946    0.2318
Balanced Accuracy      0.50000    0.7031    0.6619    0.7743
```

خروجی ۴:ارزیابی درخت به روی مجموعه اعتبار سنجی

همچنین با توجه به اجرای درخت عمیق به روی این متغیر داریم.

```
> confusionMatrix(deeper.ct.point.pred.train,train.data$SarparstSen)
Confusion Matrix and Statistics

          Reference
Prediction  24  35  50  80
        24 101   0   0   0
        35   0 448   0   0
        50   0   0 816   0
        80   0   0   0 499

Overall Statistics

               Accuracy : 1
                 95% CI : (0.998, 1)
    No Information Rate : 0.4378
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 24 Class: 35 Class: 50 Class: 80
Sensitivity            1.00000    1.0000    1.0000    1.0000
Specificity            1.00000    1.0000    1.0000    1.0000
Pos Pred Value         1.00000    1.0000    1.0000    1.0000
Neg Pred Value         1.00000    1.0000    1.0000    1.0000
Prevalence             0.05418    0.2403    0.4378    0.2677
Detection Rate         0.05418    0.2403    0.4378    0.2677
Detection Prevalence   0.05418    0.2403    0.4378    0.2677
Balanced Accuracy      1.00000    1.0000    1.0000    1.0000
```

خروجی ۵:ارزیابی عملکرد درخت عمیق به روی مجموعه آموزشی با ماتریس درهم‌ریختگی

همانطور که مشاهده میکنید میزان یادگیری درخت عمیق به روی مجموعه داده آموزشی ۱۰۰ درصد است همچنین میزان دقت آن برای رده‌بندی به روی مجموعه آموزشی ۱ است. ولی با توجه به خروجی ۶ ، این دقت به عدد ۵۳ درصد رسیده است و به طور جدی میزان دقت رده‌بندی به روی مجموعه اعتبار سنجی با مشکل بیش برازش رو به رو شده است.

یک خطر رشد دادن درخت‌های عمیق‌تر روی داده‌های آموزشی، بیش برازش است.بیش برازش به عملکرد ضعیف در داده‌های جدید منجر خواهد شد.[1]

```
> deeper.ct.point.pred.valid<-predict(deeper.ct,valid.data,type="class")
> confusionMatrix(deeper.ct.point.pred.valid,valid.data$SarparstSen)
Confusion Matrix and Statistics

          Reference
Prediction  24  35  50  80
        24   8  30  11   2
        35  28 105  94  11
        50   8  75 212  73
        80   1   7  89 178

Overall Statistics

               Accuracy : 0.5397
                 95% CI : (0.5071, 0.5721)
    No Information Rate : 0.4356
    P-Value [Acc > NIR] : 1.136e-10

                  Kappa : 0.3254

 Mcnemar's Test P-Value : 0.4837

Statistics by Class:

                     Class: 24 Class: 35 Class: 50 Class: 80
Sensitivity          0.177778    0.4839    0.5222    0.6742
Specificity          0.951522    0.8140    0.7034    0.8548
Pos Pred Value       0.156863    0.4412    0.5761    0.6473
Neg Pred Value       0.958002    0.8386    0.6560    0.8691
Prevalence           0.048283    0.2328    0.4356    0.2833
Detection Rate       0.008584    0.1127    0.2275    0.1910
Detection Prevalence 0.054721    0.2554    0.3948    0.2951
Balanced Accuracy    0.564650    0.6489    0.6128    0.7645
```

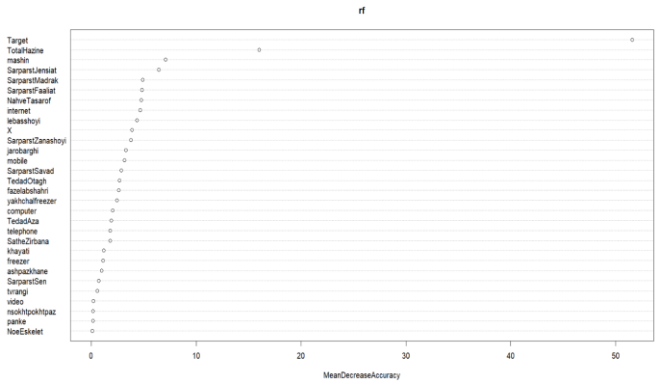
خروجی ۶:ارزیابی عملکرد درخت عمیق به روی مجموعه اعتبار سنجی

پس بنابراین برای دسترسی به مدلی بهتر روش‌های دیگر را بررسی می‌کنیم.

جنگل تصادفی

یکی دیگر از روش‌های رده‌بندی درخت‌های تصادفی است.بر خلاف یک درخت منفرد، نتایج یک جنگل تصادفی را نمی‌توان در یک نمودار درخت گونه نمایش داد، بدین ترتیب قابلیت تفسیری که یک درخت منفرد ارائه می‌کند، از دست می‌رود. در عین حال، جنگل‌های تصادفی می‌توانند امتیازهای اهمیت متغیر را که مشارکت نسبی پیشگوهای مختلف را اندازه می‌گیرد، تولید کنند.

در نمودار زیر جنگل تصادفی اجراشده به روی مجموعه آموزشی را مشاهده می‌نمایید.، همچنین ارزیابی عملکرد این جنگل تصادفی را نیز به شرح خروجی زیر داریم.



نمودار ۷۰: جنگل تصادفی اجرا شده به روی مجموعه آموزشی و اهمیت متغیرها

```
> rf.pred<-predict(rf,valid.data)
> confusionMatrix(rf.pred,valid.data$decile)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 833   4
         1   0  95

               Accuracy : 0.9957
                 95% CI : (0.989, 0.9988)
    No Information Rate : 0.8938
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.977

 Mcnemar's Test P-Value : 0.1336

            Sensitivity : 1.0000
            Specificity : 0.9596
         Pos Pred Value : 0.9952
         Neg Pred Value : 1.0000
             Prevalence : 0.8938
         Detection Rate : 0.8938
   Detection Prevalence : 0.8981
      Balanced Accuracy : 0.9798

       'Positive' Class : 0
```

خروجی ۷:ارزیابی عملکرد جنگل تصادفی به روی مجموعه اعتبار سنجی

همانطور که در نمودار ۷۰ مشهود است متغیرهای **target**، و **totalHazine** خانوار به شدت در دهک بندی تاثیر گذارند، و همچنین عملکرد جنگل تصادفی در دهک‌بندی مجموعه اعتبارسنجی نیز با دقت۰,۹۹۵۷ درصد ثبت شده است.

فصل چهارم:رگرسیون لوژستیک

در این فصل روش رده‌بندی بسیار قدرتمند و رایجی را که رگرسیون لوژستیک نامیده می‌شود را اجرا میکنیم.رگرسیون لوژستیک، ایده رگرسیون خطی را برای حالتی که متغیر برامد، y، رسته‌ای است، توسعه می‌دهد. در خروجی زیر برازش یک مدل رگرسیون لوژستیک را مشاهده می‌کنید.

```
Call:
glm(formula = decile ~ ., family = "binomial", data = train.data)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-0.00051881  -0.00000002  -0.00000002  -0.00000002   0.00064826

Coefficients: (1 not defined because of singularities)
                      Estimate        Std. Error z value
(Intercept)        758.7888247872 884130.2108969315   0.001
X                   -0.0000992591      0.2929280706   0.000
NoeKhanevar        429.9595118529 234256.0990300766   0.002
TedadAza            -2.9082348184   1266.0892132882  -0.002
SarparstJensiat    -11.1710913315  73103.3810471379   0.000
SarparstSen35      -16.9275916730  22872.6623968570  -0.001
SarparstSen50      -34.8997947540  22399.9050100665  -0.002
SarparstSen80      -40.3812627503  22524.6154149806  -0.002
SarparstSavad      -21.9862789346   6834.1358264012  -0.003
SarparstMadrak      -3.3360699503   1017.1716430431  -0.003
SarparstFaaliat     -2.0751053683   1830.2599089943  -0.001
SarparstZanashoyi   19.6671128711  72101.6081296941   0.000
NahveTasarof         1.4216216481   1232.7154499149   0.001
TedadOtagh          -9.6517822722   3224.5497703946  -0.003
SatheZirbana         0.1322979777     64.7661783808   0.002
NoeEskelet          -1.2243954255   1637.2034759108  -0.001
mashin             -28.4706560342   4654.6820093825  -0.006
tvrangi              7.8608282168   5049.1886963285   0.002
video               19.3149174731   4214.9497308810   0.005
computer            13.1255480291   2599.7357716833   0.005
mobile             -26.0998125729  29653.2462874807  -0.001
freezer              6.4487722675   7994.8573918092   0.001
yakhchalfreezer      9.7279607335   6417.3876831679   0.002
ojaghgaz           -58.5566938175  60083.9650348164  -0.001
jarobarghi        -106.4682040714  18536.8798478428  -0.006
lebasshoyi          -3.7277233736   6414.9289758174  -0.001
khayati             -1.9343738770   3767.9213785165  -0.001
panke              -28.9811450592   5658.0973016281  -0.005
lolekeshiab        -304.4675853024 163525.2626220236  -0.002
gazlolekeshi        154.2812865702 149836.9266871985   0.001
telephone             8.6423784205   3473.8590844062   0.002
internet              3.4373887329   7738.7511258033   0.000
SarparstSen80       0.999
SarparstSavad       0.997
SarparstMadrak      0.997
SarparstFaaliat     0.999
SarparstZanashoyi   1.000
NahveTasarof        0.999
TedadOtagh          0.998
SatheZirbana        0.998
NoeEskelet          0.999
mashin              0.995
tvrangi             0.999
video               0.996
computer            0.996
mobile              0.999
freezer             0.999
yakhchalfreezer     0.999
ojaghgaz            0.999
jarobarghi          0.995
lebasshoyi          1.000
khayati             1.000
panke               0.996
lolekeshiab         0.999
gazlolekeshi        0.999
telephone           0.998
internet            1.000
hamam               1.000
ashpazkhane         0.999
coolerabisabet      1.000
fazelabshahri       1.000
nsokhtpokhtpaz      1.000
nsokhtgarma            NA
nsokhtabgarm        0.999
Target              0.976
TotalHazine         0.996

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1196.8989138156  on 1863  degrees of freedom
Residual deviance:    0.0000026023  on 1824  degrees of freedom
AIC: 80

Number of Fisher Scoring iterations: 25
```

خروجی ۸: برازش مدل رگرسیون لوژستیک به روی مجموعه آموزشی

همانطور که مشهود است، معادله زیر حاصل می‌شود.

$$
\begin{aligned}
\text{Logit}(\text{decile} = 1) &= 758.78 + 429.95\ \text{NoeKhanevar} \\
&- 2.90\ \text{TedadAZA} \\
&- 11.17\ \text{SarparstJensiat} \\
&- 16.92\ \text{SarparstSen35} \\
&- 34.89\ \text{SarparstSen50} \\
&- 40.38\ \text{SarparstSen80} \\
&- 21.98\ \text{SarparstSavad} \\
&- 3.33\ \text{SarparstMadrak} \\
&- 2.07\ \text{SarparstFaaliat} \\
&+ 19.66\ \text{SarparstZanashoyi} \\
&+ 1.42\ \text{NahveTasarof} \\
&- 9.65\ \text{TedadOtagh} \\
&+ 0.13\ \text{SatheZirbana} \\
&- 1.22\ \text{NoeEskelet} \\
&- 28.47\ \text{mashin} + 7.86\ \text{tvrangi} \\
&+ 19.31\ \text{video} + 13.12\ \text{computer} \\
&- 26.09\ \ \text{mobile} + 6.44\ \text{freezer} \\
&+ 9.72\ \ \text{yakhchalfreezer} \\
&- 58.55\ \text{ojaghgaz} \\
&- 106.46\ \text{jarobarghi} \\
&- 3.72\ \text{lebasshoyi} - 1.93\ \text{khayati} \\
&- 28.98\ \text{panke} \\
&- 304.46\ \text{lolekeshiab} \\
&+ 154.28\ \ \text{gazlolekeshi} \\
&+ 8.64\ \text{telephone} + 3.43\ \text{internet} \\
&+ 150.88\ \text{hamam} \\
&- 221.26\ \text{ashpazkhane} \\
&+ 3.63\ \text{coolerabisabet} \\
&+ 0.58\ \text{fazelabshahri} \\
&+ 200.72\ \text{nsokhtpokhtpaz} \\
&+ 50.80\ \text{nsokhtgarma} \\
&- 98.5\ \text{nsokhtabgarm} \\
&+ 0.000012\ \ \text{Target} \\
&+ 0.00000029\ \text{TotalHazine}
\end{aligned}
$$

همانطور که در خروجی ۸ مشاهده می نمایید AIC=80 می باشد. و Fisher Scoring iterations به تعداد ۲۵ عدد در این مجموعه می-باشد.

ارزیابی عملکرد رده‌بندی

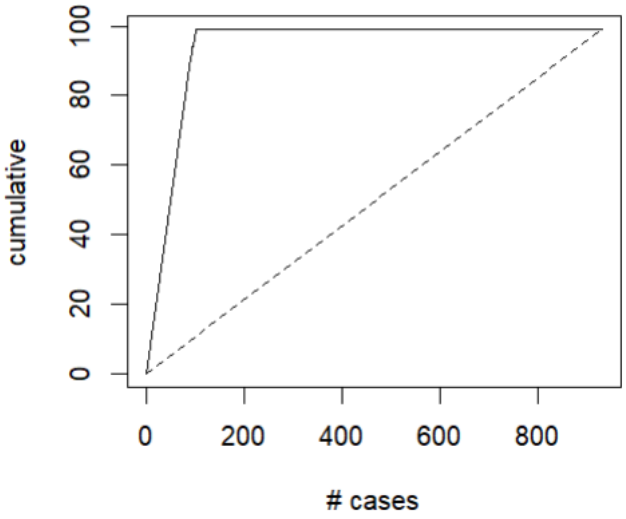
رایج ترین روش ارزیابی عملکرد، استفاده از ماتریس در هم ریختگی است.در خروجی زیر نسبت به ارزیابی عملکرد مدل می‌پردازیم.

```
> data.frame(actual=valid.data$decile[1:10],predicted=logit.reg.pred[1:1
0])
     actual                predicted
2168      0 0.00000000000000002220446
1248      0 0.00000000000000002220446
1353      0 0.00000000000000002220446
1393      0 0.00000000000000002220446
1367      1 0.99999999999999997779554
1925      0 0.00000000000000002220446
2070      1 0.99999999999999997779554
2699      0 0.00000000000000002220446
24        0 0.00000000000000002220446
541       0 0.00000000000000002220446
```

خروجی ۹: ارزیابی عملکرد مدل رگرسیون لوژستیک

همانطور که مشاهده می‌کنید به نتایج مطلوبی در ارزیابی مدل به روی داده‌های اعتبارسنجی رسیدیم.

دیگر ابزار مفید برای ارزیابی عملکرد رده‌بندی مدل، نمودارهای بالابردن و بالا بردن دهکی هستند.در زیر نمودار بالا بردن به دست آمده برای مدل لوژستیک روی مجموعه اعتبارسنجی را مشاهده می‌کنید.



نمودار ۷۱: نمودار بالابردن برای ارزیابی مدل لوژستیک

همانطور که در نمودار ۷۱ نیز مشهود است هرچه فاصله بالا بردن از خط قطری میان صفحه بیشتر باشد، نشان دهنده عملکرد خوب مدل است.پس نتیجه میگیریم که مدل رگرسیون لوژستیک بسیار عملکرد خوبی در این بخش دارد.

فصل پنجم: k –نزدیک ترین همسایه(k-NN)

در این فصل الگوریتم **k**-نزدیک‌ترین همسایه را که می‌تواند در رده-بندی (برای یک برامد رسته‌ای) یا پیشگویی (برای یک برامد عددی) استفاده شود،توصیف می‌کنیم.

الگوریتم **k**-نزدیک‌ترین همسایه، یک روش رده‌بندی است که در آن، هیچ فرضی درباره شکل رابطه بین عضویت رده **Y** و پیشگوهای $X_1,X_2,...,X_p$ وجود ندارد. پس این روش ناپارامتری است.

حال به پیاده سازی این روش به روی مجموعه آموزشی می‌پردازیم.

```
#K-NN
#initialize normalized training.valid,complate data frames to orginals
#use preprocess() from the caret package to normalize decile and sarparstsen
norm.values<-preProcess(train.data[,1:2],method =c("center", "scale"))
train.data[,1:2]<-predict(norm.values,train.data[,1:2])
valid.data[,1:2]<-predict(norm.values,valid.data[,1:2])
data.df[,1:2]<-predict(norm.values,data.df[,1:2])
#initilize a dataframe with two columns: k, and accuracy.
new.norm.df<-predict(norm.values,valid.data[,1:2])
library(class)
library(FNN)
#use knn()
nn<-knn(train=train.data[,1:2],test=new.norm.df,cl=train.data[,3],k=3)
row.names((train.data)[attr(nn,"nn.index")])
accuracy.df<-data.frame(k=seq(1,14,1),accuracy=rep(0,14))
#compute knn for different k on validation
for(i in 1:14){knn.pred<-knn(train.data[,1:2],valid.data[,1:2],
                             cl=train.data[, 3],k=i)
accuracy.df[i,2]<-confusionMatrix(knn.pred,as.factor(valid.data[, 3]))$overall[1]
```

همانطور که در تصویر فوق مشاهده می‌کنید، الگوریتم **k**-نزدیک‌ترین همسایه به شرح فوق پیاده سازی می‌گردد. تا بهترین **k** را بیابد.خروجی زیر را داریم.

```
> #use knn()
> nn<-knn(train=train.data[,1:2],test=new.norm.df,cl=train.data[,
 3],k=3)
> row.names(train.data)[attr(nn,"nn.index")]
  [1] "10"   "18"   "18"   "32"   "56"   "83"   "90"   "97"
  [9] "122"  "133"  "144"  "144"  "180"  "184"  "191"  "191"
 [17] "194"  "196"  "200"  "203"  "247"  "259"  "272"  "286"
 [25] "298"  "300"  "304"  "313"  "323"  "324"  "329"  "330"
> nn
  [1] 4 3 3 1 3 4 5 1 4 2 2 2 2 1 3 3 3 2 2 2 1 1 4 2 2 5 5 3
 [29] 4 4 4 2 2 4 3 4 3 3 3 3 2 2 2 2 3 4 2 3 2 5 5 2 2 2 3 3
 [57] 3 4 6 2 3 3 3 3 4 4 1 1 4 4 2 2 2 2 3 2 3 4 2 3 5 2 4 3
 [85] 3 3 3 4 4 5 3 4 2 2 2 3 2 5 3 2 1 5 2 4 1 1 1 3 2 4 2 3
attr(,"nn.index")
     [,1] [,2] [,3]
[1,]  816  339   29
[2,]   96   29  618
[3,]   96  618 1079
[4,]  807  732 1217
attr(,"nn.dist")
             [,1]         [,2]         [,3]
[1,] 0.0001859634 0.0001859634 5.578901e-04
[2,] 0.0003719267 0.0003719267 5.578901e-04
[3,] 0.0001859634 0.0003719267 5.578901e-04
[4,] 0.0001859634 0.0001859634 3.719267e-04
[5,] 0.0003719267 0.0005578901 5.578901e-04
[6,] 0.0001859634 0.0003719267 5.578901e-04
```

خروجی ۱۰: روش پیدا کردن بهترین k در الگوریتم k-NN

در نهایت به خروجی زیر دست میابیم.

	k	accuracy
1	1	0.6
2	2	0.6
3	3	0.7
4	4	0.7
5	5	0.8
6	6	0.9
7	7	0.9
8	8	0.9
9	9	1.0
10	10	0.9
11	11	0.8
12	12	0.8
13	13	0.4
14	14	0.5

خروجی ۱۱:بهترین K همسایه در الگوریتم K-NN

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1615  111
         1   66   72

               Accuracy : 0.905
                 95% CI : (0.8908, 0.918)
    No Information Rate : 0.9018
    P-Value [Acc > NIR] : 0.3374011

                  Kappa : 0.3978

 Mcnemar's Test P-Value : 0.0009422

            Sensitivity : 0.9607
            Specificity : 0.3934
         Pos Pred Value : 0.9357
         Neg Pred Value : 0.5217
             Prevalence : 0.9018
         Detection Rate : 0.8664
   Detection Prevalence : 0.9260
      Balanced Accuracy : 0.6771

       'Positive' Class : 0
```

خروجی ۱۲: ارزیابی عملکرد شبکه عصبی به روی مجموعه آموزشی

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 794  59
         1  39  40

               Accuracy : 0.8948
                 95% CI : (0.8734, 0.9138)
    No Information Rate : 0.8938
    P-Value [Acc > NIR] : 0.48436

                  Kappa : 0.3921

 Mcnemar's Test P-Value : 0.05495

            Sensitivity : 0.9532
            Specificity : 0.4040
         Pos Pred Value : 0.9308
         Neg Pred Value : 0.5063
             Prevalence : 0.8938
         Detection Rate : 0.8519
   Detection Prevalence : 0.9152
      Balanced Accuracy : 0.6786

       'Positive' Class : 0
```

خروجی ۱۳: ارزیابی عملکرد مدل شبکه‌عصبی به روی مجموعه اعتبار سنجی

همانطور که مشهود است عملکرد بی نظیر شبکه عصبی با ۲ لایه پنهان با ۱۰ و ۵ نورون به گونه‌ای شگفت آوری توجه ما را جلب می‌کند . ولی مشکلی نیز وجود دارد.. و آن میزان **Sensitivity** و میزان **Specificity** است. ولی در حالی است که میزان دقت آن در دو مجموعه آموزشی و اعتبار سنجی مطلوب است. ولی با توجه به هدف پروژه این دو مقدار قابل قبول است.

همانطور که مشاهده می‌کنید، بهترین K=5 است به دلیل آنکه اگر K=9 انتخاب شود، مطمئنا با مشکل بیش برازش روبه رو خواهیم شد.لذا بهترین و سریع‌ترین را انتخاب می‌کنیم. سپس بار دیگر الگوریتم را با K=5 اجرا میکنیم.

مزیت عمده روش K-NN سادگی آن و نداشتن فروض پارامتری است.در صورت وجود مجموعه آموزشی به اندازه کافی بزرگ، این روش‌ها به طرز شگفت‌آوری خوب عمل می‌کنند.

زمان یافتن نزدیک‌ترین همسایه‌ها در یک مجموعه آموزشی بزرگ می‌تواند کمرشکن باشد،این یکی از معایب این روش است، از راهکار-های حل این مشکل می‌توان به کاهش بعد اشاره کرد.

فصل ششم:شبکه عصبی

در این فصل، شبکه‌های عصبی را توصیف می‌کنیم،روشی داده‌رهنمون و انعطاف‌پذیر که می‌تواند برای رده‌بندی و پیشگویی به کار گرفته شود. شبکه‌های عصبی اگرچه از نقطه نظر تفسیرپذیری، جعبه سیاه تلقی می‌شود، از نظر درستی پیشگویانه قویا موفق است.

شبکه عصبی بر یک مدل فعالیت بیولوژیکی در مغز مبتنی است که در آن سلول‌های عصبی در اتصال با یکدیگرند و از تجربه یاد میگیرند.

تابع فعال ساز لوژستیک مورد استفاده قرار می‌گیرد.تابع لوژستیک رایج‌ترین تابع در شبکه‌های عصبی است.ارزش عملی آن از این واقعیت بر می‌خیزد که روی مقادیر بسیار کوچک یا مقادیر بسیار بزرگ اثر تخریبی دارد، اما در دامنه‌ای که مقادیر تابع بین ۰,۱ و ۰,۹ باشد، تقریبا خطی است. پس در نتیجه می‌توانیم خروجی راس j در لایه پنهان را به صورت زیر بنویسیم.

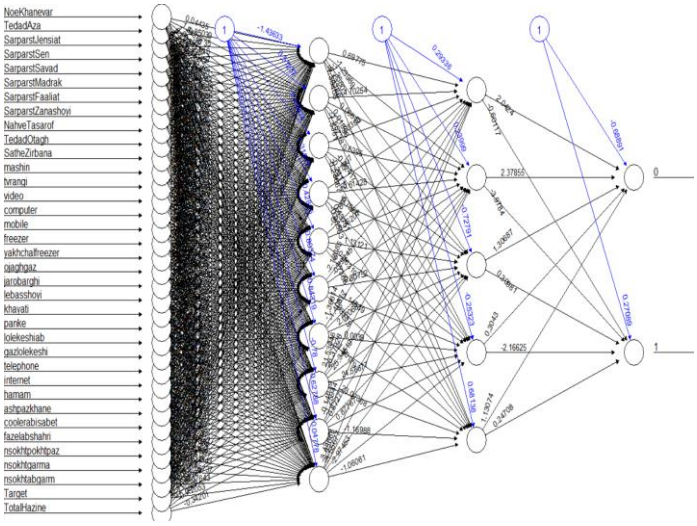
$$Output_j = g\left(\Theta_j + \sum_{i=1}^{p} w_{ij}x_i\right) = \frac{1}{1 + e^{-(\Theta_j + \sum_{i=1}^{p} w_{ij}x_i)}}$$

θ_j اریبی راس j را نشان می‌دهد.

همانطور که مشاهده می‌کنید، این معادله فرمول رگرسیون لوژستیک است!!

حال به اجرای مدل شبکه عصبی می‌پردازیم.

برای درک بهتر ساختار شبکه عصبی اجرا شده، می‌توانید به شکل زیر توجه کنید.



نمودار۷۲ : ساختارشبکه عصبی اجرا شده به روی مجموعه آموزشی

همچنین برای اجرای این شبکه عصبی بدین نحو عمل کردیم.

```
#####################################
library(neuralnet)
Nn<-neuralnet(as.factor(decile)~NoeKhanevar+TedadAza+SarparstJensiat+SarparstSen+
                SarparstSavad+SarparstMadrak+SarparstFaaliat+SarparstZanashoyi
              +NahveTasarof+TedadOtagh+SatheZirbana+mashin+tvrangi+video+
                computer+mobile+freezer+yakhchalfreezer+ojaghgaz+jarobarghi
              +lebasshoyi+khayati+panke+lolekeshiab+gazlolekeshi+telephone+
                internet+hamam+ashpazkhane+coolerabisabet+fazelabshahri+
                nsokhtpokhtpaz+nsokhtgarma+nsokhtabgarm+Target+TotalHazine,
              data = train.data,linear.output = F,hidden =c(10,5))
plot(Nn)
trai.data<-select(train.data,-decile)
training.prediction=compute(Nn,trai.data[,1:38])
training.class=apply(training.prediction$net.result,1,which.max)-1
confusionMatrix(as.factor(training.class),as.factor(train.data$decile))

valid.prediction=compute(Nn,valid.data[,1:39])
valid.class=apply(valid.prediction$net.result,1,which.max)-1
confusionMatrix(as.factor(valid.class),as.factor(valid.data$decile))
```

خروجی ۱۴: اجرای شبکه عصبی به روی مجموعه آموزشی و اعتبار سنجی

برجسته‌ترین مزیت شبکه‌های عصبی، عملکرد پیشگویانه خوب آنهاست. آنها به تحمل بالا برای داده‌های نوفه‌ای و توانایی شکارکردن ارتباط‌های بسیار پیچیده بین پیشگوها و یک برامد مشهورند. مهمترین نقطه ضعف آن‌ها به فراهم آوردن بینش درباره ساختار ارتباط و در نتیجه سوء شهرت به *جعبه سیاه* مربوط است.

نتیجه گیری کلی

همانطور که مشاهده نمودید بهترین عملکرد مربوط به شبکه‌های عصبی است پس در نهایت این مدل را به روی مجموعه آزمون با تابع فعال ساز لوژستیک برازش می‌دهیم و می‌توانیم درخروجی زیر نتیجه آن را مشاهده کنیم.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 271  19
         1  12   9

               Accuracy : 0.9003
                 95% CI : (0.8615, 0.9313)
    No Information Rate : 0.91
    P-Value [Acc > NIR] : 0.7604

                  Kappa : 0.3144

 Mcnemar's Test P-Value : 0.2812

            Sensitivity : 0.9576
            Specificity : 0.3214
         Pos Pred Value : 0.9345
         Neg Pred Value : 0.4286
             Prevalence : 0.9100
         Detection Rate : 0.8714
   Detection Prevalence : 0.9325
      Balanced Accuracy : 0.6395

       'Positive' Class : 0
```

خروجی ۱۵: ارزیابی عملکرد مدل نهایی به روی مجموعه آزمون با تابع فعال ساز لوژستیک

همانطور که مشاهده می‌کنید تعادل دقت مدل ۰,۹۰ رسیده است.

و این یعنی می‌توانیم با استفاده از مدل شبکه‌های عصبی با تابع فعال-ساز لوژستیک به بهترین نتیجه در خصوص داده‌های جدید برسیم.[۱]

سخن پایانی

در پایان ضمن تقدیر و تشکر از زحمات استاد گرانقدر درس داده‌کاوی جناب اقای دکتر *محمدرضا فقیهی حبیب آبادی* و همچنین انجمن ریاضی دانشگاه شهید بهشتی، از تمامی خوانندگان کمال تشکر را اعلام می‌دارم.

تهیه‌کننده: سجاد صحت بخش

دانشجوی کارشناسی‌ارشد دانشگاه شهید بهشتی

رشته ریاضی‌کاربردی–گرایش علوم داده

۱ - فایل اصلی کل پروژه درصفحه گیت‌هاب شخصی نویسنده قرار گرفته است. (https://github.com/sajjadsehat/DataMining-project.git)