



دانشگاه تربیت مدرس

دانشکده علوم زیستی

پایان‌نامه کارشناسی ارشد

رشته بیوانفورماتیک

**پیش­بینی پپتید­های ضد­سرطانی با استفاده از یادگیری ماشین**

نگارنده

**زینب محمدتبار**

استاد راهنما

**دکتر پرویز عبدالمالکی**

دی ۱۴۰۰

چكيده

اگر­چه تلاش­های بسیار زیادی در جهت توسعه درمان­های جدید سرطان صورت گرفته­است اما هنوز هم سرطان یکی ازعلل عمده مرگ در جهان است. شیمی درمانی با وجود عوارض جانبی شدید بر روی سلول­های طبیعی، هنوز هم روش اصلی مورد استفاده در درمان سرطان در مراحل پیشرفته یا متاستاتیک است. بنابراین، توسعه داروهای ضد سرطانی جدید با سمیت کم برای سلول های طبیعی، یک مسیر جدید برای درمان سرطان فراهم می کند. در حال حاضر یکی از روش­های موثر نسبت به شیمی­درمانی استفاده از پپتید­های ضد­سرطانی است. اما شناسایی این نوع پپتید­ها به روش­های تجربی، پرهزینه و زمانبر است. بنابراین در سال‌های اخیر استفاده از روش­های محاسباتی جهت شناسایی پپتیدهای ضد­سرطانی، مورد توجه پژوهشگران بسیاری قرار­گرفته­است. به­همین منظور در این پژوهش نیز سعی شده است تا توسط یادگیری ماشین مدلی برای تشخیص پپتید­های ضدسرطانی از پپتید­های غیر­ضد­سرطانی ایجاد شود.

در این پژوهش با جمع­آوری داده مربوط به توالی پپتید­های ضد­سرطانی و غیر­ضد­سرطانی، استخراج ویژگی از توالی این پپتیدها توسط کتابخانه پایتون iFeature، ساخت و آموزش دو طبقه­بند جنگل تصادفی و ماشین بردار پشتیبان توسط کتابخانه پایتون scikit-learn بر روی ویژگی های استخراج شده، به نتایج با صحت بالای 80% دست یافته­ایم. با بررسی نتایج مدل­های آموزش دیده بر­روی داده های تست مستقل، می­توان اعلام کرد که دو ویژگی QSOrder و APseduoAAC توسط هر دو طبقه­بند جنگل تصادفی و ماشین بردار پشتیبان بالاترین عملکرد برای تشخیص پپتید­های ضد­سرطانی از بین پپتید­های غیر­ضد­سرطانی را با 80 درصد صحت دارند.

كليد واژه: پیش‌بینی پپتید­های ضد­سرطانی، یادگیری ماشین، جنگل تصادفی، ماشین بردار پشتیبان، سرطان

فهرست مطالب

فهرست شکل‌ها و نمودارها

فهرست جدول‌ها

# فصل اول: مقدمه

## مقدمه‌ای بر پپتید­های ضد­سرطانی[[1]](#footnote-1)

سرطان یکی از علل عمده مرگ و میر میلیونها انسان در جهان است[1, 2]. برآورد آژانس بین المللی تحقیقات سرطان (IARC)[[2]](#footnote-2) نشان می دهد که در سال 2018، 18.1 میلیون مورد جدید سرطان و 9.6 میلیون مرگ ناشی از سرطان در سراسر جهان رخ داده­است. همچنین، آمار در سراسر جهان نشان می دهد که سرطان ریه، پستان و روده بزرگ بیش ازهمه انواع سرطان­ها بوده است[3].

یکی از روش­های معمول در درمان سرطان، پرتو درمانی است. اما در روش پرتو درمانی علاوه بر از بین رفتن سلول­های سرطانی، سلول­های طبیعی نیز دچار آسیب می­شوند. مشاهده شده­است، که سلولهای سرطانی در برابر داروهای شیمی درمانی که جهت مهار تکثیر DNA به کار می­روند به صورت مقاوم و خاموش به تکثیر خود ادامه می­دهند[4]. بنابراین، توسعه داروهای ضد سرطانی جدید با سمیت کم برای سلول های طبیعی، یک مسیر جدید برای درمان سرطان فراهم می کند. پپتیدها به دلیل عواملی نظیر اندازه کوچک، سنتز راحت، فعالیت و ویژگی بالا و تنوع بیولوژیکی توجه دانشمندان را به خود جلب کرده است. از جمله پپتیدهای درمانی که در سال‎های اخیر به منظور درمان سرطان مورد توجه قرارگرفته‎اند، می‎توان پپتیدهای کاتیونی ضدسرطانی و پپتیدهای نفوذپذیر سلولی را نام برد.

تحقیقات نشان می­دهد که پپتیدهای ضد­سرطانی، در مقایسه با روش­های شیمی درمانی و پرتو درمانی، عملکردی مناسب برای شناسایی و مهار سلول­های سرطانی بدون ایجاد مقاومت دارویی دارند[5].

بنابراین در حال حاضر یکی از روش­های موثردر درمان سرطان که توجه دانشمندان رابه خود جلب کرده است، استفاده از پپتیدهای ضدسرطانی است. در ادامه به بررسی ویژگی­های زیستی پپتیدهای ضدسرطانی می­پردازیم.

## ویژگی­های زیستی پپتید­های ضدسرطانی

همانطور که گفته شد یکی از راهکارهای نوین برای درمان سرطان، استفاده از پپتیدهای ضد­سرطانی است که با توجه به انتخابی بودن نسبت به سلولهای سرطانی، عوارض کمتر و اثربخشی آنها مورد توجه بسیاری از دانشمندان قرار گرفته است[6]. یکی از انواع پپتیدهای مورد استفاده، پپتید ضدمیکروبی[[3]](#footnote-3) است که بخشی از پاسخ ایمنی ذاتی در برابر میکروب­ها در بسیاری از گونه­ها است. این پپتیدها­ دارای وزن مولکولی کم (10-40 اسید­آمینه) و ساختار آمفی­پاتیک و کاتیونی هستند که به آن­ها اجازه می­دهد تا غشاهای منفی و سلول­های سرطانی (همانند باکتری­ها) را هدف قرار دهند. غشاء بیرونی سلول های سرطانی نسبت به سلول های طبیعی دارای فسفولیپیدهایی با بار منفی بیشتر نظیرفسفاتیدیل سرین و گلیکوپروتئین ها و گلیکوز آمینوگلیکان های منفی هستند[7]. پپتیدهای ضد سرطانی با توجه به ویژگی های کاتیونی و آمفی­پاتیک خود با تعاملات الکترواستاتیک به سلول های سرطانی متصل شده و از این رو این پپتیدها از طریق نکروز یا آپوپتوز سمیت سلول­های سرطانی را محدود می کنند[8-10].

پپتیدهای ضدسرطانی در درمان دیابت، سرطان و بیماری های قلبی عروقی به کار برده می­شوند[11]. اکثر پپتیدهای ضد­سرطانی به صورت محلول و بدون ساختار هستند و تنها پپتید ضد­میکروبی انسانی به نام LL37 با توجه به حضور یک یا چند پیوند دی­سولفیدی با صفحات β-sheets به شکل گلومرولی دیده­ شده­است. از نظر تنوع ساختاری چند صد پپتید­ ضد­میکروبی مورد مطاله قرارگرفته است. این پپتیدها دارای ساختار آلفاهلیکسی غنی از سیستئین(نظیر cecropins) یا دارای ساختار β-sheet(نظیر defensins) هستند، همچنین حضور اسیدآمینه آرژنین، پرولین، هیستدین و تریپتوفان در این پپتیدها(نظیر indolicidin) معمول است[12].



شکل ‏1‑: تعدادی پپتید ضد­میکروبی طبیعی با فعالیت ضد­سرطانی[6]

در دهه­های اخیر استفاده از این پپتید­ها به عنوان عوامل ضد­سرطان به عنوان یک روش درمانی نوین در نظر گرفته شده است [13-15].

## طبقه­بندی پپتیدهای ضدسرطانی

رزونانس مغناطیسی هسته ای NMR[[4]](#footnote-4) به عنوان یک روش مفید برای مطالعه جزئیات ساختاری بسیاری از پپتیدهای ضد میکروبی شناخته شده پدید آمده است. تجزیه و تحلیل ساختار سه بعدی این پپتیدها منجر به درک بهتری از عملکرد این پپتیدها می­گردد[16]. براساس ساختارهای شناخته شده توسط NMR پپتیدهای ضد­سرطانی به پنج گروه طبقه­بندی می­گردند.

### پپتید­های ضد­سرطانی با ساختار آلفا-هلیکس (*α-*HelicalACPs)

برای اولین بار در پپتید cecropins بسیاری از ویژگی­ها شناسایی شد و مطالعات اولیه با NMR نشان داد که پپتید cecropin-A دارای ساختارهلیکس است. در ساختار هلیکس این پپتید 15% الکل hexafluoro isopropyl شرکت دارد. در نتیجه پپتیدهای ضد­میکروبی دارای ساختار هلیکس، پپتیدهایی آمفی­پاتیک با سطوح آبگریز و حاوی بار خالص مثبت هستند. پپتیدهای ضدمیکروبی با ساختار هلیکس بیشتر از سایر ساختارها مشاهده می­شوند. پپتید ضدمیکروبی به نام Magainins گروه دیگری از پپتیدها با ساختارآلفا-هلیکس است. این پپتیدها از پوست نوعی قورباغه آفریقایی جدا می گردد. مطالعات NMR نشان داده­است، که cecropins، نیز مانند magainins دارای ساختارآمفی پاتیک-هلیکس در 25% از trifluoroethanol است [17].

### پپتید­های ضد­سرطانی غنی از سیستئن

پپتیدهای نوتروفیل انسان HNP-123 اولین پپتیدهای غنی از سیستئین جدا شده از گرانول انسان بودند [18]. پپتیدهای غنی از سیستئین با 30 اسیدآمینه در طیف گسترده­ای از موجودات زنده وجود دارد. این پپتیدها به صورت یک موتیف حفاظت شده حاوی شش سیستئین با سه پیوند دی­سولفید درون مولکولی است. اکثرا موقعیت پل دی­سولفیدی بین C1-C4، C2، C3-C5-C6 است. مطالعات کریستالوگرافی HNP-3 توسط XRAY، در ترکیب با سانتریفوژ تعادل رسوب، نشان می­دهد که این پپتید به صورت دایمر وجود دارد [19]. در مطالعات NMR بر روی ساختارdefensin مشخص گردید، که این پپتید حاوی سه رشته آنتی­پارالل است. پپتید Drosomycin جدا شده از مگس سرکه شامل چهار پیوندهای دی­سولفید است و ازسه رشته آنتی­پارالل ساخته شده­است. در این پپتید یک ساختار هلیکس بین دو رشته اول وجود دارد [20].

### پپتید­های ضد­سرطانی با ساختار بتا-شیت (*β-*SheetACPs)

تعدادی از پپتیدهای ضد­سرطانی با طول 20 اسیدآمینه دارای یک ساختار سنجاق سری هستند. این پپتیدها دارای یک یا دو پیوند دی­سولفیدی می باشند. پپتیدهای خرچنگ نعل اسبی، tachyplesin وpolyphemusin II که به صورت موتیف کوتاه سنجاق سری هستند، این پپتید کوتاه توسط دو پیوند دی سولفیدی پایدار می گردد [20, 21]. مطالعات NMR همراه با ساختار­های D3 نشان می­دهد که پپتید tachyplesin شباهت بسیار زیادی با پپتیدهای جدا شده از خوک دارد. مولکول پپتیدی صفحات آنتی­پارالل به یک ساختار turn متصل شده­اند و شامل دو پل دی­سولفید هستند [21].

### پپتید­های ضد­سرطانی غنی از اسید­آمینه­های منظم

تعدادی از پپتیدهای ضد­سرطانی دارای تعداد زیادی از یک اسیدآمینه خاص هستند. این پپتیدها دارای ساختار­های مختلف آلفا­-­هلیکس و بتا­-­شیت هستند. پپتید Histatin، غنی از اسید آمینه هیستیدین از بزاق انسان جدا شده و این پپتید در برابر کاندید آلبیکانس فعال است[22]. در حالی که پپتیدهای cathelicidins، غنی از پرولین و دارای ساختار نامنظم هستند. پپتیدهایtritrpticin سرشار از تریپتوفان و پپتیدهای Bactenecin Bac-5 and Bac-7 مانند پپتیدهای cathelicidins غنی از اسیدآمینه پرولین هستند؛ در حالی که پپتید39 PR-، غنی از اسیدآمینه آرژنین است [23-26].

### پپتید­های ضد­سرطانی با اسید­آمینه­های تغییر­یافته یا اصلاح­شده

تعدادی از پپتیدهای ضد­سرطانی دارای اسید­آمینه­های اصلاح شده هستند. بهترین نمونه از این پپتیدها، پپتیدهای تولید شده توسط باکتری­ها است. از این پپتیدها می­توان به نیسین[[5]](#footnote-5)، که توسط باکتری لاکتوکوکوس­لاکتیس[[6]](#footnote-6) تولید می­شود اشاره کرد. این پپتید باکتریایی از اسید­آمینه­های نادر مانند( 3-methyllationine, dyhydroalanine and dehydrobutyrine) تشکیل شده­است و این پپتیدها برعلیه باکتری­های گرم مثبت فعال هستند[27]. از دیگر پپتیدهای ضدمیکروبی که دارای اسیدآمینه اصلاح­شده هستند، می­توان به پپتید Leucocin A، که ازLeuconostoc gelidum جدا شده­است، اشاره کرد. این پپتید ساختار کانفورماسیونی آمفی­پاتیک از خود نشان می­دهد، که این ساختارها دارای نقش مهمی در تعامل با غشا هستند [28].

## روش‌های تجربی شناسایی پپتید­های ضدسرطانی

برای شناسایی پروتئین­ها یا پپتیدها به روش تجربی اول از همه نیازمند استخراج نمونه و یا آزمایش­هایی روی موجود زنده است. هم­چنین برای کشف عملکرد مورد­نظر که در این پژوهش خاصیت ضد­سرطانی پپتیدها مد­نظر است نیازمند روش­های آزمایشگاهی خاص خود است. تمام مراحل از یافت نمونه تا بررسی عملکرد پپتید نیازمند متخصص و تجهیزات زیستی است. این پر ­واضح است که انجام روش­های آزمایشگاهی زیستی برای شناسایی عملکرد یک پپتید نیازمند زمان و هزینه مالی و انسانی قابل توجهی است. به­همین دلیل همیشه امید است که با مطالعه بر روی روش­های محاسباتی بتوان محققان زیستی را تا حدی نجات داد، تا محققان زیستی برای انجام آزمایش­های خود، از بین کاندیدهای میلیونی برای مطالعه و آزمایش به کاندیدهای صدتایی یا حتی هزارتایی برسند. این پدیده باعث میشود تا محققان زیستی در زمان و هزینه­های مالی و مواد صرفه جویی کنند و این زمان را صرف تحقیقات با ارزش­تری کنند. شناسایی پپتید­های ضد­سرطانی نیز از این قاعده مستثنی نمی­شود. به­همین دلیل در این پایان­نامه نیز سعی می­شود تا با انجام یک روش محاسباتی به تشخیصی از پپتیدهای ضدسرطانی برسیم.

## روش‌های محاسباتی تشخیص پپتیدهای ضدسرطانی

همانطور که در قسمت قبل اشاره شد، روش­های تجربی نیازمند صرف انرژی انسانی و هزینه­های مالی زیاد است، در نتیجه نیاز به رویکرد­های محاسباتی در­کنار روش­های تجربی وجود دارد.

یکی از روش­های محاسباتی برای تشخیص پپتید­های ضد­سرطانی، استفاده از داده­های بدست آمده توسط محققان زیستی برای یادگیری­ماشین است. در ادامه به توضیح مختصری از الگوریتم­های به کار­رفته در این پایان­نامه می­پردازیم.

### روش‌های مبتنی بر یادگیری ماشین

یادگیری ماشین به مجموعه وسیعی از روش‌های محاسباتی برای درک داده‌ها اشاره دارد که می‌تواند الگوهای پیچیده در یک مجموعه را پیدا و بر اساس آن‌ها تصمیم‌گیری کند. این روش‌ها را می‌توان به دو دسته یادگیری با نظارت[[7]](#footnote-7) و یادگیری بدون نظارت[[8]](#footnote-8) تقسیم کرد. یادگیری ماشین با نظارت، یک مدل آماری برای پیش‌بینی برچسب کلاس[[9]](#footnote-9) بر اساس یک یا چند توصیف­کننده[[10]](#footnote-10) (ویژگی[[11]](#footnote-11)) ایجاد می‌کند. در این روش با هدف پیش‌بینی دقیق کلاس­ها برای مشاهدات آینده، مدلی[[12]](#footnote-12) تنظیم می‌شود که برچسب کلاس­ها را به ویژگی‌ها مرتبط کند.

در یادگیری ماشین بدون نظارت، به توصیف­کننده­ها کلاسی نسبت داده نشده است. در این روش با پیدا کردن ساختار و روابط بین توصیف‌کننده‌ها، داده‌ها کلاسترمی‌شوند.[[13]](#footnote-13)

در مسائلی که داده‌ها برچسب کلاس دارند؛ از روش‌های با نظارت که به آن­ها طبقه‌بندی[[14]](#footnote-14) می­گویند، استفاده می‌شود. در ادامه به اختصار طبقه­بندهای استفاده شده در این پایان­نامه توضیح داده می‌شود.

#### طبقه­بند درخت تصمیم[[15]](#footnote-15)

درخت نوع خاصی از گراف است و درخت تصمیم یک روش برای پیش‌بینی بر اساس این ساختار است. به دلیل تفسیر­پذیر بودن درخت­تصمیم، این روش از پرکاربردترین مدل‌ در مسائل یادگیری ماشین کلاسیک است. این ساختار از گره ریشه[[16]](#footnote-16) شروع می‌شود و به برگ­ها[[17]](#footnote-17) (گره پایانی) ختم می‌شود. برگ‌ است که مشخص می‌کند یک نمونه در چه کلاسی قرار خواهد گرفت. گره‌های داخلی با استفاده از قوانینی و با توجه به ویژگی‌ها، نمونه‌ها را به زیر مجموعه‌های مختلف تقسیم می‌کند، تا در نهایت مشخص شود هر نمونه به کدام کلاس تعلق دارد.[29]



شکل ‏1‑: ساختار کلی یک درخت تصمیم

#### طبقه­بند جنگل تصادفی[[18]](#footnote-18)

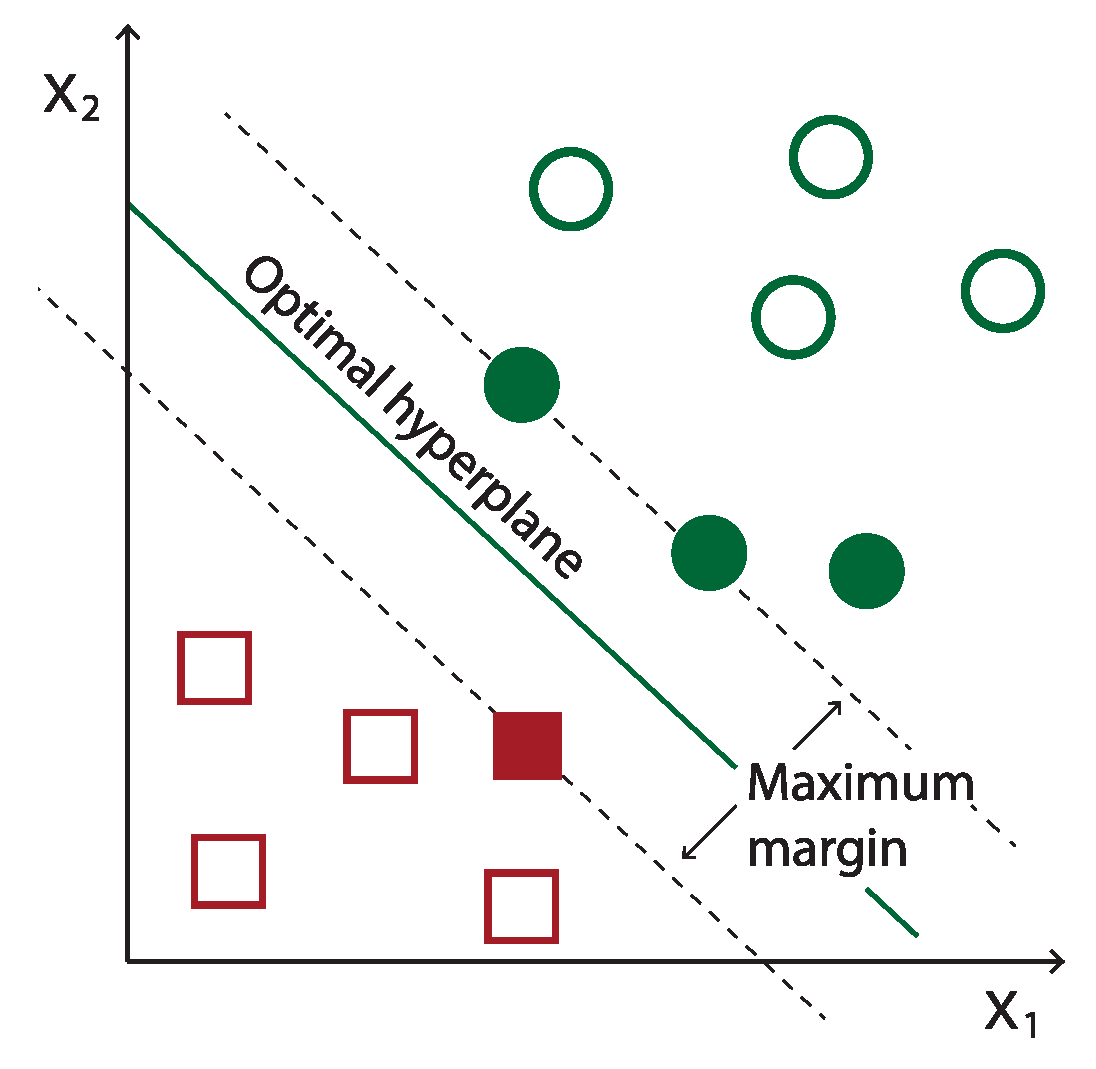
طبقه‌بند جنگل تصادفی از تعداد زیادی درخت­تصمیم استفاده می‌کند. به همین دلیل الگوریتم جنگل تصادفی در دسته الگوریتم­های یادگیری گروهی[[19]](#footnote-19) قرار می­گیرد. در این روش هر درخت­تصمیم به صورت تصادفی تعدادی از ویژگی‌ها را انتخاب می‌کند و در مورد هر نمونه پیش‌بینی انجام می‌دهد. برای تعیین برچسب یک نمونه بین تمام برچسب‌هایی که هر درخت به آن نمونه نسبت داده است؛ رأی اکثریت گرفته می‌شود[30].



شکل ‏1‑: ساختار کلی از یک طبقه بند جنگل تصادفی

#### طبقه­بند ماشین بردار پشتیبان[[20]](#footnote-20)

ماشین بردار پشتیبان یک الگوریتم یادگیری ماشین است که سعی دارد با ایجاد یک ابرصفحه داده‌ها را به نسبت تعداد برچسب­ها تقسیم کند. این طبقه‌بند در ایجاد ابر صفحه مناسب به دنبال بیشینه کردن فاصله نمونه‌ها از ابرصفحه است. به این فاصله حاشیه[[21]](#footnote-21) گفته­ می­شود.



شکل ‏1‑: ساختار یک ماشین بردار پشتیبان دوکلاس خطی

از آنجایی که همیشه و به راحتی امکان یافتن چنین ابرصفحه‌ای نیست؛ با تابعی به نام هسته[[22]](#footnote-22) نمونه‌ها به فضایی جدید منتقل می‌شوند و سپس ابرصفحه مناسب ایجاد می‌شود. همچنین این الگوریتم اجازه می‌دهد تعداد کمی از نمونه‌ها در سمت نادرست ابرصفحه قرار بگیرند[31] .



شکل ‏1‑: مثالی از فضای نمونه‌ پس از اعمال تابع هسته

## پیشینة تحقیق

برای پیش­بینی برهمکنش میزبان-پاتوژن بین گونه­های *Pseudomonas syringae* و *A.thaliana* در مطالعه‌ای دو رویکرد محاسباتی اینترولوگ و مبتنی بر دومین استفاده شده است. بر اساس رویکرد محاسباتی اینترولوگ حدود ۰٫۷۹ میلیون برهمکنش پیش­بینی شده که حاصل از برهمکنش­ ۷۷۰۰ پروتئین از *A.thaliana* و ۱۰۶۸ پروتئین از *Pseudomonas syringae* بوده ­است. بر اساس رویکرد مبتنی بر دومین ۸۵۶۵۰ برهمکنش میان ۱۱۴۳۲ پروتئین از *A.thaliana* و ۸۸۷ پروتئین از *Pseudomonas syringae* پیش‌بینی شده است. حدود ۱۱۰۰۰ برهمکنش گزارش شده در هر دو رویکرد مشترک بوده­اند.[11]

در مطالعه­ای دیگر پاتوژن *Penicillium expansum* که تعداد زیادی از محصولات کشاورزی را بعد از برداشت آلوده می­کند و باعث پوسیدگی آن‌ها می­شود، بررسی شده­ است. از جمله میزبانان این پاتوژن می­توان سیب، مرکبات، گلابی، گوجه ­فرنگی، کیوی، ذرت و برنج را نام برد. در این مطالعه برهمکنش ۲۴۵ پروتئین از *Penicillium expansum* با پروتئین­های هفت گیاه با روش اینترولوگ پیش­بینی و با بررسی معنا داری اطلاعات هستی شناسی[[23]](#footnote-23) صحت پیش­بینی­ها تایید شده ­است. تمامی پیش­بینی­ها در پایگاه داده PCPPI[[24]](#footnote-24) در دسترس است. در جدول ‏1‑2 خلاصه­ای از تعداد برهمکنش­های پیش‌بینی شده و پروتئین­های درگیر در این برهمکنش­ها آورده شده است.[18]

جدول ‑2: خلاصه تعداد برهمکنش‌ها [18]

|  |  |  |  |
| --- | --- | --- | --- |
| نام گونه | نام رایج | تعداد برهمکنش­ها | تعداد پروتئین­ها |
| *Malus domestica* Brokh. | سیب | 100120 | 2270 |
| *Actinidia chinensis* Planch. | کیوی | 51538 | 1196 |
| *Zea mays* L. | ذرت | 63922 | 1347 |
| *Pyrus communis* L. | گلابی | 75055 | 1608 |
| *Oryza sativa* L. | برنج | 65140 | 1425 |
| *Fragaria ananassa* (Duchesne ex Weston) Duchesne ex Rozier | توت فرنگی | 33240 | 740 |
| *Solanum lycopersicum* L. | گوجه فرنگی | 50889 | 1080 |
| *Penicillium expansum* | کپک کبود | 439904 | 245 |

تحقیقات نشان داده که*Xanthomonas oryzae*  پاتووار[[25]](#footnote-25) *oryzae* (Xoo) یک باکتری بیماری­زای بسیار مهم در برنج است. در مطالعه­ای با استفاده از روش­های بیوانفورماتیک و ژنوم میزبان، پروتئین­هایی که احتمال دارد با این باکتری برهمکنش داشته باشند، شناسایی شده­اند. این گروه با ساختن یک پایگاه داده به نام XooNET سعی کرده­اند شبکة برهمکنش در Xoo را پیش­بینی کنند تا بتوانند مسیرهای مهم بیوشیمیایی را به منظور تولید سموم دفع آفات درک کنند.[19]

پاتوژن *Ralstonia solanacearum* روی تعداد زیادی میزبان اثر می­گذارد. در مطالعه­ای برهمکنش بین این پاتوژن و گیاه *Arabidopsis thaliana* مورد بررسی قرار گرفته است. در این مطالعه نیز از دو رویکرد محاسباتی اینترولوگ و مبتنی بر دومین استفاده شده است. حدود ۳۰۷۴ برهمکنش پیش­بینی شده که شامل۱۴۴۲ پروتئین از *A.thaliana* و ۱۱۹ پروتئین از *Ralstonia solanacearum* بوده­است. تمامی پیش­بینی­ها در پایگاه داده PPIRA[[26]](#footnote-26)در دسترس است.[20]

## هدف از انجام طرح

در دهة اخیر مطالعات بسیاری برای پیش­بینی برهمکنش پروتئین-پروتئین درون­گونه­ای[[27]](#footnote-27) انجام شده است در حالی که مطالعات مربوط به پیش­بینی برهمکنش پروتئین-پروتئین بین­گونه­ای[[28]](#footnote-28) بسیار کمتر است. جنبة مهم هر سیستم میزبان-پاتوژن، مکانیسمی است که در آن عامل بیماری­زا میزبان را آلوده می­کند؛ بنابراین به نظر می‌رسد برهمکنش پروتئین­های میزبان-پاتوژن نقش مهمی در شروع بیماری داشته باشد. یکی از دلایل شروع و گسترش بیماری، برهمکنش بین پروتئین­ها است که در آن پروتئین پاتوژن، پروتئین میزبان را مورد هدف قرار می­دهد. توسعه روش­های محاسباتی که مشخص می­کند کدام برهمکنش عامل بیماری­زا را قادر به آلوده کردن میزبان می­کند، می­تواند در شناسایی اهداف بالقوه برای درمان­ مفید باشد. روش­های آزمایشگاهی برای تشخیص برهمکنش بین پروتئین­های میزبان-پاتوژن محدودیت­های ذاتی بسیاری دارد. از این رو رویکردهای محاسباتی برای پیش­بینی برهم­کنش­ها از اهمیت ویژه­ای برخوردارند. [21] [22]

گزارش­ها حاکی از آن است که خسارت ناشی از عوامل بیماری­زا در محصولات کشاورزی ۲۰ تا ۴۰ درصد است و تنها در ایالات متحده حدود ۴۰ میلیارد دلار در سال ضرر مستقیم اقتصادی به وجود می‌آورد. همچنین در چین یک نوع قارچ می­تواند تقریباً ۱۵ تا ۲۰ میلیون هکتار از اراضی برنج را آلوده کرده و باعث از بین رفتن سالانه ۶ میلیون تن برنج شود.[23] موارد ذکر شده اهمیت تشخیص برهمکنش بین پروتئين­های میزبان و پاتوژن در گیاهان را روشن می­کند. در این میان بررسی پاتوژن­هایی که طیف گسترده‌ای از گیاهان را تحت تاثیر قرار می‌دهند، اهمیت ویژه­ای دارد. با توجه به تحقیقاتی که تاکنون صورت گرفته گونة باکتریایی *Pseudomonas syringae* تعداد زیادی از محصولات کشاورزی از جملهسویا[[29]](#footnote-29)، گوجه‌فرنگی[[30]](#footnote-30)*،* توتون[[31]](#footnote-31)*و A.thaliana* را آلوده می­کنند و همچنین گیاه *A.thaliana* به دلیل ویژگی­های ژنتیکی شناخته شده به عنوان ارگانیسم مدل توسط پژوهشگران پذیرفته شده­­ است.[24]‌[25]‌[26]

با جست­وجو در مقالات علمی مشخص شد که مطالعات محاسباتی در رابطه با پیش­بینی برهمکنش بین پروتيئن­های گونة باکتریایی و گیاهان ذکر شده در بالا صورت نگرفته است. بنابراین در این پژوهش به پیش­بینی برهمکنش بین پروتئین­های گیاه *A.thaliana* و باکتری­ *Pseudomonas syringae* با استفاده از روش‌های یادگیری ماشین و بررسی تاثیر نوع انتخاب دادة منفی در نتایج این روش‌ها پرداخته شد.

# فصل دوم: مواد و روش‌ها

## مجموعه داده

برای ساختن مدل‌های پیش‌بینی برهمکنش پروتئین-پروتئین نیاز به دو مجموعه داده[[32]](#footnote-32) است. مجموعه دادة مثبت که شامل جفت پروتئین‌هایی است که به صورت آزمایشگاهی برهمکنش فیزیکی بین آن‌ها تایید شده ‌است. مجموعه دادة منفی که شامل جفت پروتئین‌هایی است که برهمکنش فیزیکی بین آن جفت پروتئین گزارش نشده ‌باشد.

### مجموعه دادة مثبت

در این پایان‌نامه داده‌های مثبت از مطالعه انجام شده روی باکتری *Pseudomonas syringae* و گیاه*Arabidopsis thaliana* استخراج شده‌است.[27] این مجموعه داده شامل ۱۵۳ برهمکنش بین ۴۸ پروتئین‌ پاتوژن و ۶۲ پروتئین‌ میزبان است.‌

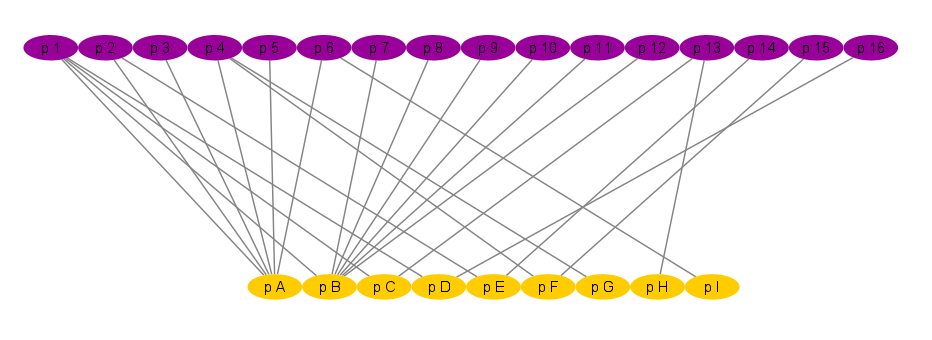
### مجموعه دادة منفی

با توجه به بزرگ‌تر بودن مجموعه دادة منفی و همچنین عدم گزارش وسیع این نوع از داده‌ها به صورت آزمایشگاهی، انتخاب دادة منفی یکی از چالش‌های ساختن مدل مناسب است. برای انتخاب داده‌های منفی روش‌های مختلفی وجود دارد که رایج‌ترین آن، انتخاب به صورت تصادفی از شبکة مکمل دادة مثبت است. نقدهای زیادی در مورد این روش وجود دارد و در بعضی از مقالات از روشی استفاده شده ‌است که دادة منفی به صورتی انتخاب شود که تعداد تکرار هر پروتئین در دو مجموعه داده یکسان باشد به این معنی که درجة متناظر با هر پروتئین در شبکة دادة مثبت و شبکة دادة منفی برابر باشد.[28]

همان‌طور که در فصل اول گفته شد، یکی از اهداف این پایان‌نامه بررسی اثر روش انتخاب دادة منفی در کارآیی مدل‌های یادگیری ماشین است. به همین منظور از دو روش برای انتخاب دادة منفی استفاده شده ‌است.

روش اول انتخاب تصادفی از بین جفت پروتئین‌های شبکة مکمل دادة مثبت است. برای این منظور ماتریس مجاورت شبکة دادة مثبت محاسبه شد، سپس به تعداد ۱۵۳ یعنی به تعداد جفت پروتئین‌ها در مجموعه دادة مثبت از درآیه‌های صفر ماتریس گفته شده به صورت تصادفی انتخاب شد.

روش دوم انتخاب جفت پروتئین از بین جفت پروتئین‌های شبکة مکمل دادة مثبت است به صورتی که درجة رأس متناظر با هر پروتئین در شبکه دادة مثبت با درجة همان رأس در شبکة دادة منفی یکسان باشد. برای روشن شدن نحوة انتخاب در این روش، برای مثال یک شبکة دوبخشی با ۲۵ رأس به عنوان شبکة دادة مثبت ساخته شد که در شکل ‏2‑1 نشان داده شده‌است. پروتئین‌های بخش یک با رنگ بنفش و پروتئین‌های بخش دو با رنگ زرد نشان داده شده‌ است.



شکل ‏2‑ : شبکة دادة مثبت با ۲۵ رأس

در مرحلة بعد ماتریس مجاورت[[33]](#footnote-33) گراف فوق ساخته شد. سپس یک سطر و یک ستون به انتهای ماتریس اضافه شد و درجة مربوط به هر رأس در آن قرار گرفت. برای انتخاب داده‌های منفی در این روش باید ابتدا از پروتئينی شروع کرد که بیشترین درجه را دارد، زیرا در غیر این صورت این احتمال وجود دارد که نتوان به تعداد لازم دادة منفی مربوط به این پروتئین انتخاب کرد وجود دارد؛ به همین دلیل سطرهای ماتریس براساس آخرین ستون که شامل درجة هر رأس بودند مرتب شد. جدول ‏2‑1 ماتریس گفته شده را نشان می‌دهد.

جدول ‏2‑1: ماتریس مجاورت شبکة دوبخشی ساخته شده با ۲۵ رأس به همراه درجة هر رأس

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| درجه | p 3 | p 12 | p 2 | p 8 | p 11 | p 14 | p 1 | p 9 | p 7 | p 13 | p 6 | p 5 | p 10 | p 16 | p 15 | p 4 |  |
| ۷ | ۰ | ۱ | ۰ | ۱ | ۱ | ۰ | ۱ | ۱ | ۱ | ۰ | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ | p B |
| ۶ | ۱ | ۰ | ۱ | ۰ | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ | ۱ | ۱ | ۰ | ۰ | ۰ | ۱ | p A |
| ۲ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱ | ۱ | p F |
| ۲ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱ | ۰ | ۰ | p D |
| ۲ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱ | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | p C |
| ۲ | ۰ | ۰ | ۱ | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | p E |
| ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱ | p G |
| ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | p I |
| ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | p H |
| ۲۴ | ۱ | ۱ | ۲ | ۱ | ۱ | ۱ | ۴ | ۱ | ۱ | ۲ | ۲ | ۱ | ۱ | ۱ | ۱ | ۳ | درجه |

درآیه‌های حاوی صفر نشان دهندة این است که برای آن جفت پروتئین برهمکنشی گزارش نشده است، یا به عبارتی درآیه‌های صفر، شبکة مکمل دادة مثبت را تشکیل می‌دهند و از آن‌ها می‌توان دادة منفی انتخاب کرد. در این روش سطر به سطر عمل شده ‌است، به این صورت که در مرحلة اول باید به تعداد درجة رأس متناظر با پروتئین B یعنی ۷ تا از صفرهای سطر اول انتخاب شود. باید توجه کرد که انتخاب به روشی صورت گیرد که امکان انتخاب دادة منفی برای پروتئین‌های ۱ تا ۱۶ به تعداد درجة رأس متناظر با آن پروتئین وجود داشته باشد.

برای حل این مشکل، انتخاب با یک احتمال صورت می‌گیرد. این احتمال به صورت تقسیم درجة رأس متناظر با پروتئین بخش یک بر درجة رأس متناظر با پروتئین بخش دو محاسبه می‌شود. به عنوان مثال برای مرحلة اول، ‏جدول ‏2‑2 در نظر گرفته می‌شود. سطر اول جدول، شامل اسم پروتئين بخش یک است که با پروتئين B برهمکنشی نداشته‌اند (متناظر با درآیه‌های صفر در ماتریس مجاورت) و سطر دوم درجة رئوس پروتئین‌هایی است که در سطر اول آمده‌است. سطر سوم مقدار احتمال این‌که هر پروتئین به صورت تصادفی انتخاب شود را نشان می‌دهد. در این مرحله احتمال انتخاب هریک از پروتئین‌های سطر اول برابر است با درجة رأس متناظر با آن پروتئین تقسیم بر درجة رأس متناظر با پروتئین B، یعنی ۷، است.

جدول ‏2‑2: احتمال محاسبه شده برای انتخاب پروتئین مقابل پروتئین B برای دادة منفی

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| p 3 | p 2 | p 14 | p 13 | p 6 | p 5 | p 16 | p 15 | p 4 |  |
| ۱ | ۲ | ۱ | ۲ | ۲ | ۱ | ۱ | ۱ | ۳ | درجه |
|  |  |  |  |  |  |  |  |  | احتمال |

به عنوان مثال فرض کنیم، پروتئین‌های p 2، p 3، p 4، p 13، p 14، p 15 و p 16 انتخاب شده‌اند. درآیه‌های متناظر با پروتئین‌هایی که انتخاب شده‌اند به تغییر داده می‌شود و جمع هر سطر و ستون دوباره محاسبه می‌شود و در سطر و ستون آخر قرار می‌گیرد. بعد از این مرحله ماتریس به شکل زیر می‌شود.

جدول ‏2‑3 : ماتریس بعد از انتخاب پروتئین مقابل پروتئین B برای دادة منفی

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| مجموع | p 3 | p 12 | p 2 | p 8 | p 11 | p 14 | p 1 | p 9 | p 7 | p 13 | p 6 | p 5 | p 10 | p 16 | p 15 | p 4 |  |
| ۰ | ۱- | ۱ | ۱- | ۱ | ۱ | ۱- | ۱ | ۱ | ۱ | ۱- | ۰ | ۰ | ۱ | ۱- | ۱- | ۱- | p B |
| ۶ | ۱ | ۰ | ۱ | ۰ | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ | ۱ | ۱ | ۰ | ۰ | ۰ | ۱ | p A |
| ۲ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱ | ۱ | p F |
| ۲ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱ | ۰ | ۰ | p D |
| ۲ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱ | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | p C |
| ۲ | ۰ | ۰ | ۱ | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | p E |
| ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱ | p G |
| ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | p I |
| ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | p H |
| ۱۷ | ۰ | ۱ | ۱ | ۱ | ۱ | ۰ | ۴ | ۱ | ۱ | ۱ | ۲ | ۱ | ۱ | ۰ | ۰ | ۲ | مجموع |

به این نکته توجه شود که، از این مرحله به بعد بعضی از درآیه‌های سطر آخر صفر می‌شوند، در واقع تعداد دفعاتی که پروتئین بخش یک در دادة مثبت و دادة منفی ظاهر شده‌است با هم برابر شده‌است، در نتیجه این پروتئین‌ها در انتخاب‌های مراحل بعد در نظر گرفته نمی‌شوند. به عنوان مثال با این‌که پروتئین A با پروتئین‌های p 14، p 15 و p 16 هیچ برهمکنشی ندارند، اما برای انتخاب در نظر گرفته نمی‌شوند. زیرا تعداد دفعاتی که این پروتئین‌‌ها در دادة مثبت و دادة منفی ظاهر شده‌اند با هم برابر شده ‌است؛ یعنی درجة رئوس متناظر با این پروتئین‌ها در دادة مثبت و منفی یکسان شده‌است. همچنین مشاهده می‌شود که احتمال انتخاب آن‌ها نیز برابر با صفر شده است. برای پروتئین A جدول ‏2‑4 در نظر گرفته می‌شود که احتمال در سطر آخر با روشی که پیش‌تر گفته شد، محاسبه شده ‌است. از پروتئین‌هایی که احتمال انتخابشان غیر صفر است؛ ۶ پروتئین، یعنی به تعداد درجة رأس متناظر با پروتئین A، انتخاب می‌شود.

جدول ‏2‑4: احتمال محاسبه شده برای انتخاب پروتئین مقابل پروتئین A برای دادة منفی

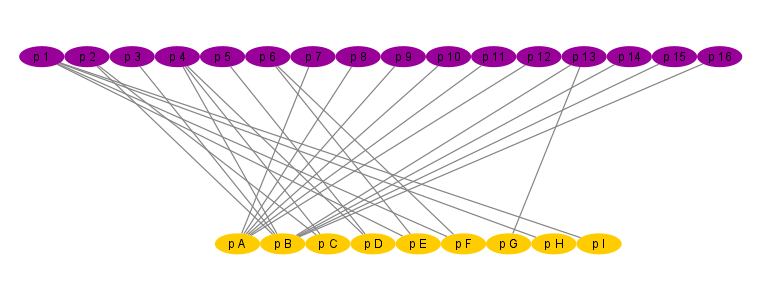
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| p 12 | p 8 | p 11 | p 14 | p 9 | p 7 | p 13 | p 10 | p 16 | p 15 |  |
| ۱ | ۱ | ۱ | ۰ | ۱ | ۱ | ۱ | ۱ | ۰ | ۰ | مجموع |
|  |  |  |  |  |  |  |  |  |  | احتمال |

با همین روش تا آخرین پروتئین‌، یعنی پروتئین H، مراحل تکرار می‌شود. در انتها ماتریس به صورت جدول ‏2‑5 می‌شود. درآیه‌هایی که شامل ۱- هستند نشان دهندة جفت پروتئین‌هایی هستند که در مجموعه دادة منفی حضور دارند. درآیه‌های سطر و ستون آخر که همگی برابر با صفر هستند نشان دهندة برابر بودن درجة رئوس در شبکة دادة مثبت و شبکة دادة منفی است.

جدول ‏2‑5 : ماتریس بعد از آخرین مرحلة انتخاب داده‌های منفی

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| مجموع | p 3 | p 12 | p 2 | p 8 | p 11 | p 14 | p 1 | p 9 | p 7 | p 13 | p 6 | p 5 | p 10 | p 16 | p 15 | p 4 |  |
| ۰ | ۱- | ۱ | ۱- | ۱ | ۱ | ۱- | ۱ | ۱ | ۱ | ۱- | ۰ | ۰ | ۱ | ۱- | ۱- | ۱- | p B |
| ۰ | ۱ | ۱- | ۱ | ۱- | ۱- | ۰ | ۱ | ۱- | ۱- | ۰ | ۱ | ۱ | ۱- | ۰ | ۰ | ۱ | p A |
| ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱- | ۰ | ۰ | ۰ | ۱- | ۰ | ۰ | ۰ | ۱ | ۱ | p F |
| ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ | ۰ | ۱- | ۰ | ۱ | ۰ | ۱- | p D |
| ۰ | ۰ | ۰ | ۱- | ۰ | ۰ | ۰ | ۱ | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱- | p C |
| ۰ | ۰ | ۰ | ۱ | ۰ | ۰ | ۱ | ۱- | ۰ | ۰ | ۰ | ۱- | ۰ | ۰ | ۰ | ۰ | ۰ | p E |
| ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱- | ۰ | ۰ | ۰ | ۰ | ۰ | ۱ | p G |
| ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱- | ۰ | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | p I |
| ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۱- | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | p H |
| ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | مجموع |

شبکة دادة منفی که با روش فوق ساخته شد در شکل ‏2‑2 نشان داده شده‌ است.



شکل ‑ : شبکة دادة منفی با ۲۵ رأس

با روش گفته شده، ۱۵۳ جفت پروتئین از شبکة مکمل دادة مثبت، انتخاب شد.

### مجموعه دادة آموزش

مجموعه دادة آموزش[[34]](#footnote-34) مجموعه‌ای است که الگوریتم‌های یادگیری ماشین با توجه به آن آموزش داده می‌شوند. 123 جفت پروتئین از مجموعه دادة مثبت به صورت کاملاً تصادفی انتخاب شد. این تعداد، ۸۰ درصد از کل مجموعه دادة مثبت است. به همین تعداد و به صورت کاملاً تصادفی از دو مجموعه دادة منفی ساخته شده در مرحلة قبل، انتخاب شد. 123 برهمکنش (جفت پروتئین) از مجموعه دادة مثبت و 123 عدم برهمکنش از مجموعه دادة منفی ساخته شده با روش اول، مجموعه آموزش اول را تشکیل می‌دهد؛ یعنی مجموعه آموزش اول شامل ۲۴۶ جفت پروتئین است. این مجموعه داده Random نام‌گذاری شده است. 123 برهمکنش (جفت پروتئین) از مجموعه دادة مثبت و 123 عدم برهمکنش از مجموعه دادة منفی ساخته شده با روش دوم، مجموعه آموزش دوم را تشکیل می‌دهد؛ یعنی مجموعه آموزش دوم شامل ۲۴۶ جفت پروتئین است. این مجموعه داده SDD (مخفف (Same Distribution Degreeنام‌گذاری شده است.

### مجموعه دادة آزمون

برای تخمین کارآیی الگوریتم یادگیری ماشین و تنظیم پارامترهای آن از مجموعه دادة آزمون[[35]](#footnote-35) که در آموزش نقشی ندارند استفاده می‌شود.30 برهمکنش باقیمانده از مجموعه دادة مثبت و ۳۰ عدم برهمکنش از مجموعه دادة منفی ساخته شده با روش اول، مجموعه آزمون Random را تشکیل می‌دهد که شامل ۶۰ جفت پروتئین است. 30 برهمکنش باقیمانده از مجموعه دادة مثبت و ۳۰ عدم برهمکنش از مجموعه دادة منفی ساخته شده با روش دوم، مجموعه آزمون SDD را تشکیل می‌دهد که شامل ۶۰ جفت پروتئین است.

### مجموعه دادة آزمون مستقل

از مجموعه دادة آزمون مستقل[[36]](#footnote-36) برای تخمین میزان کارآیی الگوریتم ماشین روی داده‌های جدید استفاده می‌شود. جفت پروتئین‌های موجود در این مجموعه با جفت پروتئین‌های مجموعة آموزش و آزمون اشتراک ندارند. در این مطالعه ۱۵ برهمکنش بینپروتئین­های گیاه *Arabidopsis thaliana* و باکتری *Pseudomonas syringae* که در پایگاه دادة IntAct گزارش شده بود و در مجموعه دادة آموزش و آزمون هم نبودند؛ به عنوان مجموعة آزمون مستقل در نظر گرفته شد. این برهمکنش‌ها بین ۱۵ پروتئین از گیاه *Arabidopsis thaliana* و ۶ پروتئین از باکتری *Pseudomonas syringae* هستند.

## استخراج ویژگی

برای استخراج ویژگی[[37]](#footnote-37) در این پایان‌نامه از ویژگی‌های مبتنی بر توالی برای پیش‌بینی برهمکنش میزبان-پاتوژن استفاده شده است. با توجه به مطالعات پیشین ویژگی‌های زیر انتخاب شد. در ادامه ویژگی‌های استخراج شده، همراه با توضیح کوتاهی در مورد آن‌ها آمده‌است. تمامی این ویژگی‌ها با بستة FtrCOOL در R استخراج شده است. [23]‌[29]‌[30]‌[31]‌[32]

### ویژگی K-Amino Acid Composition

این ویژگی فروانی -تایی‌های آمینواسیدی در پروتئین را مشخص می‌کند و بردار ویژگی به طول تولید می‌کند. در این مطالعه برابر با ۲ در نظر گرفته شده‌ و طول بردار ویژگی برای یک پروتئین برابر با ۴۰۰ شده ‌است.

### ویژگی Pseudo-Amino Acid Composition

در این ویژگی خصوصیات فیزیکوشیمیایی آمینواسیدها در نظر گرفته می‌شود. در این مطالعه سه خاصیت آبگریزی[[38]](#footnote-38)، آبدوستی[[39]](#footnote-39) و وزن زنجیرة جانبی[[40]](#footnote-40) آمینواسیدها در نظر گرفته شده‌است.

فرض کنید ، و برای به ترتیب مقدار آبگریزی، آبدوستی و وزن زنجیرة جانبی ۲۰ آمینواسید باشد. مقدار آبگریزی برای هر آمینواسید با رابطه (‏2‑1) نرمال می‌شوند.

|  |  |
| --- | --- |
| رابطه (‏2‑1) |  |

به همین ترتیب مقادیر آبدوستی و وزن زنجیرة جانبی برای هر آمینواسید نرمال می‌شود و و به دست می‌آید. سپس به ازای هر جفت آمینواسید و مقدار میانگین این سه خاصیت با رابطه (‏2‑2) محاسبه می‌شود.

|  |  |
| --- | --- |
| رابطه (‏2‑2) |  |

در ادامه مجموعه‌ای از توصیفگرها به صورت زیر تعریف می‌شود که در آن نشان دهندة طول توالی و است. در این مطالعه λ برابر با ۳۰ در نظر گرفته شده‌است.

|  |  |
| --- | --- |
| رابطه (‏2‑3) |  |

در نهایت یک بردار ویژگی به طول یعنی به طول ۵۰ برای یک پروتئین با روابط زیر ساخته می‌شود که در آن فراوانی نرمال شدة آمینواسید ام است و یک عامل وزن‌دهی است که بنا به توصیة مطالعات پیشین این مقدار برابر با ۰۵/۰ در نظر گرفته شده‌است. [33]

|  |  |
| --- | --- |
| رابطه (‏2‑4) |  |

### ویژگی Split Amino Acid Composition(SAAC)

این ویژگی توالی پروتئین را به سه قسمت N-terminal، C-terminal و قسمت میانی تقسیم می‌کند و فراوانی آمینواسیدها را در هر قسمت محاسبه می‌کند و در نهایت یک بردار ویژگی به طول ۶۰ برای یک پروتئین تولید می‌کند.

### ویژگی Auto Covariance

این ویژگی که به اختصار با AC نمایش داده می‌شود، فعل و انفعالات بین آمینواسید‌هایی که در فاصلة معینی از هم قرار دارند را در نظر می‌گیرد. در واقع این ویژگی با در نظر گرفتن اثر همسایه‌ها روی هم کشف الگوهایی که در کل توالی جریان دارد را ممکن می‌کند. در این روش باقیمانده[[41]](#footnote-41) آمینواسید به مقادیر عددی نشان دهندة خواص فیزیکوشمیایی ترجمه و سپس این توالی عددی بر اساس کوواریانس تجزیه و تحلیل می‌شود. طول این بردار ویژگی برابر با است که تعداد شاخص‌های فیزیکوشیمیایی و در این مطالعه از ۹ شاخص استفاده شده که همگی از شاخص‌های معرفی شده در پایگاه دادة AAindex هستند. این شاخص‌ها عبارتند از حجم رزیدو[[42]](#footnote-42)، سطح در دسترس رزیدو در تری‌پپتید[[43]](#footnote-43)، انرژی آزاد محلول در آب[[44]](#footnote-44)، شاخص قطبش پذیری[[45]](#footnote-45)، مقیاس متوسط نرمال شدة آبگریزی[[46]](#footnote-46)، آبدوستی، شاخص انعطاف‌پذیری متوسط[[47]](#footnote-47) و قطبیت[[48]](#footnote-48) و بار خالص[[49]](#footnote-49). نشان دهندة بیشترین فاصله بین دو آمینواسید است که در این مطالعه برابر با ۳۰ در نظر گرفته شده است. طول این ویژگی برای یک پروتئین برابر با یعنی ۲۷۰ است.

### ویژگی CDT

ویژگیCDT که مخفف سه کلمة ترکیب (Composition)، توزیع (Distribution) و انتقال (Transition) است؛ الگوهای توزیع ویژگی‌های ساختاری یا فیزیکوشیمیایی یک آمینواسید در یک توالی پروتئین یا پپتید را نشان می‌دهد. در محاسبة این ویژگی از 7 نوع ویژگی استفاده شده‌است؛ که آبگریزی، حجم نرمال شدة وندروالس[[50]](#footnote-50)، قطبیت ، قطبش پذیری، بار[[51]](#footnote-51)، ساختارهای ثانویه[[52]](#footnote-52) و قابلیت دسترسی به حلال[[53]](#footnote-53) است. هر یک از این 7 ویژگی بیست آمینواسید را بر اساس خواص فیزیکوشیمیایی به سه گروه تقسیم می‌کند و سپس توالی پروتئینی بر اساس این که هر آمینواسید در کدام گروه قرار می‌گیرند بازنویسی می‌شوند. ویژگی‌های ترکیب، انتقال و توزیع برای توالی بازنویسی شده محاسبه می‌شود و هریک به ترتیب بردارهایی به طول 21، 21 و 105 تولید می‌کنند. در نهایت بردار ویژگی CDT برای یک پروتئین از کنار هم قرار گرفتن این سه بردار ساخته می‌شود و به طول ۱۴۷ است.

### ویژگی Conjoint Triad

این ویژگی ۲۰ آمینواسید را بر اساس خواص فیزیکوشیمیایی به ۷ گروه تقسیم می‌کند و توالی پروتئینی براساس این که هر آمینواسید در کدام گروه قرار گرفته بازنویسی می‌کند. تعداد سه‌تایی‌های متوالی در این حالت برابر با ۷×۷×۷ می‌شود؛ سپس فراوانی هر یک از این سه‌تایی‌ها محاسبه و نرمال می‌شود و برای یک پروتئین یک بردار ویژگی به طول ۳۴۳ تولید می‌کند.[32]‌[34]

در نهایت بردار ویژگی از کنار هم قرار گرفتن ویژگی‌های فوق و مربوط به دو پروتئین ساخته می‌شود. با توجه به توضیحات بالا طول بردار ویژگی برای یک پروتئین برابر با ۱۲۷۰ است و برای یک برهمکنش که شامل دو پروتئین است برابر با ۲۵۴۰ است. نمایی از بردار ویژگی در شکل زیر نشان داده شده ‌است.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ویژگی۶  پروتئین۲ | ویژگی۵  پروتئین۲ | ویژگی۴  پروتئین۲ | ویژگی۳  پروتئین۲ | ویژگی۲  پروتئین۲ | ویژگی۱  پروتئین۲ | ویژگی۶  پروتئین۱ | ویژگی۵  پروتئین۱ | ویژگی۴  پروتئین۱ | ویژگی۳  پروتئین۱ | ویژگی۲  پروتئین۱ | ویژگی۱  پروتئین۱ |

شکل ‏2‑ : نمایی از بردار ویژگی

## پیش‌پردازش داده‌ها

در مرحلة پیش‌پردازش داده‌ها[[54]](#footnote-54) ویژگی‌هایی که برای تمامی نمونه‌ها یکسان است، حذف شدند؛ زیرا حاوی اطلاعاتی برای پیش‌بینی برهمکنش بین پروتئین‌ها نیستند. از آنجایی که ویژگی‌های متفاوت در محدوده‌های مختلفی تغییر می‌کنند، ممکن است ویژگی‌هایی که شامل اعداد بزرگتری هستند نسبت به بقیة ویژگی‌ها تأثیر نامناسبی روی نتیجة پیش‌بینی داشته باشند. به همین منظور تمام ویژگی‌ها باید نرمال‌سازی یا استانداردسازی شوند. در این مطالعه با استفاده از روش نرمال‌سازی min-max و با رابطه (‏2‑5) هر ویژگی‌ در مجموعه دادة آموزش به بازة نگاشت شدند.

|  |  |
| --- | --- |
| رابطه (‏2‑5) |  |

هر یک از ویژگی‌های مجموعه دادة آزمون و آزمون مستقل هم با پارامترهای مربوط به همان ویژگی در مجموعه دادة آموزش و با رابطه (‏2‑5) به اعداد جدیدی نگاشت شدند.

یکی دیگر از مسائلی که روی نتیجة اکثر طبقه‌بندی‌ها تأثیر منفی دارد، همبستگی[[55]](#footnote-55) بین ویژگی‌هاست. برای حل این مشکل در مجموعه دادة آموزش، ازبین ویژگی‌هایی که بیش از ۰٫۹ با هم همبستگی داشتند؛ تنها یکی حفظ شد و بقیه حذف شدند. ویژگی‌های حذف شده در این مرحله، از مجموعه داده‌های آزمون و آزمون مستقل نیز حذف شدند.

تمامی موارد بالا برای هر دو مجموعه دادة Random و SDD انجام شد. پس از این مرحله، طول بردار ویژگی برای مجموعه دادة Random برابر با ۲۱۵۵ و برای مجموعه دادة SDD برابر با ۲۱۳۷ شد.

## ‌ معیارهای ارزیابی

جهت سنجش کارایی الگوریتم‌های یادگیری ماشین نیاز به معیارهایی داریم و برای تعریف معیارهای ارزیابی ابتدا احتیاج به پارامترهایی داریم که بدانیم چه تعداد از نمونه‌ها به درستی پیش‌بینی شده‌اند. با توجه به این که این مطالعه در زمینة برهمکنش پروتئین-پروتئین است، پارامترها و معیارهای ارزیابی بر این اساس تعریف خواهند شد.

مثبت صحیح[[56]](#footnote-56): برهمکنش مثبت پروتئین-پروتئین که به درستی تشخیص داده شده است را گویند. تعداد این برهمکنش‌ها را با TP نشان می‌دهند.

مثبت کاذب[[57]](#footnote-57): برهمکنش منفی پروتئین-پروتئین که به اشتباه مثبت تشخیص داده شده است را گویند. تعداد این برهمکنش‌ها را با FP نشان می‌دهند.

منفی صحیح[[58]](#footnote-58): برهمکنش منفی پروتئین-پروتئین که به درستی تشخیص داده شده است را گویند. تعداد این برهمکنش‌ها را با TN نشان می‌دهند.

منفی کاذب[[59]](#footnote-59): برهمکنش مثبت پروتئین-پروتئین که به اشتباه منفی تشخیص داده شده است را گویند. تعداد این برهمکنش‌ها را با FN نشان می‌دهند.

در ادامه معیارهای ارزیابی توضیح داده شده است.

### حساسیت

توانایی یک الگوریتم یادگیری ماشین برای تشخیص مقادیر مثبت صحیح را حساسیت[[60]](#footnote-60) یا یادآوری[[61]](#footnote-61) گویند و با رابطه (‏2‑6) محاسبه می‌شود. با توجه به این رابطه می‌توان دریافت که هر چه تعداد برهمکنش‌های مثبت پروتئین-پروتئین که به اشتباه منفی پیش‌بینی شده‌اند؛ کمتر باشد، حساسیت به عدد یک نزدیک‌تر است و حساسیت الگوریتم برای پیش‌بینی بالاتر است.

|  |  |
| --- | --- |
| رابطه (‏2‑6) |  |

### اختصاصیت

توانایی یک الگوریتم یادگیری ماشین برای تشخیص مقادیر منفی صحیح را اختصاصیت[[62]](#footnote-62) گویند و با رابطه (‏2‑7) محاسبه می‌شود. با توجه به این رابطه می‌توان دریافت که هر چه تعداد برهمکنش‌های منفی پروتئین-پروتئین که به اشتباه مثبت پیش‌بینی شده‌اند؛ کمتر باشد، یعنی مثبت کاذب کمتر باشد، اختصاصیت به عدد یک نزدیک‌تر است.

|  |  |
| --- | --- |
| رابطه (‏2‑7) |  |

### صحت

یکی از مهم‌ترین معیارهای ارزیابی صحت[[63]](#footnote-63) است و با رابطه (‏2‑8) محاسبه می‌شود. این معیار نشان می‌دهد چند درصد از داده‌ها به درستی طبقه‌بندی شده‌اند.

|  |  |
| --- | --- |
| رابطه (‏2‑8) |  |

### دقت

دقت[[64]](#footnote-64) نشان دهندة نسبت تعداد برهمکنش‌های مثبت پروتئین-پروتئین که توسط الگوریتم یادگیری ماشین به درستی مثبت گزارش شده است به تعداد تمام برهمکنش‌های پروتئین-پروتئین که توسط الگوریتم یادگیری ماشین مثبت گزارش شده است؛ می‌باشد و بارابطه (‏2‑9) محاسبه می‌شود.[9]

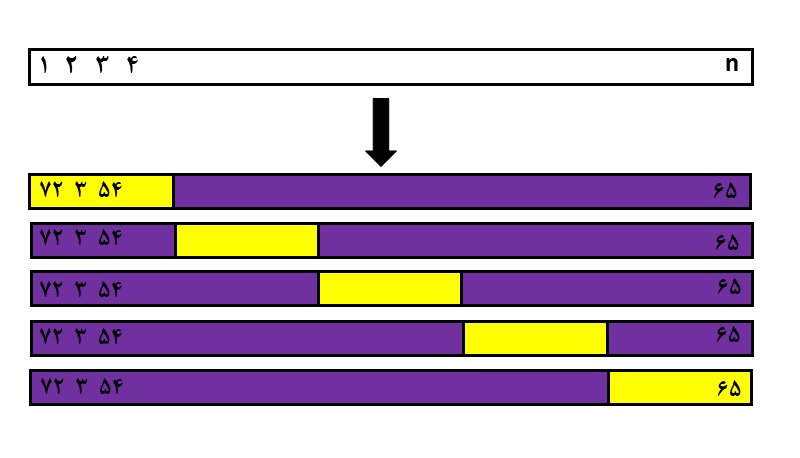
|  |  |
| --- | --- |
| رابطه (‏2‑9) |  |

## روش اعتبارسنجی متقابل

اعتبارسنجی متقابل[[65]](#footnote-65) یک روش برای ارزیابی مدل است، در این روش مجموعه دادة آموزش به چند بخش قسمت می‌شود و هر بار یک بخش به عنوان مجموعة آزمون برای ارزیابی در نظر گرفته می‌شود و مدل روی بقیة بخش‌ها اجرا می‌شود.

### اعتبارسنجی متقابل -لایه

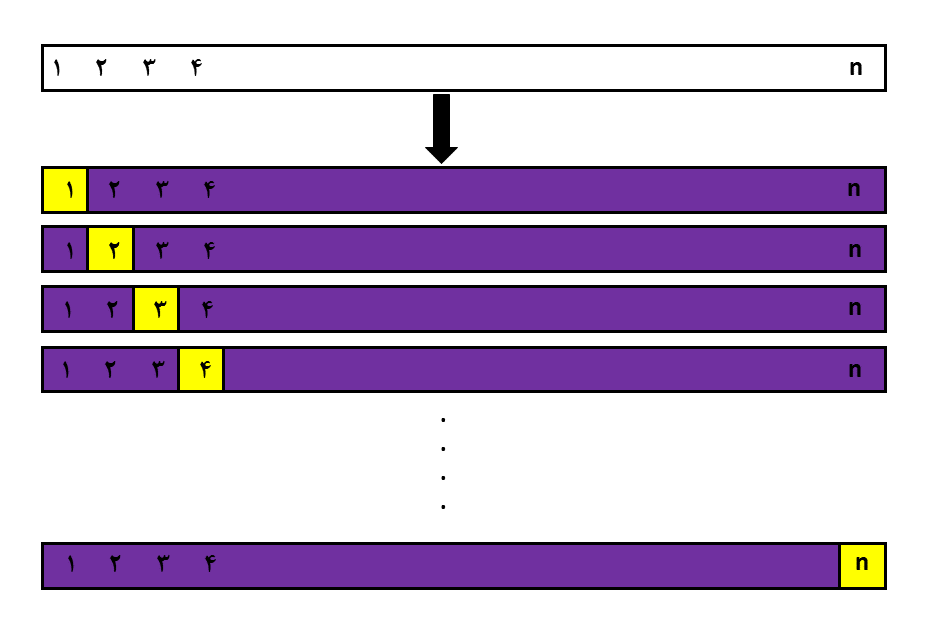
اعتبارسنجی متقابل -لایه[[66]](#footnote-66)، مجموعه دادة آموزش به افراز می‌شود. هر بار یک بخش به عنوان مجموعة آزمون در نظر گرفته می‌شود و مدل روی بخش‌ دیگر اجرا می‌شود. دقت مدل روی بخشی که به عنوان مجموعة آزمون در نظر گرفته شده بود؛ محاسبه می‌شود. در نهایت میانگین دقت محاسبه شده در مرحله فوق به عنوان دقت مدل گزارش می‌شود.



شکل ‑ : نمایی از اعتبار سنجی متقابل ۵-لایه

### اعتبارسنجی یک طرفه

اعتبارسنجی یک طرفه (LOOCV) [[67]](#footnote-67) شبیه اعتبارسنجی متقابل -لایه عمل می‌کند. اگر مجموعة آموزش شامل نمونه باشد در هر مرحله مدل برای نمونه ساخته می‌شود و سپس بر اساس مدل ساخته شده آن یک نمونه پیش‌بینی می‌شود. در نهایت دقت روی پیش‌بینی‌هایی که در مرحله صورت گرفته بود؛ به عنوان دقت مدل گزارش می‌شود.[15]



شکل ‑ : نمایی از اعتبارسنجی یک طرفه

## انتخاب ویژگی

انتخاب ویژگی[[68]](#footnote-68)، انتخاب یک زیرمجموعه از کل مجموعة ویژگی‌هاست به طوری که بیشترین و مناسب‌ترین ارتباط را با برچسب‌ها را داشته باشد. انتخاب ویژگی‌های مناسب باعث می‌شود مدل‌های یادگیری ماشین کارآیی بیشتری داشته باشند و ارتباط بین ویژگی‌ها و برچسب‌ها بیشتر قابل درک‌ شود. به طور کلی روش‌های انتخاب ویژگی را می‌توان به دو گروه بسته‌بندی[[69]](#footnote-69) و فیلتر[[70]](#footnote-70) تقسیم کرد.

### روش بسته‌بندی

در این روش، مدل را برای زیر مجموعه‌های مختلف از ویژگی‌ها که با حذف یا اضافه کردن آن‌ها ساخته می‌شود؛ اجرا می‌کند. در نهایت زیرمجموعه‌ای که موجب ‌شود مدل بالاترین کارآرایی را داشته باشد؛ انتخاب می‌شود. از روش‌های جستجو در مجموعة ویژگی‌ها می‌توان جستجوی مستقیم ترتیبی[[71]](#footnote-71) و جستجوی معکوس ترتیبی[[72]](#footnote-72) نام برد. در واقع این روش یک الگوریتم جستجو است که ویژگی‌ها به عنوان ورودی الگوریتم هستند و کارآیی مدل به عنوان خروجی است؛ که باید بهینه ‌شود.

### روش فیلتر

روش فیلتر ارتباط بین ویژگی‌ها را خارج از مدل‌های پیش‌بینی‌کننده ارزیابی می‌کند و ویژگی‌هایی را که معیارهای لازم را داشته باشند انتخاب می‌کند. به عنوان مثال در مسائل طبقه‌بندی، هر ویژگی به طور جداگانه مورد بررسی قرار می‌گیرد تا رابطة آن و برچسب‌ها مشخص شود. در نهایت ویژگی‌هایی که با برچسب‌ها ارتباط خوبی دارند؛ انتخاب می‌شوند.[35]

به دلیل این‌که این روش‌ها به صورت جداگانه روی داده‌های این مطالعه به خوبی پاسخگو نبود؛ از ترکیبی از آن‌ها استفاده شد. برای این منظور باید یک مدل یادگیری ماشین روی مجموعه داده اجرا شود و سپس تمام ویژگی‌ها بر اساس اهمیت و سهمشان در پیش‌بینی امتیاز ‌بگیرند. برای یافتن مهترین ویژگی‌ها از بستة caret در نرم‌افزار R استفاده شد.[36] در این پایان‌نامه پنج مدل k-نزدیکترین همسایه، adaBag، niave Bayes، جنگل تصادفی و treeBag روی داده‌های آموزش اجرا شد. همة مدل‌ها با ۳ بار تکرار اعتبارسنجی متقابل ۱۰-لایه اجرا شدند. برای k-نزدیکترین همسایه مدل با k برابر با ۱، ۳، ۵، ۷ و ۱۱ ساخته شد. Naïve Bayes برای Laplace correctionهای برابر با صفر تا ۵ و Bandwidth Adjustment برابر با ۱ تا ۵ اجرا شد. جنگل تصادفی برای ۱۰، ۱۵ و ۲۰ ویژگی که به صورت تصادفی انتخاب می‌شوند؛ ساخته شد. مدل adaBag برای تعداد درخت‌های ۵، ۱۰، ۱۵، ۲۰ و ۲۵ و بیشترین عمق ۳، ۵، ۷ و ۱۰ اجرا شد. مدل treeBag پارامتر تنظیمی نداشت.

سپس امتیاز هر ویژگی برابر با مجموع امتیازهایی که از هر مدل کسب کرده‌ بودند قرار داده شد و از بیشترین تا کمترین امتیاز مرتب شد. این مراحل برای هر دو مجموعه داده‌های Random و SDD انجام شد. به منظور یافتن مناسب‌ترین تعداد ویژگی برای مجموعه دادة Random، چهار مجموعه با نام‌های Random-1، Random-2، Random-3 و Random-4 که به ترتیب شامل ۵۰ ویژگی مهم اول، ۱۰۰ ویژگی مهم اول، ۱۵۰ ویژگی مهم اول و ۲۰۰ ویژگی مهم اول هستند؛ ساخته شد. به همین ترتیب مجموعه‌های SDD-1، SDD-2، SDD-3 و SDD-4 ساخته شد.

## کاهش ابعاد مجموعه داده

تعداد زیاد ویژگی‌ها باعث تحمیل هزینه محاسباتی بالا به طبقه‌بندی‌ها می‌شود و به مشقت تعداد ابعاد[[73]](#footnote-73) معروف است. برای کاهش ابعاد[[74]](#footnote-74) می‌توان از روش تحلیل مؤلفه‌های اصلی[[75]](#footnote-75) استفاده کرد.

به صورت ریاضی تحلیل مؤلفه‌های اصلی یک تبدیل خطی متعامد است که داده‌ها را به دستگاه مختصات جدید می‌برد. به ‌صورتی‌که اولین بزرگترین واریانس داده بر روی اولین محور مختصات، دومین بزرگترین واریانس بر روی دومین محور مختصات و ... قرار می‌گیرد. [15]

این روش روی هر دو مجموعه دادة آموزش Random و SDD اجرا شد. به منظور یافتن مناسب‌ترین تعداد مؤلفة اصلی برای مجموعه دادة Random، دو مجموعه با نام‌های Random-5 و Random-6، که به ترتیب شامل ۵ مؤلفة اصلی اول و ۱۰ مؤلفة اصلی اول هستند؛ ساخته شد. به همین ترتیب مجموعه‌های SDD-5 و SDD-6 ساخته شد.

## مدل‌های یادگیری ماشین اجرا شده روی مجموعه داده‌ها

در این مطالعه برای مجموعه‌های مجموعه داده‌های Random-1، Random-2، Random-3، Random-4، Random-5 و Random-6 و همچنین SDD-1، SDD-2، SDD-3، SDD-4، SDD-5 و SDD-6 مدل‌های زیر با پارامترهای تنظیمی ذکر شده در زیر، ساخته شد. برای یافتن بهترین پارامتر و تعداد ویژگی برای هر مدل از روش اعتبارسنجی یک طرفه استفاده شده است. برای اجرای تمامی مدل‌ها از بستة caret در نرم‌افزار R استفاده شد.

### طبقه‌بندی k-نزدیکترین همسایه

برای یافتن بهترین پارامتر و تعداد ویژگی‌، این مدل برای kهای برابر با ۱، ۳، ۵، ۷، ۹ و ۱۱ اجرا شد. برای محاسبة فاصله از رابطة اقلیدسی استفاده شد.

### طبقه‌بندی Naïve Bayes

این روش دو پارامتر برای تنظیم داشت. پارامتر Laplace correction برای جلوگیری از صفر شدن احتمال رخداد ویژگی‌ها به شرط دیده شدن یک برچسب است. به رابطه (‏1‑4) توجه شود. همچنین اگر به این طبقه‌بندی اجازه داده شود از توابع چگالی برای محاسبة احتمال استفاده کند؛ نیاز است پارامتر Bandwidth Adjustment نیز تنظیم شود. برای یافتن بهترین پارامتر و تعداد ویژگی‌، این مدل برای Laplace correctionهای برابر با صفر تا ۵ و Bandwidth Adjustment برابر با ۱ و ۵ اجرا شد.

### طبقه‌بندی آنالیز تشخیصی خطی

این روش پارامتری برای تنظیم جهت ساختن مدل با کارایی بهتر نداشت.

### طبقه‌بندی آنالیز تفکیک‌کننده درجه دوم

این روش پارامتری برای تنظیم جهت ساختن مدل با کارایی بهتر نداشت.

### طبقه‌بندی جنگل تصادفی

برای یافتن بهترین پارامتر و تعداد ویژگی‌، این مدل برای تعداد ۱۰، ۱۵، ۲۰، ۲۵ و ۳۰ ویژگی که به صورت تصادفی انتخاب می‌شود، اجرا شد.

### طبقه‌بندی ماشین بردار پیشتیبان با هستة خطی

برای یافتن بهترین پارامتر و تعداد ویژگی‌، این مدل برای هزینه‌های برابر با ۵، ۱۰، ۱۵ و ۲۰ اجرا شد. منظور از هزینه در این طبقه‌بندی حداکثر تعداد نمونه‌هایی است که الگوریتم اجازه می‌دهد در سمت نادرست ابرصفحه قرار گیرد.

### طبقه‌بندی ماشین بردار پیشتیبان با هستة شعاعی

برای یافتن بهترین پارامتر و تعداد ویژگی‌، این مدل برای هزینه‌های برابر با ۵، ۱۰، ۱۵ و ۲۰ اجرا شد. پارامتر هزینه در ‏6-8-2 توضیح داده شده است.

### طبقه‌بندی شبکه عصبی چند لایه

برای یافتن بهترین پارامتر و تعداد ویژگی‌، این مدل برای تعداد لایه‌های درونی‌های برابر با ۲، ۳، ۴ و ۵ اجرا شد.

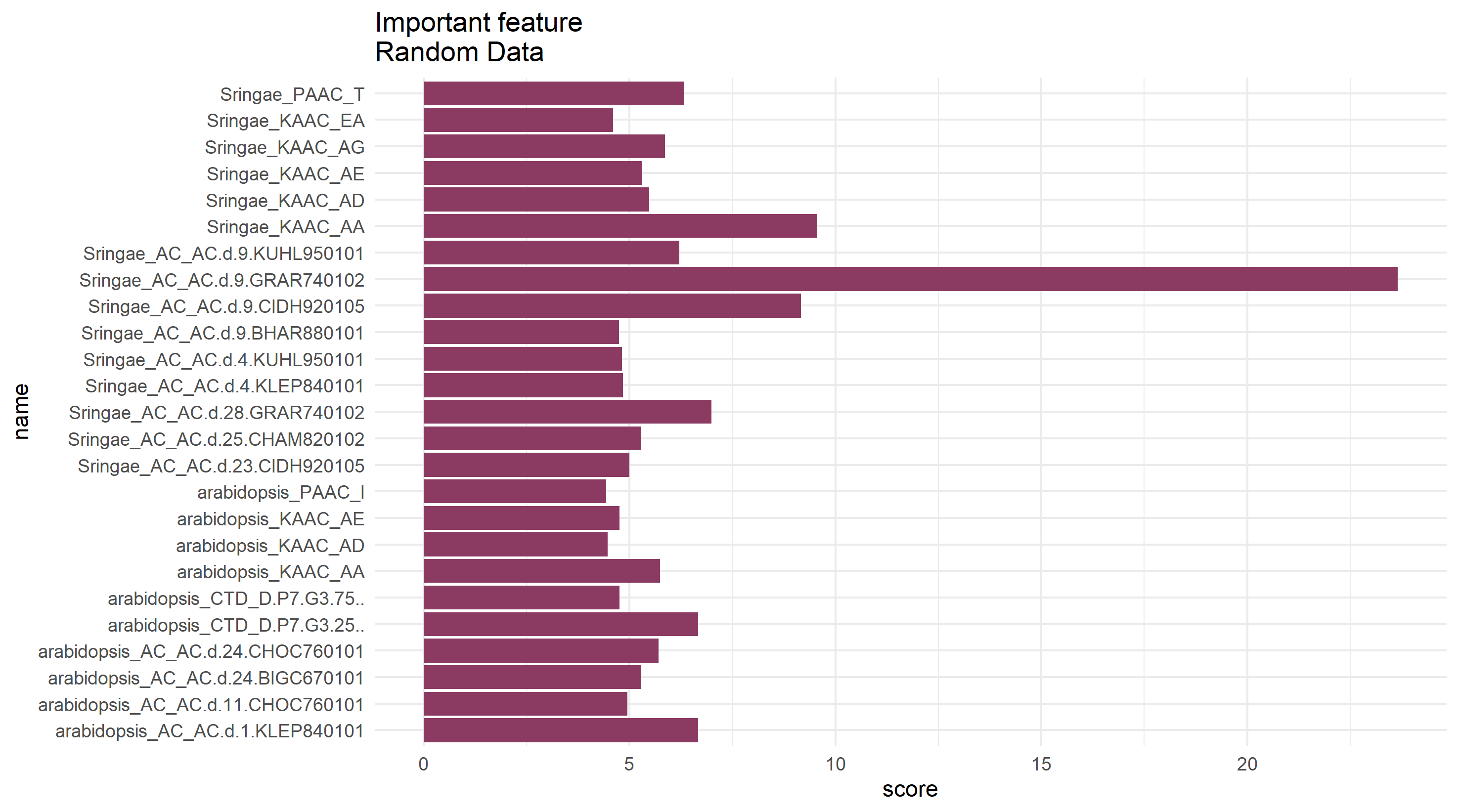
### طبقه‌بندی TreeBag

این روش پارامتری برای تنظیم جهت ساختن مدل با کارآیی بهتر نداشت.

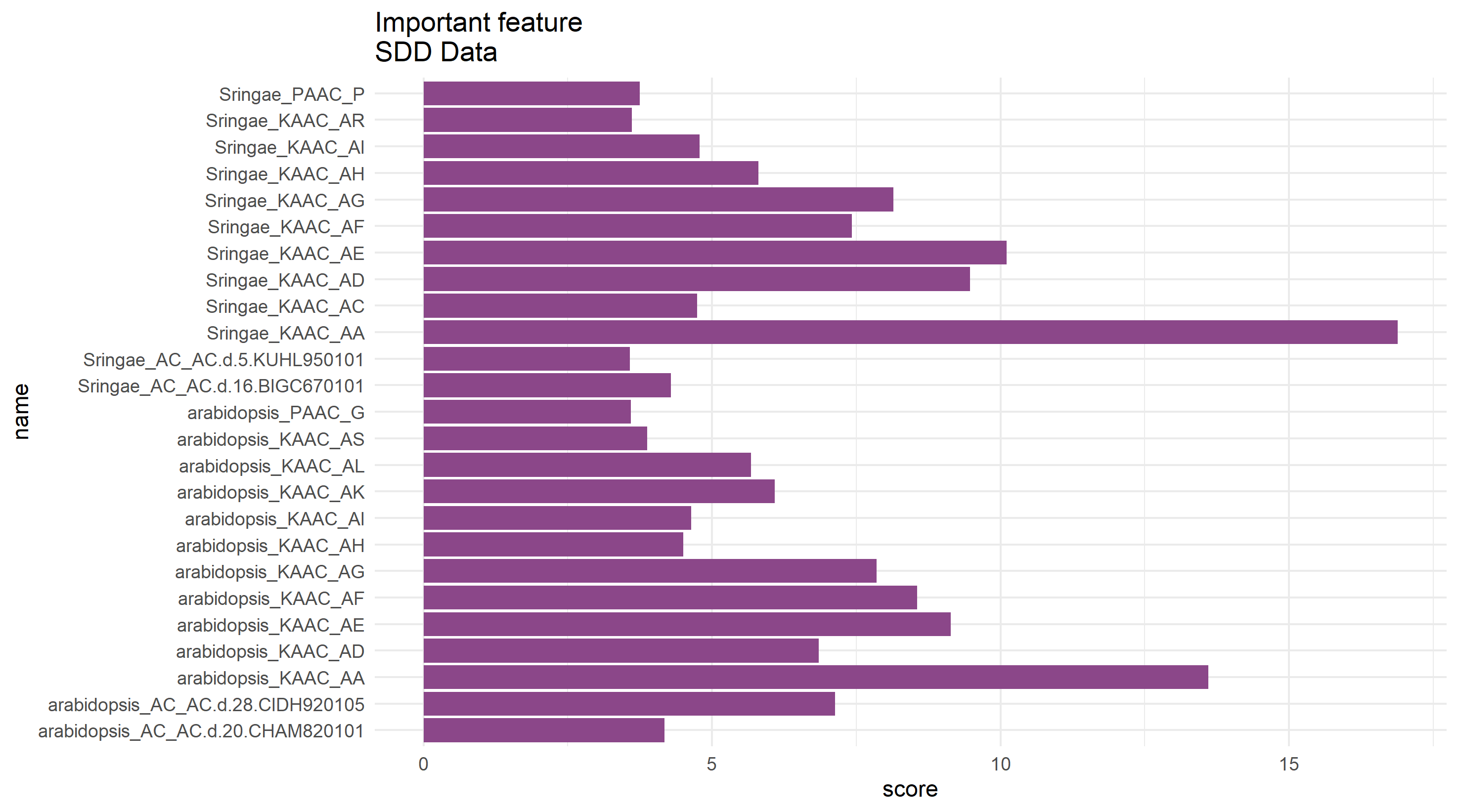
# فصل سوم: نتایج و بحث

## ویژگی‌های تأثیرگذار

جهت این که مشخص شود کدام یک از ویژگی‌های استفاده شده که در فصل دوم توضیح داده شد، اهمیت بیشتری داشته‌اند؛ باید امتیازی که ویژگی‌ها در مرحلة انتخاب ویژگی گرفته‌اند را با یکدیگر مقایسه کرد. شکل ‏3‑1 و شکل ‏3‑2 امتیاز ۲۵ ویژگی مهم اول در مجموعه داده‌های Random و SDD نشان می‌دهد. با توجه به شکل‌های گفته شده؛ مشخص می‌شود ویژگی‌هایK-Amino Acid Composition وAuto Covariance نسبت به سایر ویژگی‌ها از اهمیت بیشتری برخوردارند.



شکل ‑ : امتیاز ۲۵ ویژگی مهم اول در مجموعه دادة Random



شکل ‏3‑ : امتیاز ۲۵ ویژگی مهم اول در مجموعه دادة‌ SDD

## نتایج طبقه‌بندی‌ها

جهت ارزیابی مدل‌ها از روش اعتبارسنجی یک طرفه استفاده شد. در تمام جدول‌های مربوط به صحت مجموعه دادة آموزش، خانه‌های رنگی بهترین پارامتر را برای مدل نشان می‌دهند و با توجه به پارامترهای مدل برگزیده، صحت، حساسیت و اختصاصیت مجموعة آزمون و حساسیت مجموعة آزمون مستقل محاسبه شده است. توجه شود با وجود گزارش یکسان صحت‌های مجموعة آموزش، فقط یک پارامتر به عنوان پارامتر برگزیده انتخاب شده است و این یکسانی صحت‌ها می‌تواند در نتیجة گرد کردن اعداد صحت اتفاق افتاده باشد و یا صحت‌ها کاملاً برابر باشند ولی پارامتری برگزیده شده است که هزینة محاسباتی کمتری داشته باشد. . برای انتخاب بهترین پارامتر و تعداد ویژگی برای هر مدل، صحت مجموعه‌های آزمون در نظر گرفته شد. مجموعه آزمونی که بالاترین صحت را داشت به عنوان بهترین مدل انتخاب شد. اگر در این مرحله صحت دو یا چند مجموعه آزمون یکسان باشد، صحت مجموعة آموزش و یا حساسیت مجموعة آزمون مقایسه شد و آن که مقدار بیشتری داشت، به عنوان بهترین مدل برگزیده شد.

نتایج مربوط به طبقه‌بندی‌های ‌k-نزدیکترین همسایه، جنگل تصادفی، ماشین بردار پشتیبان با هستة شعاعی، TreeBag و شبکة عصبی چند لایه که کارآیی بالاتری داشتند؛ در این فصل و نتایج مربوط به طبقه‌بندی‌های Naïve Bayes، آنالیز تشخیصی خطی، آنالیز تفکیک‌کنندة درجه دو و ماشین بردار پشتیبان با هستة خطی در پیوست آمده است. در پایان نیز خلاصه‌ای از نتایج تمام طبقه‌بندی‌های اجرا شده، در جدول ‏3‑31 و جدول ‏3‑32 آمده است.

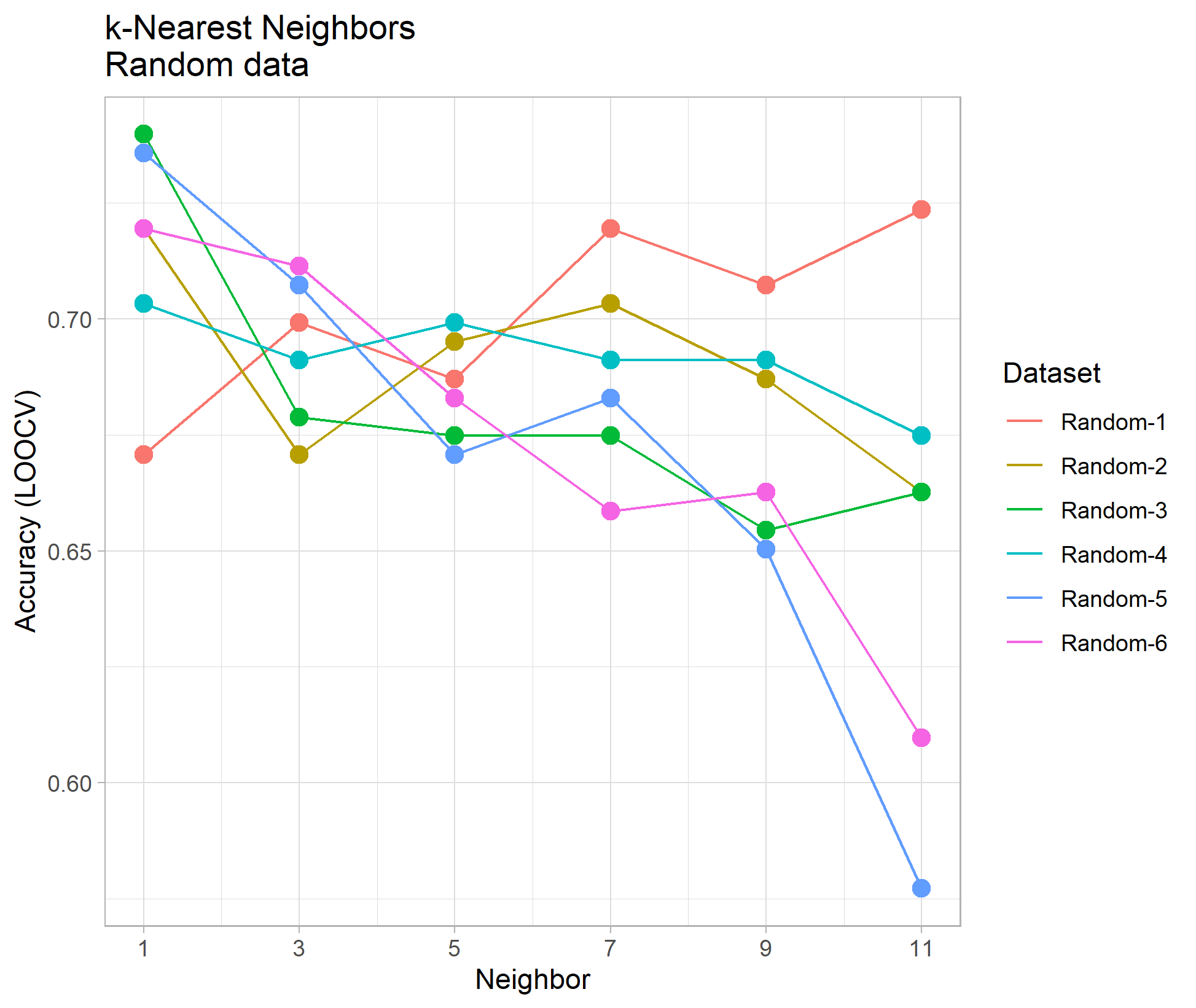
### نتایج مربوط به طبقه‌بندی k-نزدیکترین همسایه

برای یافتن بهترین پارامتر و تعداد ویژگی‌، این مدل برای kهای برابر با ۱، ۳، ۵، ۷، ۹ و ۱۱ روی مجموعه داده‌های Random-1، Random-2، Random-3، Random-4، Random-5 و Random-6 اجرا شد. همچنین این مدل با پارامترهای گفته شده روی مجموعه داده‌های SDD-1، SDD-2، SDD-3، SDD-4، SDD-5 و SDD-6 اجرا شد. نتایج به شرح زیر است.

#### نتایج مربوط به مجموعه دادة Random

جدول ‏3‑1 : صحت مجموعه داده‌های Random روی مجموعة آموزش برای طبقه‌بندی k-نزدیکترین همسایه

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| تعداد k  مجموعه داده |  |  |  |  |  |  |
| Random-1 | ۰٫۶۷ | ۰٫۷۰ | ۰٫۶۹ | ۰٫۷۲ | ۰٫۷۱ | ۰٫۷۲ |
| Random-2 | ۰٫۷۲ | ۰٫۶۷ | ۰٫۷۰ | ۰٫۷۰ | ۰٫۶۹ | ۰٫۶۶ |
| Random-3 | ۰٫۷۴ | ۰٫۶۸ | ۰٫۶۷ | ۰٫۶۷ | ۰٫۶۵ | ۰٫۶۶ |
| Random-4 | ۰٫۷۰ | ۰٫۶۹ | ۰٫۷۰ | ۰٫۶۹ | ۰٫۶۹ | ۰٫۶۷ |
| Random-5 | ۰٫۷۴ | ۰٫۷۱ | ۰٫۶۷ | ۰٫۶۸ | ۰٫۶۵ | ۰٫۵۸ |
| Random-6 | ۰٫۷۲ | ۰٫۷۱ | ۰٫۶۸ | ۰٫۶۶ | ۰٫۶۶ | ۰٫۶۱ |



شکل ‏3‑ : صحت مجموعه داده‌های Random روی مجموعة آموزش برای طبقه‌بندی k-نزدیکترین همسایه

جدول ‏3‑2 : نتایج مجموعه داده‌های Random روی مجموعة آزمون برای طبقه‌بندی k-نزدیکترین همسایه و مربوط به kهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |
| --- | --- | --- | --- |
| مجموعه داده | صحت | حساسیت | اختصاصیت |
| Random-1 | ۰٫۵۷ | ۰٫۷۰ | ۰٫۴۳ |
| Random-2 | ۰٫۶۵ | ۰٫۷۳ | ۰٫۵۷ |
| Random-3 | ۰٫۶۷ | ۰٫۷۰ | ۰٫۶۳ |
| Random-4 | ۰٫۶۳ | ۰٫۷۰ | ۰٫۵۷ |
| Random-5 | ۰٫۶۵ | ۰٫۷۰ | ۰٫۶۰ |
| Random-6 | ۰٫۷2 | ۰٫۷۷ | ۰٫۶۷ |

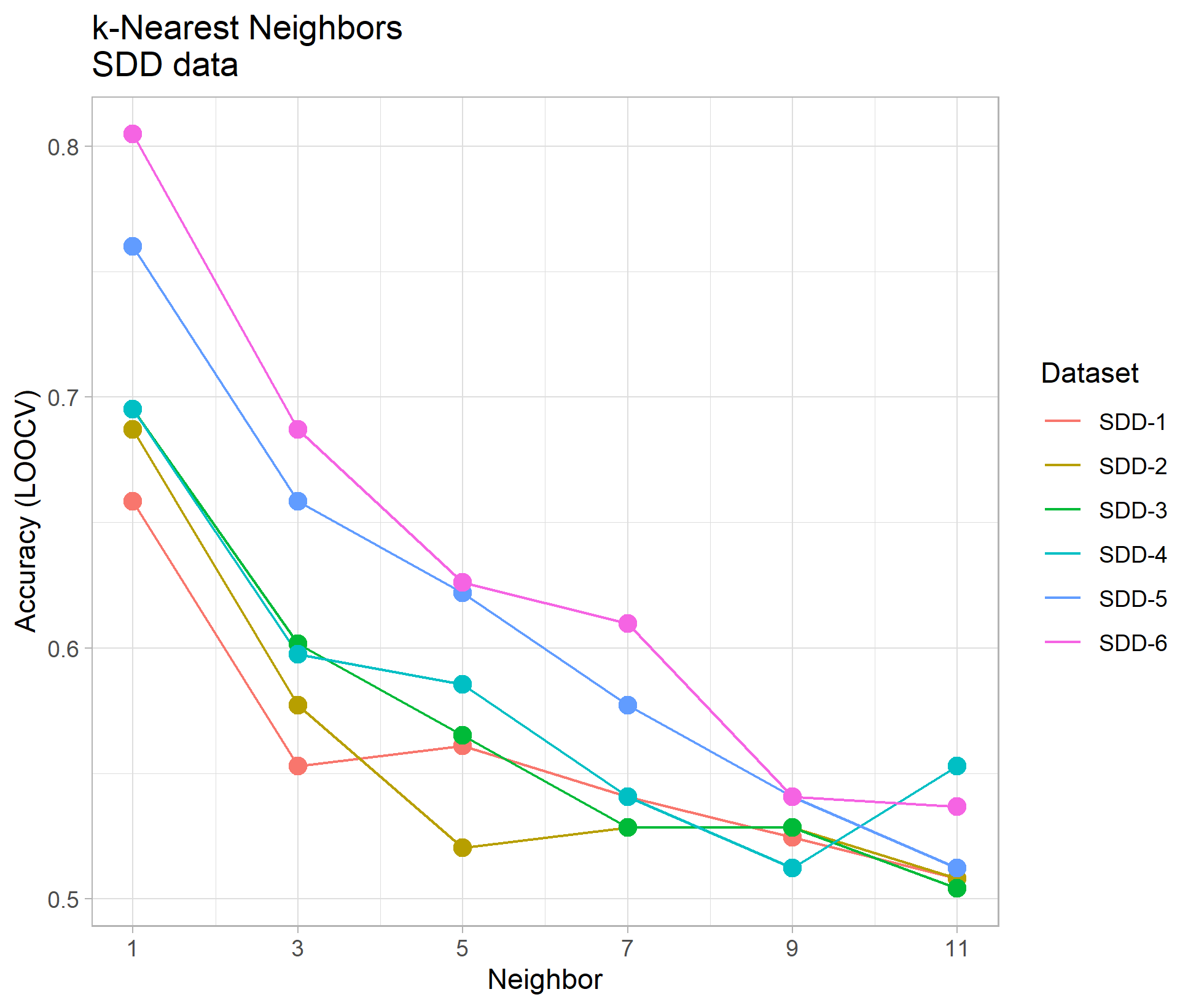
جدول ‏3‑3 : حساسیت مجموعه داده‌های Random روی مجموعة آزمون مستقل برای طبقه‌بندی k-نزدیکترین همسایه و مربوط به kهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| مجموعه داده | Random-1 | Random-2 | Random-3 | Random-4 | Random-5 | Random-6 |
| حساسیت | ۰٫۱۳ | ۰٫۲۰ | ۰٫۴۰ | ۰٫۲۰ | ۰٫۶۷ | ۰٫۱۳ |

#### نتایج مربوط به مجموعه دادة SDD

جدول ‏3‑4 : صحت مجموعه داده‌های SDD روی مجموعة آموزش برای طبقه‌بندی k-نزدیکترین همسایه

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| تعداد k  مجموعه داده |  |  |  |  |  |  |
| SDD-1 | ۰٫۶۶ | ۰٫۵۵ | ۰٫۵۶ | ۰٫۵۴ | ۰٫۵۲ | ۰٫۵۱ |
| SDD-2 | ۰٫۶۹ | ۰٫۵۸ | ۰٫۵۲ | ۰٫۵۳ | ۰٫۵۳ | ۰٫۵۱ |
| SDD-3 | ۰٫۷۰ | ۰٫۶۰ | ۰٫۵۷ | ۰٫۵۳ | ۰٫۵۳ | ۰٫۵۰ |
| SDD-4 | ۰٫۷۰ | ۰٫۶۰ | ۰٫۵۹ | ۰٫۵۴ | ۰٫۵۱ | ۰٫۵۵ |
| SDD-5 | ۰٫۷۶ | ۰٫۶۶ | ۰٫۶۲ | ۰٫۵۸ | ۰٫۵۴ | ۰٫۵۱ |
| SDD-6 | ۰٫۸۰ | ۰٫۶۹ | ۰٫۶۳ | ۰٫۶۱ | ۰٫۵۴ | ۰٫۵۴ |



شکل ‏3‑ : صحت مجموعه داده‌های SDD روی مجموعة آموزش برای طبقه‌بندی k-نزدیکترین همسایه

جدول ‏3‑5 : نتایج مجموعه داده‌های SDD روی مجموعة آزمون برای طبقه‌بندی k-نزدیکترین همسایه و مربوط به kهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |
| --- | --- | --- | --- |
| مجموعه داده | صحت | حساسیت | اختصاصیت |
| SDD-1 | ۰٫۶۲ | ۰٫۷۷ | ۰٫۴۷ |
| SDD-2 | ۰٫۵۷ | ۰٫۶۰ | ۰٫۵۳ |
| SDD-3 | ۰٫۵۸ | ۰٫۶۳ | ۰٫۵۳ |
| SDD-4 | ۰٫۵۸ | ۰٫۶۰ | ۰٫۵۷ |
| SDD-5 | ۰٫۷۲ | ۰٫۸۰ | ۰٫۶۳ |
| SDD-6 | ۰٫۷۰ | ۰٫۷۰ | ۰٫۷۰ |

جدول ‏3‑6 : حساسیت مجموعه داده‌های SDD روی مجموعة آزمون مستقل برای طبقه‌بندی k-نزدیکترین همسایه و مربوط به kهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| مجموعه داده | SDD-1 | SDD-2 | SDD-3 | SDD-4 | SDD-5 | SDD-6 |
| حساسیت | ۰٫۶۰ | ۰٫۶۰ | ۰٫۴۷ | ۰٫۸۰ | ۰٫۴۷ | ۰٫۶۰ |

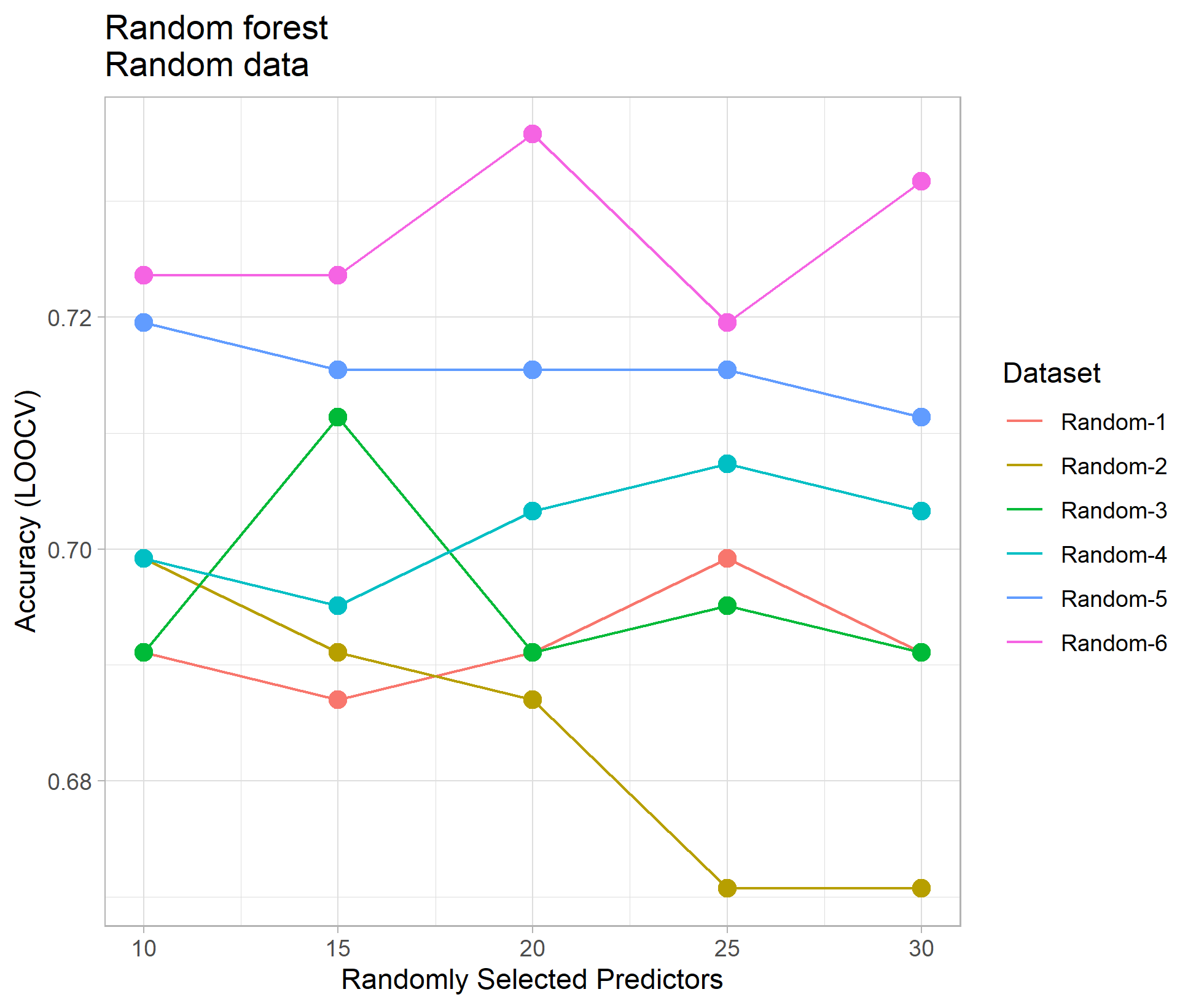
### نتایج مربوط به طبقه‌بندی جنگل تصادفی

در این طبقه‌بندی مدل برای تعداد مختلف ویژگی‌ که به صورت تصادفی انتخاب می‌شود؛ ساخته شد. این پارامتر با mtry نشان داده شده است. برای یافتن بهترین پارامتر و تعداد ویژگی‌، این مدل برای پارامتر mtry برابر با ۱۰، ۱۵، ۲۰، ۲۵ و ۳۰ روی مجموعه داده‌های Random-1، Random-2، Random-3، Random-4، Random-5 و Random-6 و همچنین روی مجموعه داده‌های SDD-1، SDD-2، SDD-3، SDD-4، SDD-5 و SDD-6 اجرا شد. نتایج به شرح زیر است.

#### نتایج مربوط به مجموعه دادة Random

جدول ‏3‑7 : صحت مجموعه داده‌های Random روی مجموعة آموزش برای طبقه‌بندی جنگل تصادفی

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| mtry  مجموعه داده | ۱۰ | ۱۵ | ۲۰ | ۲۵ | ۳۰ |
| Random-1 | ۰٫۶۹ | ۰٫۶۹ | ۰٫۶۹ | ۰٫۷۰ | ۰٫۶۹ |
| Random-2 | ۰٫۷۰ | ۰٫۶۹ | ۰٫۶۹ | ۰٫۶۷ | ۰٫۶۷ |
| Random-3 | ۰٫۶۹ | ۰٫۷۱ | ۰٫۶۹ | ۰٫۷۰ | ۰٫۶۹ |
| Random-4 | ۰٫۷۰ | ۰٫۷۰ | ۰٫۷۰ | ۰٫۷۱ | ۰٫۷۰ |
| Random-5 | ۰٫۷۲ | ۰٫۷۲ | ۰٫۷۲ | ۰٫۷۲ | ۰٫۷۱ |
| Random-6 | ۰٫۷۲ | ۰٫۷۲ | ۰٫۷۴ | ۰٫۷۲ | ۰٫۷3 |



شکل ‏3‑ : صحت مجموعه داده‌های Random روی مجموعة آموزش برای طبقه‌بندی جنگل تصادفی

جدول ‏3‑8 : نتایج مجموعه داده‌های Random روی مجموعة آزمون برای طبقه‌بندی جنگل تصادفی و مربوط به پارامترهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |
| --- | --- | --- | --- |
| مجموعه داده | صحت | حساسیت | اختصاصیت |
| Random-1 | ۰٫۶۷ | ۰٫۶۷ | ۰٫۶۷ |
| Random-2 | ۰٫۶۵ | ۰٫۶۷ | ۰٫۶۳ |
| Random-3 | ۰٫۶۸ | ۰٫۶۷ | ۰٫۷۰ |
| Random-4 | ۰٫۶۸ | ۰٫۷۳ | ۰٫۶۳ |
| Random-5 | ۰٫۶۸ | ۰٫۷۳ | ۰٫۶۳ |
| Random-6 | ۰٫۶۵ | ۰٫۷۰ | ۰٫۶۰ |

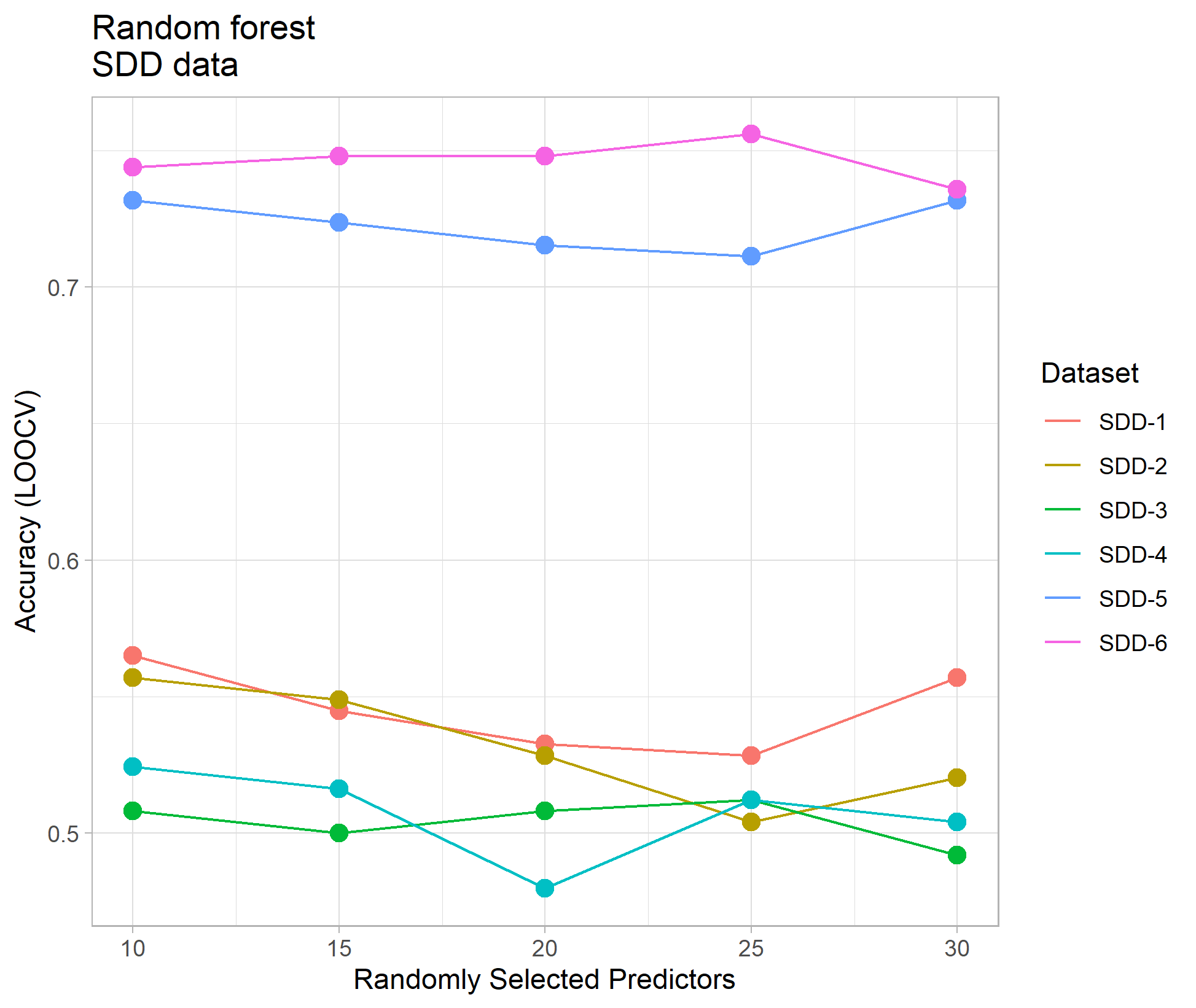
جدول ‏3‑9 : حساسیت مجموعه داده‌های Random روی مجموعة آزمون مستقل برای طبقه‌بندی جنگل تصادفی و مربوط به پارامترهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| مجموعه داده | Random-1 | Random-2 | Random-3 | Random-4 | Random-5 | Random-6 |
| حساسیت | ۰٫۱۳ | ۰٫۰7 | ۰٫۱۳ | ۰٫۰7 | ۰٫۲۰ | ۰٫۴۰ |

#### نتایج مربوط به مجموعه دادة SDD

جدول ‏3‑10 : صحت مجموعه داده‌های SDD روی مجموعة آموزش برای طبقه‌بندی جنگل تصادفی

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| mtry  مجموعه داده | ۱۰ | ۱۵ | ۲۰ | ۲۵ | ۳۰ |
| SDD-1 | ۰٫۵۷ | ۰٫۵۴ | ۰٫۵۳ | ۰٫۵۳ | ۰٫۵۶ |
| SDD-2 | ۰٫۵۶ | ۰٫۵۵ | ۰٫۵۳ | ۰٫۵۰ | ۰٫۵۲ |
| SDD-3 | ۰٫۵۱ | ۰٫۵۰ | ۰٫۵۱ | ۰٫۵۱ | ۰٫۴۹ |
| SDD-4 | ۰٫۵۲ | ۰٫۵۲ | ۰٫۴۸ | ۰٫51 | ۰٫۵۰ |
| SDD-5 | ۰٫۷۳ | ۰٫۷۲ | ۰٫۷۲ | ۰٫۷۱ | ۰٫۷۳ |
| SDD-6 | ۰٫۷۴ | ۰٫۷۵ | ۰٫۷۵ | ۰٫۷۶ | ۰٫۷۴ |



شکل ‏3‑ : صحت مجموعه داده‌های SDD روی مجموعة آموزش برای طبقه‌بندی جنگل تصادفی

جدول ‏3‑11 : نتایج مجموعه داده‌های SDD روی مجموعة آزمون برای طبقه‌بندی جنگل تصادفی و مربوط به پارامترهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |
| --- | --- | --- | --- |
| مجموعه داده | صحت | حساسیت | اختصاصیت |
| SDD-1 | ۰٫۴۳ | ۰٫۵۰ | ۰٫۳۷ |
| SDD-2 | ۰٫۳۸ | ۰٫۴۳ | ۰٫۳۳ |
| SDD-3 | ۰٫۳۳ | ۰٫۴۰ | ۰٫۲۷ |
| SDD-4 | ۰٫۳۳ | ۰٫۳۷ | ۰٫۳۰ |
| SDD-5 | ۰٫۶۲ | ۰٫۶۷ | ۰٫۵۷ |
| SDD-6 | ۰٫۶۲ | ۰٫۷۰ | ۰٫۵۳ |

جدول ‏3‑12 : حساسیت مجموعه داده‌های SDD روی مجموعة آزمون مستقل برای طبقه‌بندی جنگل تصادفی و مربوط به پارامترهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| مجموعه داده | SDD-1 | SDD-2 | SDD-3 | SDD-4 | SDD-5 | SDD-6 |
| حساسیت | ۰٫۸۰ | ۰٫۶۷ | ۰٫۸۰ | ۰٫۷۳ | ۰٫۶۰ | ۰٫۷۳ |

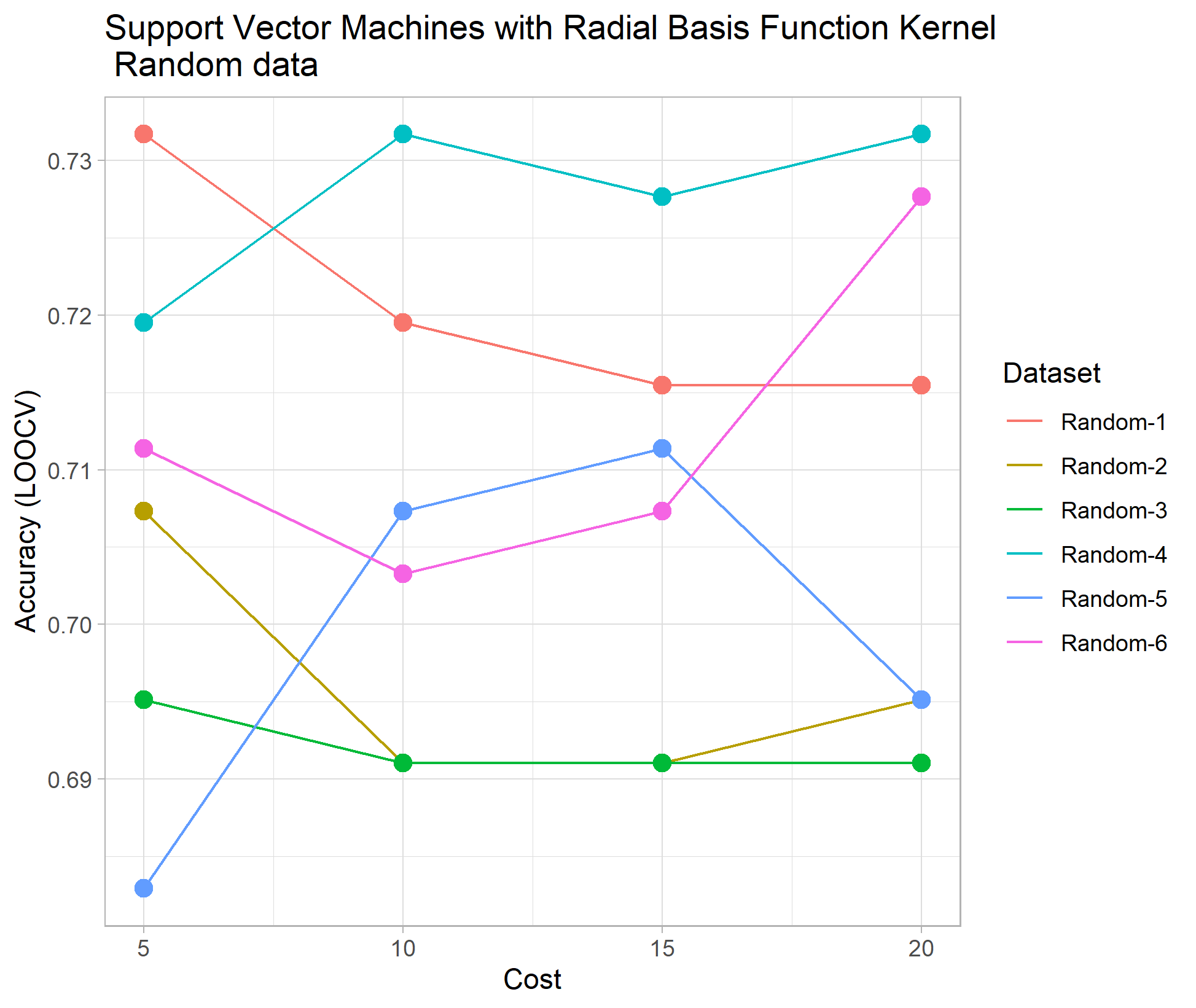
### نتایج مربوط به طبقه‌بندی ماشین بردار پیشتیبان با هستة شعاعی

برای یافتن بهترین پارامتر و تعداد ویژگی‌، این مدل برای هزینه‌های برابر با ۵، ۱۰، ۱۵ و ۲۰ روی مجموعه داده‌های Random-1، Random-2، Random-3، Random-4، Random-5 و Random-6 اجرا شد. همچنین این طبقه‌بندی با هزینه‌های گفته شده روی مجموعه داده‌های SDD-1، SDD-2، SDD-3، SDD-4، SDD-5 و SDD-6 اجرا شد. نتایج به شرح زیر است.

#### نتایج مربوط به مجموعه دادة Random

جدول ‏3‑13 : صحت مجموعه داده‌های Random روی مجموعة آموزش برای طبقه‌بندی ماشین بردار پیشتیبان با هستة شعاعی

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| هزینه  مجموعه داده | ۵ | ۱۰ | ۱۵ | ۲۰ |
| Random-1 | 0٫73 | 0٫7۲ | 0٫7۲ | 0٫7۲ |
| Random-2 | 0٫7۱ | 0٫69 | 0٫69 | 0٫۷۰ |
| Random-3 | 0٫۷۰ | 0٫69 | 0٫69 | 0٫69 |
| Random-4 | 0٫7۲ | 0٫73 | 0٫7۳ | 0٫73 |
| Random-5 | 0٫68 | 0٫7۱ | 0٫71 | 0٫۷۰ |
| Random-6 | 0٫71 | 0٫70 | 0٫7۱ | 0٫7۳ |



شکل ‏3‑ : صحت مجموعه داده‌های Random روی مجموعة آموزش برای طبقه‌بندی ماشین بردار پیشتیبان با هستة شعاعی

جدول ‏3‑14: نتایج مجموعه داده‌های Random روی مجموعة آزمون برای طبقه‌بندی ماشین بردار پیشتیبان با هستة شعاعی و مربوط به پارامترهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |
| --- | --- | --- | --- |
| مجموعه داده | صحت | حساسیت | اختصاصیت |
| Random-1 | ۰٫۶۳ | ۰٫۷۰ | ۰٫۵۷ |
| Random-2 | ۰٫۶۳ | ۰٫۷۳ | ۰٫۵۳ |
| Random-3 | ۰٫۶۳ | ۰٫۷۳ | ۰٫۵۳ |
| Random-4 | ۰٫۶۷ | ۰٫۷۷ | ۰٫۵۷ |
| Random-5 | ۰٫۶۷ | ۰٫۸۰ | ۰٫۵۳ |
| Random-6 | ۰٫۶۳ | ۰٫۷۰ | ۰٫۵۷ |

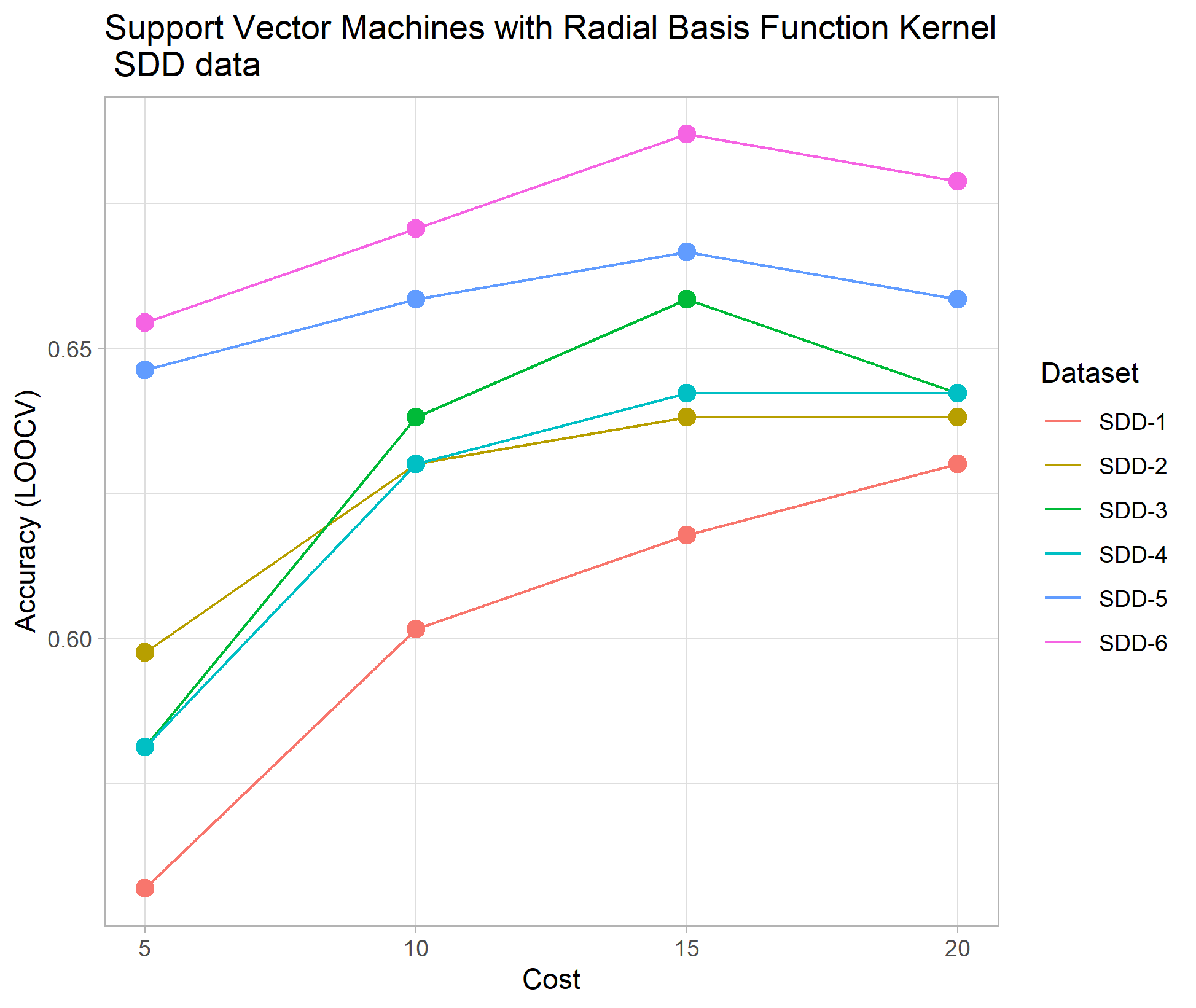
جدول ‏3‑15 : حساسیت مجموعه داده‌های Random روی مجموعة آزمون مستقل برای طبقه‌بندی ماشین بردار پیشتیبان با هستة شعاعی و مربوط به پارامترهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| مجموعه داده | Random-1 | Random-2 | Random-3 | Random-4 | Random-5 | Random-6 |
| حساسیت | 0٫2۰ | 0٫13 | 0٫07 | 0٫13 | 0٫33 | 0 |

#### نتایج مربوط به مجموعه دادة SDD

جدول ‏3‑16 : صحت مجموعه داده‌های SDD روی مجموعة آموزش برای طبقه‌بندی ماشین بردار پیشتیبان با هستة شعاعی

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| هزینه  مجموعه داده | ۵ | ۱۰ | ۱۵ | ۲۰ |
| SDD-1 | 0٫5۶ | 0٫60 | 0٫6۲ | 0٫63 |
| SDD-2 | 0٫۶۰ | 0٫63 | 0٫6۴ | 0٫6۴ |
| SDD-3 | 0٫58 | 0٫6۴ | 0٫6۶ | 0٫64 |
| SDD-4 | 0٫58 | 0٫63 | 0٫64 | 0٫64 |
| SDD-5 | 0٫6۵ | 0٫6۶ | 0٫67 | 0٫6۶ |
| SDD-6 | 0٫65 | 0٫67 | 0٫6۹ | 0٫6۸ |



شکل ‏3‑ : صحت مجموعه داده‌های SDD روی مجموعة آموزش برای طبقه‌بندی ماشین بردار پیشتیبان با هستة شعاعی

جدول ‏3‑17 : نتایج مجموعه داده‌های SDD روی مجموعة آزمون برای طبقه‌بندی ماشین بردار پیشتیبان با هستة شعاعی و مربوط به پارامترهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |
| --- | --- | --- | --- |
| مجموعه داده | صحت | حساسیت | اختصاصیت |
| SDD-1 | ۰٫۵۰ | ۰٫۴۷ | ۰٫۵۳ |
| SDD-2 | ۰٫۵۰ | ۰٫۵۷ | ۰٫۴۳ |
| SDD-3 | ۰٫۴7 | ۰٫۵۷ | ۰٫۳۷ |
| SDD-4 | ۰٫۴۸ | ۰٫۵۷ | ۰٫۴۰ |
| SDD-5 | ۰٫۶۲ | ۰٫۶۰ | ۰٫۶۳ |
| SDD-6 | ۰٫۶۵ | ۰٫۷۰ | ۰٫۶۰ |

جدول ‏3‑18 : حساسیت مجموعه داده‌های SDD روی مجموعة آزمون مستقل برای طبقه‌بندی ماشین بردار پیشتیبان با هستة شعاعی و مربوط به پارامترهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| مجموعه داده | SDD-1 | SDD-2 | SDD-3 | SDD-4 | SDD-5 | SDD-6 |
| حساسیت | 0٫47 | 0٫6۰ | 0٫67 | 0٫6۰ | 0٫27 | 0٫87 |

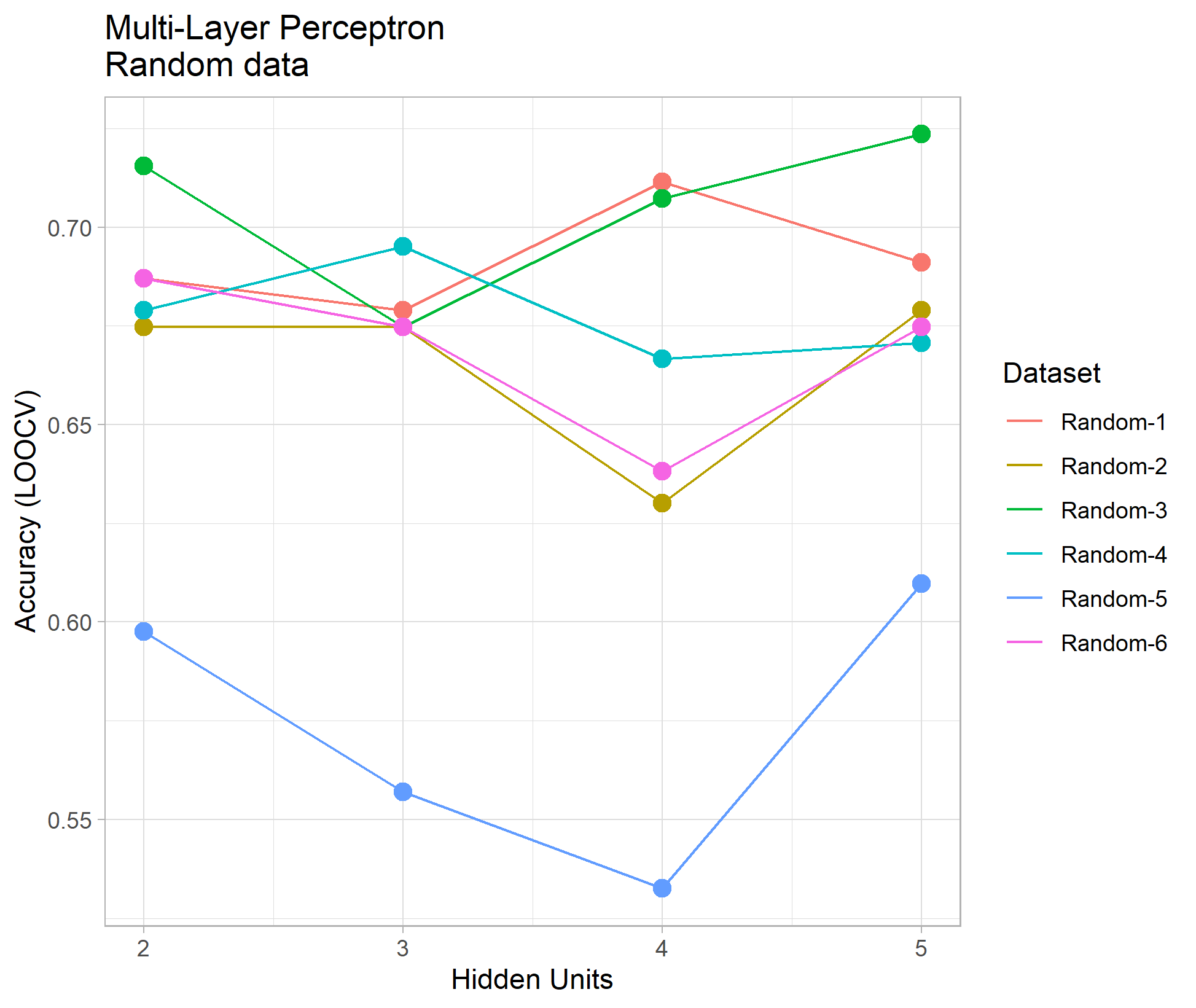
### نتایج مربوط به طبقه‌بندی شبکه عصبی چند لایه

این مدل برای تعداد لایه‌های درونی‌های برابر با ۲، ۳، ۴ و ۵ روی مجموعه داده‌های Random-1، Random-2، Random-3، Random-4، Random-5 و Random-6، برای یافتن بهترین پارامتر و تعداد ویژگی‌ اجرا شد. همچنین این طبقه‌بندی با تعداد لایه‌های درونی‌ گفته شده روی مجموعه داده‌های SDD-1، SDD-2، SDD-3، SDD-4، SDD-5 و SDD-6 اجرا شد. نتایج به شرح زیر است.

#### نتایج مربوط به مجموعه دادة Random

جدول ‏3‑19 : صحت مجموعه داده‌های Random روی مجموعة آموزش برای طبقه‌بندی شبکه عصبی چند لایه

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| لایه‌های درونی  مجموعه داده | ۲ | ۳ | ۴ | ۵ |
| Random-1 | 0٫6۹ | 0٫6۸ | 0٫71 | 0٫69 |
| Random-2 | 0٫67 | 0٫67 | 0٫63 | 0٫6۸ |
| Random-3 | 0٫7۲ | 0٫67 | 0٫7۱ | 0٫72 |
| Random-4 | 0٫6۸ | 0٫۷۰ | 0٫67 | 0٫67 |
| Random-5 | 0٫۶۰ | 0٫5۶ | 0٫53 | 0٫6۱ |
| Random-6 | 0٫6۹ | 0٫67 | 0٫6۴ | 0٫67 |



شکل ‏3‑ : صحت مجموعه داده‌های Random روی مجموعة آموزش برای طبقه‌بندی شبکه عصبی چند لایه

جدول ‏3‑20 : نتایج مجموعه داده‌های Random روی مجموعة آزمون برای طبقه‌بندی شبکه عصبی چند لایه و مربوط به پارامترهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |
| --- | --- | --- | --- |
| مجموعه داده | صحت | حساسیت | اختصاصیت |
| Random-1 | ۰٫۵۵ | ۰٫۶۷ | ۰٫۴۳ |
| Random-2 | ۰٫۶۷ | ۰٫۶۰ | ۰٫۷۳ |
| Random-3 | ۰٫۶۰ | ۰٫۶۳ | ۰٫۵۷ |
| Random-4 | ۰٫۶۵ | ۰٫۶۷ | ۰٫۶۳ |
| Random-5 | ۰٫۶۵ | ۰٫۸۰ | ۰٫۵۰ |
| Random-6 | ۰٫۶۲ | ۰٫۶۰ | ۰٫۶۳ |

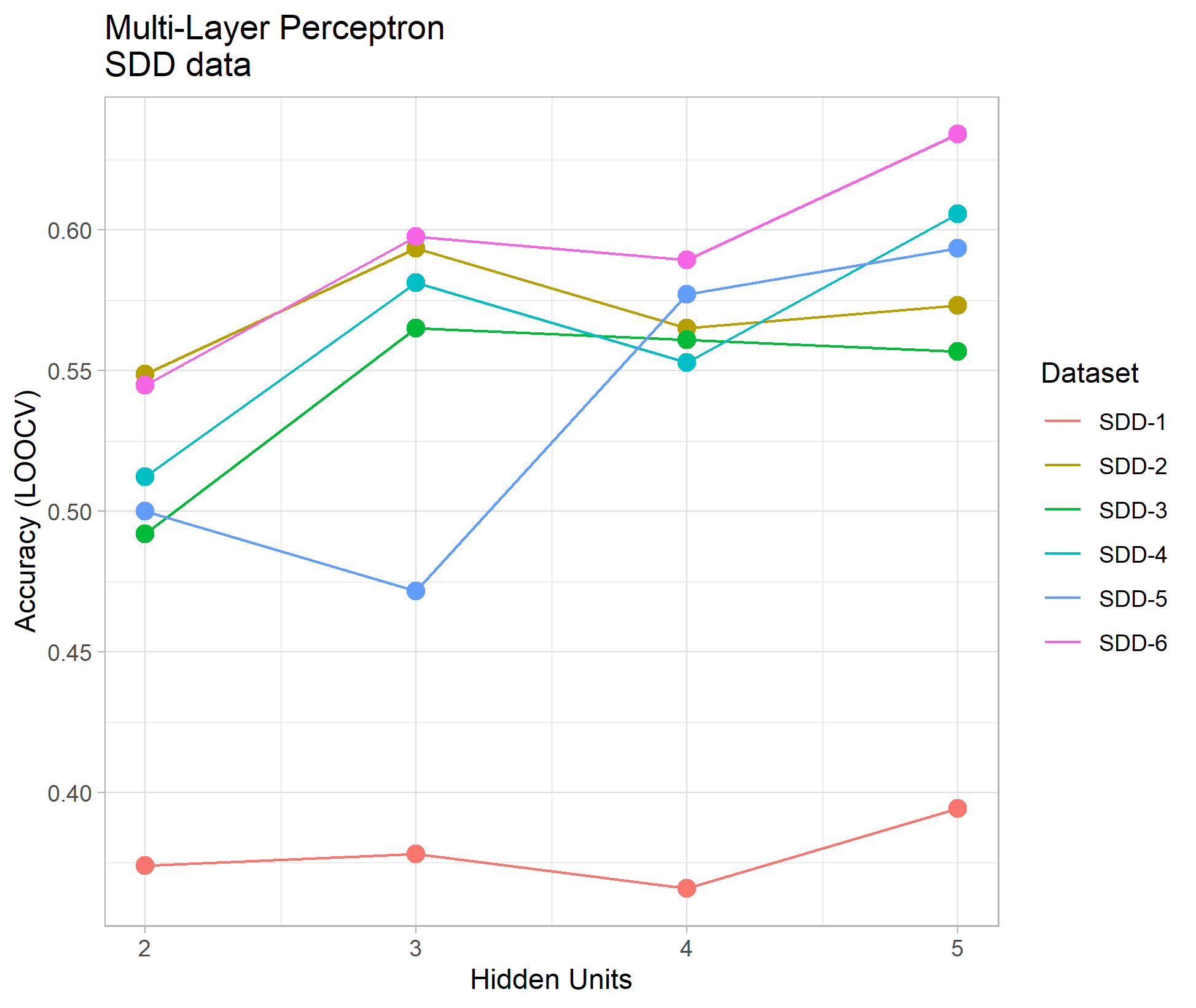
جدول ‏3‑21 : حساسیت مجموعه داده‌های Random روی مجموعة آزمون مستقل برای طبقه‌بندی شبکه عصبی چند لایه و مربوط به پارامترهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| مجموعه داده | Random-1 | Random-2 | Random-3 | Random-4 | Random-5 | Random-6 |
| حساسیت | 0٫2۰ | 0٫13 | 0٫2۰ | 0٫27 | 0 | 0٫6۰ |

#### نتایج مربوط به مجموعه دادة SDD

جدول ‏3‑22 : صحت مجموعه داده‌های SDD روی مجموعة آموزش برای طبقه‌بندی شبکه عصبی چند لایه

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| لایه‌های درونی  مجموعه داده | ۲ | ۳ | ۴ | ۵ |
| SDD-1 | 0٫37 | 0٫3۸ | 0٫3۷ | 0٫39 |
| SDD-2 | 0٫5۵ | 0٫59 | 0٫5۷ | 0٫57 |
| SDD-3 | 0٫49 | 0٫5۷ | 0٫56 | 0٫5۶ |
| SDD-4 | 0٫51 | 0٫58 | 0٫55 | 0٫61 |
| SDD-5 | 0٫5۰ | 0٫47 | 0٫5۸ | 0٫59 |
| SDD-6 | 0٫54 | 0٫۶۰ | 0٫5۹ | 0٫63 |



شکل ‏3‑ : صحت مجموعه داده‌های SDD روی مجموعة آموزش برای طبقه‌بندی شبکه عصبی چند لایه

جدول ‏3‑23 : نتایج مجموعه داده‌های SDD روی مجموعة آزمون برای طبقه‌بندی شبکه عصبی چند لایه و مربوط به پارامترهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |
| --- | --- | --- | --- |
| مجموعه داده | صحت | حساسیت | اختصاصیت |
| SDD-1 | ۰٫۳۵ | ۰٫۱۰ | ۰٫۶۰ |
| SDD-2 | ۰٫۴۵ | ۰٫۲۷ | ۰٫۶۳ |
| SDD-3 | ۰٫۴۵ | ۰٫۶۳ | ۰٫۲۷ |
| SDD-4 | ۰٫۴۷ | ۰٫۵۷ | ۰٫۳۷ |
| SDD-5 | ۰٫۵۳ | ۰٫۷۰ | ۰٫۳۷ |
| SDD-6 | ۰٫۴۲ | ۰٫۳۳ | ۰٫۵۰ |

جدول ‏3‑24 : حساسیت مجموعه داده‌های SDD روی مجموعة آزمون مستقل برای طبقه‌بندی شبکه عصبی چند لایه و مربوط به پارامترهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| مجموعه داده | SDD-1 | SDD-2 | SDD-3 | SDD-4 | SDD-5 | SDD-6 |
| حساسیت | 0٫87 | 0٫53 | 0٫6۰ | 0٫87 | 0٫67 | 0٫33 |

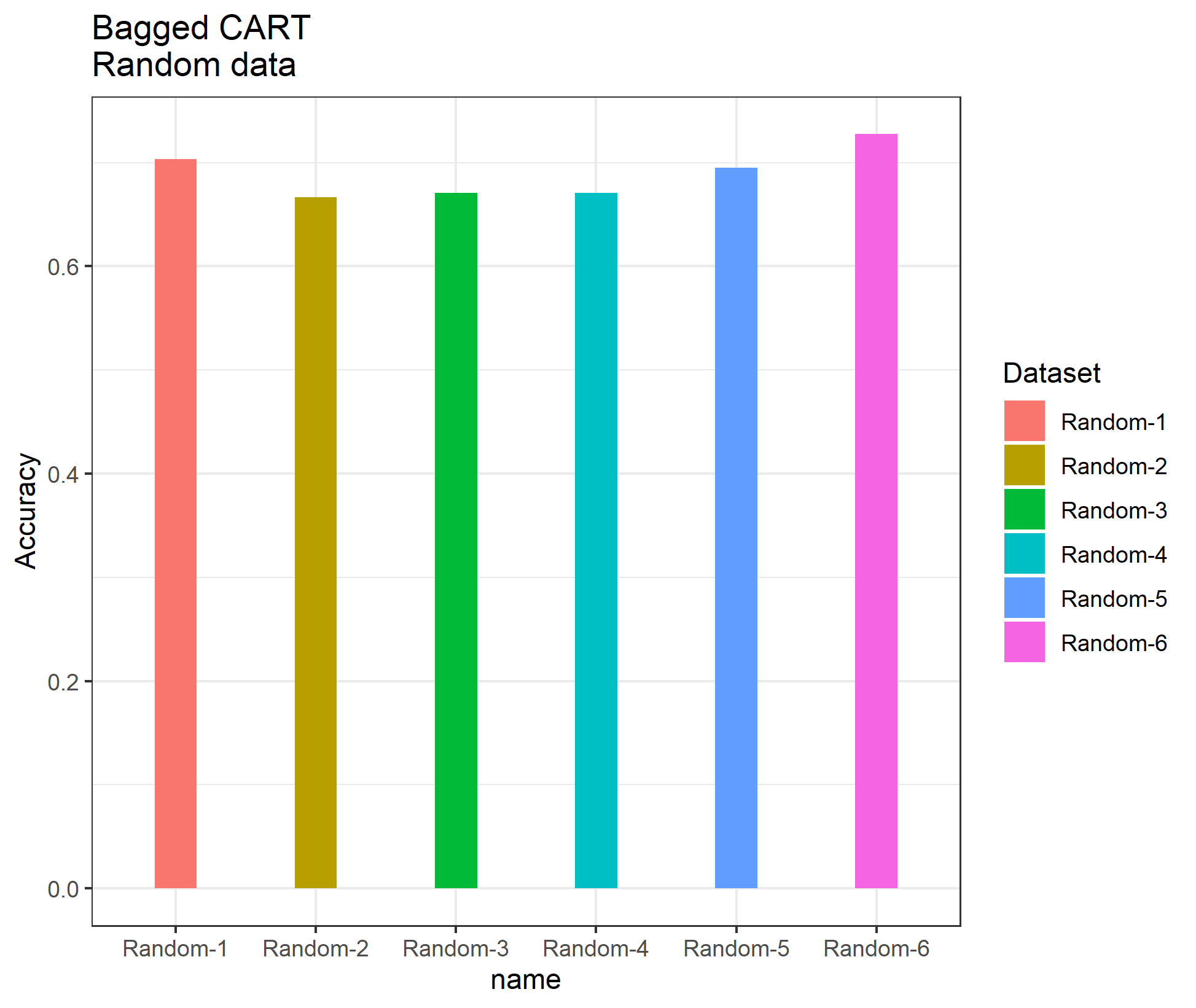
### نتایج مربوط به طبقه‌بندی TreeBag

این روش پارامتری برای تنظیم جهت ساختن مدل با کارایی بهتر نداشت؛ و آن روی مجموعه داده‌های Random-1، Random-2، Random-3، Random-4، Random-5 و Random-6 و همچنین روی مجموعه داده‌های SDD-1، SDD-2، SDD-3، SDD-4، SDD-5 و SDD-6 اجرا شد. نتایج به شرح زیر است.

#### نتایج مربوط به مجموعه دادة Random

جدول ‏3‑25 : صحت طبقه‌بندی TreeBag روی مجموعة آموزش و برای برای هر یک از مجموعه داده‌های Random

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| مجموعه داده | Random-1 | Random-2 | Random-3 | Random-4 | Random-5 | Random-6 |
| صحت | 0٫7۰ | 0٫67 | 0٫67 | 0٫67 | 0٫۷۰ | 0٫7۳ |

****

شکل ‏3‑ : صحت طبقه‌بندی TreeBag روی مجموعة آموزش و برای برای هر یک از مجموعه داده‌های Random

جدول ‑26 : نتایج مجموعه داده‌های Random روی مجموعة آزمون برای طبقه‌بندی TreeBag

|  |  |  |  |
| --- | --- | --- | --- |
| مجموعه داده | صحت | حساسیت | اختصاصیت |
| Random-1 | ۰٫۶۳ | ۰٫۶۷ | ۰٫۶۰ |
| Random-2 | ۰٫۶۷ | ۰٫۷۳ | ۰٫۶۰ |
| Random-3 | ۰٫۶۵ | ۰٫۶۳ | ۰٫۶7 |
| Random-4 | ۰٫۶۳ | ۰٫۶۳ | ۰٫۶۳ |
| Random-5 | ۰٫۶۳ | ۰٫۶۷ | ۰٫۶۰ |
| Random-6 | ۰٫۶۵ | ۰٫۶۷ | ۰٫۶۳ |

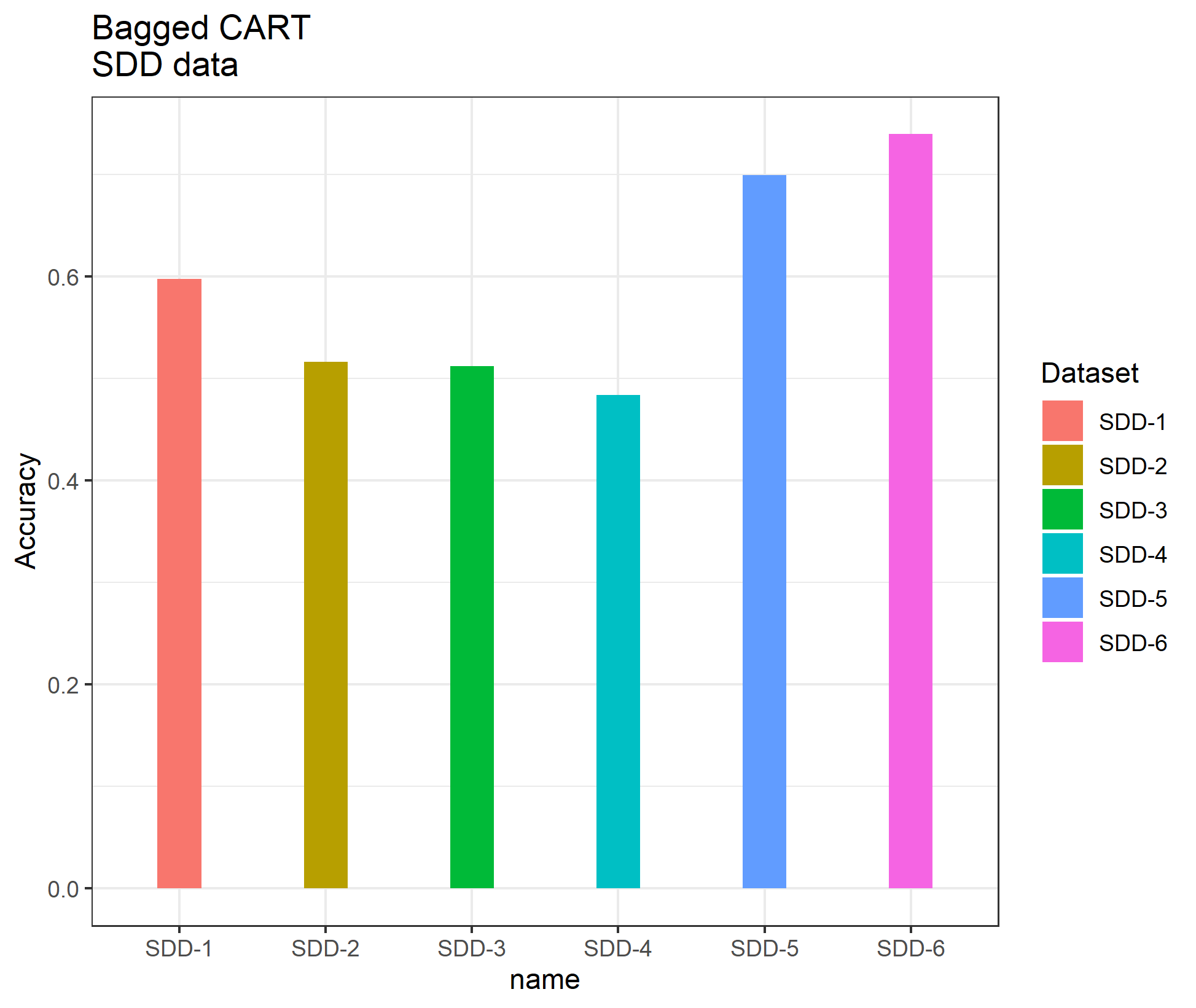
جدول ‏3‑27 : حساسیت مجموعه داده‌های Random روی مجموعة آزمون مستقل برای طبقه‌بندی TreeBag

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| مجموعه داده | Random-1 | Random-2 | Random-3 | Random-4 | Random-5 | Random-6 |
| حساسیت | 0٫13 | 0٫27 | 0٫33 | 0٫2۰ | 0٫07 | 0٫27 |

#### نتایج مربوط به مجموعه دادة SDD

جدول ‏3‑28 : صحت طبقه‌بندی TreeBag روی مجموعة آموزش و برای برای هر یک از مجموعه داده‌های SDD

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| مجموعه داده | SDD-1 | SDD-2 | SDD-3 | SDD-4 | SDD-5 | SDD-6 |
| صحت | 0٫۶۰ | 0٫5۲ | 0٫51 | 0٫48 | 0٫۷۰ | 0٫7۴ |

****

شکل ‏3‑ : صحت طبقه‌بندی TreeBag روی مجموعة آموزش و برای برای هر یک از مجموعه داده‌های SDD

جدول ‏3‑29 : نتایج مجموعه داده‌های SDD روی مجموعة آزمون برای طبقه‌بندی TreeBag

|  |  |  |  |
| --- | --- | --- | --- |
| مجموعه داده | صحت | حساسیت | اختصاصیت |
| SDD-1 | ۰٫۵۸ | ۰٫۶۳ | ۰٫۵۳ |
| SDD-2 | ۰٫۴۰ | ۰٫۳۷ | ۰٫۴۳ |
| SDD-3 | ۰٫۳۸ | ۰٫۵۰ | ۰٫۲۷ |
| SDD-4 | ۰٫۳۸ | ۰٫۴۰ | ۰٫۳۷ |
| SDD-5 | ۰٫۵۷ | ۰٫۵۷ | ۰٫۵۷ |
| SDD-6 | ۰٫۶۵ | ۰٫۷۰ | ۰٫۶۰ |

جدول ‏3‑30 : حساسیت مجموعه داده‌های SDD روی مجموعة آزمون برای طبقه‌بندی TreeBag

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| مجموعه داده | SDD-1 | SDD-2 | SDD-3 | SDD-4 | SDD-5 | SDD-6 |
| حساسیت | 0٫4۰ | 0٫47 | 0٫73 | 0٫47 | 0٫4۰ | 0٫67 |

نتایج تمام طبقه‌بندی‌ها روی دادة Random و SDD در جدول ‏3‑31 و جدول ‏3‑32 آمده است.

جدول ‏3‑31: نتایج تمام طبقه‌بندی‌های برگزیده برای مجموعه دادة Random روی مجموعه‌های آموزش، آزمون و آزمون مستقل

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| روش طبقه‌بندی | بهترین مجموعه داده (از نظر تعداد ویژگی) | بهترین پارامترها | صحت مجموعة آموزش | صحت مجموعة آزمون | حساسیت مجموعة آزمون مستقل |
| k-نزدیکترین همسایه | Random-6 |  | ۰٫۷2 | ۰٫۷۲ | ۰٫13 |
| Naïve Bayes | Random-1 |  | ۰٫۶۳ | ۰٫۶۵ | ۰٫۲۷ |
| آنالیز تشخیصی خطی | Ramdom-5 | - | ۰٫59 | ۰٫۶5 | ۰٫۲0 |
| جنگل تصادفی | Ramdom-5 |  | ۰٫۷۲ | ۰٫۶۸ | ۰٫۲۰ |
| ماشین بردار پیشتیبان با هستة خطی | Random-5 |  | ۰٫59 | ۰٫۶5 | ۰٫07 |
| ماشین بردار پیشتیبان با هستة شعاعی | Random-5 |  | ۰٫۷۱ | ۰٫۶۷ | ۰٫۳۳ |
| شبکه عصبی چند لایه | Random-2 | # | ۰٫۶۸ | ۰٫۶۷ | ۰٫۱۳ |
| treebag | Random-2 | - | ۰٫۶۷ | ۰٫۶۷ | ۰٫۲۷ |
| آنالیز تفکیک‌کننده درجه دوم | Random-6 | - | ۰٫۶۵ | ۰٫۶۵ | ۰٫۱۳ |

جدول ‏3‑32: نتایج تمام طبقه‌بندی‌های برگزیده برای مجموعه دادة SDD روی مجموعه‌های آموزش، آزمون و آزمون مستقل

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| روش طبقه‌بندی | بهترین مجموعه داده (از نظر تعداد ویژگی) | بهترین پارامترها | صحت مجموعة آموزش | صحت مجموعة آزمون | حساسیت مجموعة آزمون مستقل |
| k-نزدیکترین همسایه | SDD-5 |  | ۰٫۷۶ | ۰٫۷۲ | ۰٫۴۷ |
| Naïve Bayes | SDD-5 |  | ۰٫۵۷ | ۰٫۵۳ | ۰٫۵۳ |
| آنالیز تشخیصی خطی | SDD-5 | - | ۰٫۴۵ | ۰٫۴۰ | ۰٫۷۳ |
| جنگل تصادفی | SDD-6 |  | ۰٫۷۶ | ۰٫۶۲ | ۰٫۷۳ |
| ماشین بردار پیشتیبان با هستة خطی | SDD-5 |  | ۰٫۵۲ | ۰٫۴۲ | ۰٫۸۷ |
| ماشین بردار پیشتیبان با هستة شعاعی | SDD-6 |  | ۰٫۶۹ | ۰٫۶۵ | ۰٫۸۷ |
| شبکه عصبی چند لایه | SDD-5 |  | ۰٫۵۹ | ۰٫۵۳ | ۰٫۶۷ |
| treebag | SDD-6 | - | ۰٫۷۴ | ۰٫۶۵ | ۰٫۶۷ |
| آنالیز تفکیک‌کننده درجه دوم | SDD-5 | - | ۰٫۵۲ | ۰٫۴۲ | ۰٫۵۳ |

# فصل چهارم: بحث و پیشنهادات

با توجه به نتایج مربوط به طبقه‌بندی k-نزدیکترین همسایه و صحت گزارش شده برای داده‌های آزمون Random در جدول ‏3‑2، می‌توان نتیجه گرفت؛ این طبقه‌بندی برای k برابر با یک و مجموعة Random-6 بهترین صحت را دارد. همچنین با توجه به صحت گزارش شده برای داده‌های آزمون SDD در جدول ‏3‑5، می‌توان نتیجه گرفت؛ این طبقه‌بندی برای k برابر با یک و مجموعة SDD-5 بهترین صحت را دارد.

نتایج مربوط به طبقه‌بندی جنگل تصادفی و صحت گزارش شده برای داده‌های آزمون Random در جدول ‏3‑8 نشان می‌دهد؛ این طبقه‌بندی برای مجموعه‌های Random-4 و Random-5 بهترین صحت و حساسیت را دارد. اما چون صحت مجموعة آموزش Random-5 بالاتر بود؛ این مجموعه با تعداد ده ویژگی که به صورت تصادفی انتخاب شده است؛ برگزیده شد. همچنین با توجه به صحت گزارش شده برای داده‌های آزمون SDD در جدول ‏3‑11، می‌توان نتیجه گرفت؛ این طبقه‌بندی با تعداد ۲۵ ویژگی که به صورت تصادفی انتخاب شده است و برای مجموعة SDD-6 بهترین صحت را دارد.

با توجه به نتایج مربوط به ماشین بردار پیشتیبان با هستة شعاعی و صحت گزارش شده برای داده‌های آزمون Random در جدول ‏3‑14، می‌توان نتیجه گرفت؛ این طبقه‌بندی برای مجموعة Random-5، با هزینة برابر با ۱۵، بهترین صحت و حساسیت را دارد. همچنین صحت گزارش شده برای داده‌های آزمون SDD در جدول ‏3‑17نشان می‌دهد؛ این طبقه‌بندی برای مجموعة SDD-6 و با هزینة برابر با ۱۵، بهترین صحت را دارد.

از نتایج مربوط به شبکه عصبی چند لایه و صحت گزارش شده برای داده‌های آزمون Random در جدول ‏3‑20، می‌توان نتیجه گرفت؛ این طبقه‌بندی برای مجموعة Random-2 و با تعداد لایه‌های درونی برابر با ۵، بهترین صحت و حساسیت را دارد. همچنین صحت گزارش شده برای داده‌های آزمون SDD در جدول ‏3‑23، نتیجه می‌دهد؛ این طبقه‌بندی برای مجموعة SDD-5 و با تعداد لایه‌های درونی برابر با ۵، بهترین صحت را دارد.

نتایج مربوط به روش TreeBag و صحت گزارش شده برای داده‌های آزمون Random در جدول ‏3‑26، نتیجه می‌دهد؛ این طبقه‌بندی برای مجموعة Random-2 بهترین صحت را دارد. همچنین از صحت گزارش شده برای داده‌های آزمون SDD در جدول ‏3‑29، می‌توان نتیجه گرفت؛ این طبقه‌بندی برای مجموعة SDD-6 بهترین صحت را دارد.

با توجه به خلاصه تمام نتایج مربوط به مجموعه دادة Random در جدول ‏3‑31 و صحت مجموعه‌های آزمون می‌توان نتیجه گرفت؛ برای مجموعه دادة Random طبقه‌بندی k-نزدیکترین همسایه با k برابر با یک و با ده مولفة اصلی اول بهترین مدل خواهد بود. اما با توجه به حساسیت مجموعه دادة آزمون مستقل مشخص می‌شود؛ نه تنها این مدل بلکه سایر مدل‌ها توانایی پیش‌بینی خوبی روی مجموعه داده‌های جدید ندارند. با توجه به خلاصه تمام نتایج مربوط به مجموعه دادة SDD جدول ‏3‑32 و صحت مجموعه‌های آزمون می‌توان نتیجه گرفت؛ برای مجموعه دادة SDD طبقه‌بندی k-نزدیکترین همسایه با k برابر با یک و با پنج مولفة اصلی اول بهترین مدل خواهد بود.

همچنین نتایج جدول ‏3‑31 و جدول ‏3‑32 نشان می‌دهد، اگرچه صحت مدل‌هایی که روی مجموعة آموزش و مجموعة آزمون Random و SDD اجرا شد؛ تقریباً یکسان است و حتی در بعضی موارد روی مجموعة Random بهتر است؛ اما مدل‌های ساخته شده روی مجموعة SDD روی مجموعه دادة مستقل بهتر عمل می‌کنند. در نتیجه مدل‌های ساخته شده روی مجموعه SDD برهمکنش پروتئین-پروتئین داده‌های جدید را بهتر و مطمئن‌تر پیش‌بینی می‌کنند.

نتایج نشان می‌دهد نوع انتخاب داده‌های منفی، روی کارآیی مدل‌ها تأثیر زیادی دارد و در بعضی موارد انتخاب داده‌های منفی به صورت تصادفی به اشتباه کارآیی مدل را بیشتر تخمین می‌زند. در صورتی که انتخاب داده‌های منفی به صورتی که درجة هر رأس، در شبکة مثبت و منفی یکسان باشد، باعث می‌شود پیش‌بینی برهمکنش پروتئین-پروتئین داده‌های جدید بهتر شود. همچنین روش تحلیل مؤلفه‌های اصلی برای این داده‌های این مطالعه، بهتر از روش انتخاب ویژگی بر اساس اهمیت و سهمشان در پیش‌بینی، عمل می‌کند.

اجرای طبقه‌بندی‌های مختلف روی هر دو نوع مجموعه داده نشان می‌دهد؛ طبقه‌بندی‌هایی که از قضیة بیز برای طبقه‌بندی داده‌ها استفاده می‌کنند؛ مانند Naïve Bayes، آنالیز تشخیص خطی و آنالیز تفکیک کنندة درجه دو؛ توانایی خوبی برای طبقه‌بند‌ی مجموعه دادة این مطالعه ندارند. اما طبقه‌بندی k-نزدیکترین همسایه و طبقه‌بندی‌هایی که از درخت تصمیم استفاده می‌کنند؛ مانند جنگل تصادفی و TreeBag؛ داده‌های این مطالعه را بهتر می‌توانند طبقه‌بندی کنند.

نتایجی که از این مطالعه به دست آمده است نشان می‌دهد روش‌های یادگیری ماشین برای پیش‌بینی برهمکنش بین گیاه *Arabidopsis thaliana* و باکتری *Pseudomonas syringae* می‌تواند بسیار مفید باشد؛ اما از کمبود داده‌های آزمایشگاهی رنج می‌برد. دسترسی به داده‌های آزمایشگاهی بیشتر برهمکنش گیاه-پاتوژن به بالا بردن کارآیی مدل‌های یادگیری ماشین و در نتیجه پیش‌بینی‌های با صحت بالاتر کمک کند. پیش‌بینی برهمکنش پروتئین-پروتئین با صحت بالاتر می‌تواند به انجام کارهای آزمایشگاهی برای یافتن مکانسیم بیمار شدن گیاه و روش‌های درمانی سرعت ‌بخشد.

##### مراجع

##### پیوست‌ها

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

**چهار: نتایج مربوط به ماشین بردار پیشتیبان با هستة خطی**

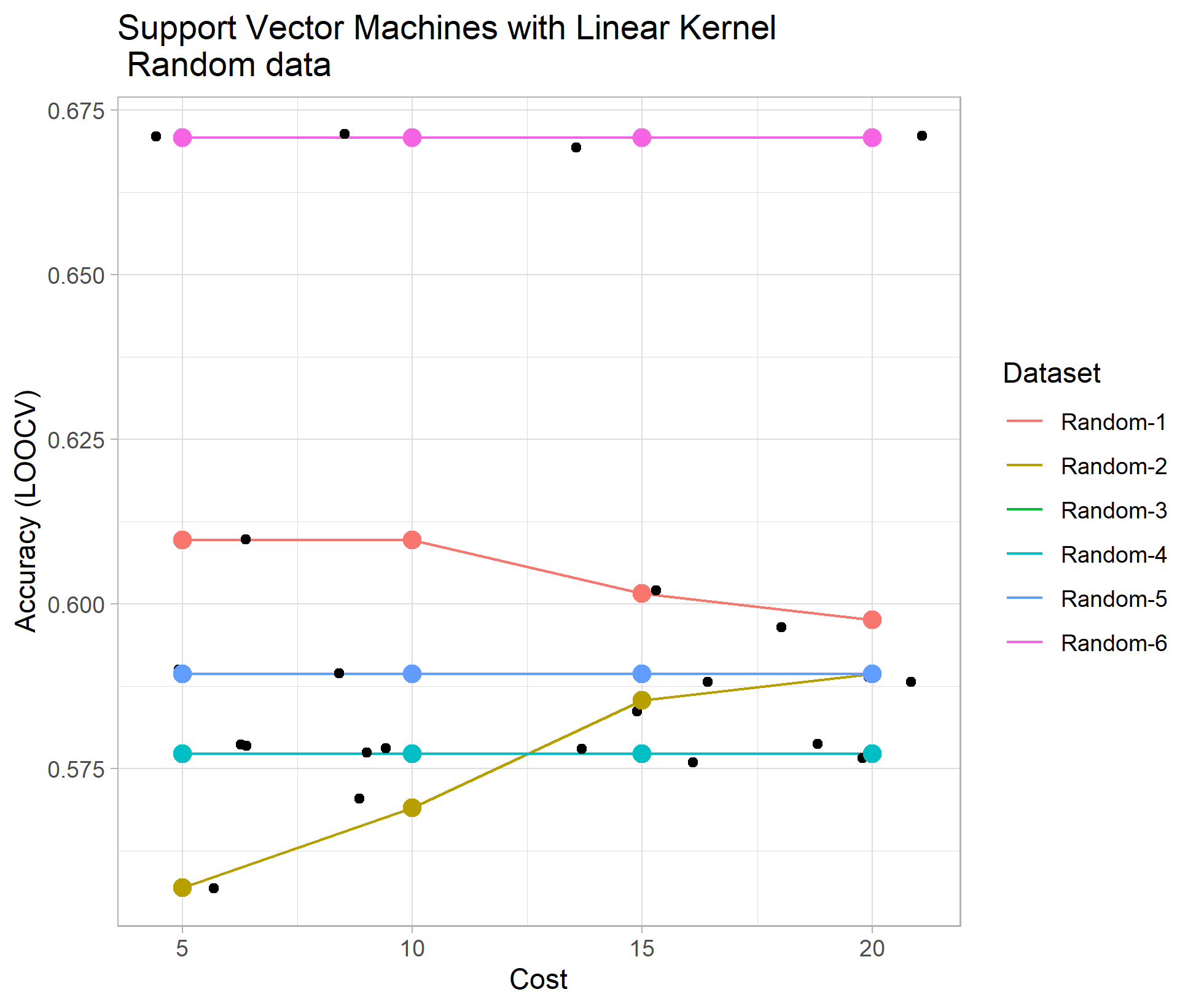
برای یافتن بهترین پارامتر و تعداد ویژگی‌، این مدل برای هزینه‌های برابر با ۵، ۱۰، ۱۵ و ۲۰ روی مجموعه داده‌های Random-1، Random-2، Random-3، Random-4، Random-5 و Random-6 اجرا شد. همچنین این طبقه‌بندی با هزینه‌های گفته شده روی مجموعه داده‌های SDD-1، SDD-2، SDD-3، SDD-4، SDD-5 و SDD-6 اجرا شد. نتایج به شرح زیر است.

با توجه به صحت گزارش شده برای داده‌های آزمون Random در جدول الف- 20، می‌توان نتیجه گرفت؛ این طبقه‌بندی برای مجموعة Random-5 و هزینة برابر با ۵، بهترین صحت را دارد. همچنین با توجه به صحت گزارش شده برای داده‌های آزمون SDD در جدول الف- 23، می‌توان نتیجه گرفت؛ این طبقه‌بندی برای مجموعة SDD-5 و هزینة برابر با ۵، بهترین صحت را دارد.

**چهار-یک: نتایج مربوط به مجموعه دادة Random**

جدول الف- 19: صحت مجموعه داده‌های Random روی مجموعة آموزش برای طبقه‌بندی ماشین بردار پیشتیبان با هستة خطی

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| هزینه  مجموعه داده | ۵ | ۱۰ | ۱۵ | ۲۰ |
| Random-1 | 0٫6۱ | 0٫6۱ | 0٫60 | 0٫۶۰ |
| Random-2 | 0٫5۶ | 0٫5۷ | 0٫5۹ | 0٫5۹ |
| Random-3 | 0٫5۸ | 0٫5۸ | 0٫5۸ | 0٫5۸ |
| Random-4 | 0٫5۸ | 0٫5۸ | 0٫5۸ | 0٫5۸ |
| Random-5 | 0٫5۹ | 0٫5۹ | 0٫5۹ | 0٫5۹ |
| Random-6 | 0٫67 | 0٫67 | 0٫67 | 0٫67 |



شکل الف- 5: صحت مجموعه داده‌های Random روی مجموعة آموزش برای طبقه‌بندی ماشین بردار پیشتیبان با هستة خطی (Random-3 و Ramdom-4 روی هم قرار گرفته است؛ به همین دلیل نقاط با کمی تکان با دایره های مشکی رسم شده اند.)

جدول الف- 20: نتایج مجموعه داده‌های Random روی مجموعة آزمون برای طبقه‌بندی ماشین بردار پیشتیبان با هستة خطی و مربوط به پارامترهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |
| --- | --- | --- | --- |
| مجموعه داده | صحت | حساسیت | اختصاصیت |
| Random-1 | ۰٫۵۳ | ۰٫۵۷ | ۰٫۵۰ |
| Random-2 | ۰٫۵۵ | ۰٫۶۷ | ۰٫۴۳ |
| Random-3 | ۰٫۵۷ | ۰٫۷۰ | ۰٫۴۳ |
| Random-4 | ۰٫۵۵ | ۰٫۷۰ | ۰٫۴۰ |
| Random-5 | ۰٫۶۵ | ۰٫۵۳ | ۰٫۷۷ |
| Random-6 | ۰٫۶۰ | ۰٫۶۳ | ۰٫۵۷ |

جدول الف- 21: حساسیت مجموعه داده‌های Random روی مجموعة آزمون مستقل برای طبقه‌بندی ماشین بردار پیشتیبان با هستة خطی و مربوط به پارامترهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| مجموعه داده | Random-1 | Random-2 | Random-3 | Random-4 | Random-5 | Random-6 |
| حساسیت | 0٫13 | 0٫53 | 0٫4۰ | 0٫53 | 0٫07 | 0٫2۷ |

**چهار-دو**الف- 23 : نتایج مجموعه داده‌های SDD روی مجموعة آزمون برای طبقه‌بندی ماشین بردار پیشتیبان با هستة خطی و مربوط به پارامترهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |
| --- | --- | --- | --- |
| مجموعه داده | صحت | حساسیت | اختصاصیت |
| SDD-1 | ۰٫۲۲ | ۰٫۳۰ | ۰٫۱۳ |
| SDD-2 | ۰٫۱۷ | ۰٫۲۳ | ۰٫۱۰ |
| SDD-3 | ۰٫۱۸ | ۰٫۲۷ | ۰٫۱۰ |
| SDD-4 | ۰٫۲۰ | ۰٫۳۰ | ۰٫۱۰ |
| SDD-5 | ۰٫۴۲ | ۰٫۳۰ | ۰٫۵۳ |
| SDD-6 | ۰٫۳۳ | ۰٫۲۷ | ۰٫۴۰ |

جدول الف- 24 : حساسیت مجموعه داده‌های SDD روی مجموعة آزمون مستقل برای طبقه‌بندی ماشین بردار پیشتیبان با هستة خطی و مربوط به پارامترهای برگزیده با توجه به مجموعة آموزش

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| مجموعه داده | SDD-1 | SDD-2 | SDD-3 | SDD-4 | SDD-5 | SDD-6 |
| حساسیت | 0٫87 | 0٫73 | 0٫6۰ | 0٫67 | 0٫87 | 0٫87 |

Abstract

Cancer is one of the major causes of death worldwide. To treat cancer, the use of anticancer peptides (ACPs) has attracted increased attention in recent years. ACPs are a unique group of small molecules that can target and kill cancer cells fast and directly. However, identifying ACPs by wet-lab experiments is time-consuming and labor-intensive. Therefore, it is significant to develop computational tools for ACPs prediction. Hence, this study tries to build a machine learning model for distinguishing ACPs from non-ACPs. This study utilizes data of ACPs and non-ACPs, uses iFeature python package to extract features from peptides sequence, builds and trains random forest models on extracted features.

Comparing random forest models in this work, random forests with QSOrder and DDE features show better performance (accuracy: 80% for QSORrder model and accuracy: 81% for QSOrder+DDE model) on independent test set.

Keywords: Anticancer Peptides Prediction, Machine Learning, Random Forest, Cancer



Prediction of *Anticancer Peptides*

*Using Random Forest*

Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science (M.Sc.) in Bioinformatics

Department of Biophysics

Faculty of Bioinformatics

Tarbiat Modares University

By:

**Mohammadtabar Zeynab**

Supervisor:

**Dr. Abdolmaleki Parviz**

January 2022

1. Xi, J., et al., *Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication.* Bioinformatics, 2020. **36**(6): p. 1855-1863.

2. Yue, Z., et al., *dbCID: a manually curated resource for exploring the driver indels in human cancer.* Briefings in bioinformatics, 2019. **20**(5): p. 1925-1933.

3. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA: a cancer journal for clinicians, 2018. **68**(6): p. 394-424.

4. Wijdeven, R.H., et al., *Old drugs, novel ways out: Drug resistance toward cytotoxic chemotherapeutics.* Drug Resistance Updates, 2016. **28**: p. 65-81.

5. Huang, Y., et al., *Alpha-helical cationic anticancer peptides: a promising candidate for novel anticancer drugs.* Mini reviews in medicinal chemistry, 2015. **15**(1): p. 73-81.

6. Hoskin, D.W. and A. Ramamoorthy, *Studies on anticancer activities of antimicrobial peptides.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2008. **1778**(2): p. 357-375.

7. Giuliani, A., G. Pirri, and S. Nicoletto, *Antimicrobial peptides: an overview of a promising class of therapeutics.* Open Life Sciences, 2007. **2**(1): p. 1-33.

8. Gaspar, D., A.S. Veiga, and M.A. Castanho, *From antimicrobial to anticancer peptides. A review.* Frontiers in microbiology, 2013. **4**: p. 294.

9. Ting, C.-H., et al., *The mechanisms by which pardaxin, a natural cationic antimicrobial peptide, targets the endoplasmic reticulum and induces c-FOS.* Biomaterials, 2014. **35**(11): p. 3627-3640.

10. Buri, M.V., et al., *Resistance to degradation and cellular distribution are important features for the antitumor activity of gomesin.* PLoS One, 2013. **8**(11): p. e80924.

11. Vijayakumar, S. and P. Lakshmi, *ACPP: A web server for prediction and design of anti-cancer peptides.* International Journal of Peptide Research and Therapeutics, 2015. **21**(1): p. 99-106.

12. Reddy, K., R. Yedery, and C. Aranha, *Antimicrobial peptides: premises and promises.* International journal of antimicrobial agents, 2004. **24**(6): p. 536-547.

13. Ejtehadifar, M., et al., *Anti-cancer effects of Staphylococcal Enterotoxin type B on U266 cells co-cultured with Mesenchymal Stem Cells.* Microbial pathogenesis, 2017. **113**: p. 438-444.

14. Wang, Z. and G. Wang, *APD: the antimicrobial peptide database.* Nucleic acids research, 2004. **32**(suppl\_1): p. D590-D592.

15. Schweizer, F., *Cationic amphiphilic peptides with cancer-selective toxicity.* European journal of pharmacology, 2009. **625**(1-3): p. 190-194.

16. Wüthrich, K., *NMR with proteins and nucleic acids.* Europhysics News, 1986. **17**(1): p. 11-13.

17. Marion, D., M. Zasloff, and A. Bax, *A two-dimensional NMR study of the antimicrobial peptide magainin 2.* FEBS letters, 1988. **227**(1): p. 21-26.

18. Ganz, T., et al., *Defensins. Natural peptide antibiotics of human neutrophils.* Journal of Clinical Investigation, 1985. **76**(4): p. 1427.

19. Hill, C.P., et al., *Crystal structure of defensin HNP-3, an amphiphilic dimer: mechanisms of membrane permeabilization.* Science, 1991. **251**(5000): p. 1481-1485.

20. Landon, C., et al., *Solution structure of drosomycin, the first inducible antifungal protein from insects.* Protein Science, 1997. **6**(9): p. 1878-1884.

21. Tamamura, H., et al., *A comparative study of the solution structures of tachyplesin I and a novel anti-HIV synthetic peptide, T22 ([Tyr 5, 12, Lys 7]-polyphemusin II), determined by nuclear magnetic resonance.* Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology, 1993. **1163**(2): p. 209-216.

22. Xu, T., et al., *Anticandidal activity of major human salivary histatins.* Infection and immunity, 1991. **59**(8): p. 2549-2554.

23. Selsted, M.E., et al., *Indolicidin, a novel bactericidal tridecapeptide amide from neutrophils.* Journal of Biological Chemistry, 1992. **267**(7): p. 4292-4295.

24. Lawyer, C., et al., *Antimicrobial activity of a 13 amino acid tryptophan-rich peptide derived from a putative porcine precursor protein of a novel family of antibacterial peptides.* FEBS letters, 1996. **390**(1): p. 95-98.

25. Gennaro, R., B. Skerlavaj, and D. Romeo, *Purification, composition, and activity of two bactenecins, antibacterial peptides of bovine neutrophils.* Infection and immunity, 1989. **57**(10): p. 3142-3146.

26. AGERBERTH, B., et al., *Amino acid sequence of PR‐39.* European Journal of Biochemistry, 1991. **202**(3): p. 849-854.

27. De Vos, W.M., et al., *Properties of nisin Z and distribution of its gene, nisZ, in Lactococcus lactis.* Applied and environmental microbiology, 1993. **59**(1): p. 213-218.

28. Fregeau Gallagher, N.L., et al., *Three-dimensional structure of leucocin A in trifluoroethanol and dodecylphosphocholine micelles: spatial location of residues critical for biological activity in type IIa bacteriocins from lactic acid bacteria.* Biochemistry, 1997. **36**(49): p. 15062-15072.

29. Rokach, L. and O. Maimon, *Decision trees*, in *Data mining and knowledge discovery handbook*. 2005, Springer. p. 165-192.

30. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.

31. Osuna, E.E., *Support vector machines: Training and applications*. 1998, Massachusetts Institute of Technology.

1. Anticancer Peptides (ACP) [↑](#footnote-ref-1)
2. International Agency for Research on Cancer [↑](#footnote-ref-2)
3. Antimicrobial Peptides (AMP) [↑](#footnote-ref-3)
4. NMR spectroscopy [↑](#footnote-ref-4)
5. Nissen [↑](#footnote-ref-5)
6. Lactococcus Lactis [↑](#footnote-ref-6)
7. Supervised [↑](#footnote-ref-7)
8. Unsupervised [↑](#footnote-ref-8)
9. Class Label [↑](#footnote-ref-9)
10. Descriptor [↑](#footnote-ref-10)
11. Feature [↑](#footnote-ref-11)
12. Predictors [↑](#footnote-ref-12)
13. Clustering [↑](#footnote-ref-13)
14. Classification [↑](#footnote-ref-14)
15. Decision Tree Classifier [↑](#footnote-ref-15)
16. Root Node [↑](#footnote-ref-16)
17. Leaves [↑](#footnote-ref-17)
18. Random Forest Classifier [↑](#footnote-ref-18)
19. Ensemble Learning [↑](#footnote-ref-19)
20. Support Vector Machine [↑](#footnote-ref-20)
21. Margin [↑](#footnote-ref-21)
22. Kernel [↑](#footnote-ref-22)
23. Gene Ontology [↑](#footnote-ref-23)
24. *Penicillium*-Crop Protein–Protein Interactions [↑](#footnote-ref-24)
25. Pathovar [↑](#footnote-ref-25)
26. Protein–protein interactions between *Ralstonia solanacearum* and *Arabidopsis thaliana* [↑](#footnote-ref-26)
27. Intra-species [↑](#footnote-ref-27)
28. Inter-species [↑](#footnote-ref-28)
29. *Glycine max* [↑](#footnote-ref-29)
30. *Lycopersicum esculentum* [↑](#footnote-ref-30)
31. *Nicotiana tabacum* [↑](#footnote-ref-31)
32. Dataset [↑](#footnote-ref-32)
33. adjacency matrix [↑](#footnote-ref-33)
34. Training set [↑](#footnote-ref-34)
35. Test Set [↑](#footnote-ref-35)
36. Independent test set [↑](#footnote-ref-36)
37. Feature extraction [↑](#footnote-ref-37)
38. hydrophobicity [↑](#footnote-ref-38)
39. hydrophilicity [↑](#footnote-ref-39)
40. side chain mass [↑](#footnote-ref-40)
41. Residue [↑](#footnote-ref-41)
42. Residue volume [↑](#footnote-ref-42)
43. Residue accessible surface area in tripeptide [↑](#footnote-ref-43)
44. Free energy of solution in water [↑](#footnote-ref-44)
45. Polarizability parameter [↑](#footnote-ref-45)
46. Normalized average hydrophobicity scale [↑](#footnote-ref-46)
47. Average flexibility index [↑](#footnote-ref-47)
48. Polarity [↑](#footnote-ref-48)
49. Net charge [↑](#footnote-ref-49)
50. normalized Van der Waals Volume [↑](#footnote-ref-50)
51. charge [↑](#footnote-ref-51)
52. secondary structures [↑](#footnote-ref-52)
53. solvent accessibility [↑](#footnote-ref-53)
54. Data pre-processing [↑](#footnote-ref-54)
55. Correlation [↑](#footnote-ref-55)
56. True positive [↑](#footnote-ref-56)
57. False positive [↑](#footnote-ref-57)
58. True negative [↑](#footnote-ref-58)
59. False negative [↑](#footnote-ref-59)
60. Sensitivity [↑](#footnote-ref-60)
61. Recall [↑](#footnote-ref-61)
62. Specificity [↑](#footnote-ref-62)
63. Accuracy [↑](#footnote-ref-63)
64. Precision [↑](#footnote-ref-64)
65. Cross validation (CV) [↑](#footnote-ref-65)
66. K-fold cross validation [↑](#footnote-ref-66)
67. Leave one out cross validation (LOOCV) [↑](#footnote-ref-67)
68. Feature selection [↑](#footnote-ref-68)
69. Wrapper [↑](#footnote-ref-69)
70. Filter [↑](#footnote-ref-70)
71. Sequential Forward Search [↑](#footnote-ref-71)
72. Sequential Backward Search [↑](#footnote-ref-72)
73. Curse of dimensionality [↑](#footnote-ref-73)
74. Dimension reduction [↑](#footnote-ref-74)
75. Principal Component Analysis [↑](#footnote-ref-75)