بسم الله الرحمن الرحیم

تحلیل آماری کاربردی

عنوان:

" پیش بینی قیمت و پست بازیکن فوتبال
با استفاده از رگرسیون و طبقه بندی"

گردآورنده:

امیرعلی خطیب

۹۷۲۲۷۴۳۱۴۹

استاد:

جناب آقای دکتر علیرضا شادمان

تاریخ تحویل:

تیر ۱۴۰۱

فهرست مطالب

مقدمه:

در این گزارش ما قصد داریم به بررسی عملکرد الگوریتم‌های یادگیری ماشین [1]برای داده‌های بازیکنان فوتبال [2]بپردازیم. در مجموعه داده اشاره شده ما حدود ۱۹۲۳۹ بازیکن داریم که تمام مشخصات آن‌ها نظیر سن، ملیت، تیم باشگاهی، و توانایی‌های فردی آن‌ها ذخیره شده است. در این پروژه ما قصد داریم تو گروه از الگوریتم‌های رگرسیون [3]و طبقه‌بندی[4] را با اهداف متفاوت بر روی داده‌ها پیاده‌سازی کنیم تا پیش‌بینی‌های مورد نظرمان را انجام دهیم. به طور کلی ما ۲ هدف را در این پروژه دنبال می‌کنیم:

۱) **هدف مسئله رگرسیون:** هدف از پیاده‌سازی الگوریتم‌های رگرسیون، پیش بینی ارزش بازیکن در بازار نقل و انتقالات با در نظر گرفتن پارامترهایی نظیر سن، پتاسیل رشد، شهرت و توانایی‌های فردی می‌باشد. با کمک گرفتن از این الگوریتم‌ها باشگاه‌ها می‌توانند مدیریت بهتری روی خرید و فروش خود داشته باشند؛ زیرا می‌توانند تا حدی قیمت بازیکن موردنظر خود را با وارد کردن اطلاعات شخصی و توانایی‌های فردی بازیکن تخمین بزنند.

۲) **هدف مسئله طبقه بندی:** هدف از پیاده‌سازی الگوریتم‌های طبقه‌بندی، پیش‌بینی پست بازیکن است. به این صورت که با دیدن توانایی‌های فردی بازیکن و اندازه‌گیری آن‌ها، می‌تواند دید خوبی به مربی جهت به کارگیری آن‌ها در پست مناسب بدهد. این روش بیشتر در زمینه استعدادیابی بازیکنان در رده‌های پایه می‌تواند مفید واقع شود. زیرا بازیکنان را به بازی در پستی که در آن مستعدتر هستند، سوق می‌دهد.

تشریح داده‌ها:

در این قسمت ابتدا تمامی ویژگی‌ها را به اختصار توضیح می‌دهیم (در اسامی ویژگی‌ها به لاتین نوشته می‌شود؛ زیرا معادل دقیقی برای بعضی از ویژگی‌ها موجود نمی‌باشد):

۱) Acceleration: بیانگر این است که چقد زمان صرف می‌شود تا پس از شروع به دویدن، بازیکن به حداکثر سرعتی که دارد برسد. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۲) Sprint speed: بیانگر بالاترین سرعت بازیکن در هنگام دویدن. (در قالب امتیاز از ۰ تا ۹۹ درجه‌بندی می‌شود.)

۳) Finishing: قابلیت تمام کنندگی گویند که بیانگر دقت شوت در محوطه پنالتی می‌باشد. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

[1] Machine learning
[2] FIFA 22 dataset
[3] Regression
[4] Classification

۴) Long shot: دقت شوت در خارج از محوطه جریمه. (در قالب امتیاز از ۰ تا ۹۹ درجه‌بندی می‌شود.)

۵) Penalties: دقت در ضربات پنالتی. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۶) Positioning: قابلیت جایگیری بازیکن در فضـــاهای خالی برای دریافت توپ و خلق موقعیت خطرناک. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۷) Shot power: بیانگر میزان قدرتی ست که بازیگن هنگا زدن شوت با حفظ دقت شوت به توپ منتقل می‌کند. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۸) Volleys: توانایی در زدن شوت هنگامی توپ در هوا ست. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۹) Crossing: کیفیت سـانتر چه در جریان بازی و در هنگام کرنر یا ضـربه آزاد. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۱۰) Curve: توانایی دادن کات به توپ هنگام پاس و شوت. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۱۱) Free kick: دقت در زدن ضربات آزاد. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۱۲) Long pass: توانایی و دقت دادن پاس بلند. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۱۳) Short pass: توانایی و دقت دادن پاس کوتاه. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۱۴) Vision: میزان آگاهی بازیکن از جایگاه دقیق هم تیمی‌هایش در زمین بازی. بالا بودن این قابلیت احتمال موفقیت پاس‌های بلند رو افزایش می‌دهد. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۱۵) Agility: بیانگر سرعت بازیکن در کنترل توپ می‌باشد. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۱۶) Balance: توانایی حفظ کردن تعادل در تقابل‌های فیزیکی در زمین. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۱۷) Ball control: تولنایی کنترل توپ در هنگام دریافت پاس از هم تیمی‌ها. (در قللب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۱۸) Composure: توانایی تحمل فشار از جانب بازیکن تیم حریف. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۱۹) Dribbling: توانایی دریبل زدن بازیکن تیم حریف. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۲۰) Aggression: جنگندگی. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۲۱) Jumping: توانایی پرش. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۲۲) Stamina: بازیکنی که این مقدار را به میزان بالا در قابلیت‌های فردی خود دارد، انتظار میرود تایم زیادی را بدون خستگی در زمین بازی کند. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۲۳) Strength: قدرت از لحاظ فیزیکی. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۲۴) Defensive awareness: شامل قابلیت جا گیری مدافع و ندادن فرصت با حمله‌کننده در ایجاد خلق موقعیت با پاس یا شوت ناگهانی. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۲۵) Heading accuracy: دقت ضربه سر برای پاس و شوت. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۲۶) Interception: قابلیت قطع توپ. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۲۷) Sliding tackle: بیانگر دقت و زمان‌بندی مناسب در زدن تکل. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۲۸) Gk_diving: قدرت شیرجه دروازه بان. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۲۹) Gk_handling: قدرت گرفت توپ و در دست گرفتن آن. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۳۰) Gk_kicking: بیانگیر قدرت و دقت ضربه دروازه. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۳۱) Gk_positioning: جایگیری دروازه بان در چهارچوب دروازه. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۳۲) Gk_speed: سرعت دروازه بان در مصاف های تک به تک. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۳۳) Gk_reflexes: انعطاف پذیری دروازه بان رد شیرجه زدن برای دفع توپ. (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

۳۴) Age: سن بازیکن

۳۵) Potential growth: میزان رشد احتمالی در آینده

۳۶) Overall: معدل کل توانایی‌های فردی بازیکن (در قالب امتیاز از ۰ تا ۹۹ درجه بندی می‌شود.)

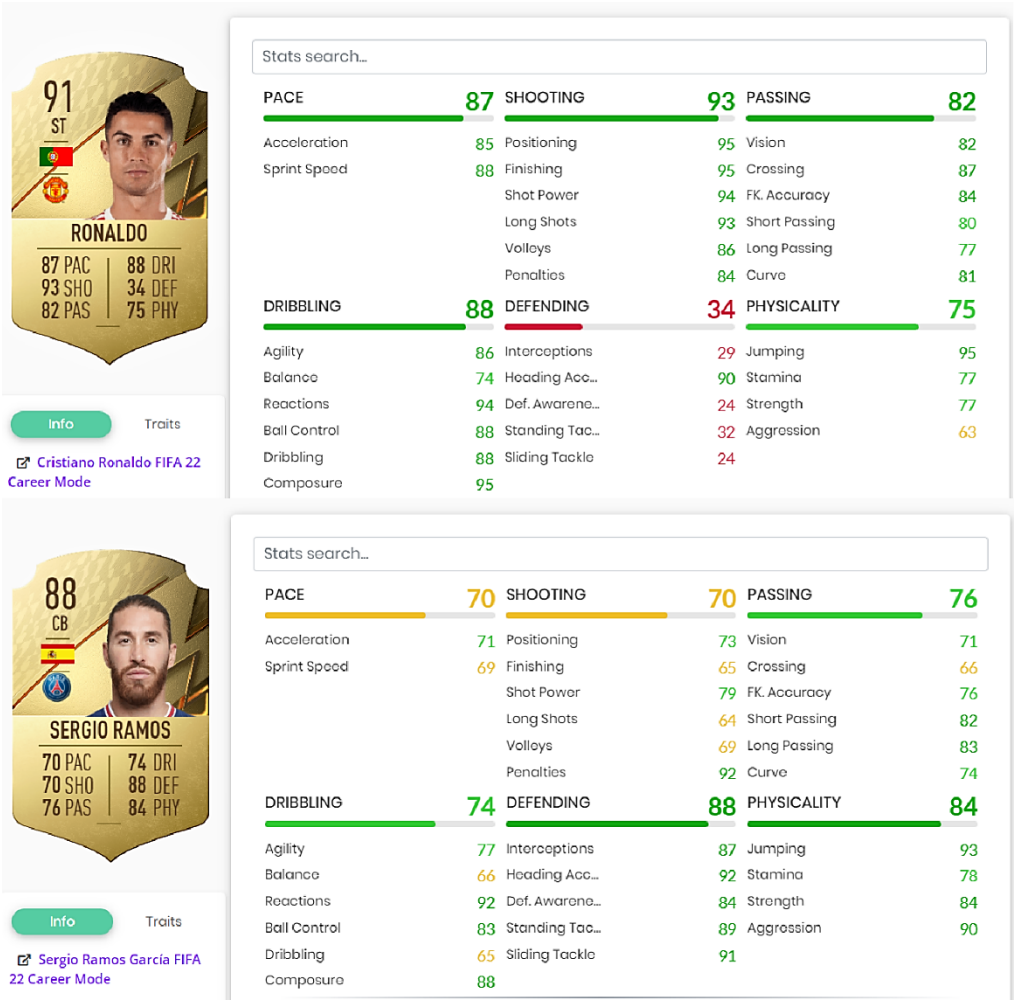
۳۷) International reputation: درجه معروف بودن بازیکن

دو ویژگی دیگر نیز وجود دارند که به عنوان متغیر پاسخ [5]در مسائل رگرسیون و طبقه بندی مورد استفاده قرار می‌گیرند:

۳۸) Position: پست بازیکن در زمین. که به ۴ کلاس دروازه‌بان، دفاع، هافبک، و مهاجم دسته‌بندی می‌شدند، ولی به دلیل محدودیت‌های موجود در روش‌هایی که به آن‌ها مسلط هستیم و نابالانس بودن داده‌ها، کلاس‌دفاع و دروازبان تحت عنوان دفاع- دروازه‌بان، و کلاس مهاجم و هافبک تحت عنوان هافبک-مهاجم در نظر گرفته شد.

۳۹) Value euro: ارزش بازیکنان در بازار نقل انتقالات به یورو.

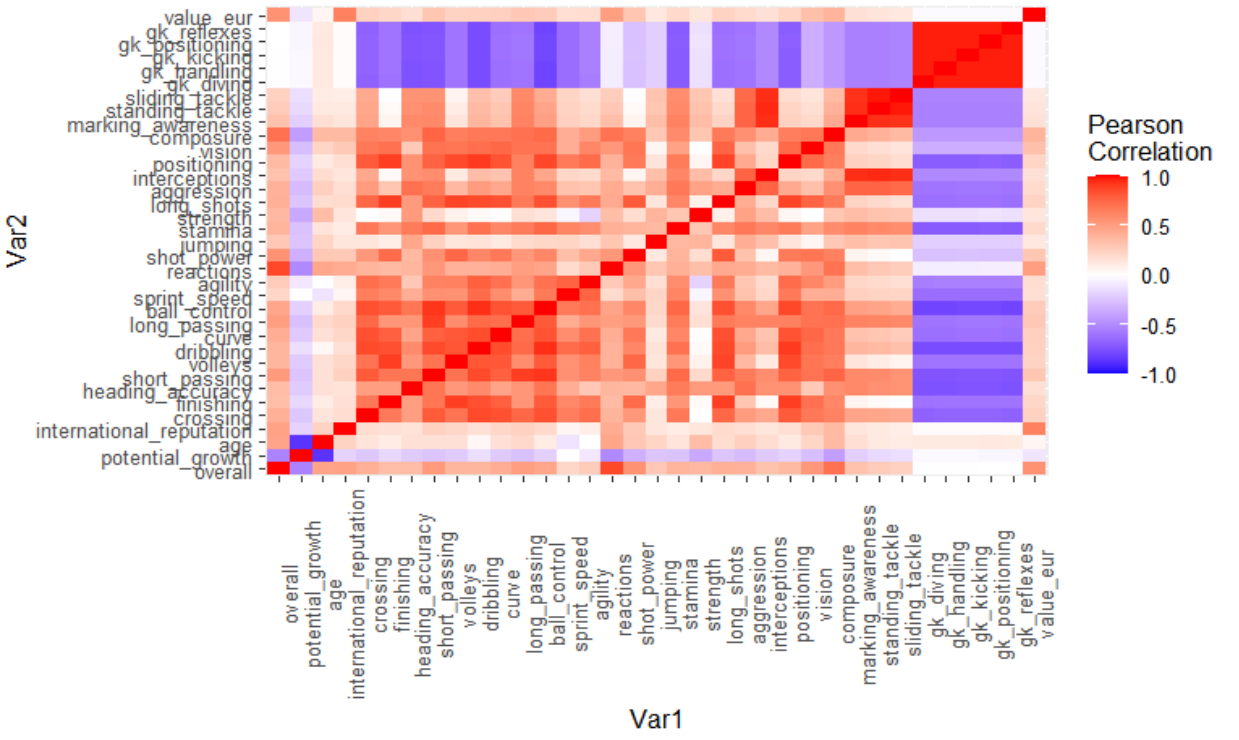
به عنوان مثال توانایی های یک مدافع و یک مهاجم در شکل ۱ [6] آمده است.



شکل ۱: نمونه‌ای از توانایی‌ها دو بازیکن در پست دفاع و مهاجم

[5] Output variable
[6] https://www.futbin.com/22/player/589/sergio-ramos-garcia
https://www.futbin.com/22/player/426/cristiano-ronaldo

به دلیل تعداد بالای ویژگی‌ها و وابستگی تعداد قابل توجهی از آن‌ها به یکدیگر، نمی‌توان تصویر جامعی با کمک روش‌های مبتنی بر تحلیل اکتشافی [7] بدست آورد. تنها تصویر جامع ممکن از این مجموعه داده، مصورسازی دو به دو ویژگی‌ها در مقادیر همبستگی[8] می‌باشد، تا بتوان بدان وسیله در انتخاب متغیرهای مناسب در مدل‌های پیش‌بینی به ما کمک کند.



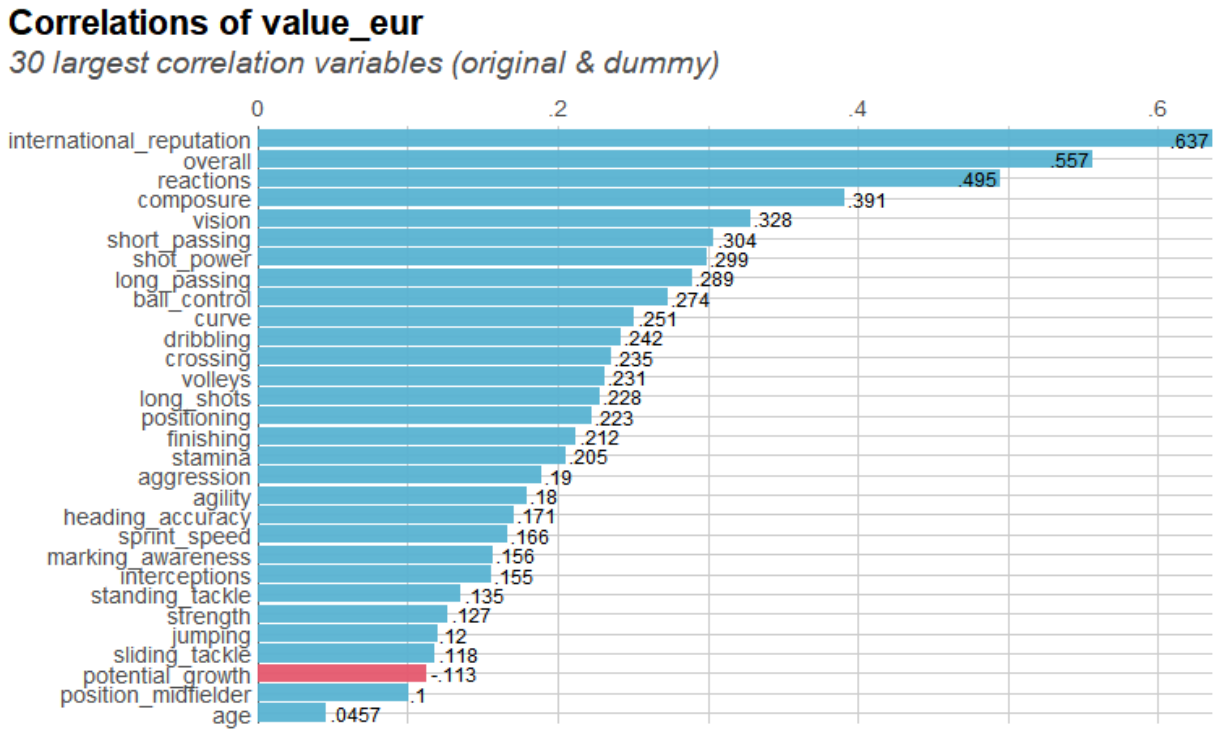
نمودار ۱: نمودار شدت همبستگی بین ویژگی‌ها

آنچه در نمودار ۱ مشـــهود اســـت همبســـتگی تعداد قابل توجهی از ویژگی ها با یکدیگر اســـت. برای مثال قابلیت های دروازه‌بانی رابطه مســـتقیمی با یکدیگر دارند این در حالی اســـت که با ســـایر توانایی ها رابطه عکس دارند زیرا دروازه بانی تفاوت زیادی با سایر پست ها در بازی فوتبال دارد که قابلیت‌های آن نیز تفاوت قابل توجهی با دیگر مهارت ها دارد.

برای ذکر مثال دیگر می‌توان به رابطه سن با پتانسیل رشد اشاره کرد زیرا سن مفید در فوتبال به طور متوسط ۳۳ سال تا ۳۴ سال تخمین زده شده و بازیکنان از این سن به بعد دچار افول می‌شود پس معمولا فرصتی برای رشد مهارت‌های خود نخواهند داشت.

[7] Exploratory data analysis
[8] Correlation

مورد دیگری که می‌توان به آن پرداخت به طور خاص رتبه‌بندی میزان همبستگی ویژگی‌ها با قیمت بازیکن می‌باشد که می‌تواند به طور خاص در مسئله رگرسیون راهگشا باشد.



نمودار ۲: رتبه بندی ویژگی‌ها در همبستگی برای انتخاب ویژگی

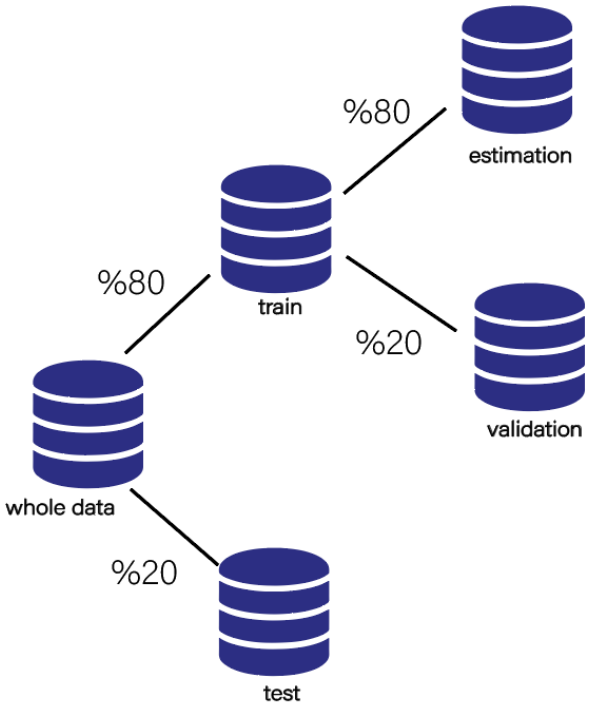
آماده‌سازی داده‌ها:

داده‌های موجود دارای مقداری داده گمشده می‌باشند. که در این مرحله باید پاکسازی شوند. با پاکسازی صورت گرفته تعداد مشاهدات به ۱۹۱۶۵ بازیکن می‌رسد.

همانطور که در شکل ۲ قابل مشاهده است، مرحله بعدی در تقسیم‌بندی داده‌ها، داده‌ها به سه گروه داده اعتبارسنجی[۹]، برآورد [۱۰] و آزمایشی[۱۱] تقسیم‌بندی می‌شوند. به این صورت که ۸۰ درصد داده‌ها به داده‌ها آموزشی[۱۲] و ۲۰ درصد باقی‌مانده به داده‌های آزمایشی تخصیص می‌یابد. همچنین ۸۰ درصد داده‌های تخصیص داده شده به داده‌های آموزشی،

[9] Validation set
[10] Estimation set
[11] Test set
[12] Train set

خود نیز به دو گروه داده تقسیم‌بندی می‌شود که ۸۰ درصد آن به مجموعه داده برآورد و ۲۰ درصد آن به مجموعه داده اعتبارسنجی تخصیص می‌یابد.



شکل ۲: شماتیک تقسیم‌بندی داده‌ها

همچنین در مراحل پیش رو نیاز به استفاده از داده‌های نرمال سازی شده داریم (در روش k نزدیک ترین همسایه[۱۳]). به همین جهت داده‌های برآورد را با استفاده از میانگین و انحراف معیار بدست آمده از خود آن داده‌ها نرمال‌سازی می‌کنیم و داده‌های اعتبارسنجی را با کمک میانگین و انحراف معیار داده‌های برآورد نرمال‌سازی[۱۴] می‌کنیم.

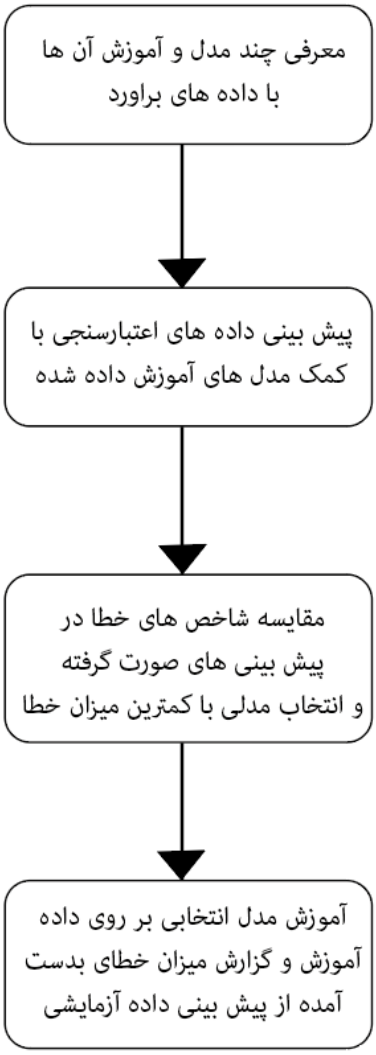
برای نرمال‌سازی داده‌های آزمایشی باید ابتدا داده‌های آموزشی را نرمال سازی کرده سپس داده‌های آزمایشی را با استفاده از میانگین و انحراف معیار داده‌های آموزشی نرمال‌سازی می‌کنیم.

و همانطور که در توضیح ویژگی موقعیت بازیکن گفتیم، مجبور به کاهش تعداد کلاس‌های مان برای مسئله طبقه‌بندی شدیم.

[13] K-nearest-neighbor
[14] Normalization

پیاده‌سازی روش‌ها و ارزیابی:

به طور کلی در بدست آوردن مدل بهینه در مسائل یادگیری ماشین، گام‌های ذکر شده در شکل ۳ طی می‌شود.



شکل ۳: فلوچارت مراحل گزینش مدل بهینه با کمترین میزان خطا

این گام‌ها به تفصیل در دو پروژه رگرسیون و طبق بندی انجام شده.

معیار ارزیابی در مسئله رگرسیون:

مدل‌های رگرسیون مجذور مربعات خطا [15]می‌باشد. که به صورت زیر می‌باشد.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y_j})^2}$$

معیار ارزیابی [16]در مسئله طبقه بندی:

در مسائل طبقه‌بندی مختلف زمانی که با دو کلاس مواجه هستیم شاخص‌های ارزیابی مختلفی برای مدل مطرح می‌گردد نظیر sensitivity، specificity،precision، accuracy و غیره. موضوعی که در رابطه با این شاخصه‌های ارزیابی مطرح می‌شود این است که این‌ها زمانی مفید واقع می‌شود که اشتباه یادرست پیش‌بینی کردن یک کلاس نسبت به کلاس دیگر ارجحیت داشته باشد. برای مثال در پیش‌بینی بیماری سرطان به اشتباه سالم تشخیص دادن بیمار به مراتب تأثیرات مخرب‌تری دارد نسبت به حالتی که به اشتباه فرد سالم رو بیمار تشخیص بدهیم و می‌تواند موجب وقوع اتفاقات جبران ناپذیری شود. در نتیجه شاخص ارزیابی ما sensitivity می‌باشد. و سعی می‌شود با پایین بردن آستانه مقدار این شاخص را افزایش دهیم.

اما در اینجا ما دو کلاس داریم یک کلاس مدافع-دروازه بان و کلاس هافبک-مهاجم اهمیت هر دوی این دو کلاس در پیش‌بینی یکسان می‌باشد و ارجحیتی نسبت به یکدیگر ندارند. لذا ترجیح ما این است که شاخص accuracy را ملاک ارزیابی قرار دهیم.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}$$

رگرسیون:

۱) مدل رگرسیون خطی:

جزئیات مدل‌های آموزش داده شده در جدول ۱ آمده است.

[15] Root mean squared error
[16] Evaluation metric

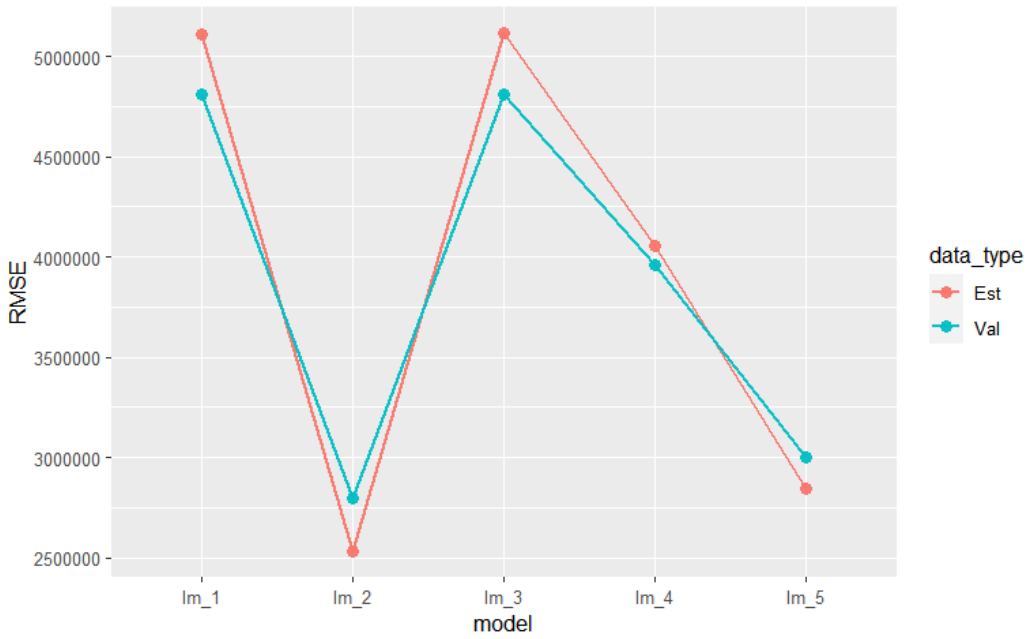
جدول ۱: توضیحات مدل‌های برازش داده شده‌ی رگرسیون خطی روی داده‌های برآورد

شماره مدل	متغیر های مورد استفاده	توضیحات مدل
۱	همه متغیرها	-
۲	همه متغیرها	اثرات متقابل
۳	همه متغیرها	دستور گام برای انتخاب متغیر با روش دو طرفه
۴	تعدادی از متغیرها	استفاده از دستور poly(x,2) برای تمای متغیرهای انتخاب شده
۵	تعدادی از متغیرها	استفاده از دستور poly(x,2) برای تمامی متغیرهای انتخاب شده

مقادیر مجذور مربعات خطا مدل‌ها بر روی داده‌های اعتبارسنجی مطابق جدول ۲ و نمودار ۳ می‌باشد.

جدول ۲: میزان خطای مجذور مربعات خطای مدل‌های رگرسیون در خطی در بیش‌بینی داده‌های اعتبارسنجی

شماره مدل	RMSE-validation
۱	۴۸۱۳۸۸۵
۲	۲۷۹۷۶۰۶
۳	۴۸۱۱۷۶۳
۴	۳۹۶۰۹۳۹
۵	۳۰۰۰۲۶۶



نمودار ۳: نمودار میزان خطای مدل‌های رگرسیون خطی روی داده‌های اعتبارسنجی و برآورد

۲) مدل K نزدیک‌ترین همسایه:

مدل را به ازای مقادیر مختلف k روی داده‌های برآورد استاندارد شده و غیر استاندارد شده برازش دادیم. که جزئیات مدل‌های آموزش داده شده در جدول ۳ آمده است.

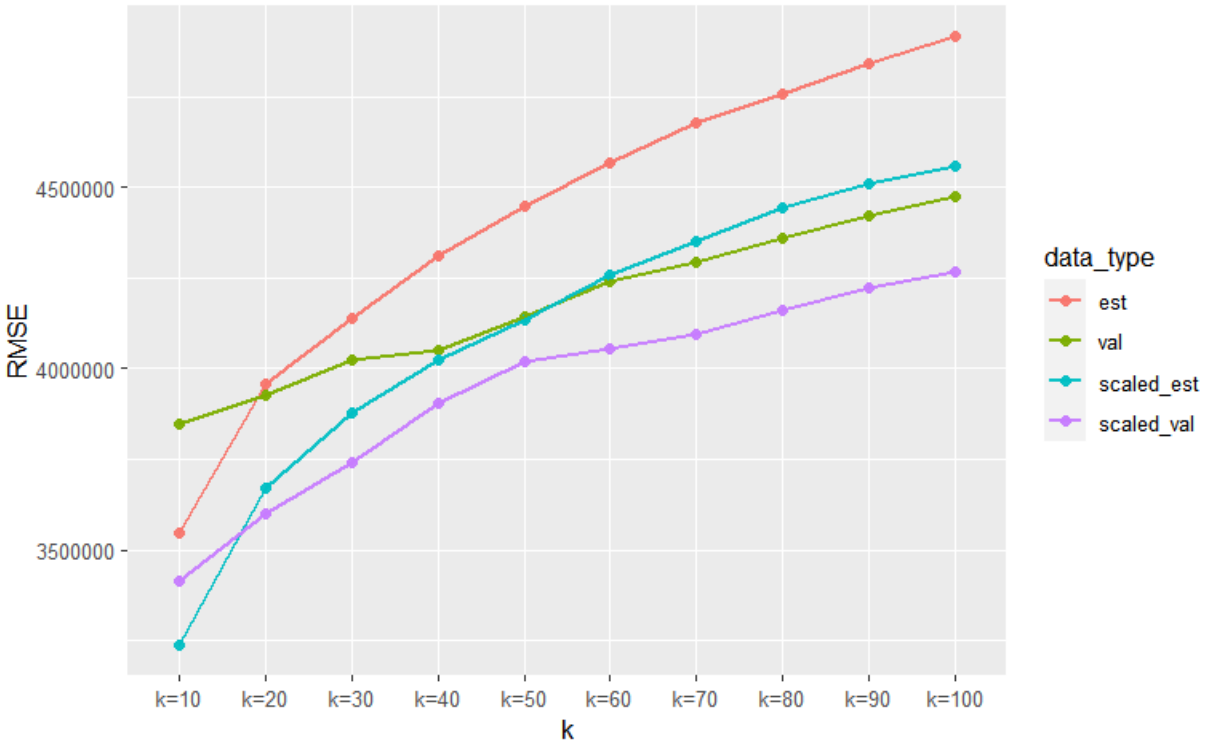
جدول ۳: توضیحات مربوط به مدل‌های برازش داده شده‌ی کا نزدیک‌ترین همسایه روی داده‌های برآورد

K	متغیرهای مورد استفاده	توضیحات مدل
۱۰	همه متغیرها	داده استاندارد شده
۲۰	همه متغیرها	داده استاندارد شده
۳۰	همه متغیرها	داده استاندارد شده
۴۰	همه متغیرها	داده استاندارد شده
۵۰	همه متغیرها	داده استاندارد شده
۶۰	همه متغیرها	داده استاندارد شده
۷۰	همه متغیرها	داده استاندارد شده
۸۰	همه متغیرها	داده استاندارد شده
۹۰	همه متغیرها	داده استاندارد شده
۱۰۰	همه متغیرها	داده استاندارد شده
۱۰	همه متغیرها	داده غیر استاندارد
۲۰	همه متغیرها	داده غیر استاندارد
۳۰	همه متغیرها	داده غیر استاندارد
۴۰	همه متغیرها	داده غیر استاندارد
۵۰	همه متغیرها	داده غیر استاندارد
۶۰	همه متغیرها	داده غیر استاندارد
۷۰	همه متغیرها	داده غیر استاندارد
۸۰	همه متغیرها	داده غیر استاندارد
۹۰	همه متغیرها	داده غیر استاندارد
۱۰۰	همه متغیرها	داده غیر استاندارد

مقادیر مجذور مربعات خطا مدل‌ها بر روی داده‌های اعتبارسنجی استاندارد شده و غیر استاندارد مطابق جدول ۴ و نمودار ۴ می‌باشد.

جدول ۴: مقادیر مجذور مربعات خطای مدلهای کا نزدیک ترین همسایه روی دادهای اعتبارسنجی استاندارد شده و غیر استاندارد

RMSE (validation-scaled)	RMSE (validation-non-scaled)	k
۳۴۱۵۴۶۲	۳۸۴۷۷۱۱	۱۰
۳۵۹۹۹۶۲	۳۹۲۸۱۴۳	۲۰
۳۷۴۰۹۷۵	۴۰۲۴۴۰۴	۳۰
۳۹۰۴۸۸۸	۴۰۵۲۳۳۵	۴۰
۴۰۱۹۱۲۸	۴۱۴۳۶۰۱	۵۰
۴۰۵۶۸۵۷	۴۲۳۹۰۵۶	۶۰
۴۰۹۴۸۸۴	۴۲۹۳۶۵۶	۷۰
۴۱۶۲۸۳۶	۴۳۵۹۵۱۶	۸۰
۴۲۲۴۷۶۱	۴۴۲۰۴۰۹	۹۰
۴۲۶۷۸۷۲	۴۴۷۲۴۶۱	۱۰۰



نمودار ۴: مقادیر مجذور مربعات خطا برای مدل مدل‌های کا نزدیک‌ترین همسایه بر روی داده های برآورد و اعتبارسنجی استاندارد شده و غیر استاندارد

۳) مدل درخت رگرسیون[17]:

مدل درخت رگرسیون را به ازای مقادیر cp و minsplit روی داده برآورد آموزش دادیم. در مدل درخت رگرسیون انشعاباتی با برش روی ویژگی‌های مختلف ایجاد می‌شود. برای متوقف شدن هر یک از این برش‌ها باید یکی از دو شرط ذیل برقرار کردد. اولا این که با ایجاد برش آماره R2 کمتر از مقدار cp افزایش یابد. دوما اینکه تعداد داده‌هایی که می‌خواهیم روی آن برش بزنیم تعدادشان کوچکتر مساوی minsplit باشد.

ویژگی مدل‌های برازش[18] داده شده مطابق جدول ۵ می‌باشد.

جدول ۵: توضیحات مربوط به مدل‌های برازش داده شده‌ی درخت رگرسیون روی داده‌های برآورد

minsplit	cp	متغیرهای استفاده شده
۱۰	۰.۱	همه متغیرها
۴۰	۰.۱	همه متغیرها
۷۰	۰.۱	همه متغیرها
۱۰۰	۰.۱	همه متغیرها
۱۰	۰.۰۱	همه متغیرها
۴۰	۰.۰۱	همه متغیرها
۷۰	۰.۰۱	همه متغیرها
۱۰۰	۰.۰۱	همه متغیرها
۱۰	۰.۰۰۱	همه متغیرها
۴۰	۰.۰۰۱	همه متغیرها
۷۰	۰.۰۰۱	همه متغیرها
۱۰۰	۰.۰۰۱	همه متغیرها

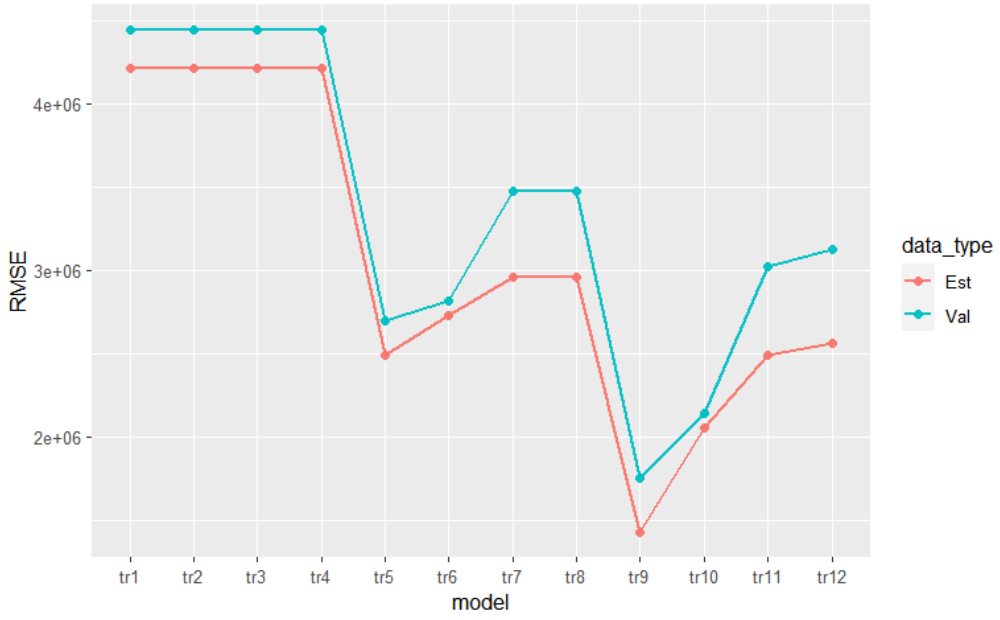
مقادیر مجذور مربعات خطا مطابق جدول ۶ و نمودار ۵ می‌باشد.

جدول ۶: مقادیر مجذور مربعات خطا برای مدل‌های درخت رگرسیون در بیش‌بینی داده‌های اعتبارسنجی

minsplit	cp	RMSE-validation
۱۰	۰.۱	۴۴۳۸۷۳۵
۴۰	۰.۱	۴۴۳۸۷۳۵
۷۰	۰.۱	۴۴۳۸۷۳۵
۱۰۰	۰.۱	۴۴۳۸۷۳۵
۱۰	۰.۰۱	۲۶۹۴۴۴۳

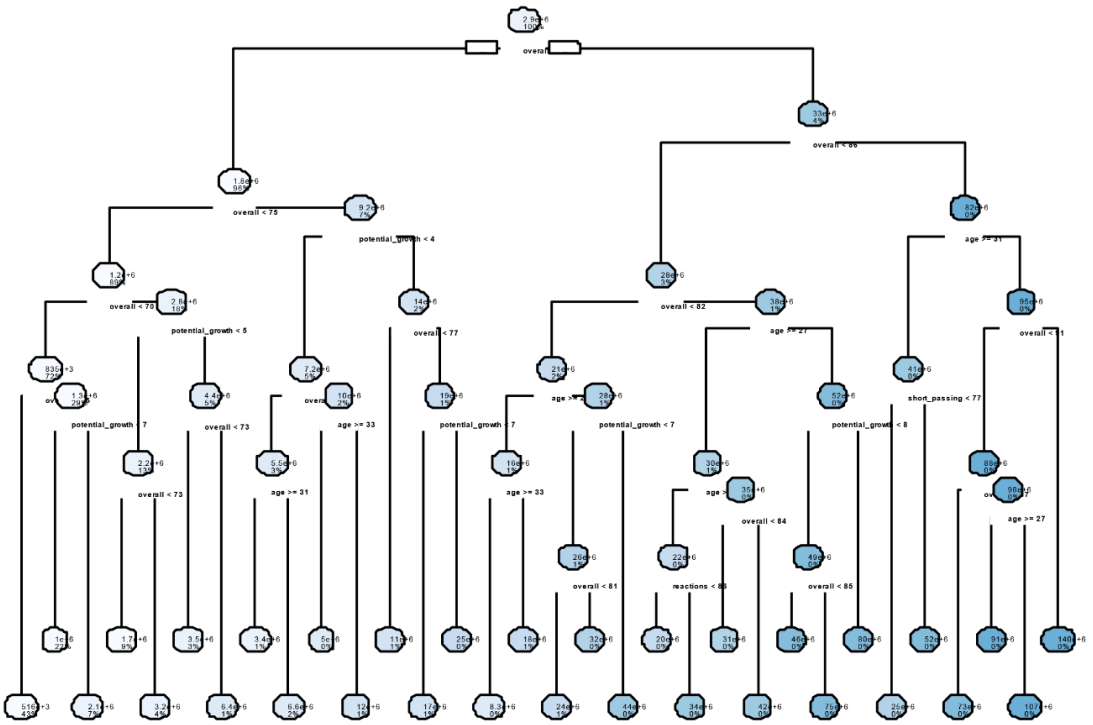
[17] Regression tree
[18] fit

۴۰	۰.۰۱	۲۸۱۳۳۳۵
۷۰	۰.۰۱	۳۴۷۸۸۲۰
۱۰۰	۰.۰۱	۳۴۷۸۸۲۰
۱۰	۰.۰۰۱	۱۷۵۴۵۴۲
۴۰	۰.۰۰۱	۲۱۴۰۳۷۷
۷۰	۰.۰۰۱	۳۰۲۲۴۲۸
۱۰۰	۰.۰۰۱	۳۱۲۷۸۷۸



نمودار ۵: مقادیر مجذور مربعات خطا برای مدل‌های درخت رگرسیون در بیش‌بینی داده‌های اعتبارسنجی و برآورد

با توجه به نمودار ۵ و جدول ۶، مدل درخت شماره ۹ (cp = 0.001 , minsplit = 10) با کمترین مجذور مربعات خطا، مدل انتخابی ما از بین درخت‌ها می‌باشد. نمودار انشعابات مدل درخت رگرسیون انتخابی مطابق نمودار ۷ می‌باشد.

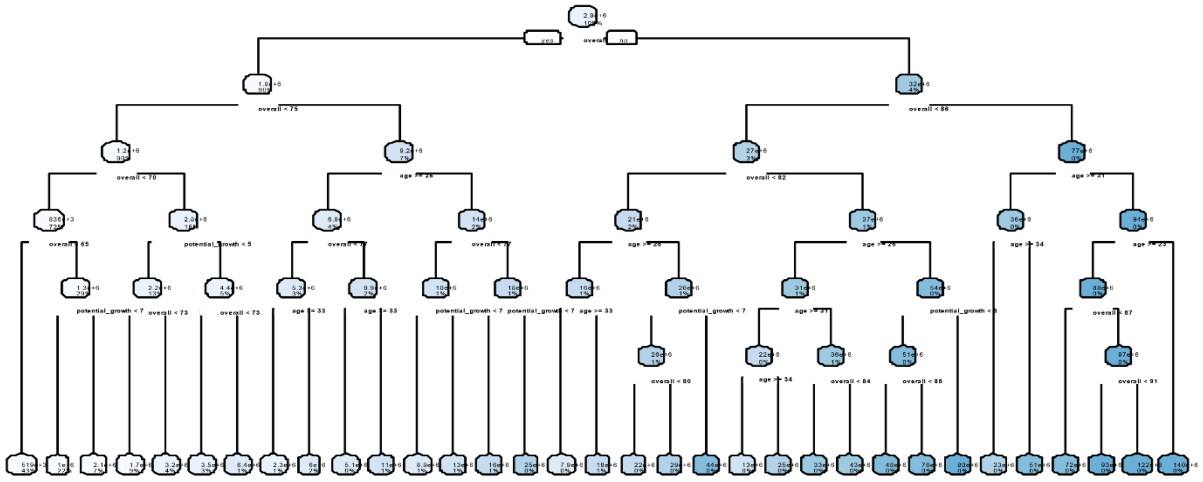


نمودار ۷: نمودار انشعابات درخت برازش داده‌شده (روی داده‌های برآورد) انتخابی از بین مدل‌های مدل‌های درخت رگرسیون

جمع‌بندی نهایی برای مدل‌های رگرسیون:

از بین تمامی مدل‌های انتخابی، مدل درخت رگرسیون (cp = 0.001 , minsplit = 10) با میزان مجذور مربعات خطای ۱۷۵۴۵۴۲ به عنوان بهترین مدل انتخاب گردید.

حال مدل انتخابی را روی مجموعه داده آموزش برازش می‌دهیم و مقدار مجذور مربعات خطای بدست آمده برابر با ۲۱۴۳۰۷۹ می‌باشد. درخت برازش داده شده مطابق نمودار ۸ است.



نمودار ۸: نمودار انشعابات درخت رگرسیون (مربوط به مدل منتخب که روی داده‌های آموزش برازش داده شده است.)

طبقه بندی:

۱) رگرسیون لجستیک:

جزئیات مدل‌های آموزش داده شده در جدول ۷ آمده است.

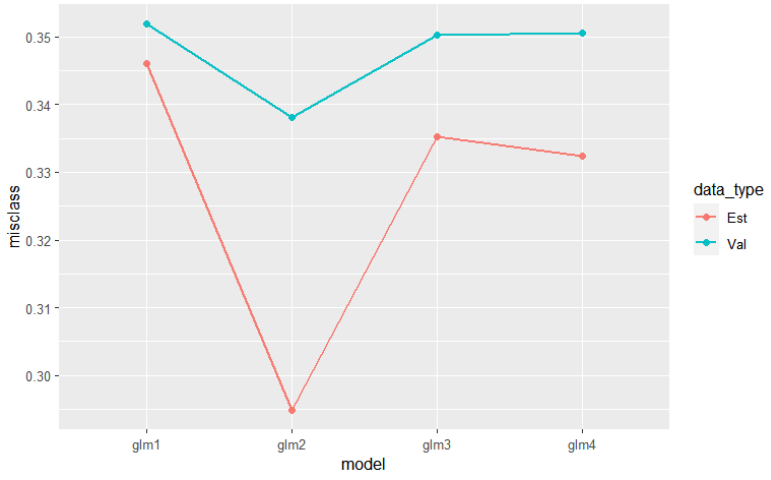
جدول ۷: توضیحات مدل‌های رگرسیون لجستیک برازش داده شده روی داده‌های برآورد

شماره مدل	متغیرهای مورد استفاده	توضیحات مدل
۱	همه متغیرها	-
۲	همه متغیرها	اثرات متقابل
۳	تعدادی از متغیرها	استفاده از دستور poly(x,2) برای تمای متغیرهای انتخاب شده
۴	تعدادی از متغیرها	استفاده از دستور poly(x,3) برای تمای متغیرهای انتخاب شده

مقادیر خطای طبقه‌بندی[19] مدل‌های برازش داده شده در پیش‌بینی داده‌های اعتبارسنجی مطابق جدول ۸ و نمودار ۹ آمده است.

جدول ۸: مقادیر خطای طبقه بندی مدل‌های رگرسیون لجستیک در پیش‌بینی داده‌های اعتبارسنجی

شماره مدل	خطای طبقه بندی (اعتبار سنجی)
۱	۰.۳۵۱۸۰۹۶
۲	۰.۳۳۸۱۱۵۴
۳	۰.۳۵۰۱۷۹۳
۴	۰.۳۵۰۵۰۵۴



نمودار ۹: مقادیر خطای طبقه بندی مدل‌های رگرسیون لجستیک در پیش‌بینی داده‌های برآورد و اعتبارسنجی

[19] Misclassification error

۲) مدل تحلیل تفکیکی خطی[۲۰]:

جزئیات مدل‌های آموزش داده شده به شرح زیر در جدول ۹ بیان شده است.

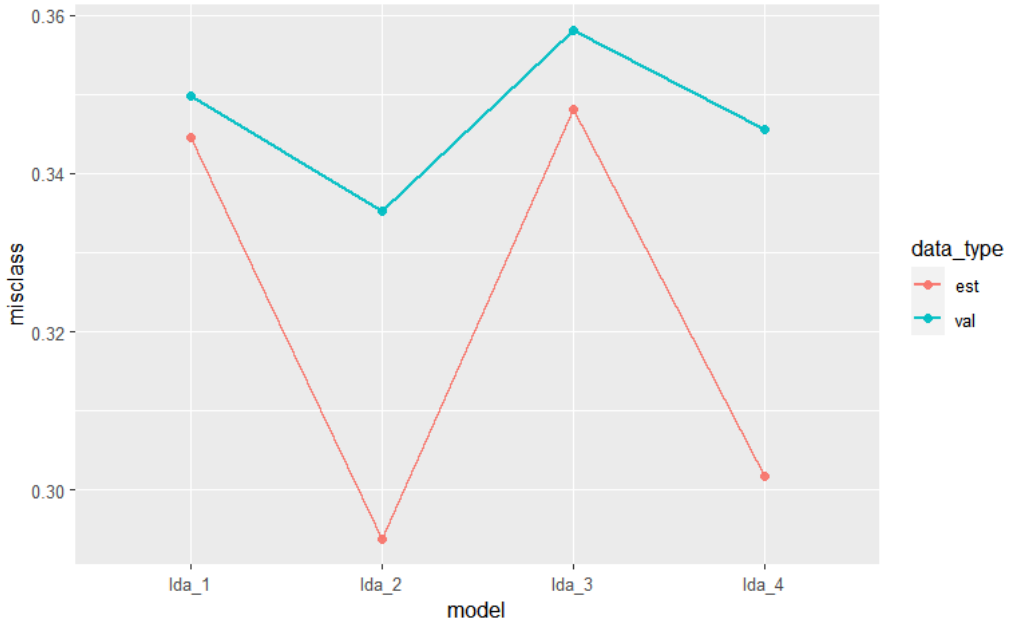
جدول ۹: توضیحات مدل‌های تحلیل تفکیکی خطی برازش داده شده روی داده‌های برآورد

شماره مدل	متغیرهای مورد استفاده	توضیحات مدل
۱	همه متغیرها	برآورد احتمال پیشین از روی داده برآورد
۲	همه متغیرها	اثرات متقابل، poly(x,2) و برآورد احتمال پیشین از روی داده برآورد
۳	همه متغیرها	احتمال پیشین یکنواخت
۴	همه متغیرها	اثرات متقابل، poly(x,2) و احتمال پیشین یکنواخت

مقادیر خطای طبقه‌بندی مدل‌های برازش داده شده در پیش‌بینی داده‌های اعتبار سنجی مطابق جدول ۱۰ و نمودار ۱۰ است.

جدول ۱۰: : مقادیر خطای طبقه بندی مدل‌های تحلیل تفکیکی خطی در پیش‌بینی داده‌های اعتبارسنجی

شماره مدل	خطای طبقه بندی(اعتبارسنجی)
۱	۰.۳۴۹۸۵۳۳
۲	۰.۳۳۵۱۸۱۰
۳	۰.۳۵۸۰۰۴۶
۴	۰.۳۴۵۶۱۴۶



نمودار ۱۰: مقادیر خطای طبقه بندی مدل‌های تحلیل تفکیکی خطی در پیش‌بینی داده‌های برآورد و اعتبارسنجی

[20] Linear discriminant analysis

۳) مدل تحلیل تفکیکی کوآدراتیک[21]:

جزئیات مدل‌های آموزش داده شده به شرح جدول ۱۱ است.

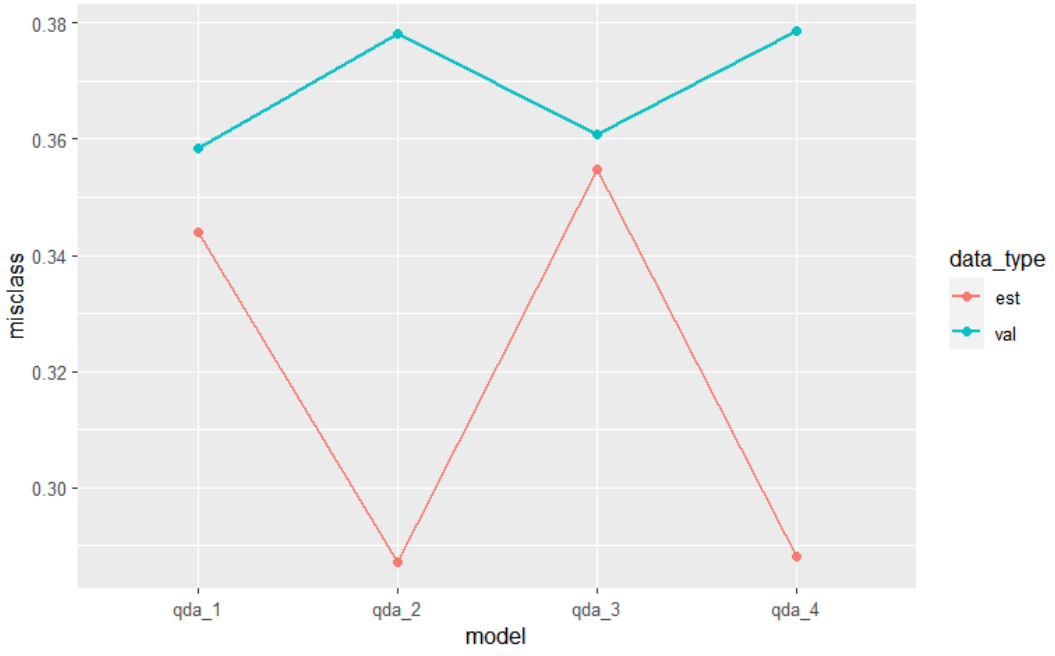
جدول ۱۱: توضیحات مدل‌های تحلیل تفکیکی کوآدراتیک در پیش‌بینی داده‌های اعتبارسنجی

شماره مدل	متغیرهای مورد استفاده	توضیحات مدل
۱	همه متغیرها	برآورد احتمال پیشین از روی داده برآورد
۲	همه متغیرها	اثرات متقابل، poly(x,2) و برآورد احتمال پیشین از روی داده برآورد
۳	همه متغیرها	احتمال پیشین یکنواخت
۴	همه متغیرها	اثرات متقابل، poly(x,2) و احتمال پیشین یکنواخت

مقادیر خطای طبقه‌بندی مدل‌های برازش داده شده در پیش‌بینی داده‌های اعتبارسنجی مطابق جدول ۱۲ و نمودار ۱۱ است.

جدول ۱۲: مقادیر خطای طبقه بندی مدل‌های تحلیل تفکیکی کوآدراتیک در پیش‌بینی داده‌های اعتبارسنجی

شماره مدل	خطای طبقه بندی(اعتبار سنجی)
۱	۰.۳۵۸۳۳۰۶
۲	۰.۳۷۸۲۱۹۸
۳	۰.۳۶۰۹۳۹۰
۴	۰.۳۷۸۵۴۵۸



نمودار ۱۱: مقادیر خطای طبقه بندی مدل‌های تحلیل تفکیکی کوآدراتیک در پیش‌بینی داده‌های اعتبارسنجی و برآورد

[21] Quadratic discriminant analysis

مشـاهده می‌شـود که مدل تحلیل تفکیکی کوآدراتیک دچار پیش برازش شـده اسـت. و مدل مناسـبی برای طبقه‌بندی داده‌های ما نیست.

۴) مدل بیز ساده[۲۲]:

با مسـتقل در نظر گرفتن ویژگی‌ها نسـبت به یکدیگر مدل تحلیل تفکیکی خطی به مدل بیز سـاده تغییر فرم می‌دهد در اینجا ما یک مدل بیز سـاده را با درنظر گرفتن تمام ویژگی‌ها روی داده برآورد برازش دادیم. که مقدار خطای طبقه‌بندی آن روی مجموعه داده اعتبارسنجی ۰.۴۴۴۷۳۴۳ شده است.

۵) مدل درخت تصمیم[۲۳]:

مدل‌های درخت طبقه بندی را با مقادیر cp برابر با ۰.۱ و ۰.۰۱ و ۰.۰۰۱ با minsplit های بین ۱۰ تا ۲۰۰ با گام ۱۰ بر روی داده‌های برآورد برازش داده‌ایم که در مجموع ۶۰ مدل به دسـت آمده که در جدول ۱۳ قابل مشـاهده می باشـد. همچنین نمودار ها به تفکیک cp در نمودار ۱۲ قابل مشاهده است.

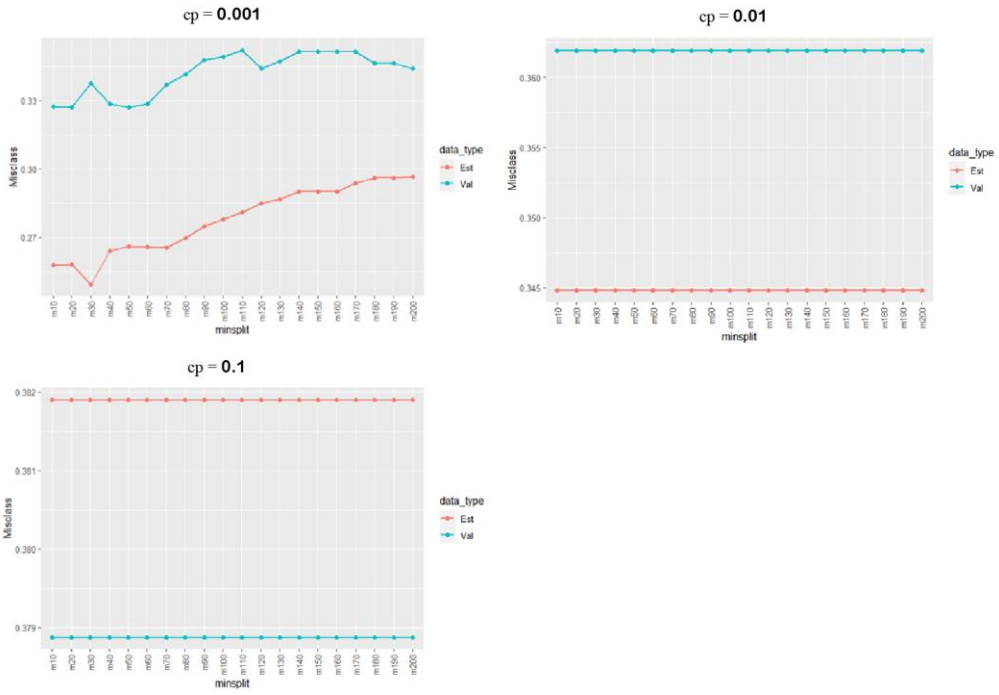
جدول ۱۳: مقادیر خطای طبقه بندی مدل‌های درخت تصمیم در پیش‌بینی داده‌های اعتبارسنجی

minsplit	cp	خطای طبقه بندی
۱۰	۰.۱	۰.۳۷۸۸۷۱۹
۲۰	۰.۱	۰.۳۷۸۸۷۱۹
۳۰	۰.۱	۰.۳۷۸۸۷۱۹
۴۰	۰.۱	۰.۳۷۸۸۷۱۹
۵۰	۰.۱	۰.۳۷۸۸۷۱۹
۶۰	۰.۱	۰.۳۷۸۸۷۱۹
۷۰	۰.۱	۰.۳۷۸۸۷۱۹
۸۰	۰.۱	۰.۳۷۸۸۷۱۹
۹۰	۰.۱	۰.۳۷۸۸۷۱۹
۱۰۰	۰.۱	۰.۳۷۸۸۷۱۹
۱۱۰	۰.۱	۰.۳۷۸۸۷۱۹
۱۲۰	۰.۱	۰.۳۷۸۸۷۱۹
۱۳۰	۰.۱	۰.۳۷۸۸۷۱۹
۱۴۰	۰.۱	۰.۳۷۸۸۷۱۹
۱۵۰	۰.۱	۰.۳۷۸۸۷۱۹

[22] Naïve Bayes
[23] Decision tree

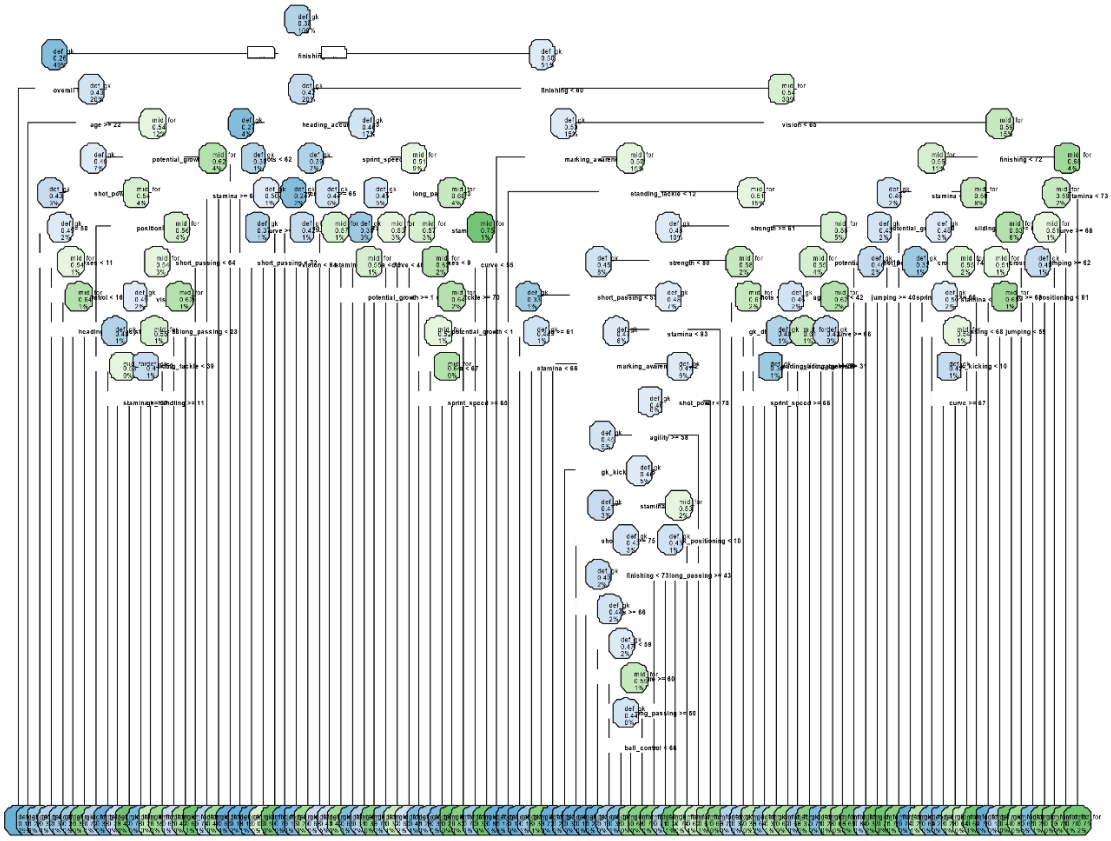
۱۶۰	۰.۱	۰.۳۷۸۸۷۱۹
۱۷۰	۰.۱	۰.۳۷۸۸۷۱۹
۱۸۰	۰.۱	۰.۳۷۸۸۷۱۹
۱۹۰	۰.۱	۰.۳۷۸۸۷۱۹
۲۰۰	۰.۱	۰.۳۷۸۸۷۱۹
۱۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۲۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۳۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۴۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۵۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۶۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۷۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۸۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۹۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۱۰۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۱۱۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۱۲۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۱۳۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۱۴۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۱۵۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۱۶۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۱۷۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۱۸۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۱۹۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۲۰۰	۰.۰۱	۰.۳۶۱۹۱۷۲
۱۰	۰.۰۰۱	۰.۳۲۷۳۵۵۷
۲۰	۰.۰۰۱	۰.۳۲۷۰۲۹۷
۳۰	۰.۰۰۱	۰.۳۳۷۴۶۳۳
۴۰	۰.۰۰۱	۰.۳۲۸۶۵۹۹
۵۰	۰.۰۰۱	۰.۳۲۷۰۲۹۷
۶۰	۰.۰۰۱	۰.۳۲۸۶۵۹۹
۷۰	۰.۰۰۱	۰.۳۳۷۱۳۷۳
۸۰	۰.۰۰۱	۰.۳۴۱۷۰۲۰
۹۰	۰.۰۰۱	۰.۳۴۷۸۹۷۰
۱۰۰	۰.۰۰۱	۰.۳۴۹۲۰۱۲
۱۱۰	۰.۰۰۱	۰.۳۵۲۱۳۵۶

۱۲۰	۰.۰۰۱	۰.۳۴۳۹۸۴۳
۱۳۰	۰.۰۰۱	۰.۳۴۷۲۴۴۹
۱۴۰	۰.۰۰۱	۰.۳۵۱۴۸۳۵
۱۵۰	۰.۰۰۱	۰.۳۵۱۴۸۳۵
۱۶۰	۰.۰۰۱	۰.۳۵۱۴۸۳۵
۱۷۰	۰.۰۰۱	۰.۳۵۱۴۸۳۵
۱۸۰	۰.۰۰۱	۰.۳۴۶۲۶۶۷
۱۹۰	۰.۰۰۱	۰.۳۴۶۲۶۶۷
۲۰۰	۰.۰۰۱	۰.۳۴۳۹۸۴۳



نمودار ۱۲: مقادیر خطای طبقه بندی مدل‌های درخت تصمیم در پیش‌بینی داده‌های اعتبارسنجی و برآورد

درخت بدست آمده از برازش بر روی داده‌های برآورد با (cp = 0.001 و minsplit = 50) در نمودار ۱۳ آمده است.



نمودار ۱۳: نمودار انشعابات درخت تصمیم برازش داده‌شده (روی داده‌های برآورد) انتخابی از بین مدل‌های مدل‌های درخت تصمیم

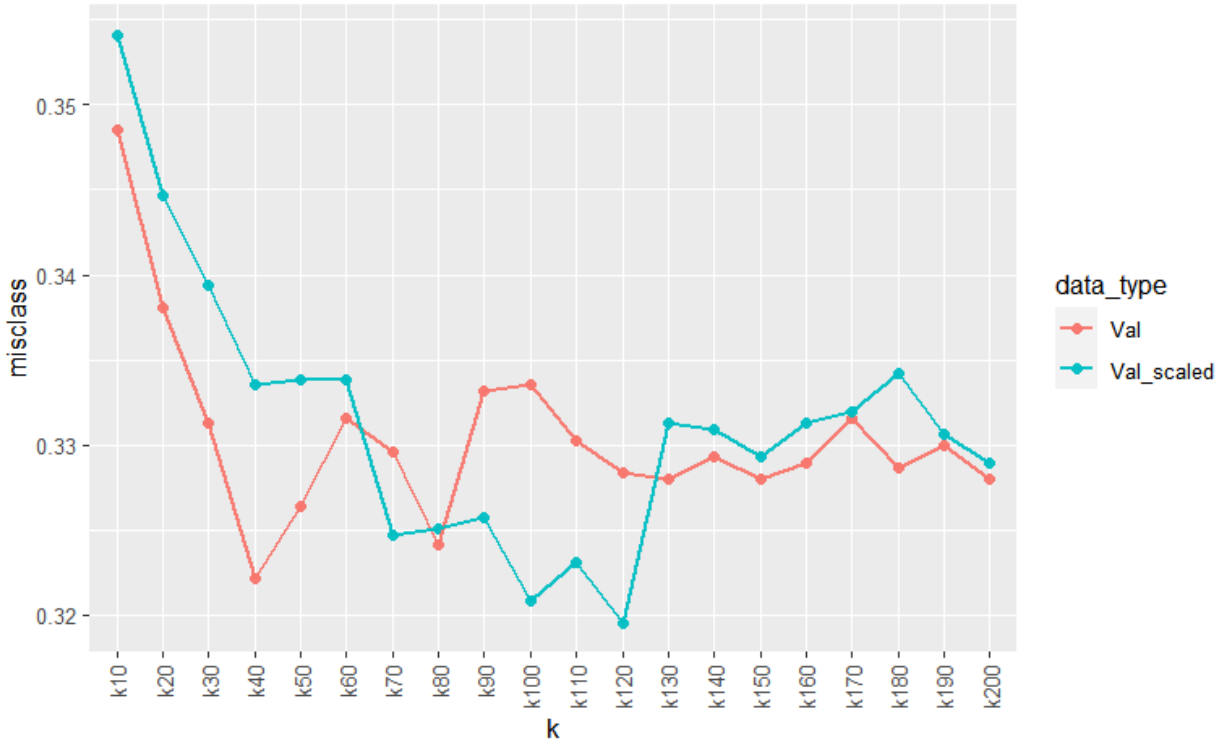
۵) مدل K نزدیک ترین همسایه:

مدل های k نزدیکترین همسایه را روی داده‌های برآورد با مقادیر k ۱۰ تا ۲۰۰ با در نظر گرفتن تمامی ویژگی‌ها برازش داده‌ایم. مقدار خطای طبقه بندی برای هر یک از این مدل ها در جدول ۱۴ و نمودار ۱۴ برای دو گروه داده اعتبارسنجی استاندارد شده و غیر استاندارد به دست آمده که به شرح زیر است.

جدول ۱۴: مقادیر خطای طبقه بندی مدل‌های درخت تصمیم در پیش‌بینی داده‌های اعتبارسنجی استاندارد شده و غیر استاندارد

k	خطای طبقه بندی اعتبار سنجی غیر استاندار	خطای طبقه بندی اعتبار سنجی استاندارد شده
۱۰	۰.۳۴۸۵۴۹۱	۰.۳۵۴۰۹۱۹
۲۰	۰.۳۳۸۱۱۵۴	۰.۳۴۴۶۳۶۵
۳۰	۰.۳۳۱۲۶۸۳	۰.۳۳۹۴۱۹۶
۴۰	۰.۳۲۲۱۳۸۹	۰.۳۳۳۵۵۰۷
۵۰	۰.۳۲۶۳۷۷۶	۰.۳۳۳۸۷۶۸

۶۰	۰.۳۳۱۵۹۴۴	۰.۳۳۳۸۷۶۸
۷۰	۰.۳۲۹۶۳۸۱	۰.۳۲۴۷۴۷۳
۸۰	۰.۳۲۴۰۹۵۲	۰.۳۲۵۰۷۳۴
۹۰	۰.۳۳۳۲۲۴۶	۰.۳۲۵۷۲۵۵
۱۰۰	۰.۳۳۳۵۵۰۷	۰.۳۲۰۸۳۴۷
۱۱۰	۰.۳۳۰۲۹۰۲	۰.۳۲۳۱۱۷۱
۱۲۰	۰.۳۲۸۳۳۳۹	۰.۳۱۹۵۳۰۵
۱۳۰	۰.۳۲۸۰۰۷۸	۰.۳۳۱۲۶۸۳
۱۴۰	۰.۳۲۹۳۱۲۰	۰.۳۳۰۹۴۲۳
۱۵۰	۰.۳۲۸۰۰۷۸	۰.۳۲۹۳۱۲۰
۱۶۰	۰.۳۲۸۹۸۶۰	۰.۳۳۱۲۶۸۳
۱۷۰	۰.۳۳۱۵۹۴۴	۰.۳۳۱۹۲۰۴
۱۸۰	۰.۳۲۸۶۵۹۹	۰.۳۳۴۲۰۲۸
۱۹۰	۰.۳۲۹۹۶۴۱	۰.۳۳۰۶۱۶۲
۲۰۰	۰.۳۲۸۰۰۷۸	۰.۳۲۸۹۸۶۰



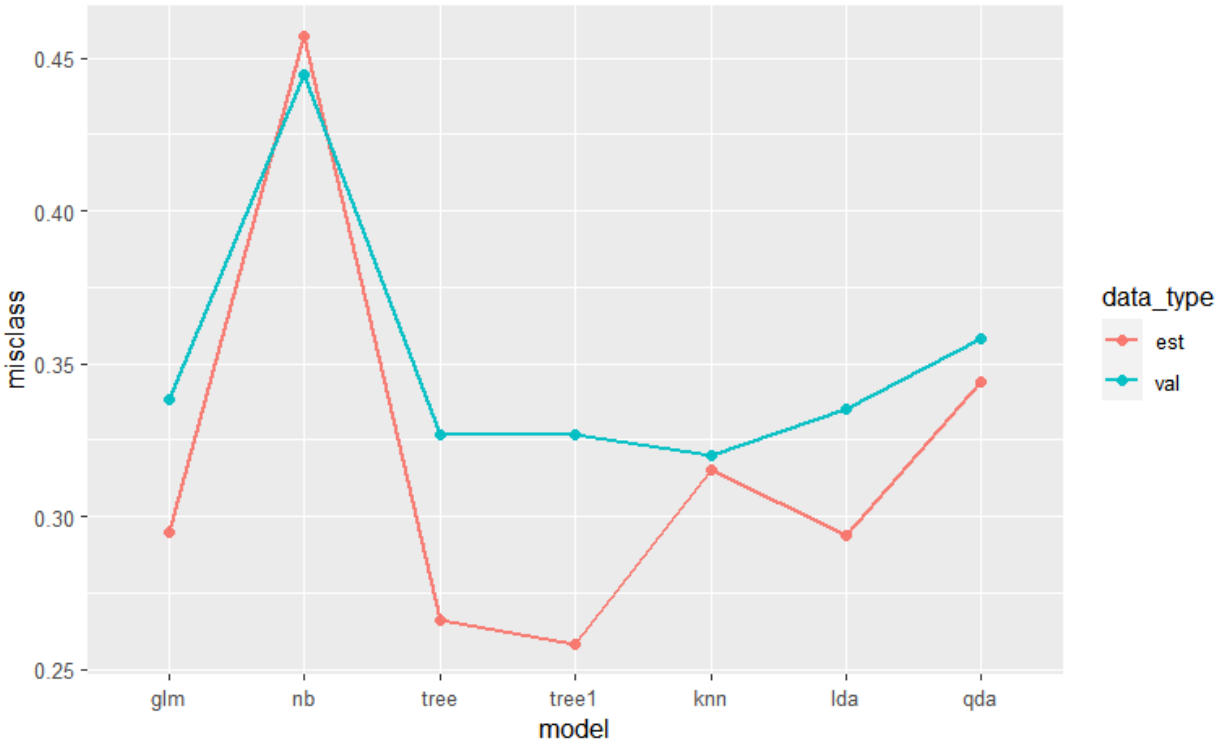
نمودار ۱۴: مقادیر خطای طبقه بندی مدل‌های درخت تصمیم در پیش‌بینی داده‌های اعتبارسنجی استاندارد شده و غیر استاندارد

جمع‌بندی نهایی برای مدل‌های طبقه‌بندی:

در این مرحله مدل‌های منتخب از هر یک از مدل‌های طبقه‌بندی در اینجا جمع می‌شوند و کوچک‌ترین خطای طبقه‌بندی (روی داده‌های اعتبارسنجی) انتخاب شده و مدل منتخب روی داده‌های آموزش برازش داده می‌شود و سپس مقدار خطای طبقه‌بندی روی داده‌های آزمایشی محاسبه می‌شود. (خطای طبقه بندی مدل k نزدیک‌ترین همسایه روی داده‌های اعتبارسنجی استاندارد شده بدست آمده)

جدول ۱۵: جمع بندی خطای طبقه بندی برای مدل‌های مختلف انتخابی در پیش‌بینی داده‌های اعتبار سنجی

مدل	خطای طبقه بندی
رگرسیون لجستیک	۰.۳۳۸۱۱۵۴
تحلیل تفکیکی کوآدراتیک	۰.۳۵۸۳۳۰۶
تحلیل تفکیکی خطی	۰.۳۳۵۱۸۱۰
کا نزدیک‌ترین همسایه	۰.۳۱۹۵۳۰۵
درخت تصمیم ۱	۰.۳۲۷۰۲۹۷
درخت تصمیم ۲	۰.۳۲۷۰۲۹۷
بیز ساده	۰.۴۴۴۷۳۴۳



نمودار ۱۵: جمع بندی خطای طبقه بندی برای مدل‌های مختلف انتخابی در پیش‌بینی داده‌های اعتبار سنجی و برآورد

با مقایسه میزان خطای طبقه بندی بین مدل‌ها، مدل کا نزدیک‌ترین همسایه که روی داده‌های استاندارد شده‌ی برآورد نیز برازش داده شده بود، به عنوان بهترین مدل شناسایی گردید. حال مقدار این مدل روی داده‌های آموزشی استاندارد شده برازش داده شده است که مقدار خطای پیش‌بینی آن روی داده‌های آزمایشی استاندارد شده برابر با ۰.۳۲۶۶۳۷ می‌باشد.

جمع‌بندی نهایی:

در این پروژه در هر دو قسمت رگرسیون و طبقه‌بندی مدل های انتخابی برازش شد و مدل نهایی برای هر دو مسئله انتخاب گردید و رویدادهای آموزش برازش گردید.

اما با این وجود هنوز میزان خطا مقدار قابل توجهی است دلیل این امر این است که تعداد ویژگی‌های دخیل در برازش داده‌ها بسیار زیاد است و باید از روش‌های پیشرفته‌تر و انعطاف‌پذیرتری در حوزه یادگیری ماشین استفاده کرد. که می‌توان به الگوریتم‌های کاهش ابعاد [۲۴]و شبکه‌های عصبی [۲۵] اشاره داشت که شاید بتوانیم از این روش‌ها میزان خطای بسیار کمتری به دست آوریم.

[24] Principal component analysis and dimensionality reduction
[25] Artificial Neural Network