Prediction of blood donations using machine learning techniques based on Decision tree, KNN, SVM, and MLP algorithms

Conference Paper · January 2019

CITATIONS
0

READS
384

3 authors:

Arash Fahmi Hassan
Kharazmi University
5 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Mohammadreza Moghari
Kharazmi University
3 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Omid Mahdi Ebadati E.
Kharazmi University
43 PUBLICATIONS **76** CITATIONS
SEE PROFILE

4th Conference on Contemporary Issues in Computer Information and Sciences
23-25 January 2019
چهارمین کنفرانس
موضوعات نوین در علوم کامپیوتر و اطلاعات
۳ تا ۵ بهمن ۱۳۹۷
CICIS 2019
دانشگاه خوارزمی

پیش بینی اهداء خون با استفاده از داده کاوی بر پایه الگوریتم های درخت تصمیم، KNN، SVM و MLP

آرش فهمی حسن[۱]، محمدرضا مغاری[۲]، امیدمهدی عبادتی[۳*]

۱. دانشجوی کارشناسی ارشد تحقیق در عملیات، دانشگاه خوارزمی، تهران.
std_fahmihassan@khu.ac.ir

۲. دانشجوی کارشناسی ارشد تحقیق در عملیات، دانشگاه خوارزمی، تهران.
mr_moghari@yahoo.com

۳. استادیار گروه مدیریت فناوری اطلاعات، دانشگاه خوارزمی، تهران.
ebadati@khu.ac.ir

چکیده

اهدای خون به دلیل نقش حیاتی و حساسی که در امر حفظ سلامت و بقاء زندگی انسان دارد مورد توجه می‌باشد. در جهان امروز علیرغم تحول عظیم علمی و با وجود پیشرفت‌های بزرگی که در علوم پزشکی رخ داده است، هنوز تامین کافی خون سالم یکی از چالش‌ها و دغدغه‌های مجامع پزشکی جهان است. حفظ و تامین حجم خون مورد نیاز در بانک‌های خون هر مرکز انتقال خون در هر منطقه، گروه‌های متنوع خونی و ارتباطاتی که بین آن‌ها وجود دارد و با فرض اینکه یکسری گروه‌های خونی کمیاب‌تر می‌باشند، پیش بینی و برنامه ریزی اهداء خون را در طول زمان مهم‌تر و پیچیده‌تر می کند. استفاده از داده کاوی در پایگاه‌های داده بیمارستان‌ها و مراکز انتقال خون به کشف روابط کمک می‌کند تا آن‌ها بتوانند بر مبنای گذشته یک پیش بینی از آینده داشته باشند، و بتوانند به بهترین شکل برای کمک، تشخیص و درمان‌های پزشکی موفق بیماری‌های مختلف را شناسایی کرده و الگوهای جراحات جدید را نشان دهند. در این مقاله سعی شده است تا در سطوح تصمیم گیری مربوط به حوزه مذکور، از تکنیک‌های داده کاوی و یادگیری ماشین برای پیش بینی اهداء خون استفاده شود تا با استفاده از این مکانیزم بتوانیم پیش بینی کنیم که در بازه‌های زمانی مختلف، چه میزان خون به بانک‌ها و مراکز انتقال خون اهداء خواهد شد که در این صورت بتوانیم حجم خون مورد نیاز بانک‌های خون مناطق مختلف را تخمین و تامین نمائیم. در همین راستا از چند الگوریتم طبقه بندی در یادگیری با نظارت از جمله الگوریتم‌های درخت تصمیم، KNN، SVM و MLP که یکی از انواع شبکه‌های مصنوعی عصبی (ANN) می باشد، برای پیش بینی استفاده شده و نتایج میزان دقت هر کدام ارائه شده است.

کلمات کلیدی

داده کاوی، یادگیری ماشین، درخت تصمیم، K- نزدیکترین همسایه، ماشین بردارپشتیبان، شبکه عصبی مصنوعی.

۱- مقدمه

اهدای خون به دلیل نقش حیاتی و حساسی که در امر حفظ سلامت و بقاء زندگی انسان دارد مورد توجه می‌باشد و می‌طلبد تا سازمان دهی‌هایی در این حوزه در سطح خرد و کلان کشور صورت بپذیرد. در جهان امروز علیرغم تحول عظیم علمی و با وجود پیشرفت‌های بزرگی که در علوم پزشکی رخ داده است، هنوز تامین کافی خون سالم یکی از چالش‌ها و دغدغه‌های مجامع پزشکی جهان است. بشر تاکنون هیچ جایگزین مناسبی برای این ماده حیاتی نیافته است و لذا یکی از مهم ترین نیازهای مراکز درمانی در جهان برای نجات جان آسیب دیدگان، خون و فراورده‌های خونی سالم است.

بسیاری از اجزای سازنده خون عمر مفید کوتاهی دارند و نگهداری و عرضه مداوم آن‌ها همواره با مشکلاتی همراه است. امور مربوط به اهدای خون در کشورهای مختلف توسط سازمان‌های متفاوتی انجام می پذیرد، به عنوان مثال در استرالیا توسط سرویس خون صلیب سرخ این کشور و در ایران توسط سازمان انتقال خون انجام می‌گیرد.

اهدای خون هنگامی رخ می‌دهد که یک فرد سالم به طور داوطلبانه مقدار مشخصی از خون خود را در یک مرکز انتقال خون هدیه می‌کند. انتقال خون در پزشکی جنبه حیاتی دارد و در موارد خاصی توسط پزشک معالج تجویز می‌گردد. با توجه به وظایف حیاتی خون، کمبود و یا وقفه طولانی در خونرسانی هر فرد می‌تواند منجر به آسیب‌های وسیع در اجزای بدن شخص شود که در نهایت به مرگ یا معلولیت‌های غیر قابل برگشت منجر خواهد شد. از هر سه نفر مردم دنیا، یک نفر در طول زندگی احتیاج به تزریق خون و فرآورده‌های خونی را پیدا می‌کند. بارزترین مثال برای موقعیت‌هایی که در آن نیاز واجب به خون پیدا می‌شود عبارت است از زمان بروز حوادث و سوانح گوناگونی نظیر تصادفات رانندگی، سوختگی‌ها و اعمال جراحی، همچنین خانم‌های باردار در حین زایمان، نوزادان و بخصوص نوزادان نارسی که به زردی دچار می‌شوند، بیماران سرطانی که تحت شیمی درمانی یا اشعه درمانی قرار دارند و مواردی دیگر از جمله‌ی نیازمندان به خون سالم می‌باشند.

در مراکز انتقال خون بیشتر کشورها، اطلاعات افرادی که برای اهداء خون به آنجا مراجعه می‌کنند بر اساس چندین مشخصه، جمع‌آوری و در یک پایگاه داده[1] ذخیره می شود. دسترسی به این اطلاعات آن هم در سطح کلان و به صورت یکپارچه، فواید و استفاده‌های بسیار زیادی در سطوح مختلف از جمله تشخیص، درمان و تصمیم‌گیری‌های کلان دارد. به طوری که این اطلاعات و سوابق باعث افزایش سرعت و کیفیت در ارائه خدمات درمانی به افراد و بیماران به خصوص در مواقع حساس و اورژانسی می‌شود. در سطوح مختلف تصمیم‌گیری نیز به منظور حفظ و تامین حجم خون مورد نیاز در بانک‌های خون هر مرکز انتقال خون در هر منطقه، با در نظر گرفتن میزان تقاضای جاری بر اساس اطلاعات گذشته و در نظر گرفتن ظرفیت احتیاطی برای مواقع بحرانی و اتفاقات غیر مترقبه، این اطلاعات به منظور پیش بینی و تصمیم‌گیری مورد استفاده قرار می‌گیرند. یکی دیگر از مواردی هم که این حوزه را مهم‌تر و پیچیده‌تر می کند گروه‌های متنوع خونی و ارتباطاتی می باشد که بین آن‌ها وجود دارد و با فرض اینکه یکسری گروه‌های خونی کمیاب‌تر می باشند، برنامه ریزی اهداء کنندگان در طول زمان مهم‌تر و پیچیده‌تر می شود.

با توجه به توضیحات ارائه شده، در این مقاله سعی شده است تا در سطوح تصمیم‌گیری مربوط به حوزه مذکور، از تکنیک‌های داده‌کاوی و یادگیری ماشین برای پیش بینی اهداء خون استفاده کنیم. و در همین راستا از چند الگوریتم طبقه بندی در یادگیری با نظارت از جمله الگوریتم‌های درخت تصمیم، KNN، SVM و MLP که یکی از انواع شبکه‌های مصنوعی عصبی (ANN) می‌باشد، استفاده شده و نتایج هر کدام ارائه شده است.

۲- پیشینه تحقیق

پژوهش [۳۲] با استفاده از داده‌های آماری و مدل سازی و آنالیز به کمک الگوریتم درخت تصمیم صورت گرفته است. مدل‌سازی انجام شده و نتیجه مدل این پژوهش، قادر به شناسایی اهدا کنندگانی است که برای اولین بار خون اهدا می‌کنند و پتانسیل تبدیل شدن به یک اهداء کننده متعهد را که دوباره برای اهداء خون مراجعه می‌کنند، دارد. این اطلاعات برای توسعه استراتژی‌های حفظ اهداء کننده به عنوان هدف مفید است. یک یافته جالب این مقاله این است که مکان اهداء خون نیز یک معیار مهم در تعیین و تشخیص اهداء مجدد خون است. مطالعه [۲۷] از تکنیک‌های داده کاوی، روش خوشه‌ای دو مرحله‌ای و روش طبقه‌بندی درخت رگرسیون برای شناسایی الگوی ورود روزانه و ساعتی در مرکز خون یک بیمارستان استفاده کرد.

پژوهش [۶] با کمک روش داده کاوی و استفاده از داده‌های یک بانک خون، رسیدن به اهداف این بانک خون را تسهیل می‌کند. اهداف سیستم بانک اهداء خون عبارت است از: افزایش نرخ اهدای خون و موارد مرتبط، استفاده مفیدتر از خون‌های اهدا شده، سیاست‌های استفاده از اهدا کنندگان و تاسیس بانک خون‌های جدید. برای مثال تکنیک تحلیل پیش‌بینی می‌تواند برای اهداء کنندگان خون در پیش بینی رفتار آینده آن‌ها مورد استفاده قرار گیرد. در این مقاله اهدا کننده متغیر مستقل می باشد و خون متغیر وابسته می شود. سپس بر اساس داده‌های تاریخی، می توانیم منحنی رگرسیون متناسب را که برای پیش‌بینی رفتار اهدا کننده استفاده می‌شود، رسم کنیم. این مقاله نتیجه می‌گیرد می‌توان از تکنیک‌های داده کاوی در راستای اهداف سیستم بانک اهدای خون استفاده کرد.

۲-۱- داده کاوی[2]

داده‌کاوی عبارت است از فرایند استخراج اطلاعات معتبر، از پیش ناشناخته، قابل فهم و قابل اعتماد از پایگاه داده‌های بزرگ و استفاده از آن در تصمیم گیری در فعالیت‌های تجاری مهم. اصطلاح داده کاوی به فرایند نیمه خودکار تجزیه و تحلیل پایگاه داده‌های بزرگ به منظور یافتن الگوهای مفید اطلاق می‌شود. [۱]

به گفته Brunassi و همکاران [۸]، اغلب عملیات و فعالیت‌های نهادهای دولتی و خصوصی در پایگاه‌داده‌های بزرگ ثبت و جمع‌آوری می‌شوند، تکنیک داده کاوی (DM) یکی از موثرترین گزینه‌ها برای استخراج دانش از حجم بالای داده‌ها، کشف روابط مخفی، الگوها و ایجاد قواعد برای پیش‌بینی و ارتباط دادن داده‌ها است که می‌تواند به موسسات در تصمیم‌گیری سریع‌تر کمک کند و یا حتی به درجه بیشتری از اعتماد برسد. داده کاوی به معنی جستجو برای الگوهای خاص درون مجموعه داده‌های بزرگ است، که بسیاری از احتمالات برای مدیران کسب وکار و تصمیم گیرندگان را ایجاد می‌کند. این روزها اطلاعات و دانش، امتیازات قانونی، برای شرکت‌های سلامت و کنترل اجتماعی که در جستجوی استقلال بیشتر در اقدامات خود و کاهش زمان تصمیم گیری

هستند، استراتژیک و ضروری محسوب می‌شوند. به همین دلیل، شرکت‌های مختلف ملی و بین‌المللی در زمینه تولید، مصرف، بازار مالی، موسسات آموزشی و کتابخانه‌ها پیش از این در امور عادی خود، داده کاوی را برای نظارت بر بودجه، مصرف مشتری، جلوگیری و کشف تقلب و پیش‌بینی ریسک‌های بازار در میان دیگران، به کار گرفته‌اند. [۲۲]

در بخش بهداشت عمدتا بخش عمومی، کاربرد آن به عنوان روشی برای تسریع جستجوی دانش پذیرفته شده‌است. علاوه بر این، استفاده از داده کاوی در پایگاه‌های داده بیمارستان‌های بزرگ و یا حتی در سیستم‌های اطلاعاتی سلامت عمومی به کشف روابط کمک می‌کند تا آن‌ها بتوانند بر مبنای گذشته یک پیش بینی از آینده داشته باشند، تا بتوانند به بهترین شکل برای کمک، تشخیص و درمان‌های پزشکی موفق بیماری‌های مختلف را شناسایی کرده و الگوهای جراحات جدید را نشان دهند. [۹]

۳- روش پژوهش

این پژوهش از نوع توصیفی با رویکرد کاربردی می‌باشد. طبقه بندی یکی از تکنیک‌های یادگیری ماشین است که به منظور پیش بینی کلاس داده‌ها به کار می رود. پیش بینی یعنی، آنچه که انتظار میرود در آینده بر اساس دانش و تجربه اتفاق بیفتد، اما نه همیشه [۱۳] به عبارت دیگر طبقه بندی به پیاده سازی ساختار شناخته شده بر داده‌های آزمایشی می‌پردازد. [۵] در پژوهش حاضر، از روش‌های طبقه‌بندی پیشرفته‌ی الگوریتم‌های درخت تصمیم، KNN، SVM و MLP که یکی از انواع شبکه‌های‌مصنوعی عصبی (ANN) می‌باشد، استفاده شده است.

در این پژوهش از زبان برنامه نویسی پایتون استفاده شده، زبان برنامه نویسی پایتون (Python) یک زبان برنامه نویسی پویا و همه منظوره است که در طیف وسیعی از برنامه‌های نرم افزاری از جمله در توسعه‌ی برنامه‌های تحت وب و برنامه‌های با قابلیت واسط گرافیکی کاربر (GUI) قابل استفاده می‌باشد. علاوه بر این، Python یکی از ابزارهای اصلی برای توسعه پلتفرم‌های در مقیاس BigData می‌باشد.

الگوریتم‌های فوق الذکر را با استفاده از زبان پایتون بر روی مجموعه داده‌های مرکز انتقال خون که شامل ۷۴۸ نمونه و ویژگی‌های تازگی، تناوب، حجم خون اهدائی، زمان اولین مراجعه افراد و یک متغیر صفر و یک می‌باشد، پیاده سازی و نتایج هر کدام ارائه شده است.

٤- تجزیه و تحلیل روش‌ها

٤-١- مجموعه داده[٣]

مجموعه داده مورد استفاده در این پژوهش از سایت UCI گرفته شده و مربوط به اطلاعات پایگاه‌داده یک مرکز خدمات انتقال خون در شهر Hsin-Chu در تایوان می‌باشد. مجموعه داده شامل اطلاعات ۷۴۸ اهدا کننده خون‌می باشد و شامل پنج مشخصه R (تازگی- آخرین زمان اهداء خون)، F (تناوب- تعداد دفعات اهداء خون)، M (حجم خون اهداء شده) و T(زمان- زمان اولین مراجعه فرد برای اهداء خون) و یک متغییر باینری که نشان دهنده آن است که هر فرد در مارس ۲۰۰۷، خون اهداء کرده یا خیر (اهداء نموده برابر مقدار یک و در صورت عدم اهداء خون، برابر صفر).

مشخصه‌های انتخاب شده براساس مدل RFM می باشد که به تحلیل رفتار و بیان تفاوت مشتریان (که در اینجا اهداء کنندگان) با استفاده از سه متغیر تازگی، تکرار و مبلغ خرید (در اینجا حجم خون اهدائی) می‌پردازد که توسط Hughes [۱۸] ارائه شده است. بر طبق نظر Reinartz و Kumar [۲۵]، Chang و Tsay [۱۰] مدل RFM نمی‌تواند مشتریان دارای ارتباط بلند مدت و مشتریان دارای ارتباط کوتاه مدت با سازمان را مشخص نماید. آنها در تحقیق خود ایده طول ارتباط مشتری را پیشنهاد می‌دهند و به بررسی تاثیر آن بر وفاداری و سود آوری مشتری می پردازند. آنها بیان می کنند که افزایش طول ارتباط با مشتری، وفاداری مشتری را بهبود خواهد بخشید. و این متغیر را که نشان دهنده فاصله زمانی بین اولین و آخرین مراجعه مشتری در بازه مورد مشاهده است تعریف کرده‌اند. بنابراین بُعد طول ارتباط مشتری (L) به مدل RFMاضافه می‌شود که در اغلب متون RFML، و در برخی از متون از آن به عنوان RFMT یاد می‌کنند. این مدل RFML یا RFMT روشی است که برای خوشه‌بندی مشتریان در مدیریت ارتباط با مشتری (CRM) استفاده می‌شود.

در این پژوهش با به کارگیری تعدادی از الگوریتم‌های یادگیری ماشین در حوزه یادگیری با ناظر، و پیاده سازی آن‌ها بر روی اطلاعات به دست آمده با استفاده از مدل RFML از مرکز انتقال خون، عمل اهداء خون را در افراد، بر اساس اطلاعات در هر مشخصه، پیش بینی کنیم و دقت پیش بینی هر کدام را متناسب با این مجموعه داده مشخص نمائیم. الگوریتم‌های مورد استفاده در این پژوهش شامل الگوریتم درخت تصمیم ، نزدیک ترین همسایه (KNN)، ماشین بردار پشتیبان (SVM) و همچنین پرسپترون چند لایه (MLP) که از الگوریتم‌های پیشخور شبکه عصبی مصنوعی می باشد.

الگوریتم‌های مذکور با زبان برنامه نویسی پایتون، در نرم افزار spyder و با مشخصات پردازنده و ورژن زبان برنامه ریزی (Python ۳,۶,۴ - ۱۹۰۰ MSC v. - ۶۴ bit AMD۶۴)، پیاده سازی شده و در ادامه نحوه اجرای هر کدام از الگوریتم‌ها توضیح داده می شود.

٤-٢- الگوریتم درخت تصمیم[٤](ID3)

درخت تصمیم در داده کاوی مدلی است که جهت نمایش طبقه بندی‌ها و رگرسیون‌ها استفاده می‌شود. همانطور که از نام آن مشخص است، این درخت از تعدادی گره و شاخه تشکیل شده است. در درخت تصمیمی که عمل طبقه بندی را انجام می‌دهد، برگ‌ها بیانگر کلاس‌ها هستند. در هر یک از گره‌های دیگر (گره‌های غیربرگ) با توجه به یک یا چند صفت خاص تصمیم گیری

صورت می‌گیرد. درخت تصمیم به دلیل سادگی و قابل فهم بودن تکنیک محبوبی در داده کاوی محسوب می‌شود. به عبارت دیگر درخت تصمیم خود به تنهایی همه‌ی مطالب را توصیف می‌کند و نیاز به فرد خبره‌ای نیست تا خروجی را تفسیر کند. در واقع این یک روش گرافیکی است و بدین دلیل تفسیر آن شاید ساده تر از تکنیک‌های دیگر طبقه بندی باشد. اما به خاطر داشته باشید که داشتن تعداد گره‌های زیاد در درخت می‌تواند نمایش گرافیکی درخت تصمیم را با مشکل روبرو سازد. [۴]

درخت‌های تصمیم‌گیری در بین رویکردهای یادگیری ماشین، به عنوان روش کارا و اثربخش شناخته شده‌اند و آن‌ها با موفقیت برای حل مشکلات دنیای واقعی در حوزه هوش مصنوعی به کار گرفته شده‌اند. این موفقیت به دلیل توانایی عالی آن‌ها برای حل مشکلات پیچیده از طریق نمایش‌های گرافیکی قابل خواندن توسط انسان و توسط کامپیوتر است.[۲۰، ۲۱، ۲۴، ۲۹]

در این الگوریتم پس از فراخوانی داده‌ها در محیط برنامه، داده‌ها را به دو بخش داده‌های آموزشی و داده‌های تست تقسیم نموده که معمولا ۷۰٪ درصد داده‌ها را به عنوان داده‌های آموزشی و ۳۰٪ آن را به عنوان داده‌های تست در نظر می‌گیرند. در اینجا ما داده‌های تست را در چهار اندازه مختلف به کار گرفتیم و دقت پیش بینی را با توجه به هر کدام بدست آوردیم که در جدول (۱) قابل مشاهده است. در ادامه با فراخوانی کتابخانه scikit-learn ، ساب پکیج sklearn.tree و طبقه بندی کننده DecisionTreeClassifier، مدل درخت تصمیم را ساخته و داده‌های آموزشی (x_train, y_train) را وارد مدل کرده تا مدل آموزش ببیند. در ادامه برای مشخص نمودن دقت مدل، داده های تست (x_test) را وارد مدل کرده تا پیش بینی کند و در مقایسه با برچسب‌های دادهای تست (y_test) دقت پیش بینی را ارزیابی نماید.

این فرایند را بار دیگر انجام داده اما این بار به جای استفاده از داده های اصلی، برای هم مقیاس شدن داده‌ها، آن‌ها را نرمالایز کرده و سپس داده‌ها را به دو دسته داده‌های آموزشی و تست تقسیم کرده و دقت مدل را برای هر یک از مقادیر داده‌های تست، ارزیابی کردیم. نتایج ارزیابی مدل دسته بندی کننده درخت تصمیم در جدول (۱) قابل مشاهده است.

جدول (۱): نتایج ارزیابی مدل درخت تصمیم

	Data type	Test size	Accuracy
Decision Tree	Original	0.3	0.7688
		0.25	0.7540
		0.2	0.7666
		0.15	0.8053
	Normalize	0.3	0.7644
		0.25	0.7914
		0.2	0.7866
		0.15	0.7876

٤-٣- K- نزدیکترین همسایگی [۵](KNN)

الگوریتم KNN یکی از متداولترین الگوریتم‌های طبقه بندی است. این الگوریتم مبتنی بر نمونه است و بر اساس k همسایه نزدیک، طبقه بندی را انجام می‌دهد. الگوریتم KNN به عنوان الگوریتم تنبل شناخته می شود، زیرا مبتنی بر تقریب محلی است و همه‌ی محاسبات تا انجام طبقه بندی معوق می‌ماند. [۳، ۲۶] این روش بر اساس شباهت داده ها طبقه بندی را انجام می دهد. در واقع برای هر داده‌ی آزمایشی جدید، فواصل k همسایه نزدیک را محاسبه کرده و برچسبی مشابه برچسب غالب این k همسایه برای نقطه مورد نظر را تعیین می‌کند. [۳] طبقه‌بندی کننده k–نزدیک‌ترین همسایه، یکی از الگوریتم‌های طبقه‌بندی شناخته‌شده و ساده می‌باشد. این اولین بار توسط Fix و Hodges به عنوان یک الگوریتم ناپارامتری معرفی شد که هیچ فرضی بر توزیع داده‌های ورودی ایجاد نمی‌کند؛ بنابراین به طور گسترده در کاربردهای مختلف استفاده می‌شود. [۱۱]

در طبقه‌بندی‌کننده KNN، یک نمونه ناشناخته براساس شباهت بین نمونه‌های شناخته‌شده آموزش‌دیده یا برچسب دار برمبنای محاسبه فاصله بین نمونه‌های ناشناس با نمونه‌های برچسب دار، شناخته می‌شود. سپس k نزدیک‌ترین نمونه‌ها به عنوان پایه برای طبقه‌بندی انتخاب می‌شوند و نمونه نامشخص (*xtest*) به کلاسی اختصاص می‌یابد که بیش‌ترین نمونه‌ها را در میان نزدیک‌ترین نمونه‌ها دارد. به همین منظور، الگوریتم طبقه‌بندی کننده KNN بستگی دارد به: (۱) تعداد k همسایه عدد صحیح و تغییر مقدار پارامتر k که ممکن است نتایج طبقه‌بندی را تغییر دهد. (۲) مجموعه داده‌های برچسب دار؛ بنابراین اضافه کردن یا حذف هر گونه نمونه به نمونه‌های آموزشی، بر تصمیم نهایی طبقه‌بندی کننده KNN ، تاثیر می‌گذارد. و (۳) معیار فاصله. در KNN، از فاصله اقلیدسی معمولاً به عنوان معیار فاصله برای اندازه‌گیری فاصله بین دو نمونه استفاده می‌شود. طبقه بندی کننده KNN به صورت تحلیلی قابل‌ردیابی است و به سادگی پیاده‌سازی می‌شود، اما یکی از مشکلات اصلی الگوریتم KNN این است که به همه نمونه‌های آموزشی نیاز دارد که در زمان اجرا در حافظه باشند؛ به همین دلیل، طبقه بندی مبتنی بر حافظه نامیده می‌شود. [۱۱، ۲۸]

این الگوریتم نیز همانند الگوریتم درخت تصمیم، پس از فراخوانی داده ها در محیط برنامه، داده‌ها را به دو بخش داده‌های آموزشی و داده‌های تست تقسیم نموده، در ادامه با فراخوانی کتابخانه scikit-learn، ساب پکیج sklearn.neighbors و طبقه بندی کننده KNeighborsClassifier، مدل K- نزدیکترین همسایه را ساخته و داده‌های آموزشی (x_train, y_train) را وارد مدل کرده تا مدل آموزش ببیند. در ادامه برای مشخص نمودن دقت مدل، داده‌های تست (x_test) را وارد مدل کرده تا پیش بینی کند و در مقایسه با برچسب‌های داده‌های تست (y_test) دقت پیش بینی را ارزیابی نماید.

این فرایند را بار دیگر انجام داده اما این بار به جای استفاده از داده های اصلی، برای هم مقیاس شدن داده‌ها، آن‌ها را نرمالایز کرده و سپس داده‌ها را به دو دسته داده‌های آموزشی و تست تقسیم کرده و دقت مدل را برای هر یک از مقادیر داده‌های تست، ارزیابی کردیم. نتایج ارزیابی مدل دسته بندی کننده KNN در جدول (۲) قابل مشاهده است.

جدول (۲): نتایج ارزیابی مدل K- نزدیکترین همسایه

	Data type	Test size	k- nearest neighbor	Accuracy
KNN	Original	0.3	13	0.8488
		0.25	13	0.8502
		0.2	13	0.8666
		0.15	6-13-14	0.8849
	Normalize	0.3	19-20	0.8622
		0.25	25	0.8609
		0.2	21	0.8666
		0.15	25	0.8938

٤-٤- ماشین بردارپشتیبان[٦](SVM)

SVM یک ابزار ریاضی است که مبتنی بر اصل حداقل سازی خطای عملیاتی است و سابقه آن به سال ۱۹۶۰ برمیگردد. SVM براساس نظریه یادگیری آماری بنا نهاده شده و یک روش آماری غیرپارامتریک نظارت شده است. [۲۳] در کاربردهای امروزی یادگیری ماشین، ماشین بردار پشتیبان به عنوان یکی از قدیمی‌ترین و دقیق‌ترین روش‌ها در میان الگوریتم‌های معروف شناخته می‌شود. الگوریتم SVM جزء الگوریتم‌های تشخیص الگوی دسته بندی می‌باشد. از الگوریتم SVM، در هر جایی که نیاز به تشخیص الگو یا دسته بندی اشیا در کلاس‌های خاص باشد می‌توان استفاده کرد. همچنین ماشین بردار پشتیبان یکی از روش های یادگیری با ناظر است که از آن برای طبقه بندی و رگرسیون استفاده می‌کنند. مبنای کاری دسته بندی کننده ایـن مدل، دسته بندی خطی داده می‌باشد و در تقسیم خطی داده‌ها سعی بر آن است خطی انتخاب شود که حاشیه اطمینان بیشتری را داشته باشد. البته ماشین بردار پشتیبان در دسته بندی غیرخطی هم کاربرد دارد. به طور کلی این الگوریتم از یک نگاشت غیرخطی برای تبدیل داده های اصلی به ابعاد بالاتر استفاده می کند و سپس در این بعد جدید به دنبال ابرصفحه‌ای است که نمونه‌های یک کلاس را از کلاس های دیگر جداکند. با یک نگاشت غیرخطی مناسب، مجموعه داده‌های دو کلاسی می توانند توسط یک ابر صفحه جدا شوند. در واقع ایده اصلی ماشین بردار پشتیبان رسم ابرصفحه هایی در فضا است که عمل تمایز نمونه‌های مختلف داده‌ها را به طور بهینه انجام می دهند و ابرصفحه هایی را که بیشترین حاشیه جداسازی را دارند پیدا می‌کند و نزدیکترین داده‌های آموزشی به ابرصفحه، جداکننده بردارهای پشتیبان نامیده می شوند. این روش تا حدودی پیچیده است و ویژگی مثبت آن در این است که به تعداد نمونه‌های آموزش وابسته نمی‌باشد، و با تعداد ویژگی‌های بالا و تعداد نمونه‌های کم می‌تواند به خوبی کار کند. از جمله محدودیت‌های این الگوریتم این است که فقط بر روی داده‌هایی با مقدار واقعی کار می‌کند و انواع دیگر داده‌ها باید به داده‌های عددی تبدیل شوند. [۱۴] در حقیقت تابع هسته از شباهت بین داده‌ها در فضای اولیه برای یافتن شباهت بین بردارها در فضایی با ابعاد بالاتر استفاده می‌کند. تابع هسته [۷]می‌تواند تابع چند جمله ای، تابعRBF، تابع تانژانت هایپربولیک و یا توابع مناسب دیگری انتخاب شود.[۲]

این الگوریتم نیز همانند الگوریتم‌های قبلی، پس از فراخوانی داده‌ها در محیط برنامه، داده‌ها را به دو بخش داده‌های آموزشی و داده‌های تست تقسیم نموده، در ادامه با فراخوانی کتابخانه scikit-learn، ساب پکیج sklearn.svm و مدل SVC، مدل ماشین بردارپشتیبان را ساخته و داده‌های آموزشی (x_train, y_train) را وارد مدل کرده تا مدل آموزش ببیند. همانطور که گفت شد، تابع کرنل یا هسته انواع مختلفی دارد که در این تحقیق از یک تابع پایه‌ای شعاعی گوسی (RBF)[۸] استفاده شده است. در ادامه برای مشخص نمودن دقت مدل، داده‌های تست (x_test) را وارد مدل کرده و تا پیش بینی کند و در مقایسه با برچسب‌های داده‌های تست (y_test) دقت پیش بینی را ارزیابی نماید.

این فرایند را بار دیگر انجام داده اما این بار به جای استفاده از داده های اصلی، برای هم مقیاس شدن داده‌ها، آن‌ها را نرمالایز کرده و سپس داده‌ها را به دو دسته داده‌های آموزشی و تست تقسیم کرده و دقت مدل را برای هر یک از مقادیر داده‌های تست، ارزیابی کردیم. نتایج ارزیابی مدل SVM در جدول (۳) قابل مشاهده است.

جدول (۳): نتایج ارزیابی مدل SVM

	Data type	Test size	Accuracy
SVM	Original	0.3	0.8088
		0.25	0.8074
		0.2	0.8333
		0.15	0.8672
	Normalize	0.3	0.8133
		0.25	0.8021
		0.2	0.82
		0.15	0.8318

٤-٥- پرسپترون چند لایه[۹](MLP)

شبکه عصبی مصنوعی[۱۰] مشابه مغز انسان است زیرا هر دو آن‌ها شامل تعداد زیادی پردازش و واحدهای هوشمند هستند که نورون‌ها یا سلول‌های مغزی نامیده می‌شوند .هدف توسعه شبکه عصبی مصنوعی، یافتن رابطه بین داده‌های ورودی و داده‌های خروجی است. نورون‌ها مانند سلول‌های مغزی بیولوژیکی عمل می‌کنند تا لایه‌هایی را بسازند که عملکرد مدل را ارزیابی می‌کنند. شبکه عصبی مصنوعی به عنوان یک سیستم توزیع‌شده‌ی موازی شناخته می‌شود که شامل نورون‌ها محاسباتی ساده است. با استفاده از آزمون و خطا تعداد نورون‌ها در لایه‌های پنهان و تعداد لایه‌های پنهان را می توان محاسبه کرد. [۱۲]

مزیت اصلی روش شبکه عصبی مصنوعی رسیدن به راه حل مشکلات پیچیده است که حل آن با سایر تکنیک‌های متعارف دشوار است و سرعت پردازش آن بسیار سریع است. [۱۵] شبکه عصبی تکنیکی است که توانایی ضبط و نمایش روابط پیچیده ورودی/ خروجی را دارد. [۷] یکی از متداول‌ترین مدل‌های شبکه عصبی، شبکه عصبی پیشخور (MLP) نامیده می شود. [۳۰]

ساختار یک شبکه عصبی مصنوعی دارای سه نوع لایه بوده و لایه‌های ورودی، پنهان و خروجی لایه‌های مذکور هستند. تعداد لایه‌های پنهان و نورون‌های هر لایه، با اعمال الگوریتم‌های بهینه‌سازی، قابل‌محاسبه است. جزئیات ساختار شبکه عصبی MLP و اتصالات، بسیار وابسته به متغیرهای مسئله هستند و برای ایجاد ارتباطات، گام آموزشی بکار گرفته می‌شود. به منظور دستیابی به بهترین مدل، ساختار بهینه باید با استفاده از الگوریتم‌های بهینه‌سازی مناسب انتخاب شود. [۱۹]

در این الگوریتم نیز همانند الگوریتم‌های قبلی، پس از فراخوانی داده‌ها در محیط برنامه، داده‌ها را به دو بخش داده‌های آموزشی و داده‌های تست تقسیم نموده، سپس x_train, و x_test را نرمالایز می‌نماییم.

در ادامه با فراخوانی کتابخانه scikit-learn، ساب پکیج sklearn.neural_network و طبقه بندی کننده MLPClassifier، مدل MLP را ساخته و داده‌های آموزشی (x_train, y_train) را وارد مدل کرده (که x_train نرمالایز شده هستند) تا مدل آموزش ببیند. در این تحقیق تعداد لایه‌های پنهان این مدل سه لایه در نظر گرفته شده و در هر لایه تعداد ۱۰ نرون را قرار دادیم (۱۰، ۱۰،۱۰). در ادامه برای مشخص نمودن دقت مدل، داده های تست (x_test) که نرمالایز شده هستند را وارد مدل کرده و تا پیش بینی کند و در مقایسه با برچسب‌های داده‌های تست (y_test) دقت پیش بینی را ارزیابی نماید. نتایج ارزیابی مدل MLP در جدول (۴) قابل مشاهده است.

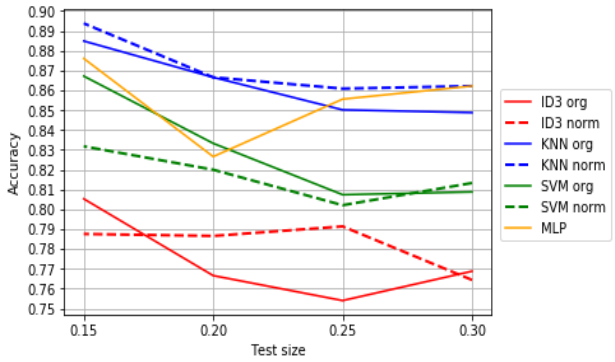
جدول (۴): نتایج ارزیابی مدل MLP

	Test size	Accuracy
MLP	0.3	0.8622
	0.25	0.8556
	0.2	0.8266
	0.15	0.8761

۵- نتیجه گیری وکارهای آتی

اهدای خون به دلیل نقش حیاتی و حساسی که در امر حفظ سلامت و بقاء زندگی انسان دارد مورد توجه می‌باشد. در جهان امروز علیرغم تحول عظیم علمی و با وجود پیشرفت‌های بزرگی که در علوم پزشکی رخ داده است، هنوز تامین کافی خون سالم یکی از چالش‌ها و دغدغه‌های مجامع پزشکی جهان است. در این مقاله سعی شد تا از تکنیک‌های داده‌کاوی و یادگیری ماشین برای پیش بینی اهداء خون استفاده کنیم تا با استفاده از این مکانیزم بتوانیم پیش بینی کنیم که در بازه‌های زمانی مختلف، چه میزان خون به بانک ها و مراکز انتقال خون اهداء خواهد شد که در این صورت بتوانیم حجم مورد نیاز بانک‌های خون مناطق مختلف را تخمین و تامین نمائیم. در همین راستا از چند الگوریتم طبقه بندی در یادگیری با نظارت از جمله الگوریتم‌های درخت تصمیم، KNN، SVM و MLP برای پیش بینی اهداء خون استفاده شد و نتایج میزان دقت هر کدام در قالب جداول مختلف ارائه شد.

در اجرای الگوریتم درخت تصمیم، بیشترین میزان دقت در داده اصلی با اندازه داده های تست ۰,۱۵ برابر ۰,۸۰۳۵ بوده و زمانی که داده ها نرمالایز شدند، با اندازه داده های تست ۰,۲۵، دقت برابر ۰,۷۹۱۴ بوده است. در پیاده سازی الگوریتم KNN، بیشترین دقت در داده ای اصلی با اندازه نمونه ۰,۱۵ و با تعداد همسایه (۱۴-۱۳-۶)، برابر ۰,۸۸۴۹ بوده و زمانی که داده ها نرمالایز شدند، با اندازه داده تست ۰,۱۵، با تعداد همسایه ۲۵، دقت برابر با ۰,۸۹۳۸ بوده است. در اجرای الگوریتم SVM با تابع کرنل RBF، در داده های اصلی و در اندازه داده تست ۰,۱۵، دقت برابر ۰,۸۶۷۲ ثبت شده و در داده های نرمال با اندازه داده تست ۰,۱۵، دقت ۰,۸۳۱۸ بوده است. و در آخر با پیاده سازی الگوریتم MLP، بیشترین میزان دقت با داده های تست ۰,۱۵، برابر ۰,۸۷۶۱ بوده است. در شکل (۱) نیز نتایج ارزیابی هر کدام از الگوریتم‌ها جهت مقایسه بصری با توجه به اندازه‌ی داده‌های تست هرکدام، با استفاده از کتابخانه matplotlib.pyplot، رسم شده و به نمایش درآمده است. همانطور که ملاحظه می شود در کل، الگوریتم KNN و MLP در ارزیابی‌ها از دقت بیشتری برخوردار هستند.



شکل (۱): مقایسه نتایج ارزیابی الگوریتم ها

در کارهای آتی می‌توان با بدست آوردن ویژگی‌های دیگر در مجموعه داده و بکارگیری روش‌های ترکیبی در پیاده سازی الگوریتم ها، میزان دقت پیش بینی را افزایش داد.

مراجع

[۱] خمری، ندا و هادی بارانی، **"داده کاوی، مفاهیم و کاربرد آن (شهر الکترونیک)"**، *دومین همایش بین المللی مهندسی برق،علوم کامپیوتر و فناوری اطلاعات*، همدان، ۱۳۹۷.

[۲] شکیبا, زینب؛ مهدیه خدری و فایقه فقیه موسوی، **"مقایسه ی عملکردی الگوریتم های KNN و SVM در دسته بندی متون"**، *چهارمین کنفرانس بین المللی تحقیقات دانش بنیان در مهندسی کامپیوتر و فناوری اطلاعات*، تهران، دانشگاه ابرار، ۱۳۹۶.

[۳] عقیقی, فرزانه؛ حسین عقیقی و امیدمهدی عبادتی، **"بررسی کارایی روشهای طبقه بندی SVM و KNN در استخراج عوارض شهری از ابر نقاط لیدار"**، *دومین کنفرانس بین المللی پژوهش های دانش بنیان در مهندسی کامپیوتر و فناوری اطلاعات*، تهران، دانشگاه مجلسی، ۱۳۹۶.

[۴] نوروزی طیولا, ساره؛ مرتضی موسوی و منوچهر کاظمی، **"تشخیص نفوذ با استفاده از الگوریتم خوشه بندی ترکیبی و knn "**، *چهارمین کنفرانس ملی فناوری اطلاعات، کامپیوتر و مخابرات*، مشهد، دانشگاه تربت حیدریه، ۱۳۹۶.

[5] Balakrishnan, J.M.D., Significance of classification techniques in prediction of learning isabilities. arXiv preprint arXiv:1011.0628, 2010.

[6] Bhardwaj, Ankit, Arvind Sharma, and V. Shrivastava. "Data mining techniques and their implementation in blood bank sector–a review." *International Journal of Engineering Research and Applications (IJERA)* 2.4: 1303-1309, 2012.

[7] Bishop CM. Neural networks for pattern recognition. New York, NY, USA: Oxford University Press; 1995.

[8] Brunassi, L. D., D. J. de Moura, I. D. Naas, M. M. do Vale, S. R. L.de Souza, K. A. O. de Lima, T. M. R. de Carvalho, and L. G. D.Bueno. Improving detection of dairy cow estrus using fuzzy logic. Sci. Agric. 67:503– 509, 2012

[9] Cardoso, H.F.V. & Marchado k. Sample-specific (universal) metric approaches for determining the sex of immature human skeletal remains using permanent tooth dimensions. *Journal of Archaeological Science*, 35(1): 158-168, 2008.

[10] Chang, H. H., & Tsay, S. F, Integrating of SOM and K-mean in data mining clustering: An empirical study of CRM and profitability evaluation, 2004.

[11] Duda, Richard O., Peter E. Hart, and David G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

[12] D.L. Sparks, R. Hernandez, L.A. Estévez, Evaluation of density-based models for the solubility of solids in supercritical carbon dioxide and formulation of a new model, Chem. Eng. Sci. 63 (17), 4292–4301, 2008.

[13] Elmamouz, G. and M. Nadimi, A review of methods for prediction of type 2 diabetes based on Bayesian theory. National Conference on Science and Computer Engineering, 2012.

[14] Fazli H, Momeni H. Comparison and evaluation of data mining algorithms, decision tree and SVM application for intrusion detection. In: Proceedings of 8th Symposium progress in science and technology 2013, Mashhad. Iran; 2013.

[15] Ghritlahre, Harish Kumar, and Radha Krishna Prasad. "Exergetic performance prediction of solar air heater using MLP, GRNN and RBF models of artificial neural network technique." *Journal of environmental management* 223: 566-575,2018.

[16] Goldschmidt, Ronaldo; Passos, Emmanuel. Data Mining – Um Guia Prático. Rio de Janeiro, editora Campus, (2005).

[17] Grilli ,E., F. Menna, and F. Remondino, a Review of Point Clouds Segmentation and Classification Algorithms. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences: p. 339-344, 2017.

[18] Hughes, A. M, Strategic database marketing. IL: Probus Publishing Company, 1994.

[19] J. Mendez-Santiago, A.S. Teja, Solubility of solids in supercritical fluids: consistency of data and a new model for cosolvent systems, Ind. Eng. Chem. Res. 39 (12), 4767–4771, 2000.

[20] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.

[21] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1(1), 81–106, 1986.

[22] Khamis, H. S., Cheruiyot, K. W., & Kimani, S, Application of k-nearest neighbour classification in medical data mining. *International Journal of Information and Communication Technology Research*, *4*(4), 2014.

[23] Lambda, A., & Kumar, D. Survey on KNN and Its Variants. *International Journal of Advanced Research in Computer and Communication Engineering*, *5*(5), 2016.

[24] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth, 1984.

[25] Reinartz, W. J., & Kumar, V. On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of marketing*, *64*(4), 17-35, 2000.

[26] Richards, J.A. and J. Richards, Remote sensing digital image analysis. Vol. 3: Springer, 1999.

[27] Testik, Murat Caner, et al. "Discovering blood donor arrival patterns using data mining: A method to investigate service quality at blood centers." *Journal of medical systems* 36.2: 579-594, 2012.

[28] Tharwat, Alaa, Ahmed M. Ghanem, and Aboul Ella Hassanien. "Three different classifiers for facial age estimation based on k-nearest neighbor." *Computer Engineering Conference (ICENCO), 2013 9th International*. IEEE, 2013.

[29] Trabelsi, Asma, Zied Elouedi, and Eric Lefevre. "Decision tree classifiers for evidential attribute values and class labels." *Fuzzy Sets and Systems* (2018).

[30] van Eck, N. J., & van Wezel, M. Application of reinforcement learning to the game of Othello. *Computers & Operations Research*, *35*(6), 1999-2017, (2008).

[31] Yeh, I. C., Yang, K. J., & Ting, T. M. Knowledge discovery on RFM model using Bernoulli sequence. *Expert Systems with Applications*, *36*(3), 5866-5871, 2009.

[32] Yu, P. L. H., et al. "Predicting potential drop-out and future commitment for first-time donors based on first 1· 5-year donation patterns: the case in Hong Kong Chinese donors." *Vox sanguinis* 93.1: 57-63, 2007.

Prediction of blood donations using machine learning techniques based on Decision tree, KNN, SVM, and MLP algorithms

Arash Fahmihassan[1], Mohammadreza Moghari[2], Omid Mahdi Ebadati[3*]

1. Masters student of Operations Research, Kharazmi University, Tehran.
std_fahmihassan@khu.ac.ir

2. Masters student of Operations Research, Kharazmi University, Tehran.
mr_moghari@yahoo.com

3. Department of Management Information Technology, Kharazmi University, Tehran.
ebadati@khu.ac.ir

***Abstract*:** Blood donation has an important and critical role to preserve the health and survival of human life. In today's world, despite the enormous scientific advancements and the great developments in medical sciences, adequate supply of healthy blood is one of the challenges and concerns of the medical community in the world. Preserving and supplying the volume of blood required in blood banks of each region, and the diverse blood groups with the connections between them, with assuming that the number of blood groups are rarer; makes the prediction and planning of blood donation more and more complicated and important during the time. The use of data mining in hospitals and blood transfer centers databases helps in the discovery of relations, so that they can have a future prediction based on the past information. Accordingly, they have better diagnosed and successful cure various illnesses and show the patterns of new injuries. In this paper, we try to use data mining and machine learning techniques in decision making levels at mentioned field, to use this mechanism for prediction that how much blood will be donate to blood transfusion centers and blood banks in different period time, to estimate and supply the required blood volume of blood banks in different areas. In this regard, we use several classification algorithms in supervised learning for the prediction, including decision tree algorithms, KNN, SVM and MLP, these algorithms are implemented to predict and results of accuracy are presented.

Keywords: Machine learning, Decision tree, *k*-nearest neighbor, Support Vector Machine, Multi Layer Perceptron

Corresponding Author *

[1] Data base
[2] Data Mining
[3] Data Set
[4] Decision tree
[5] k-nearest neighbor
[6] Support Vector Machine
[7] Kernel
[8] Gaussian Radial Basis Function
[9] Multi Layer Perceptron
[10] Artificial Neural Network