**به­نام خدا**

عنوان:رده­بندی خوانوارها برحسب دهک دهم

نام استاد:جناب آقای دکترفقیهی

نام دانشجو:سناآریا

شماره دانشجویی:400422217

دانشگاه شهید بهشتی

دانشکده علوم ریاضی

**فصل 1**

**پیش پردازش داده­ها**

**نام مجموعه داده**: آمار هزینه و درامد خانوارهای شهری

توضیحی در مورد مجموعه:

داده­های این پروژه مربوط به طرح آمارگیری هزینه ودرآمد خانوار­های در استاد های آذربایجان شرقی و آذربایجان غربی و زنجان و اردبیل می­باشد

**تعداد ثبت ها:** 2081

**تعداد متغیر ها**:90

**نام متغیر ها:**

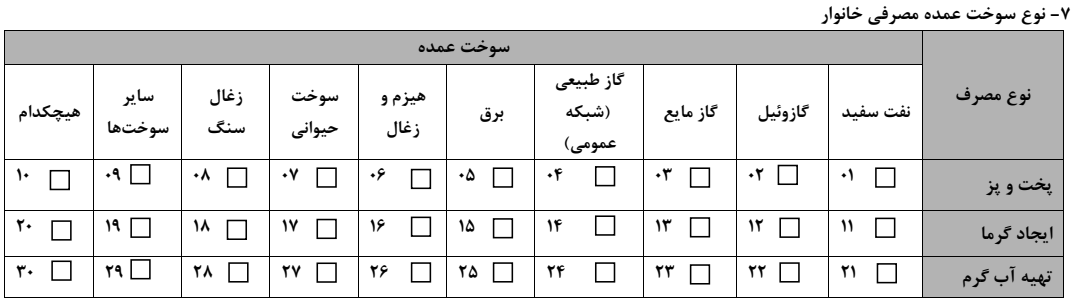
1. نام استان: متغیری رسته ای که شامل اذربایجان شرقی،آذربایجان غربی ، اردبیل وزنجان است
2. آدرس: آدرس خانوار (یک عدد 11 رقمی)
3. نوع خانوار: متغیر دودوییِ نشان دهنده ی سکونت معمولی ساکن یا گروهی می باشد
4. تعداداعضا: تعداد اعضا خانوار(متغیر گسسته)
5. جنسیت سرپرست: متغیری دودویی ِنشان دهنده ی جنسیت سرپرست (1و2)
6. سن سرپرست: سن سرپرست متغیر پیوسته (عددی کم تراز 100)
7. سواد سرپرست: متغیری دودوییِ نشان دهنده ی باسواد بودن سرپرست(1و2)
8. تحصیل سرپرست:متغیری دودوییِ نشان دهنده ی درحال تحصیل بودن یا نبودن سرپرست
9. مدرک سرپرست: متغیری رسته ای(1تا9) که نشان دهنده سطح مدرک سرپرست
10. فعالیت سرپرست: متغیری رسته ای از عدد 1 تا 6 که نشان دهنده 1. شاغل 2.جویای کار 3. دارای درآمد بدون کار 4.محصل 5.خانه دار 6.سایر
11. زناشویی سرپرست: متغیری رسته ای از عدد 1 تا 4 که نشان دهنده 1.همسردارد 2.ندارد(فوت) 3. ندارد(طلاق) 4.ازدواج نکرده
12. نحوه تصرف: متغیری رسته ای از عدد 1 تا 7 که نشان دهنده 1.ملکی عرصه و اعیان 2.ملکی اعیان 3. اجاری 4.رهن 5.دربرابرخدمت 6.رایگان 7.سایر
13. تعداد اتاق: متغیری گسسته نشان دهنده ی تعداد اتاق های خانه
14. سطح زیربنا: متغیری پیوسته نشان دهنده ی متراژ زیر بنای خانه
15. نوع اسکلت: متغیری رسته ای از اعداد 1 تا 3 نشان دهنده ی 1.فلزی 2.بتون 3.سایر
16. ماشین: متغیری رسته ای از 1 و 0 نشان دهنده داشتن یا نداشتن ماشین
17. موتور: متغیری رسته ای از 1 و 0 نشان دهنده داشتن یا نداشتن موتور

از متغیر 18 تا 51 نشان دهنده داشتن و یا نداشتن وسایلی مانند دوچرخه، رادیو، رادیو ضبط، تلویزیون، تلویزیون رنگی و .... تا فاضلاب شهری که متغیرهایی دودویی(0و1) هستند می­باشد.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Radio | radiozabt | tv | Tvrangi | video | computer | mobile |
| Freezer | yakhchal | yakhchalfreezer | Ojaghgaz | jarobarghi | lebasshoyi | khayati |
| Panke | coolerabimoteharek | coolergazimoteharek | Zarfshoyi | microfer | hichkodam | lolekeshiab |
| Bargh | gazlolekeshi | telephone | Internet | hamam | ashpazkhane | coolerabisabet |
| borodatmarkazi | hararatmarkazi | package | coolergazisabet | fazelabshahri |  |  |

ستون های 52 و 53 و 54 نشان دهنده نوع سوخت مصرفی با توجه به شکل زیر است:

نوع سوخت­گرما ،نوع سوخت آب­گرم ،نوع سوخت­پخت و پز



از ستون 55 تا 66 مربوط به هزینه های خوراکی،نوشیدنی،پوشاک،مسکن،لوازم­خانگی ،درمانی ،حمل­ونقل ،ارتباطات ، تفریحات فرهنگی ،غذای­آماده ،متفرقه ، کالای­بادوام که متغیرهایی پیوسته هستند می­باشد.

از ستون 67 تا 90 هم مربوط به متغیرهای درآمد (یارانه دریافتی،یارانه ماه و...که در نمودار حرارتی ای که در ادامه آورده شده است.مشاهده می­شوند) که متغیرهایی پیوسته می­باشند هست.

**هدف پروژه:**

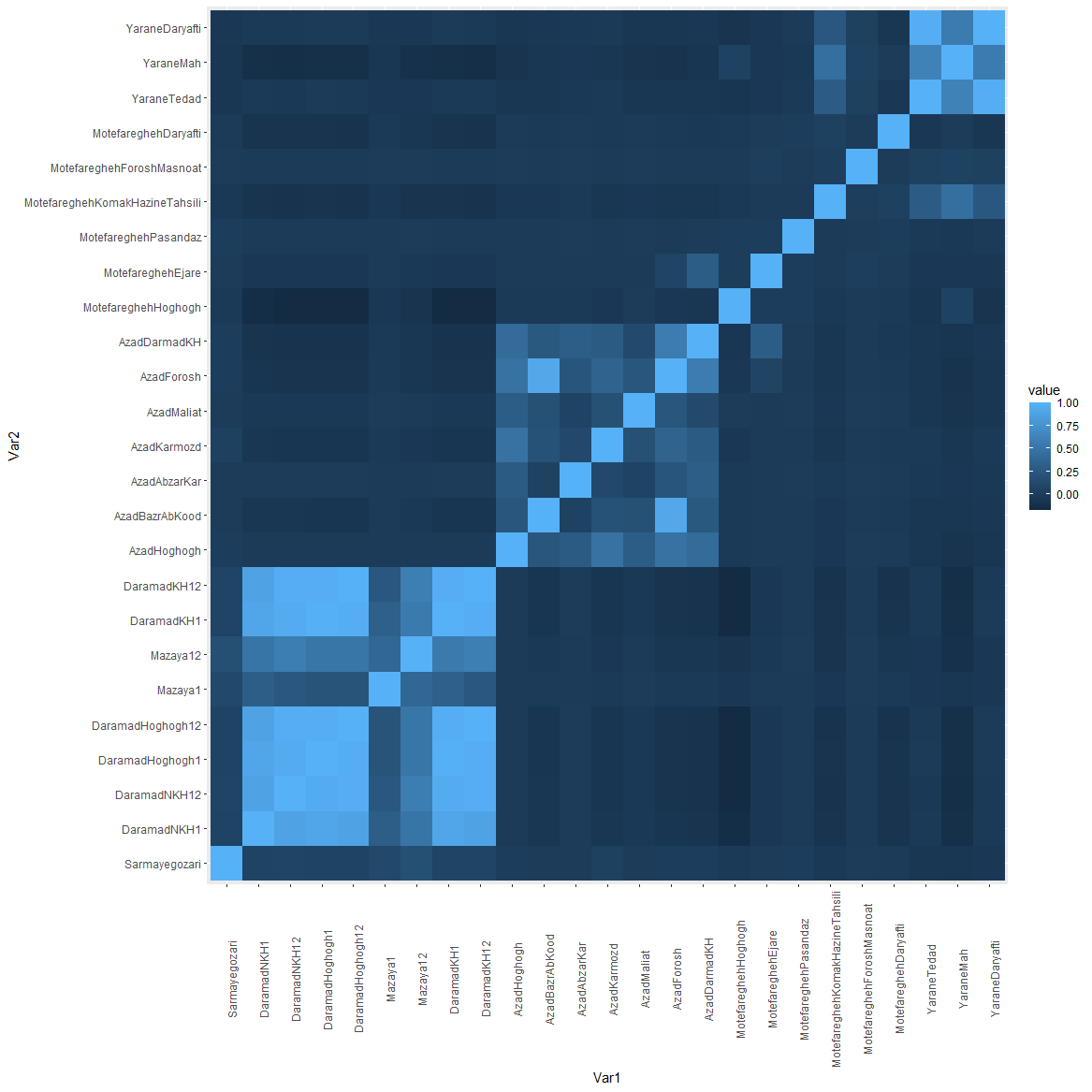
طرح جدید تخصیص یارانه به خانوار ها بدین شکل است که به 9 دهک پایین یارانه تعلق میگیرد ولی به 1 دهک بالای جامعه،یارانه­ای تعلق نمی­گیرد.

هدف تشخیص یک دهک بالای جامعه که یارانه ای به آن تعلق نمی­گیرد می­باشد.

**ضرورت پروژه:**

واضح است که اهمیت آنکه به چه افرادی درجامعه یارانه تعلق نمی­گیرد برای دولت بسیار بالاست.

برای این امر با استفاده از نمودار حرارتی بر روی متغیر های درآمد متغیر جدید تحت عنوان Target ساخته شد تا با استفاده از آن درآمدهای فرد را متوجه شده و از آن برای یافتن یک دهک بالای جامعه استفاده شود.



سپس با استفاده از این نمودار ستونی برای آنکه درآمد نهایی فرد را به­دست آوریم واز آن برای یافتن دهک دهم استفاده کنیم و آن را Target نامیدیم.

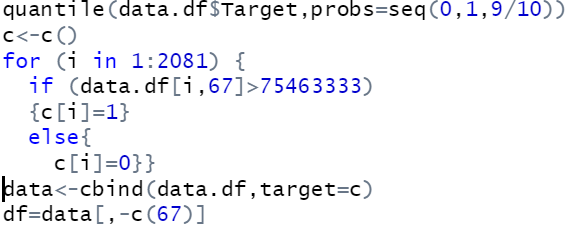
Target = (sarmayegozari + daramadNKH12 + azaddaramadKH) + motefaregheha

و ستون های درآمد را حذف کردیم.

حال 69 ستون داریم که با آنها ادامه پروژه مان را پیش خواهیم برد.

دهک بندی دیتاهای موجود:

ما برای آنکه بتوانیم اختصاص یا عدم اختصاص یارانه به هر فرد را تعیین کنیم نیاز داریم ستون Target مان را ده قست کرده و به یک دهم بالایی عدد 1 و به سایر آنها عدد 0 را نسبت دهیم؛ به همین منظور ستونی جدید تحت عنوان target با کد دستوری زیر بنا کردیم و ستون Target قبلی را نیز از مجموعه مان حذف کردیم.

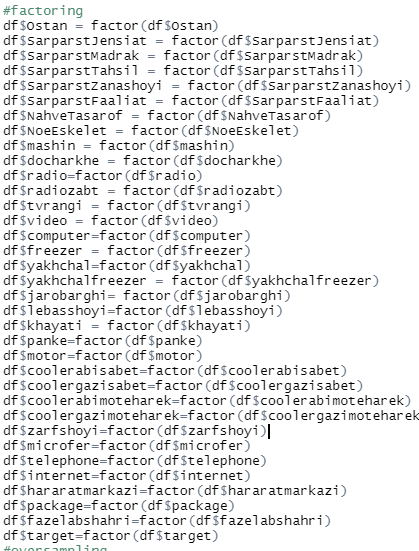


با مشاهده نتیجه­ی خط اول کد بالا مشخص شد که اگر در ستون Target مقداری بیش از 75463333 باشد متعلق به دهک دهم است و به آن عدد 1 و اگر ازین مقدار کمتر بود عدد 0 را به آن نسبت داده شد.

حال با نگاهی کلی به مجموعه داده ها میتوان دریافت برخی از متغیر­ها برای این هدف اهمیتی ندارند و از مجموعه­ داده حذف شدند که شامل متغیر های گروه(همه ستون آن یک عدد بود و بی اثر میشد) ، آدرس(کمکی به تعیین دهک شخص نمیکرد) ونوع خانوار(همه ستون آن یک عدد ثابت بود).

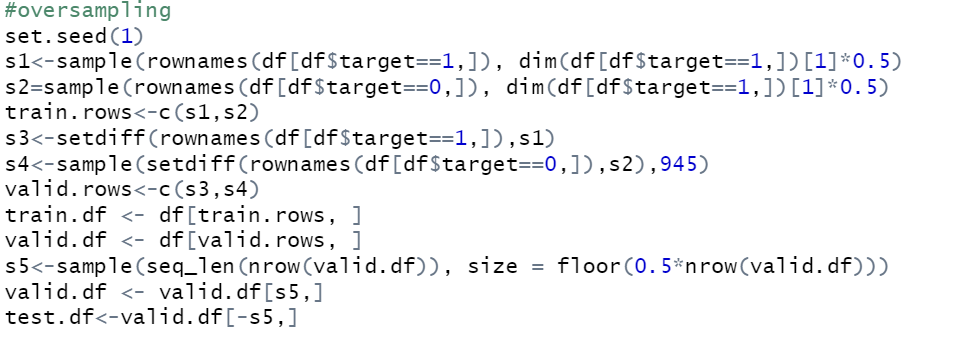
با دقت به سه متغیر سرپرست سواد ،سرپرست تحصیل وسرپرست مدرک میتوان فهمید که اگر سرپرست سواد نداشته باشد(مقدار 2 باشد) مقدار متغیر تحصیل و مدرک خالی است پس در جای خالی آن 0 جایگذاری شد و به دلیل اینکه واضحا این سه متغیر وابسته بهم بودند متغیر سواد حذف شد.

از طرفی برخی متغیرها ظاهری هستند و باید با استفاده از دستورات زیربه اینکه مثلا 0و1 بودن در ماشین به معنای عدد 0و1 نیست بلکه ظاهری هست توجه لازم را به عمل آورده شده است.



**بخش بندی داده ها به داده های آموزشی و اعتبارسنجی وآزمون:**

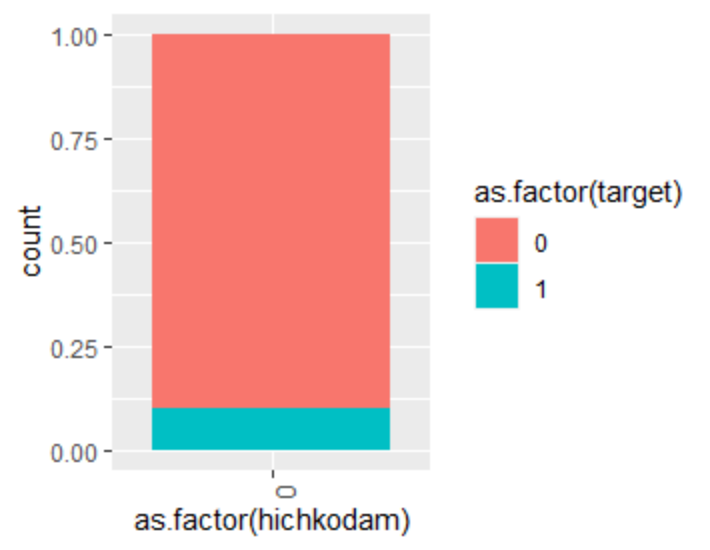
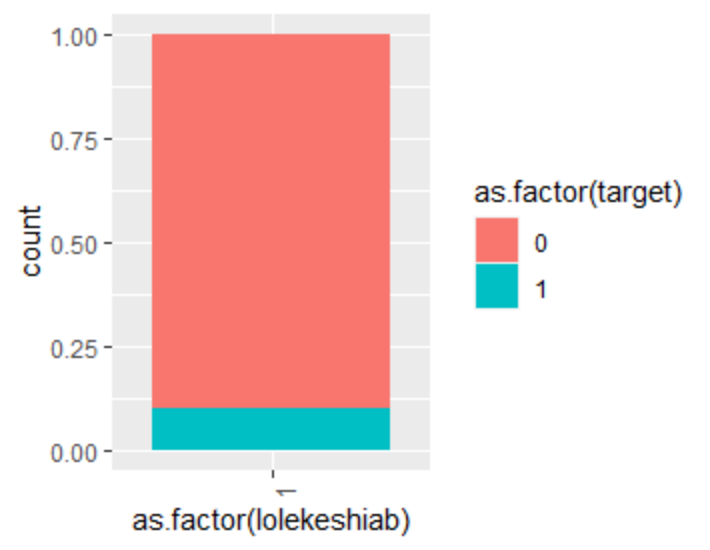
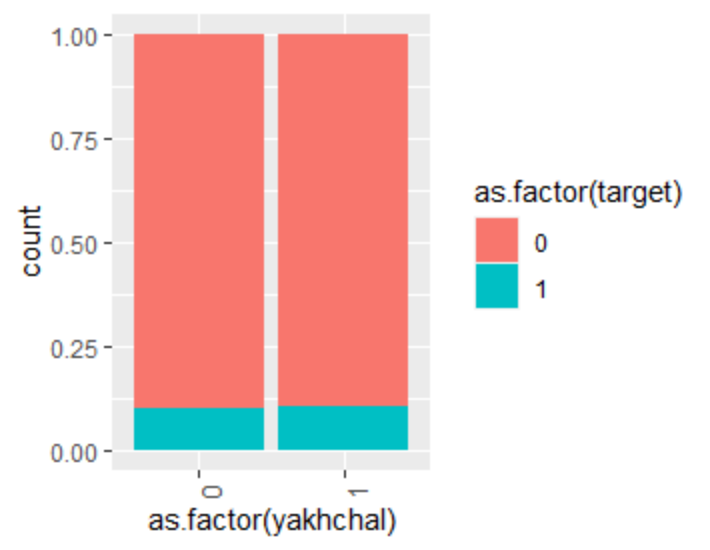
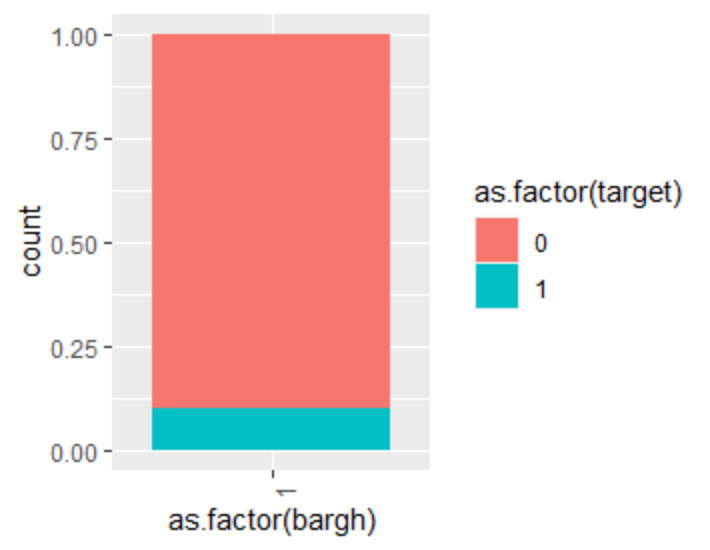
باتوجه به آنکه تعداد 0 و1 ها در مجموعه داده بسیار نامتوازن هستند اگر به صورت عادی و رندوم به دسته بندی شوند مدل هایمان فریب خورده و تواناییشان برای تشخیص درست1 ها بسیارکم می­شود؛ بنابراین از روش بیش­نمونه­گیری(oversampling) برای بخش بندی داده­ها به داده­های آموزشی و اعتبار­سنجی استفاده شد که در ادامه کد آن را مشاهده می­شود.

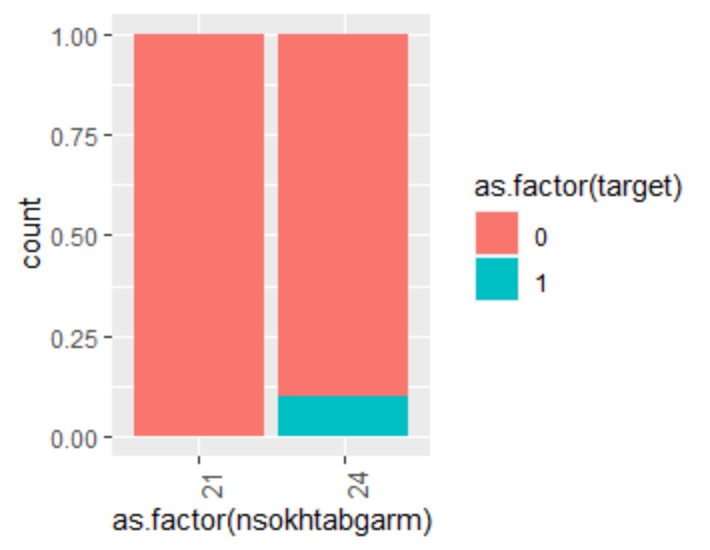
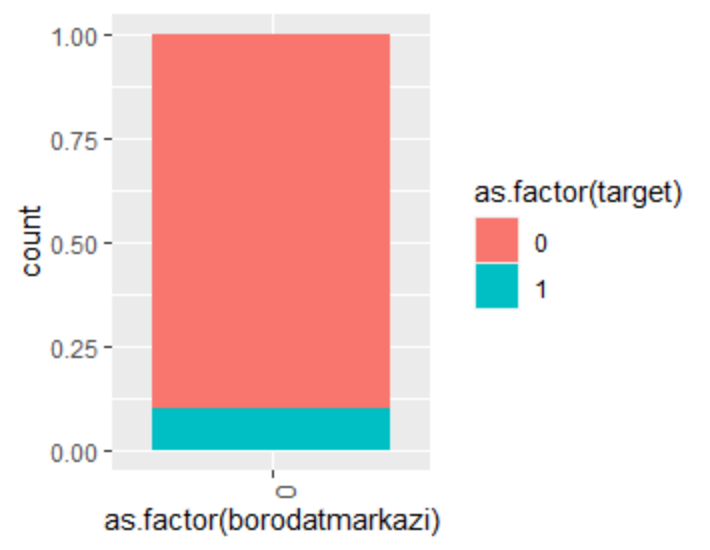


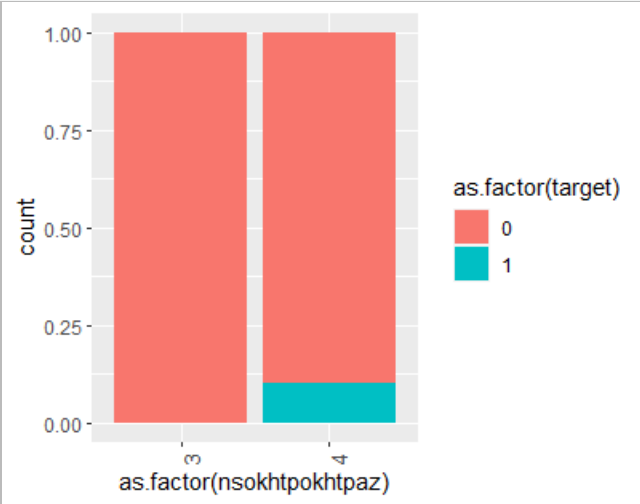
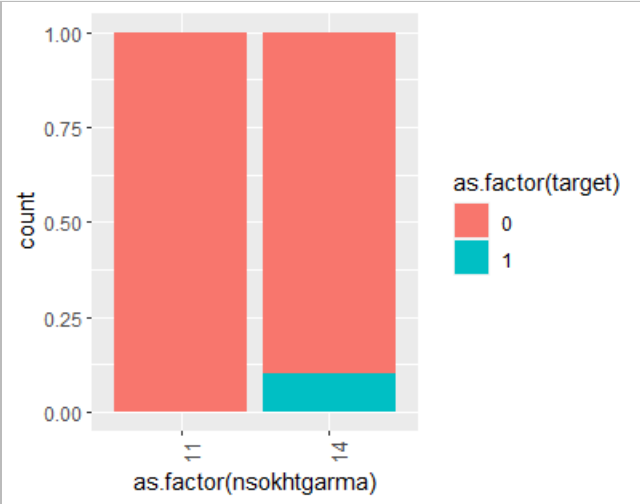
**فصل دوم**

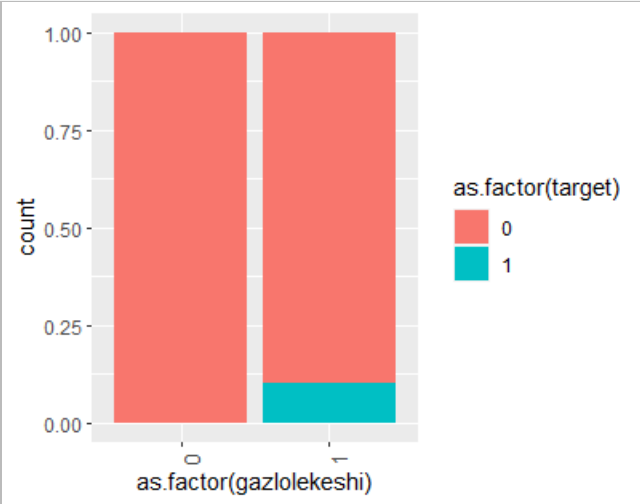
**تصویری­سازی**

در ادامه نمودارهایی بین متغیرهای باقی­مانده و متغیرtarget رسم کردم ؛برای متغیرهای گسسته نمودار­میله­ای و برای متغیر­های پیوسته نمودار جعبه­ای رسم کردم ولی از آوردن تک تک آن­ها در گزارش به دلیل زیاد بودنشان صرف نظر کرده و تنها آن­هایی را که نتیجه­شان سبب شد تا حذف گردند می­آورم و این بدین معناست که مابقی رفتارشان با متغیر برآمد درست بوده ☺ وفعلا نگاه­داشتن آن­ها برای مدل­مان خوب است.

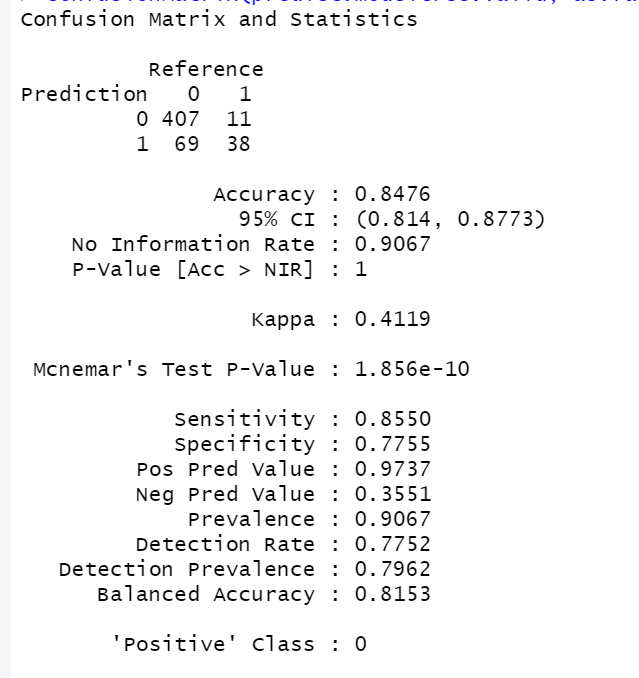
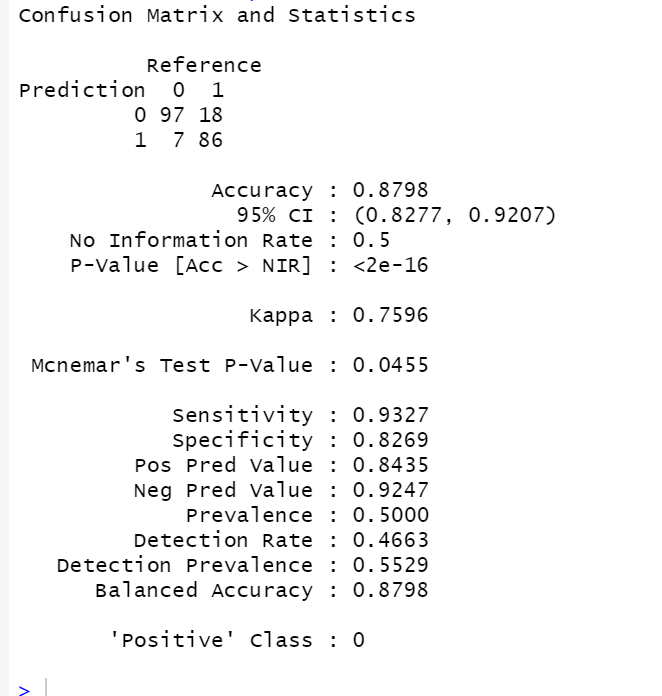
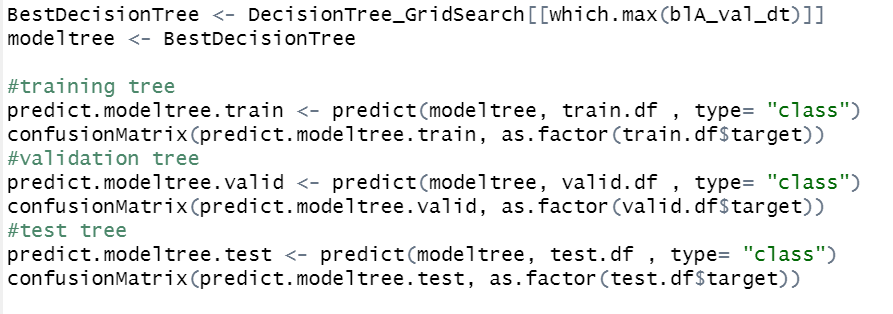


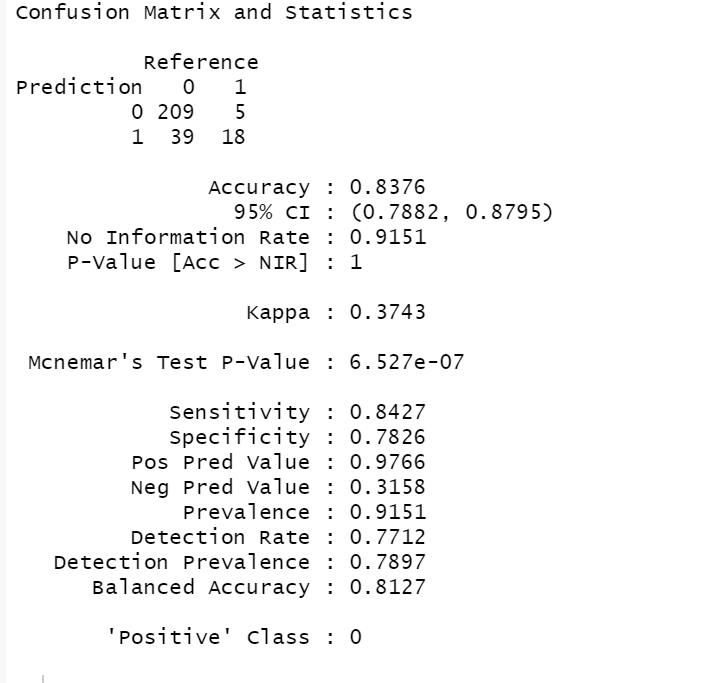


همانطور که در نمودارهای بالا مشخص است متغیرهای یخچال(0و1 بودن آن­ اثری در 0و1 شدن برآمد ندارد) ،هیچکدام(همه ستونش 0 است )،لوله کشی آب و برق (همه­ی ستونش 1 است یعنی تمام خانوارها لوله کشی­آب و برق دارند)،برودت مرکزی(همه ستونش 0 است یعنی خانواری ازین متغیر بهره­مند نیستند) ولوله کشی گاز ونوع سوخت گرما و آب گرم وپخت وپز (نشان می دهد تقریبا همه از یک نوع سوخت استفاده می­کنند) اثری بر روی متغیر برآمد(target) ندارند پس می­توانند حذف شوند.

**فصل سوم**

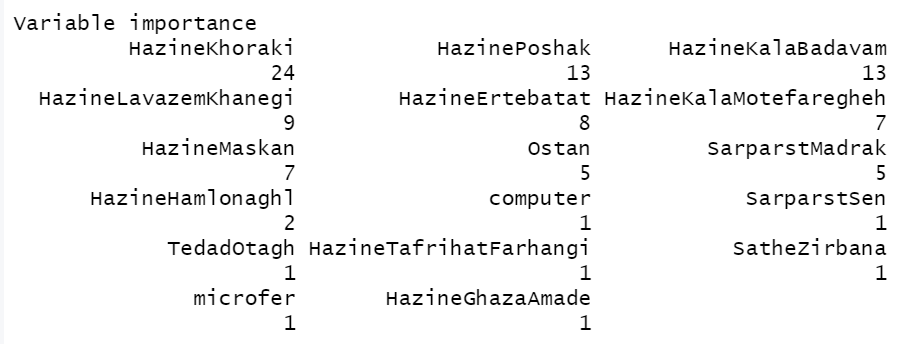
**برازش مدل درخت**

مدل درخت را روی مجموعه داده­ی آموزشی و سپس مجموعه داده­ی اعتبار سنجی با استفاده از گرید سرچ برازش داده شد که نتیجه را در زیر مشاهده می­شود: 

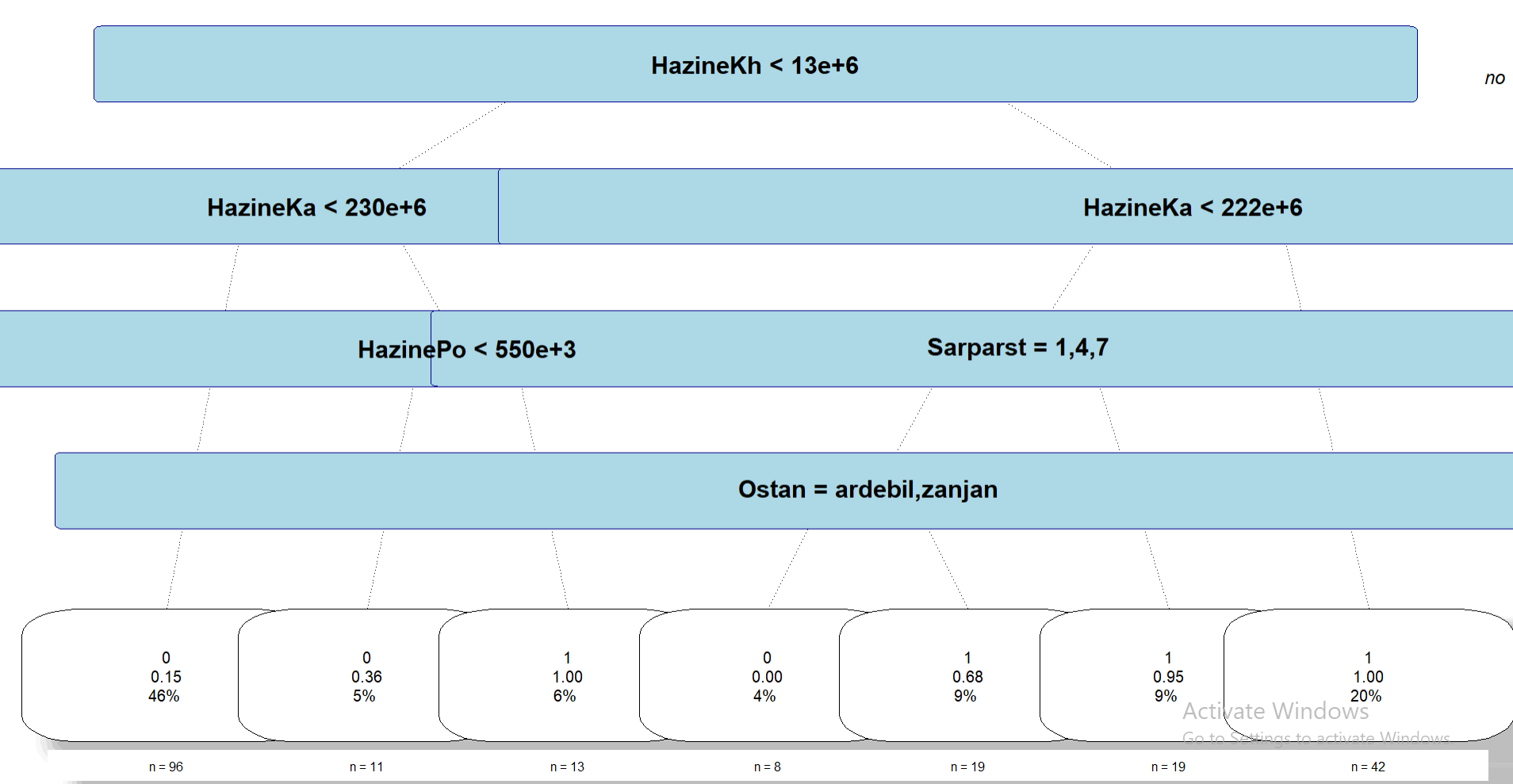


تصویر بالا سمت چپ نتایج به­دست آمده از برازش مدل بر روی مجموعه آموزشی وتصویر بالا سمت راست نتایج به دست آمده بر روی مجموعه ی اعتبار سنجی ونتایح به دست آمده در تصویر پایین سمت راست مربوط به برازش مدل بر روی مجموعه ی آزمون است که به ما می­گوید دقت این مدل 83 درصد می­باشد.

با summary گرفتن از مدل به متغیرهای پراهمیت تر در مدل دست می­یابیم:



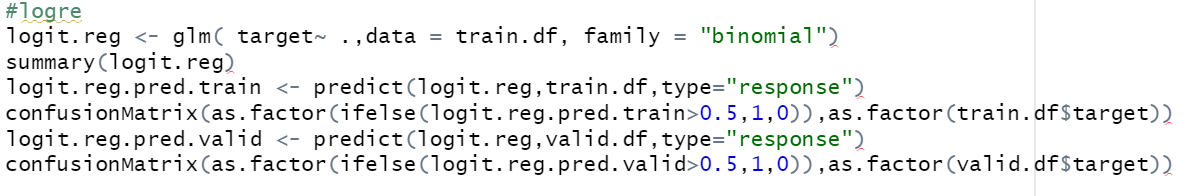
نمودار درخت والگوریتم آن نیز براساس شکل زیر می­باشد.

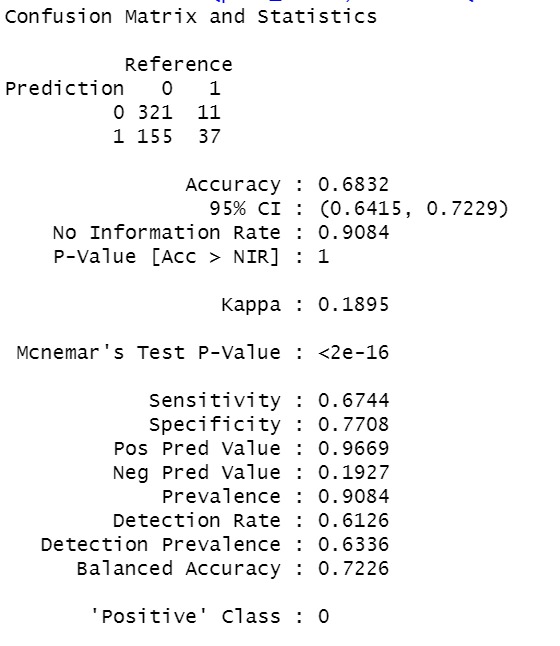
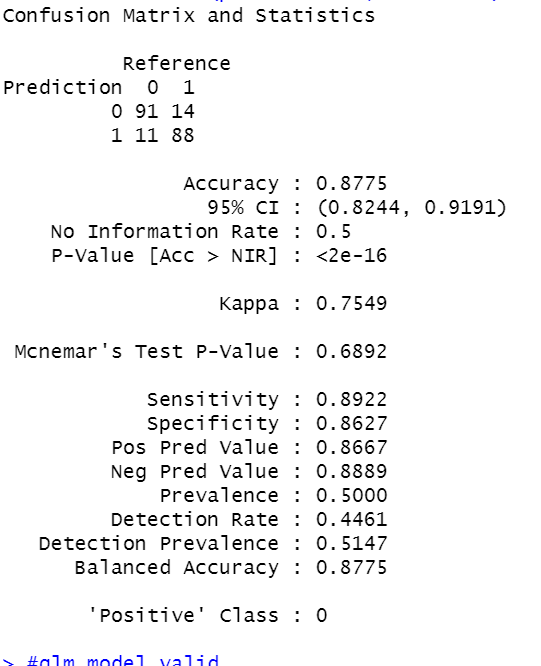


**فصل چهارم**

**برازش مدل رگرسیون لجستیک**

مدل رگرسیون لجستیک را بر روی داده­های آموزشی و اعتبارسنجی می­برازانیم و نتیجه را مشاهده می­کنیم:





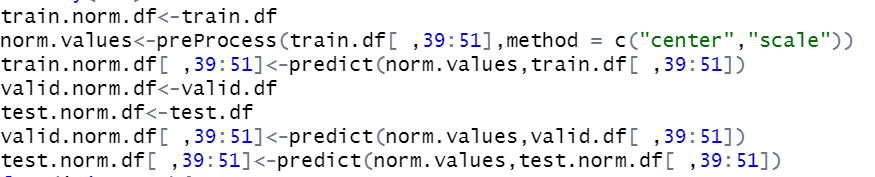
تصویر سمت راست دقت حاصل شده از برازش مدل بر روی داده های آموزشی و تصویر سمت چپ بر روی داده های اعتبارسنجی است.

همانگونه که میبینیم دقت بر روی داده های آموزشی بالا بوده(87%) و با دقت بر روی داده های اعتبارسنجی(68%) اختلاف زیادی دارد پس می­توان نتیجه گرفت که این مدل دچار بیش­برازش شده است احتمال دادم چون داده ها نرمال شده نیستند این اتفاق رخ داده و برای رفع این مشکل داده هامان را نرمال سازی کردم ولی همچنان مشکل بیش برازش وجود داشت و نتیجه گرفتم مدلمان مناسب نیست.

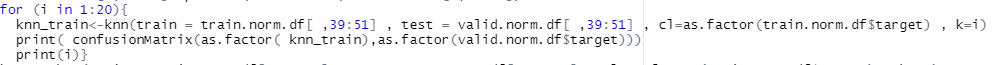
**فصل پنجم**

**برازش مدل knn**

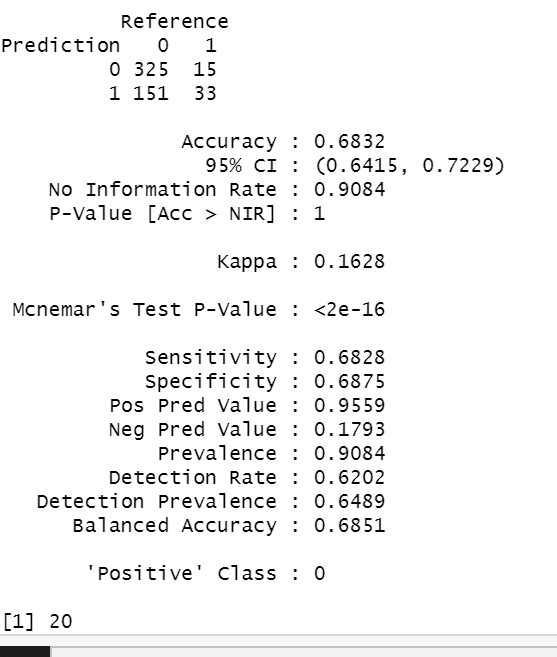
مجموعه ها را نرمال سازی کرده و داده های گمشده را خارج شده است.



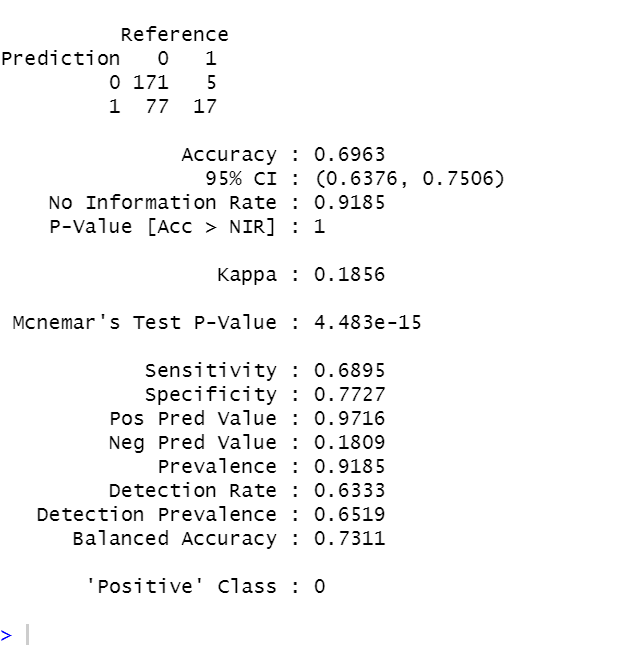
سپس مدل را بر روی مجموعه آموزشی برازانده و بهترین k را با استفاده از مجموعه اعتبارسنجی یافت شد.



20= k بهترین kمی باشد که نتیجه اش به شکل زیراست.



حال با 20=k نتیجه را برروی داده­آزمون مشاهده می­شود:

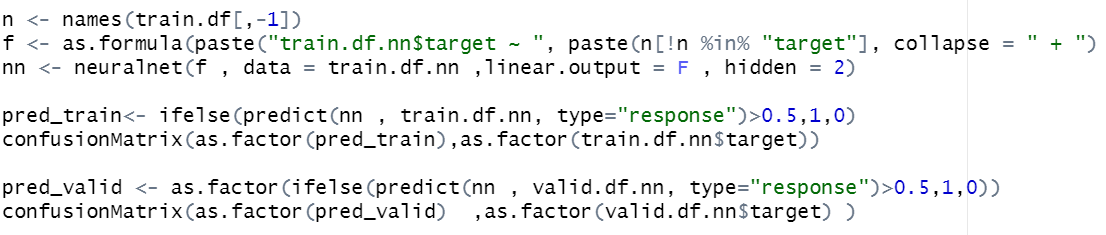


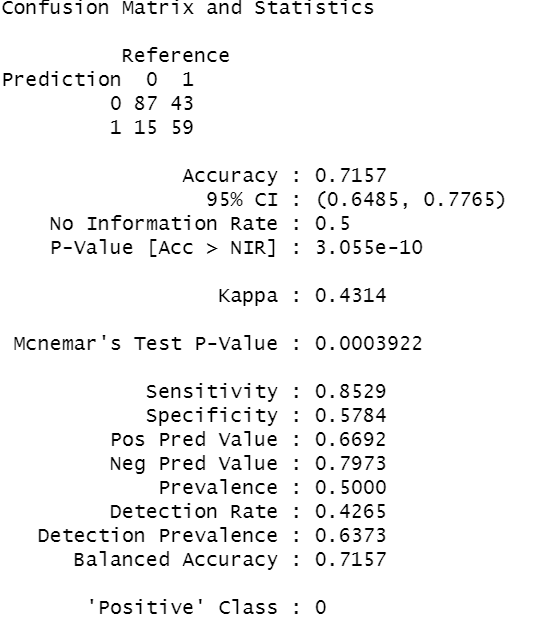
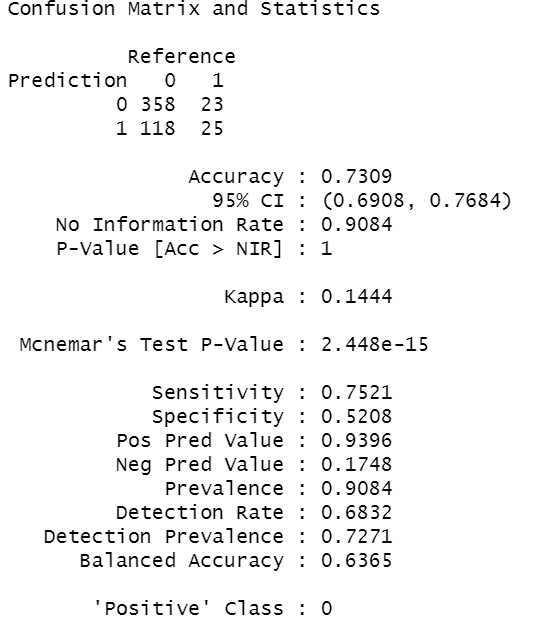
همانطور که مشاهده می­ شود دقت این مدل 69% می­باشد.

**فصل ششم**

**برازش مدل شبکه عصبی(nn)**

مدل را می­برازانیم:



و نتایج را مشاهده مینماییم:  
 

تصویرسمت راست مربوط به مجموعه آموزشی و تصویر سمت چپ مربوط به مجموعه اعتبار سنجی می­باشد و نشان دهنده آن است که دقت در این مدل73% می­باشد.

**فصل هفتم**

**نتیجه گیری**

|  |  |
| --- | --- |
| **دقت** | **مدل** |
| 83% | **درخت** |
| بیش برازش | **رگرسیون لجستیک** |
| 69% | **knn** |
| 73% | **nn** |

با توجه به جدول بالا مدل درخت در این پروژه بهترین مدل است.