Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume *, Number *, Spring 20**, pp. **-**
Received: **/**/**, Revised: **/**/**, Accepted: **/**/**

A Stacking Model for Predicting Diabetes using Classification Algorithms

Solaleh Pasandideh[1], Amineh Amini[2*]
1- Faculty of Mechatronics, Islamic Azad University of Karaj, Karaj, Iran.
2*- Faculty of Mechatronics, Islamic Azad University of Karaj, Karaj, Iran.
[1] solaleh.pasandideh@gmail.com, [2*] aamini@kiau.ac.ir

Corresponding author's address: Amineh Amini, Faculty of Mechatronics, Islamic Azad University of Karaj, Karaj, Iran.

Abstract- The amount of data collected each day is enormous, and health care communities naturally produce large amounts of information on a daily basis. Although the field of health care is rich in information, it requires the discovery of hidden relationships and patterns in the data. On the other hand, data mining techniques support the medical decision to correctly diagnose and treat the disease and reduce the workload of specialists. This article is proposed a method for predicting diabetes using classification algorithms in data mining techniques. Therefore, popular classification algorithms including decision tree, support vector machine, Naive Bayes and also the K-nearest neighbor are investigated to predict diabetes. Also, in this paper, a method using stacking algorithm is presented for PIMA Indian Diabetes dataset in four combinations of mentioned classification methods. The results show that the second combination of this stacking model (Stacking 2 model) has 75.43% accuracy, 58.38% specificity and 84.63% sensitivity, which in comparison with other combinations of stacking method as well as the decision tree classification algorithms, the Naive Bayes classification and the K-nearest neighbor classification are more accurate.

Keywords- Data Mining, Diabetes Prediction, Classification Algorithm, Stacking Algorithm, Support Vector Machine Classification, Naive Bayes Classification, K-Nearest Neighbor.

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد *، شماره *، فصل **۱۳ ، صفحه **-**
دریافت: **/**/**، بازنگری: **/**/**، پذیرش: **/**/**

مدل پشته ای پیش بینی دیابت با استفاده از الگوریتم‌های طبقه‌بندی

سلاله پسندیده[۱]، امینه امینی*[۲]
۱- دانشکده مکاترونیک، دانشگاه آزاد اسلامی واحد کرج، کرج، ایران.
*۲- دانشکده مکاترونیک، دانشگاه آزاد اسلامی واحد کرج، کرج، ایران.
[1] solaleh.pasandideh@gmail.com, [2*] aamini@kiau.ac.ir

* نشانی نویسنده مسئول: امینه امینی، کرج، دانشگاه آزاد اسلامی واحد کرج، دانشکده مکاترونیک، گروه کامپیوتر.

چکیده- حجم داده‌های جمع آوری شده در هر روز بسیار زیاد است و جوامع مراقبت‌های بهداشتی به‌طبع روزانه حجم زیادی از اطلاعات را تولید می‌کنند. اگرچه حوزه ی مراقبت‌های بهداشتی غنی از اطلاعات است اما نیاز به کشف روابط پنهان و الگوها در داده‌ها دارد. از طرفی تکنیک‌های داده کاوی نیز از تصمیم پزشکی برای تشخیص صحیح و درمان بیماری حمایت می‌کنند و موجب کاهش حجم کار متخصصان می‌شوند. در این مقاله روشی برای پیش بینی دیابت با استفاده از الگوریتم‌های طبقه‌بندی از تکنیک‌های داده کاوی پیشنهاد شده است. از این رو معروف ترین الگوریتم‌های طبقه بندی شامل درخت تصمیم، ماشین بردار پشتیبان، نایو بیز و نیز K نزدیکترین همسایگی برای پیش بینی دیابت مورد بررسی قرار گرفته‌اند. همچنین در این مقاله روش پیشنهادی با استفاده از الگوریتم پشته ای برای مجموعه داده دیابت هندی PIMA در چهار ترکیب از روش‌های طبقه‌بندی نام برده ارائه شده است. نتایج نشان داده شده است که ترکیب دوم از این مدل پشته ای دارای دقت ۷۵/۴۳٪، خاصیت ۵۸/۳۸٪ و حساسیت ۸۴/۶۳٪ است که در مقایسه با بقیه ترکیب‌های روش پشته‌ای و همچنین الگوریتم‌های طبقه بندی درخت تصمیم، طبقه بندی نایو بیز و نیز طبقه بندی K نزدیکترین همسایگی، دارای دقت بیشتری است.

واژه‌های کلیدی: داده کاوی، پیش بینی دیابت، الگوریتم طبقه بندی، الگوریتم پشته ای، طبقه بندی ماشین بردار پشتیبان، طبقه بندی نایو بیز، طبقه بندی K نزدیکترین همسایگی

۱- مقدمه

دیابت یا بیماری قند[۱]، بر اساس تعریف سازمان بهداشت جهانی[۲] (WHO)، یک بیماری مزمن دژنراتیو (تحلیل برنده) است که ناشی از تولید انسولین ناکافی در لوزالمعده یا عدم توانایی بدن در استفاده مؤثر از انسولین تولید شده، افزایش قند خون به عنوان شاخص اصلی در نظر گرفته می‌شود [۱]. چهارمین بیماری پیشرو در جهان امروز دیابت است زیرا طبق گزارش آماری، ۷۹٪ از مرگ و میر در افراد زیر ۶۰ سال به علت دیابت رخ داده است و شماری از چالش‌ها برای پیش بینی و شناسایی این بیماری وجود دارد. داده کاوی روش‌هایی مؤثر برای شناسایی بیماران دیابتی پیشنهاد داده است. در این نوع مطالعه‌ها، متغیرهای فیزیکی و خونی عده ای از بیماران دیابتی و افراد معمولی به الگوریتم‌های رده بندی داده کاوی داده می‌شود. این الگوریتم‌ها می توانند مدل‌هایی را برای رده بندی بیماران به دو رده «بیمار دیابتی» و «افراد سالم»، ایجاد نمایند [۲].

یادگیری ماشین را می‌توان به‌عنوان زیر گروهی از هوش مصنوعی برای حل مشکلات دنیای واقعی با «فراهم آوردن توانایی یادگیری برای کامپیوتر بدون برنامه نویسی اضافی» تعریف کرد. با افزایش پیشرفت در یادگیری ماشین، همراه با آن استفاده از رایانه‌ها در پزشکی افزایش یافت. برای طبقه بندی و پیش بینی وقوع دیابت، روش‌های مختلف محاسباتی ایجاد و استفاده شده است. استفاده از تکنیک‌های یادگیری ماشین در پیش بینی ثابت شده است که

بسیار مفید است زیرا باعث افزایش دقت در تشخیص، کاهش هزینه‌ها و افزایش نرخ درمان‌های موثر می‌شود [۳].

روش‌های مختلف داده کاوی که برای پیش بینی و تشخیص بیماری‌های مختلف از جمله دیابت استفاده می‌شوند عبارتند از روش طبقه بندی که فرآیند پیدا کردن یک مدل است که کلاس-ها یا مفاهیم داده‌ها را بر اساس برچسب کلاس توصیف و تفکیک می‌کند، روش خوشه بندی که فرآیند تجزیه و تحلیل اشیاء داده بدون در نظر گرفتن برچسب کلاس است. این فرآیندها گروه بندی کلاس‌های جدید بر اساس به حداکثر رساندن شباهت درون کلاسی و به حداقل رساندن شباهت برون کلاسی هستند. و نیز روش یادگیری قاعده انجمنی یک روش یادگیری ماشین است که برای پیدا کردن الگوهای مکرر استفاده می‌شود [۴] .

استفاده از این روش‌ها در پیش بینی و تشخیص بیماری‌ها از جمله دیابت به این صورت است که در ابتدا روی داده‌ها یکسری محاسبه‌ها صورت می‌گیرد سپس نتایج این محاسبه‌ها ذخیره می-شود و این نتایج مورد تجزیه و تحلیل قرار می‌گیرد و این تجزیه و تحلیل به این صورت است که در ابتدا روی داده‌ها عمل پیش پردازش صورت می‌گیرد، سپس ویژگی‌های مورد نظر استخراج می‌شوند و این ویژگی‌ها به الگوریتم‌های داده کاوی مانند الگوریتم‌های طبقه بندی داده می‌شوند تا بیماری مورد نظر پیش بینی شود که نتیجه این پیش بینی نیز تحت نظارت پزشکان صورت می‌گیرد [۵].

پشته سازی[۳] یکی از روش‌های ترکیبی است که مشابه Boosting و Bagging است. Boosting یک الگوریتم ترکیبی یادگیری ماشین است که برای کاهش واریانس و بایاس مورد استفاده قرار می‌گیرد. این امر بر مبنای تبدیل مجموعه ای از یادگیرنده‌های ضعیف به یادگیرنده‌های قوی است. از طرف دیگر، Bagging به-منظور بهبود پایداری و دقت الگوریتم‌های یادگیری ماشین طراحی شده است، که در طبقه بندی آماری و رگرسیون کاربرد دارند. نه تنها باعث کاهش واریانس می‌شود بلکه به جلوگیری از بیش برازش[۴] کمک می‌کند. در واقع دو روش برای ترکیب مدل‌ها وجود دارد. اولین رأی گیری است، که در آن کلاس پیش بینی شده توسط اکثر مدل‌ها انتخاب شده است، در حالی که در پشته سازی پیش بینی‌ها توسط هر مدل متفاوت سطح پایه به‌عنوان ورودی برای طبقه بندی کننده سطح متا که خروجی آن کلاس نهایی است، داده می‌شود. پشته سازی همچنین به نتایج عالی در هر دو کار یادگیری نظارت شده مانند رگرسیون، طبقه بندی و یادگیری فاصله و یادگیری بدون نظارت مانند شبکه‌های عصبی و تخمین چگالی منجر شده است [۶].

در این مقاله ابتدا الگوریتم‌های طبقه بندی درخت تصمیم[۵]، الگوریتم طبقه بندی ماشین بردار پشتیبان (SVM)[۶]، الگوریتم طبقه بندی نایو بیز[۷] و نیز الگوریتم طبقه بندی K نزدیکترین همسایگی (KNN)[۸] برای پیش بینی دیابت بررسی شده است. سپس روشی تحت عنوان الگوریتم پشته‌ای[۹] در چهار ترکیب معرفی شده است و در آخر سه معیار دقت[۱۰]، خاصیت[۱۱] و حساسیت[۱۲] برای همه‌ی الگوریتم‌های گفته شده محاسبه شده و توسط نمودارهایی ارزیابی و با هم مقایسه شده اند.

ادامه این مقاله به شرح زیر است: در ابتدا در بخش ۲ مروری از کارهای مرتبط ارائه شده است، در بخش ۳ روش پیشنهادی را توضیح داده ایم، در بخش ۴ چگونگی انجام ارزیابی‌ها که شامل معرفی مجموعه داده، استخراج ویژگی، معیارهای ارزیابی و ابزار پیاده سازی می‌شود را توضیح داده ایم، در بخش ۵ نتایج را با استفاده از نمودار مورد ارزیابی قرار داده ایم، در بخش ۶ راجع به نتایج به دست آمده در این مقاله بحث کرده ایم و در نهایت در بخش ۷ از کارمان در این مقاله نتیجه گیری کرده ایم.

۲- کارهای مرتبط

این بخش، کارهای اخیر ادبیات موجود را مرور می‌کند و بینشی در درک چالش‌ها ارائه می‌دهد و تلاش می‌کند تا شکاف‌های روش-های موجود را پیدا کند.

سایسودیا و همکاران [۷] با ارزیابی عملکرد الگوریتم‌های مختلف مانند ماشین بردار پشتیبان، درخت تصمیم و نایو بیز با استفاده از مجموعه داده‌های دیابت هندی Pima، مدلی را ارائه دادند. نتایج با استفاده از منحنی ROC[۱۳] تأیید شد. نایو بیز با ۷۶/۳۰٪ دقت بهتری به‌دست آورد. میر و دهیگ [۸] برای ارائه الگوریتم بهینه بر اساس نتایج تجربی آنها با استفاده از ابزار WEKA، مجموعه داده-های Pima را تجزیه و تحلیل کردند. برخی از کارهای قبلی با نتایج و محدودیت‌ها همراه بود. برای بررسی عملکرد هر الگوریتم از ماتریس درهم ریختگی استفاده می‌شود. الگوریتم‌های ماشین بردار پشتیبان، نایو بیز، درخت تصادفی[۱۴] و الگوریتم ساده CART[۱۵] برای آزمایشی که ماشین بردار پشتیبان با ۷۹/۱۳٪ دقت بهتری به‌دست آورد، استفاده شد. ولدمایکل و مناریا [۴] از مجموعه داده‌های Pima برای طبقه بندی افراد دیابتی و غیر دیابتی با استفاده از الگوریتم‌های ماشین بردار پشتیبان، نایو بیز، درخت تصمیم J48 و پس انتشار خطا[۱۶] استفاده کردند. مجموعه داده‌های PIMA با استفاده از ابزار Rstudio و آزمون مجذور کای[۱۷] برای انتخاب ویژگی مجموعه داده‌ها مورد ارزیابی قرار می‌گیرد. نتیجه به‌دست آمده در جایی مقایسه می‌شود که الگوریتم پس

انتشار خطا دقت بالاتر ۸۳/۱۱٪ نسبت به سایر طبقه بندی‌ها ارائه می‌دهد. نیرمالادوی و همکاران [۹] توسعه یک مدل آمالگام را برای طبقه بندی پایگاه داده دیابتی هندی PIMA ارائه دادند. این مدل آمالگام ترکیب k-means با k نزدیکترین همسایه با پیش پردازش چند بخشی است. از خوشه بندی k-means برای شناسایی و حذف موارد طبقه بندی شده اشتباه استفاده می‌شود. مقادیر از دست رفته با میانگین‌ها و میانه‌ها جایگزین می‌شوند. یک طبقه بندی دقیق تنظیم شده با استفاده از k نزدیکترین همسایه با در نظر گرفتن نمونه خوشه ای صحیح با زیر مجموعه پیش پردازش شده به‌عنوان ورودی‌های k نزدیکترین همسایه انجام می-شود. نتایج تجربی نشان می‌دهد که آمالگام KNN [۱۸]پیشنهادی همراه با پیش پردازش بهترین نتیجه را برای مقادیر مختلف k تولید می‌کند. اگر مقدار k بیشتر باشد، مدل پیشنهادی دقت طبقه بندی ۹۷/۴٪ را به‌دست می‌آورد. برهت و کولکارنی [۱۰] الگوریتم‌های مختلف طبقه بندی را بر اساس سابقه سلامتی بیمار تجزیه و تحلیل کردند تا به پزشکان کمک کنند تا وجود بیماری را تشخیص دهند و همچنین تشخیص و درمان به موقع را ارتقا دهند. این آزمایش‌ها روی مجموعه داده‌های دیابت هندی Pima انجام شده است. طبقه بندی‌های مختلفی که استفاده می‌شود شامل K نزدیکترین همسایگان، رگرسیون لجستیک، درختان تصمیم، جنگل تصادفی، تقویت گرادیان، ماشین بردار پشتیبان و شبکه عصبی است. نتایج نشان می‌دهد که جنگل‌های تصادفی عملکرد خوبی در مجموعه داده‌ها با دقت ۷۹/۷٪ داشتند. سریواستاوا و همکاران [۱۱] از یادگیری ماشین، شاخه ای از هوش مصنوعی برای تحلیل و ساخت مدل پیش بینی دیابت استفاده کردند. در این کار تحقیقاتی، یک نمونه داده از مجموعه داده دیابت هندی Pima برای پیش بینی احتمال دیابت گرفته شد. از بین چندین الگوریتم یادگیری ماشین، شبکه عصبی مصنوعی[۱۹] (ANN) برای ساخت مدل برای پیش بینی دیابت انتخاب شد. و نتیجه گرفته شد که این مدل برای پیش بینی احتمال دیابت با دقت ۹۲٪ ضمن آزمایش با داده‌های نمونه آزمایش، ایده آل است.

همه این مطالعه‌ها از یک مجموعه داده مشترک (مجموعه داده دیابت هندی Pima) از پایگاه داده یادگیری ماشین (UCI) استفاده می‌کردند. با در نظر گرفتن نیاز به الگوریتم پیش بینی مؤثر، بهبود الگوریتم‌های پیش بینی موجود هدف اصلی مقاله خواهد بود در حالی که از مجموعه داده مشابه سایر محققان استفاده می‌شود. بنابراین در این مقاله، برای دستیابی به دقت بهتر طبقه بندی، روشی تحت عنوان الگوریتم پشته‌ای در چهار مدل معرفی شده و با بهترین الگوریتم‌های طبقه بندی یادگیری ماشین [۱۲] موجود مقایسه شده است.

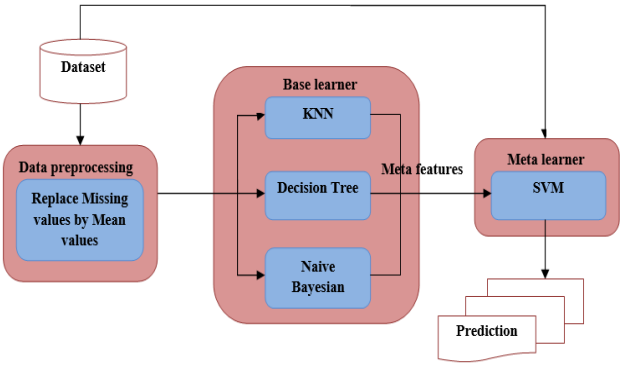
۳- روش پیشنهادی

افزایش دقت یک ترکیب، که ناشی از کاهش واریانس مدل و بایاس آن است، مبتنی بر روند ساده اما قدرتمند میانگین گیری گروهی یا رأی اکثریت است [۱۳]. به عبارت دیگر، این یک سیستم تصمیم گیری جمعی است که قادر است پیش بینی‌های طبقه بندی کننده‌های آموخته شده را به‌منظور ایجاد پیش بینی نمونه‌های جدید ترکیب کند. Stacking یا همان پشته سازی که گاهی اوقات تعمیم پشته‌ای نامیده می‌شود، روشی برای ترکیب چندین تکنیک یادگیری ماشین در یک مدل پیش بینی برای بهبود نیروی پیش بینی است. این یک طبقه بندی کننده سراسری را با آموزش یک یادگیرنده سطح متا[۲۰] برای ترکیب پیش بینی‌های طبقه بندی کننده‌های سطح پایه[۲۱] تولید می‌کند [۱۴]. در ابتدا از داده‌های موجود برای آموزش همه الگوریتم‌های دیگر استفاده می‌شود، سپس یک الگوریتم ترکیبی آموزش داده می‌شود تا پیش بینی نهایی را انجام دهد. در این مرحله از همه پیش بینی‌های الگوریتم‌های دیگر به‌عنوان ورودی‌های افزوده شده استفاده می‌شود [۶].

در این مقاله مدل پشته‌ای[۲۲] برای مدل سازی مجموعه داده‌ها برای پیش بینی بیماری دیابت پیشنهاد شده است. با توجه به شکل (۱) این مدل پشته‌ای دارای سه یادگیرنده پایه[۲۳] از جمله الگوریتم‌های طبقه بندی K نزدیکترین همسایه، درخت تصمیم گیری و نایو بیز است و نیز الگوریتم طبقه بندی ماشین بردار پشتیبان یادگیرنده متا[۲۴] در مدل پشته‌ای است. شکل (۱) مدل پشته‌ای را برای پیش بینی بیماری دیابت نشان می‌دهد. در مدل پشته‌ای پیشنهادی، ویژگی‌های متا، که همان ویژگی‌های استخراج شده و نتایج پیش بینی سه الگوریتم طبقه بندی مورد استفاده هستند، به ویژگی‌های اصلی نمونه‌ها در مجموعه داده ما اضافه می‌شوند. در نتیجه یادگیرنده متا، که همان الگوریتم طبقه بندی ماشین بردار پشتیبان (SVM) است نمونه‌هایی را با ویژگی‌هایی به تعداد (ویژگی‌های مجموعه داده PIMA) ۸ + (ویژگی‌های متا) ۳ تا مدل می‌کند و در نهایت بیماری دیابت پیش بینی می‌شود.

به‌طور کلی برای الگوریتم پشته‌ای چهار مدل به شرح زیر پیشنهاد شده است و شکل مدلی که دقت آن از همه بالاتر است در شکل (۱) مشاهده می‌شود:

1) Stacking 1:
   A. یادگیرنده پایه: SVM ،درخت تصمیم ،KNN
   B. یادگیرنده متا: نایو بیز
2) Stacking 2:
   A. یادگیرنده پایه: نایو بیز ،درخت تصمیم ،KNN
   B. یادگیرنده متا: SVM
3) Stacking 3:
   A. یادگیرنده پایه: نایو بیز ،درخت تصمیم ،SVM
   B. یادگیرنده متا: KNN
4) Stacking 4:
   A. یادگیرنده پایه: SVM ،KNN ،نایو بیز
   B. یادگیرنده متا: درخت تصمیم



شکل ۱: نمودار بلوکی حالت دوم مدل پشته‌ای (Stacking 2 model)

شبه کد الگوریتم پشته ای نیز برای مدل دوم آن یعنی Stacking 2 به شرح زیر است:

Input: pima_indians_diabetes //Input Dataset

Output: prediction of diabetes //1 = yes, 0 = no

Step:

1. Read the pima_indians_diabetes;
2. Calculate mean value for each attribute;
3. Replace missing value by mean value;
4. Calculate x_train, and y_train;
5. Calculate x_test, and y_test;
6. Function KNN (method): // Base Learner
   a. Get the x_train, and y_train;
   b. Calculate Cross Validate using 10-fold;
   c. Method= “KNN”; //train the KNN
   d. Get the x_test; //test the KNN
   e. Return predicted_k; //return label of the class variable
7. Function Decision Tree (method): //Base Learner
   a. Get the x_train, and y_train;
   b. Calculate Cross Validate using 10-fold;
   c. Method= “rpart”; //train the Decision Tree
   d. Get the x_test; //test the Decision Tree
   e. Return predicted_d; //return label of the class variable
8. Function Naive Bayesian (method): //Base Learner
   a. Get the x_train, and y_train;
   b. Calculate Cross Validate using 10-fold;
   c. Method= “nb”; //train the Naive Bayesian
   d. Get the x_test; //test the Naive Bayesian
   e. Return predicted_n; //return label of the class variable
9. Function SVM (method, c, γ): //Meta Learner
   a. Get the x_test, and y_test, and predicted_k, and predicted_d, and predicted_n;
   b. x_train_new ← update the x_train, and y_train_new ← update the y_train;
   c. x_test _new ← update the x_test, and y_test _new ← update the y_test;
   d. Get the x_train_new, and y_train_new;
   e. Calculate Cross Validate using 10-fold;
   f. Method= “svmLinear”, and c ←4, and γ ←0.5; //train the SVM
   g. Get the x_test_new; //test the SVM
   h. Return predicted_new; //return label of the class variable

۴- تنظیمات ارزیابی

۴-۱- توضیحات مجموعه داده

این مطالعه از مجموعه داده هندی PIMA استفاده می‌کند که از مخزن یادگیری ماشین UCI بارگیری می‌شود [۱۵]. مجموعه داده هندی PIMA شامل ۷۶۸ نمونه با ۸ ویژگی مستقل و یک ویژگی کلاس وابسته با دو برچسب کلاس است. ویژگی ها در زیر در جدول (۱) ذکر شده است:

جدول ۱: توضیحات مجموعه داده PIMA

صفت	نوع داده	توضیح
Npreg	عددی	Number of pregnancy
Glu	عددی	Glucose concentration in the body
Bp	عددی	Pressure Diastolic blood pressure
Skin	عددی	Skin fold thickness(mm) insulin
Serum	عددی	2-hour serum insulin
Bmi	عددی	Body mass index
Ped	عددی	Pedigree diabetic function
Age	عددی	Age of patient
Type	اسمی	Class variable of diabetes

۴-۲- پیش پردازش داده

مرحله پیش پردازش داده‌ها یک مرحله مهم در فرآیند کشف دانش است. بیشتر داده‌های مراقبت‌های بهداشتی حاوی داده‌های گمشده[۲۵]، نویزدار[۲۶] و ناسازگار[۲۷] هستند. تجزیه و تحلیل آماری روی مجموعه داده‌های هندی Pima وجود داده‌های گمشده را نشان می‌دهد. در واقع از نتایج آماری مشاهده می‌شود که غلظت گلوکز پلاسما، فشار خون دیاستولیک، ضخامت چین خوردگی پوست، انسولین سرم ۲ ساعته و شاخص توده بدنی دارای مقدار حداقل ۰ می‌باشند. ولی دانش پزشکی توضیح می‌دهد که چنین ویژگی‌هایی (نتیجه پزشکی) نمی‌توانند ۰ باشند [۱۶]. بنابراین نشان می‌دهد که مجموعه داده حاوی یک مقدار گمشده است که در صورت عدم رسیدگی می‌تواند کیفیت نتیجه و دقت مدل ما را مختل کند. روش‌های مختلفی برای رسیدگی به مقادیر از دست رفته در مجموعه داده‌ها پیشنهاد شده است، در مورد ما داده‌های گمشده مشخص شده و سپس با مقدار میانگین صفت یا همان ویژگی جایگزین شده یا اداره می‌شوند.

۴-۳- معیارهای ارزیابی

از معیارهای زیر برای ارزیابی مدل پیشنهادی استفاده می‌شود:

دقت: درصدی از پیش بینی‌های صحیح است که یک طبقه بندی کننده در مقایسه با مقدار واقعی برچسب در مرحله آزمایش انجام داده است. همچنین، می‌توان به‌عنوان نسبت تعداد ارزیابی‌های صحیح به تعداد کل ارزیابی‌ها گفت. دقت را می‌توان با استفاده از رابطه (۱) محاسبه کرد.

$$\text{Accuracy} = \frac{(TN+TP)}{(TN+TP+FN+FP)} \qquad (۱)$$

در جایی که، TP مثبت صحیح است، TN منفی صحیح است، FP مثبت کاذب است و FN منفی کاذب است. اگر برچسب کلاس یک رکورد در یک مجموعه داده مثبت باشد و طبقه بندی کننده برچسب کلاس را برای آن رکورد مثبت پیش بینی کند، از آن به‌عنوان مثبت صحیح نام برده می‌شود. اگر برچسب کلاس یک رکورد در یک مجموعه داده منفی باشد، و طبقه بندی کننده برچسب کلاس را برای آن رکورد منفی پیش بینی کند، از آن به عنوان منفی صحیح نام برده می‌شود. اگر برچسب کلاس یک رکورد در یک مجموعه داده مثبت باشد، اما طبقه بندی کننده برچسب کلاس را برای آن رکورد منفی پیش بینی کند، آن را منفی کاذب می‌نامند. اگر برچسب کلاس یک رکورد در یک مجموعه داده منفی باشد، اما طبقه بندی کننده برچسب کلاس را برای آن رکورد مثبت پیش بینی کند، آن را مثبت کاذب می‌نامند. که این مقادیر در رابطه با پیش بینی بیماری دیابت به شرح زیر می‌باشند:

TP: افراد دیابتی که به درستی دیابتی تشخیص داده می‌شوند،
TN: افراد سالمی که به درستی سالم تشخیص داده می‌شوند،
FP: افراد سالمی که به غلط دیابتی تشخیص داده می‌شوند و
FN: افراد دیابتی که به غلط سالم تشخیص داده می‌شوند.

حساسیت: به درصد مثبت صحیح گفته می‌شود که به‌طور صحیح توسط طبقه بندی کننده در حین آزمایش مشخص می‌شود. با استفاده از رابطه (۲) محاسبه می‌شود.

$$\text{Sensitivity} = \frac{(TP)}{(TP+FN)} \qquad (۲)$$

خاصیت: درصد منفی‌های صحیح است که به‌طور صحیح توسط طبقه بندی کننده در حین آزمایش مشخص می‌شود. با استفاده از رابطه زیر محاسبه می‌شود.

$$\text{Specificity} = \frac{(TN)}{(TN+FP)} \qquad (۳)$$

۴-۴- ابزار پیاده سازی

در این مطالعه از زبان برنامه نویسی R با استفاده از ابزار داده کاوی R-Studio برای پیاده سازی الگوریتم‌های طبقه بندی درخت تصمیم، نایو بیز، K نزدیکترین همسایگی، ماشین بردار پشتیبان و الگوریتم پشته‌ای پیشنهادی که در چهار ترکیب معرفی شده است و به‌طور کلی برای انجام آزمایش ها استفاده می‌شود.

۵- نتایج تجربی

در این مقاله نتایج شبیه سازی با ۷۰٪ داده آموزش و ۳۰٪ داده آزمون صورت گرفت، همچنین از روش اعتبارسنجی متقابل ۱۰ برابر[۲۸] برای آموزش و ارزیابی عملکرد پیش بینی کننده مدل و در واقع برای افزایش قابلیت اطمینان عملکرد طبقه بندی استفاده شد و در نهایت نتایج ارزیابی اجرای همه‌ی الگوریتم‌ها با ۱۰۰ بار تکرار برای دستیابی به نتیجه ثابت و سازگار به صورت میانگین دقت، میانگین حساسیت و میانگین خاصیت در نمودارهای (۱) تا (۵) نشان داده شده است.

۵-۱- انتخاب ویژگی

انتخاب ویژگی فرآیند انتخاب ویژگی‌های مفید از مجموعه داده است. برای انتخاب ویژگی‌های مهم برای پیش بینی بیماری دیابت از مجموعه داده‌های هندی Pima، از روش انتخاب ویژگی آزمون مجذور کای استفاده شده است. از بین ۸ ویژگی مستقل در روش انتخاب ویژگی آزمون مجذور کای، همه‌ی آن‌ها یعنی ویژگی‌های Serum، Skin، Glu، Age، Bmi، Npreg، Ped و Bp انتخاب شده‌اند که در جدول (۲) با استفاده از مقادیر X-squared، df و p-value نشان داده شده است که همه‌ی این ۸ ویژگی به ویژگی کلاس وابسته هستند و برای پیش بینی بیماری دیابت مهم می‌باشند.

جدول ۲: نتایج مربوط به آزمون مجذور کای بر روی مجموعه داده PIMA

صفت	X-squared	df	p-value
Npreg	64.595	16	8.648e-08
Glu	269.73	135	5.105e-11
Bp	54.934	46	0.1722
Skin	73.563	50	0.01668
Serum	227.77	185	0.0176
Bmi	286.47	247	0.04282
Ped	533.02	516	0.2929
Age	140.94	51	2.307e-10

با توجه به جدول بالا چون برای همه‌ی ویژگی‌ها مقدار X-squared با توجه به مقدار df بیشتر از مقدار p-value می‌باشد، در نتیجه در بررسی وابستگی همه‌ی ویژگی‌ها نسبت به ویژگی کلاس فرضیه رد می‌شود و یعنی همه‌ی ویژگی‌های ذکر شده در جدول مستقل نیستند و به ویژگی کلاس وابسته هستند. بنابراین معیار X^2 به صورت رابطه (۴) محاسبه می‌شود :

$$X^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \qquad (۴)$$

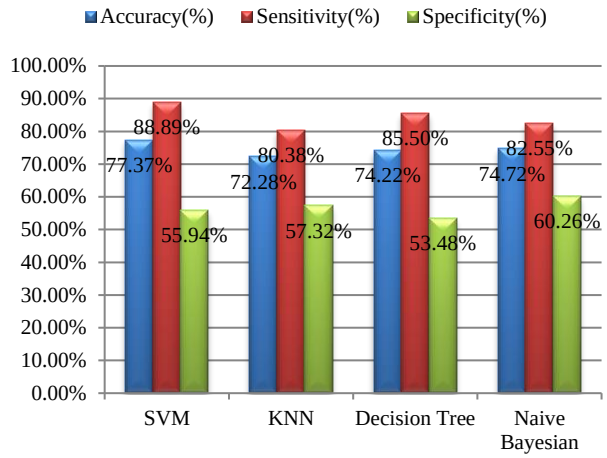
o_{ij} فراوانی مشاهده شده (واقعی) از رویداد مشترک (A_i,B_j) است و e_{ij} فراوانی مورد انتظار از (A_i,B_j) است که به‌صورت رابطه (۵) محاسبه می شود :

$$e_{ij} = \frac{count(A=a_i)\times count(B=b_j)}{n} \qquad (۵)$$

n تعداد تاپل‌های داده‌ای است، $(A = a_i)$ تعداد تاپل‌هایی است که مقدار a_i برای A دارند و $(B = b_j)$ تعداد تاپل‌هایی است که مقدار b_j برای B دارند. در این تست فرضیه این است که A و B مستقل هستند یعنی اینکه هیچ همبستگی بین آن‌ها وجود ندارد. علاوه‌بر این تست بر اساس significance level با (r-1)*(c-1) درجه آزادی df[۲۹] است. مقدار احتمال[۳۰] که گاهی به آن p-value نیز گفته می‌شود اغلب در مبا حث مربوط به آزمون فرض آماری مورد استفاده قرار می‌گیرد و ابزاری در اختیار ما قرار می‌دهد تا نسبت به رد یک فرضیه اقدام کنیم.

۵-۲- مقایسه معیارهای دقت، حساسیت و خاصیت برای الگوریتم‌های SVM، KNN، درخت تصمیم و نایو بیز

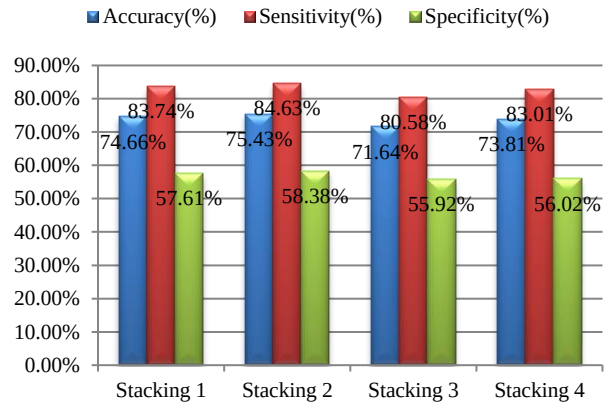
با توجه به نمودار (۱) در بین الگوریتم‌های زیر مشاهده می‌شود که الگوریتم SVM دارای بیشترین دقت و همچنین دارای بیشترین حساسیت و الگوریتم نایو بیز دارای بیشترین خاصیت می‌باشند.



نمودار ۱: ارزیابی کلی بهترین الگوریتم‌های طبقه بندی بر روی مجموعه داده PIMA

۵-۳- مقایسه معیارهای دقت، حساسیت و خاصیت برای الگوریتم‌های Stacking 1، Stacking 2، Stacking 3 و Stacking 4

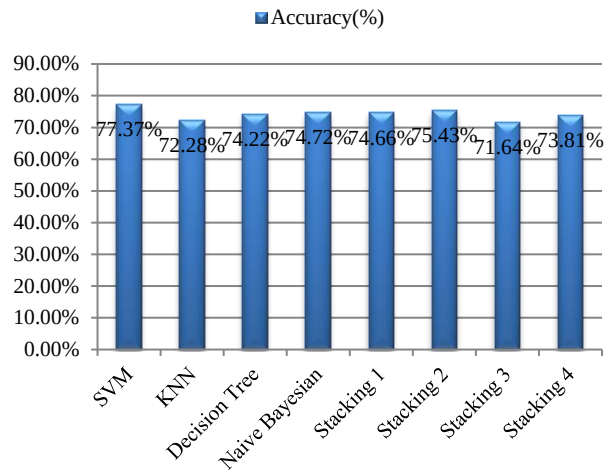
با توجه به نمودار (۲) در بین الگوریتم‌های زیر مشاهده می‌شود که الگوریتم Stacking 2 دارای بیشترین دقت، حساسیت و خاصیت می‌باشد.



نمودار ۲: ارزیابی کلی چهار مدل الگوریتم پشته‌ای بر روی مجموعه داده PIMA

۵-۴- مقایسه معیار دقت برای الگوریتم‌های SVM، KNN، درخت تصمیم، نایو بیز، Stacking 1، Stacking 2، Stacking 3 و Stacking 4

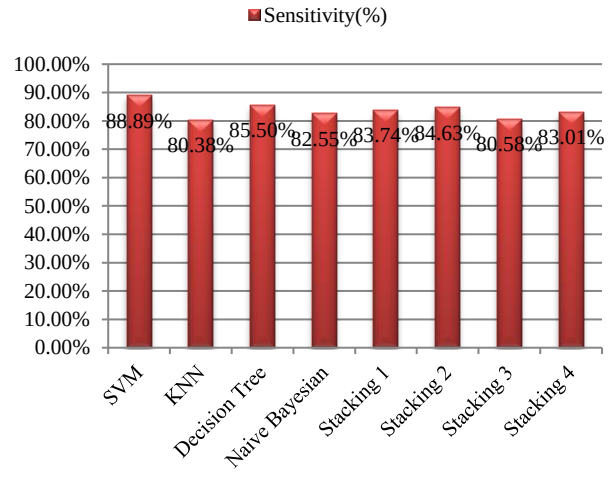
با توجه به نمودار (۳) در بین الگوریتم‌های زیر مشاهده می‌شود که الگوریتم Stacking 2 بعد از الگوریتم SVM دارای بیشترین دقت می‌باشد.



نمودار ۳: ارزیابی دقت بهترین الگوریتم‌های طبقه بندی و چهار مدل الگوریتم پشته‌ای بر روی مجموعه داده PIMA

۵-۵- مقایسه معیار حساسیت برای الگوریتم‌های SVM، KNN، درخت تصمیم، نایو بیز، Stacking 1، Stacking 2، Stacking 3 و Stacking 4

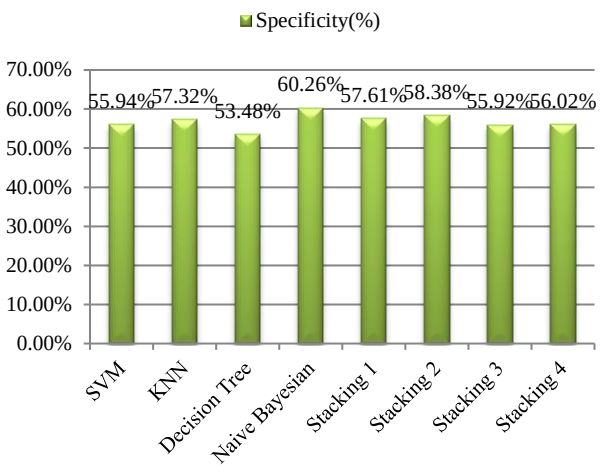
با توجه به نمودار (۴) در بین الگوریتم‌های زیر مشاهده می‌شود که الگوریتم Stacking 2 بعد از الگوریتم SVM و درخت تصمیم دارای بیشترین حساسیت می‌باشد.



نمودار ۴: ارزیابی حساسیت بهترین الگوریتم‌های طبقه بندی و چهار مدل الگوریتم پشته‌ای بر روی مجموعه داده PIMA

۵-۶- مقایسه معیار خاصیت برای الگوریتم‌های SVM، KNN، درخت تصمیم، نایو بیز، Stacking 1، Stacking 2، Stacking 3 و Stacking 4

با توجه به نمودار (۵) در بین الگوریتم‌های زیر مشاهده می‌شود که الگوریتم Stacking 2 بعد از الگوریتم نایو بیز دارای بیشترین خاصیت می‌باشد.



نمودار ۵: ارزیابی خاصیت بهترین الگوریتم‌های طبقه بندی و چهار مدل الگوریتم پشته‌ای بر روی مجموعه داده PIMA

۶- بحث

با توجه به این که از نمودارهای ارزیابی دریافتیم که الگوریتم Stacking 2 در مقایسه با بقیه مدل‌های الگوریتم پشته‌ای و همچنین الگوریتم‌های طبقه بندی درخت تصمیم، الگوریتم طبقه بندی نایو بیز و نیز الگوریتم طبقه بندی K نزدیکترین همسایگی دارای دقت و عملکرد بهتری می‌باشد، می‌توان فهمید که این مسأله که کدام الگوریتم‌ها به‌عنوان یادگیرنده‌های پایه و یادگیرنده متا الگوریتم پشته‌ای قرار بگیرند و همچنین جایگزینی آن‌ها با هم در دقت الگوریتم پشته‌ای تأثیر به‌سزایی دارد و دقت‌های متفاوتی را به ما می‌دهد، از طرفی با توجه به این که الگوریتم SVM با دقت ۷۷/۳۷٪، نایو بیز با دقت ۷۴/۷۲٪، درخت تصمیم با دقت ۷۴/۲۲٪ و الگوریتم KNN با دقت ۷۲/۲۸٪ درصد می‌باشد، در الگوریتم‌های پشته‌ای نیز زمانی که این الگوریتم‌ها به همین ترتیب به‌عنوان یادگیرنده متا قرار می‌گیرند، دقت الگوریتم‌های پشته‌ای نیز همین سیر نزولی را دارند یعنی الگوریتم Stacking 2 با دقت ۷۵/۴۳٪، Stacking 1 با دقت ۷۴/۶۶٪، Stacking 4 با دقت ۷۳/۸۱٪ و الگوریتم Stacking 3 با دقت ۷۱/۶۴٪ می‌باشند و در واقع می‌توان به‌طور کلی نتیجه گرفت که علاوه‌بر دقت پیش بینی الگوریتم‌های یادگیرنده پایه، دقت پیش بینی الگوریتم یادگیرنده متا نیز ارتباط مستقیم و تأثیر گذاری با دقت الگوریتم پشته‌ای دارد. البته الگوریتم پشته‌ای برای این که بتواند از همه‌ی الگوریتم-های طبقه بندی یادگیری ماشین دقت بالاتر و عملکرد خیلی بهتری داشته باشد هنوز هم جای کار دارد که می‌تواند در کارهای آینده مورد توجه قرار گیرد.

۷- نتیجه گیری

داده کاوی فرآیند استخراج الگوهای مفید و ناشناخته پیشین از پایگاه داده بزرگ یا انبار داده است. در حال حاضر داده کاوی امروزه نقش مهمی در بسیاری از بخش‌ها ایفا می‌کند، برخی از این موارد در بخش سلامت، بانک، بخش مالی، بخش آموزش و سایر بخش‌ها می‌باشند. تحقیقات مختلفی نیز درباره پیش بینی دیابت با استفاده از الگوریتم‌های مختلف انجام شده است. در این مقاله نیز ما با هدف پیش بینی این که فرد مورد نظر به بیماری دیابت مبتلا هست یا خیر، ابتدا الگوریتم‌های طبقه بندی درخت تصمیم، ماشین بردار پشتیبان، نایو بیز و نیز الگوریتم طبقه بندی K نزدیکترین همسایگی را برای پیش بینی دیابت مطالعه و پیاده سازی کردیم. سپس روشی را تحت عنوان الگوریتم پشته‌ای برای مجموعه داده دیابت هندی PIMA در چهار مدل معرفی کردیم و دریافتیم که حالت دوم این مدل پشته‌ای با دقت ۷۵/۴۳٪، در مقایسه با بقیه مدل‌های الگوریتم پشته‌ای و همچنین الگوریتم‌های طبقه بندی درخت تصمیم، الگوریتم طبقه بندی نایو بیز و نیز الگوریتم طبقه بندی K نزدیکترین همسایگی که در این مقاله بررسی شده اند، دارای دقت و عملکرد بهتری است.

مراجع

[1] A. Viloria, Y. Herazo-Beltran, D. Cabrera, and O. B. Pineda, “Diabetes Diagnostic Prediction Using Vector Support Machines,” *Procedia Comput. Sci.*, vol. 170, pp. 376–381, 2020.

[2] P. M. Shakeel, S. Baskar, V. R. S. Dhulipala, and M. M. Jaber, “Cloud based framework for diagnosis of diabetes mellitus using K-means clustering,” *Heal. Inf. Sci. Syst.*, vol. 6, no. 1, pp. 1–7, 2018.

[3] A. Hussain and S. Naaz, “Prediction of diabetes mellitus: Comparative study of various machine learning models,” in *International Conference on Innovative Computing and Communications*, 2021, vol. 2, pp. 103–115.

[4] F. G. Woldemichael and S. Menaria, “Prediction of Diabetes Using Data Mining Techniques,” *Proc. 2nd Int. Conf. Trends Electron. Informatics, ICOEI 2018*, no. Icoei, pp. 414–418, 2018.

[5] S. Jatav and V. Sharma, “An Algorithm for Predictive Data Mining Approach in Medical Diagnosis,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 10, no. 1, pp. 11–20, 2018.

[6] S. K. Dehkordi and H. Sajedi, “Prediction of disease based on prescription using data mining methods,” *Health Technol. (Berl).*, vol. 9, no. 1, pp. 37–44, 2019.

[7] D. Sisodia and D. S. Sisodia, “Prediction of Diabetes using Classification Algorithms,” *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1578–1585, 2018.

[8] A. Mir and S. N. Dhage, “Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare,” *Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018*, pp. 1–6, 2018.

[9] M. Nirmaladevi, S. A. Alias Balamurugan, and U. V. Swathi, "An amalgam KNN to predict diabetes mellitus," *2013 IEEE Int. Conf. Emerg. Trends Comput. Commun. Nanotechnology, ICE-CCN 2013*, no. Iceccn, pp. 691–695, 2013.

[10] R. Barhate and P. Kulkarni, "Analysis of Classifiers for Prediction of Type II Diabetes Mellitus," *Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018*, pp. 1–6, 2018.

[11] S. Srivastava, L. Sharma, V. Sharma, A. Kumar, and H. Darbari, "Prediction of Diabetes Using Artificial Neural Network Approach," in *Engineering Vibration, Communication and Information Processing*, Springer, 2019, pp. 679–687.

[12] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*, 2nd ed. John Wiley & Sons, 2011.

[13] M. A. King, A. S. Abrahams, C. T. Ragsdale, L. A. Matheson, A. Wang, and C. W. Zobel, "Ensemble Learning Techniques for Structured and Unstructured Data," dissertation for the degree of doctor of philosophy in business information technology, chapter 1, Virginia, United States, 2015.

[14] J. N. Sulzmann and J. Fürnkranz, "Rule stacking: An approach for compressing an ensemble of rule sets into a single classifier," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6926 LNAI, pp. 323–334, 2011.

[15] "Pima Indians Diabetes Database." [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database.

[16] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," *Informatics Med. Unlocked*, vol. 17, no. January, p. 100179, 2019.

پاورقی‌ها:

[1] Diabetes mellitus
[2] World Health Organization
[3] Stacking
[4] Over-fitting
[5] Decision Tree
[6] Support vector machine
[7] Naive Bayes
[8] K-nearest neighbor
[9] Stacking algorithm
[10] Accuracy
[11] Specificity
[12] Sensitivity
[13] Receiver Operating Characteristic
[14] Random Forest
[15] Classification And Regression Trees
[16] Back propagation
[17] Chi-square test
[18] An amalgam KNN
[19] Artificial neural network
[20] Meta-level
[21] Base-level
[22] Stacking model
[23] Base learner
[24] Meta learner
[25] Missing
[26] Noise
[27] Inconsistent
[28] 10-fold cross-validation
[29] Degrees of freedom
[30] Probability Value