فصلنامهٔ علمی - پژوهشی گردشگری و توسعه، سال هشتم، شمارهٔ چهارم، زمستان ۹۸
صفحهٔ ۲۱۸-۲۰۲

مقایسهٔ دو الگوریتم درخت تصمیم‌گیری و ماشین بردار پشتیبان برای طبقه‌بندی مکان‌های جاذب گردشگری براساس اطلاعات زمینه‌ای کاربر

سهیل رضایی[۱]، ابوالقاسم صادقی نیارکی[۲]، مریم شاکری[۳]

تاریخ دریافت: ۱۳۹۷/۰۸/۱۳ تاریخ پذیرش: ۱۳۹۷/۱۱/۱۵

چکیده

امروزه گردشگری و جذب گردشگر، به‌منزلهٔ یکی از منابع اقتصادی، و همچنین بررسی داده‌های گردشگری، با توجه به اهمیت روزافزون صنعت گردشگری و تجارتی و رقابتی‌شدن آن، اهمیت ویژه‌ای یافته است. در صنعت گردشگری شناخت ویژگی‌ها و اطلاعات زمینه‌ای کاربر سبب اتخاذ تصمیمات هدفمندتر و ارائهٔ خدمات رضایت‌بخش‌تر به کاربر می‌شود که این امر بدون استفاده از ابزارها و تکنیک‌های داده‌کاوی میسر نمی‌شود. روش‌های گوناگونی برای طبقه‌بندی و بررسی داده‌ها ارائه شده است. با توجه به اهمیت بالای شناخت رفتار و ویژگی‌های گردشگران در انتخاب مکان جاذب گردشگری و درنتیجه جلب رضایت آنان، هدف این مطالعه مقایسهٔ دو الگوریتم «درخت تصمیم‌گیری» و «ماشین بردار پشتیبان» برای طبقه‌بندی مکان‌های جاذب گردشگری براساس اطلاعات زمینه‌ای کاربر در نرم‌افزار Weka است. به این منظور، از اطلاعات زمینه‌ای کاربر، ازجمله سن، جنسیت، میزان تحصیلات، نوع مکان گردشگری و امتیازی که کاربران به مکان گردشگری داده‌اند برای طبقه‌بندی مکان‌های جاذب گردشگری استفاده شده است. با این هدف، اطلاعات زمینه‌ای و اطلاعات مکان‌های گردشگری درمورد جاذبه‌های گردشگری تهران از ۲۲۰ کاربر جمع‌آوری و برای آموزش و تست دو الگوریتم استفاده شد. با بررسی نتایج این تحقیق با معیارهای مختلف، مشخص شد که درخت تصمیم‌گیری در مقایسه با روش ماشین بردار پشتیبان بر روی داده‌های استفاده‌شده عملکرد بهتری دارد.

واژه‌های کلیدی: داده‌کاوی، طبقه‌بندی، زمینه، درخت تصمیم‌گیری، ماشین بردار پشتیبان، گردشگری

---

۱. دانش آموخته کارشناسی ارشد سیستم اطلاعات مکانی، دانشکده مهندسی نقشه برداری، دانشگاه خواجه نصیرالدین طوسی، تهران، ایران

۲. نویسنده مسئول: استادیار گروه سیستم اطلاعات مکانی، دانشکده مهندسی نقشه برداری، دانشگاه خواجه نصیرالدین طوسی، تهران، ایران (a.sadeghi@kntu.ac.ir)

۳. دانشجوی دکتری سیستم اطلاعات مکانی، دانشکده مهندسی نقشه برداری، دانشگاه خواجه نصیرالدین طوسی، تهران، ایران

مقدمه

با توجه به اهمیت روزافزون گردشگری و افزایش رقابت در بازار گردشگری در سرتاسر جهان، میان سازمان‌ها و مقاصد گردشگری، برای جذب گردشگر، رقابت شدیدی پیدا شده است. همچنین رشد گردشگری سبب شده است که توجه به صنعت گردشگری روزبه‌روز افزایش یابد و این صنعت به یکی از صنایع مهم و درآمدزا تبدیل شود. با توجه به ابزارها، سخت‌افزارها، نرم‌افزارها و پیشرفت فناوری، شناسایی و تقسیم‌بندی گردشگران از اهداف مهم فناوری اطلاعات به‌شمار می‌آید (انصاری و اسدی، ۱۳۹۵). امروزه با گسترش و توسعهٔ شهرها، پراکندگی مکان‌های گردشگری در سطح شهرها افزایش یافته و این امر موجب افزایش اهمیت پیش‌بینی تقاضای گردشگری شده است؛ زیرا ذی‌نفعان را قادر می‌سازد که برنامه‌ها و سیاست‌ها را تعدیل کنند (Liu et al., 2018). در صنعت گردشگری، شناخت ویژگی‌ها، رفتار و خصوصیات گردشگران و شناخت عواملی که هریک از این گروه‌ها را به سمت یک نوع مکان گردشگری سوق می‌دهد اهمیت زیادی دارد. با شناخت نیازها، رفتار، خصوصیات و اطلاعات زمینهٔ[1] گردشگران می‌توان محصولاتی با بازده بالاتر طراحی کرد و نیز در زمینهٔ گردشگری سیاست‌های مؤثرتری اتخاذ کرد. در چنین صنعت رقابتی، حفظ سهم بازار برای کسب‌وکار بدون استفاده از ابزارها و تکنیک‌های داده‌کاوی[2] به‌منظور توسعه و مدیریت خدمات گردشگری امکان‌پذیر نخواهد بود (انصاری و اسدی، ۱۳۹۵). داده‌کاوی با استفاده از تکنیک‌ها و روش‌های گوناگون، به کشف دانش مفید گردشگران از حجم زیاد داده‌های ذخیره‌شده در پایگاه‌های داده می‌پردازد. هدف اصلی از استخراج داده‌ها، استخراج اطلاعات مفید از داده‌های خام بزرگ و تبدیل آن به شکل قابل‌فهم برای استفادهٔ مؤثر و کارآمد از آن است (Jadhav et al., 2016; Nikam, 2015).

در سال‌های اخیر، پژوهش‌هایی به طبقه‌بندی و تکنیک‌های طبقه‌بندی در دو حوزهٔ گردشگری و داده‌کاوی توجه کرده‌اند. لیو و همکاران ابزاری را برای پیش‌بینی تقاضای صنعت گردشگری توسعه دادند (Liu et al., 2018). این مطالعه از ترکیب روش‌های انتخاب ویژگی، رگرسیون بردار پشتیبان،[3] بهینه‌سازی ازدحام ذرات[4] به نام FS–PSOSVR استفاده کرده است. رافیدا و همکاران (2017) از مدل ماشین بردار پشتیبان موجک (WSVM)[5] برای پیش‌بینی‌های سری زمانی توریستی ماهانه استفاده کردند. مدل WSVM ترکیبی از تحلیل موجک و ماشین بردار پشتیبانی است. لیم و همکارانش (2015) روشی برای توصیهٔ تور براساس علاقه‌مندی کاربر از تاریخچهٔ بازدیدهای قبلی معرفی کردند. آن‌ها روش خود را با استفاده از یک مجموعهٔ داده از فلیکر، شامل سه شهر ارزیابی کردند و مشخص شد که روش آن‌ها می‌تواند تورهایی را پیشنهاد دهد که محبوب‌ترند و مکان‌ها و نقاط علاقه‌مندی[6] بیشتری را دربر می‌گیرند. آماراگانا و بوخالیس (2015) در مطالعه‌ای دربارهٔ گردشگری هوشمند به این

1. Context
2. Data Mining
3. Support vector Regression
4. Particle Swarm Optimization
5. Wavelet Support Vector Machine
6. Point of Interest

موضوع پرداختند که با استفاده از اطلاعات کاربر، هم مستقیماً و از پروفایل کاربر و هم غیرمستقیم و از تاریخچهٔ بازدیدهای قبلی، سرویس‌های شخصی‌سازی‌شده را برای پیش‌بینی مقصد گردشگری به کاربر ارائه کنند. باتت و همکارانش (2012) سیستم توصیه‌گر جدیدی به نام Turist@ را ارائه کردند که طراحی عامل مبنا بر آن اجازه می‌دهد تا انواع گوناگون فعالیت‌ها را به شیوهٔ انعطاف‌پذیری مدل‌سازی کند. توصیه‌ها در این سیستم براساس اطلاعات پروفایل کاربر ـ برای مثال سن و تحصیلات ـ ارائه می‌شود که پس از تجزیه و تحلیل فعالیت کاربر به‌روزرسانی می‌شود. چن و وانگ (2007) از تکنیک شبکهٔ عصبی جدیدی، رگرسیون بردار پشتیبان (SVR)، به‌منظور پیش‌بینی تقاضای گردشگری استفاده کردند. هدف از این مطالعه بررسی امکان‌سنجی SVR در پیش‌بینی تقاضای گردشگری بود. زو و همکاران (2007) روش داده‌کاوی مبتنی‌بر ماشین بردار پشتیبان برای هزینهٔ گردشگر را توسعه دادند.

در سال‌های اخیر در ایران، مطالعات کمتری به طبقه‌بندی داده‌ها در زمینهٔ اطلاعات گردشگری پرداخته‌اند. برزمینی (۱۳۹۶) ضمن بررسی اجمالی داده‌کاوی، به نقش آن در شناسایی و تعیین بازار هدف در صنعت گردشگری پرداخته است. صفدری و همکاران (۱۳۹۶) در پژوهشی توصیفی با استفاده از مجموعه داده‌های جمع‌آوری‌شده دربارهٔ زردی نوزادان در شهر قاهرهٔ مصر به این موضوع پرداختند. آن‌ها از تکنیک‌های داده‌کاوی، از قبیل درخت تصمیم‌گیری،[1] بیز ساده و نزدیک‌ترین همسایه استفاده کردند. مسلمی نجار کلائی و همکاران (۱۳۹۴) از تکنیک‌های داده‌کاوی مکانی خوشه‌بندی k-means و همین‌طور روش شبکهٔ عصبی با هدف برآورد زمان سفر استفاده کردند. محمدی و پیرمحمدیانی (۱۳۹۴) با استفاده از رویکرد داده‌کاوی و فرایند تحلیل سلسله‌مراتبی،[2] مدلی برای امتیازبندی رفتاری مشتریان حقیقی به‌منظور ارزیابی ریسک اعتباری و تولید دانش سازمانی درخصوص اعطای تسهیلات اعتباری ارائه کردند.

براساس مطالعات بررسی‌شده، اطلاعات زمینه‌ای کاربر (اطلاعات توصیف‌کنندهٔ کاربر ازجمله سن، جنسیت و تحصیلات) کمتر در ارائهٔ سرویس‌های گردشگری، به‌ویژه در انتخاب مکان جاذب گردشگری، مدنظر قرار گرفته است؛ حال آنکه شناخت ویژگی‌ها، رفتار و خصوصیات در صنعت گردشگری و رضایت گردشگر اهمیت زیادی دارد. هدف این مطالعه بررسی اطلاعات گردشگران و مکان‌های گردشگری و طبقه‌بندی این مکان‌ها براساس اطلاعات زمینه‌ای کاربر با دو روش درخت تصمیم‌گیری و ماشین بردار پشتیبان است. اطلاعات زمینهٔ استفاده‌شده در این مطالعه شامل سن، جنسیت، میزان تحصیلات، نوع مکان گردشگری و امتیازی است که هر کاربر به مکان گردشگری داده است. داده‌های جمع‌آوری‌شده مربوط به شهر تهران است و درنهایت با بررسی دقت هر روش، نتایج ارزیابی شده است.

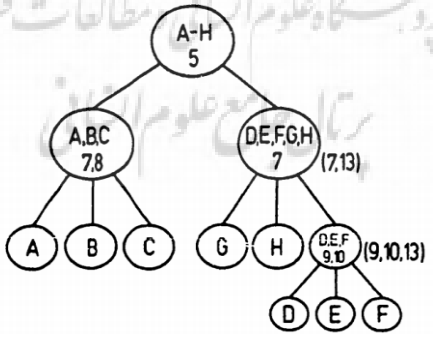
1. Decision Tree
2. Analytical Hierarchy process

مبانی تحقیق

درخت تصمیم‌گیری

درخت تصمیم‌گیری، همانند ساختار درختان، از گره‌های متفاوتی مانند گرهٔ ریشه، گرهٔ میانی و گرهٔ برگ تشکیل شده است (شکل ۱). درخت تصمیم‌گیری پراستفاده‌ترین تکنیک در داده‌کاوی برای طبقه‌بندی حجم زیادی از داده‌ها و استخراج مجموعه داده‌ای است که الگوهای مشابه دارند (Ramya et al., 2018). یکی از ابزارهای طبقه‌بندی، درخت تصمیم‌گیری است که در این حوزه بسیار استفاده شده است. این ابزار علاوه‌بر متغیرهای کمّی، متغیرهای کیفی را نیز طبقه‌بندی می‌کند (Golbandi et al., 2011). درخت تصمیم‌گیری تکنیک داده‌کاوی است که به‌صورت بازگشتی داده‌ها را ـ تا زمانی که هر داده به یک کلاس تعلق بگیرد ـ طبقه‌بندی می‌کند. ساختار درخت تصمیم‌گیری از گره‌های درونی، برگ و ریشه تشکیل شده است (Jadhav et al., 2016). هر فلوچارت مانند ساختار درخت است و هر گرهٔ داخلی یک شرایط آزمایشی را درمورد یک ویژگی نشان می‌دهد. هر شاخه نتیجهٔ شرایط آزمایشی را نشان می‌دهد و هر گرهٔ برگ (گرهٔ پایانه) با یک برچسب کلاس تخصیص داده می‌شود. بالاترین گره، گرهٔ ریشه است. درخت تصمیم‌گیری با رویکرد تقسیم و تسخیر ساخته شده است. گفتنی است درخت تصمیم‌گیری از رویکرد بالا به پایین بهره می‌برد (ibid).

طبقه‌بندی درخت تصمیم‌گیری در دو فاز اجرا می‌شود: ساخت درخت و هرس درخت. ساخت درخت با استفاده از روش بالا به پایین اجرا می‌شود. در زمان اجرای این فاز، درخت به‌صورت بازگشتی تقسیم می‌شود تا زمانی که تمام داده‌ها برچسب‌گذاری شوند. این محاسبات بسیار فشرده است؛ زیرا مجموعه داده‌های آموزشی بارها و بارها این مسیر را طی می‌کنند. هرس درخت با روش پایین به بالا انجام می‌شود واز آن برای بهبود پیش‌بینی و دقت طبقه‌بندی الگوریتم با استفاده از کمینه‌کردن مشکلِ تناسب بیش‌ازحد درخت استفاده می‌شود. تناسب بیش‌ازحد در درخت تصمیم‌گیری، نتیجهٔ خطای طبقه‌بندیِ ازدست‌رفته است (ibid).



شکل ۱: نمونهٔ درخت تصمیم‌گیری (Swain and Swain, 1977)

الگوریتم درخت تصمیم به‌گونه‌ای عمل می‌کند که گوناگونی یا تنوع (از نظر ویژگی هدف) را در گره‌ها به حداقل ممکن برساند. این عدم یکنواختی در گره‌ها با استفاده از معیارهای عدم خلوص[1]

1. Measure Impurity

اندازه‌گیری می‌شود (Yoneyama, et al., 2002). اغلب تفاوت انواع درخت‌های تصمیم در همین معیار اندازه‌گیری عدم خلوص، شیوهٔ شاخه‌بندی[1] و هرس‌کردن گره‌های درخت است. مزایای درخت تصمیم‌گیری به این قرار است:

۱. درخت تصمیم‌گیری بسیار ساده و سریع است؛
۲. نتیجهٔ دقیق به‌دست می‌آورد؛
۳. بیان قابل فهم دارد؛
۴. حافظهٔ کمتری می‌گیرد؛
۵. می‌تواند با داده‌های نویزدار مقابله کند؛
۶. مقایسه‌های غیرضروری را حذف می‌کند (Jadhav et al., 2016; Stein et al., 2005).

ماشین بردار پشتیبان

SVM دسته‌بندی‌کننده‌ای است که جزو شاخهٔ کرنل[2] در یادگیری ماشین[3] به‌شمار می‌رود و بر پایهٔ تئوری یادگیری آماری[4] بنا شده است. اساس SVM را واپینک در سال ۱۹۹۵ توسعه داد و به علت ویژگی‌های جذاب عملکرد آزمایشی امیدوارکننده محبوبیت بسزایی کسب کرد. در SVM، هدف پیداکردن تابعی است که بتواند داده‌ها را از هم متمایز کند. این تابع، که با استفاده از نمونه‌های آزمایشی به‌دست می‌آید، می‌تواند مانند شکل ۲ حالت‌های گوناگونی داشته باشد، اما فقط یک صفحه وجود دارد که فاصلهٔ بین ابرصفحه[5] و نزدیک‌ترین دادهٔ هر کلاس را بیشینه کند که به آن ابرصفحهٔ جداکنندهٔ بهینه[6] می‌گویند (Gunn, 1998). آموزش نسبتاً ساده است. برخلاف شبکه‌های عصبی، در ماکزیمم‌های محلی گیر نمی‌افتد و برای داده‌های با ابعاد بالا تقریباً خوب جواب می‌دهد. همچنین مصالحهٔ بین پیچیدگی دسته‌بندی‌کننده و میزان خطا را به‌وضوح کنترل می‌کند.

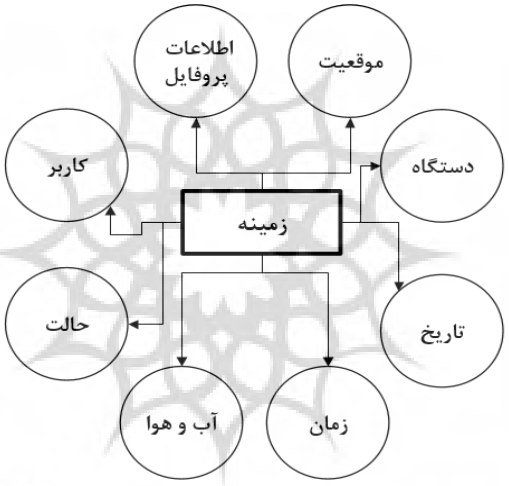
شکل ۲: ابرصفحهٔ جداکنندهٔ بیشینه

1. Splitting
2. Kernel
3. Machine Learning
4. Statistical Learning Theory
5. Hyperplane
6. Optimal Separating Hyperplane

زمینه

هرگونه اطلاعاتی که برای توصیف وضعیت یک هستنده استفاده می‌شود زمینه نام دارد. هستندهٔ فرد، مکان یا شیئی است که با یک کاربر و یک برنامهٔ کاربردی در تعامل است (Dey & Computing, 2001). زمینه به دو دستهٔ کلی تقسیم می‌شود: زمینهٔ اولیه و زمینهٔ ثانویه (Perera et al., 2014). زمینهٔ اولیه هرگونه اطلاعاتی است که از سنسورها به‌دست می‌آید و بدون استفاده از زمینه‌های موجود و هرگونه داده‌یابی بازیابی می‌شود. زمینهٔ ثانویه هرگونه اطلاعاتی است که ازطریق دستکاری در زمینهٔ اولیه به‌دست می‌آید؛ مثلاً موقعیت یک کاربر از نوع زمینهٔ اولیه است، ولی فاصلهٔ بین دو کاربر، که از موقعیت هر کاربر به‌دست می‌آید، زمینهٔ ثانویه است.

انواع زمینه در شکل ۳ نشان داده شده است. در شکل ۳ زمینه به نه بخش تقسیم شده است. در این مطالعه از زمینهٔ کاربر استفاده شده است که دربرگیرندهٔ اطلاعات هر کاربر است.



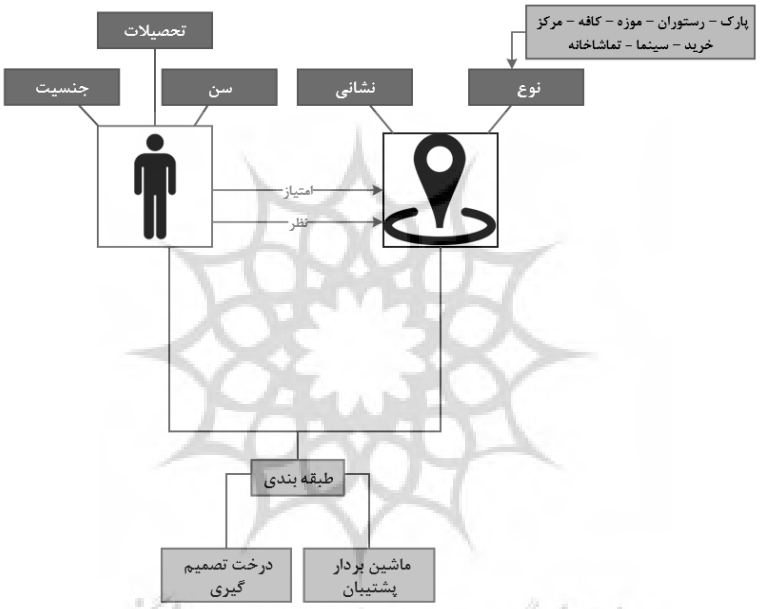
شکل ۳: انواع زمینه

روش‌شناسی

انتخاب جاذبه‌های گردشگری، که با ویژگی‌ها و علایق کاربر بیشترین شباهت را دارند، همواره از اهمیت بالایی برخوردارند؛ با توجه به اینکه افرادی که از یک مکان گردشگری بازدید می‌کنند اغلب زمینه‌هایی مشابه دارند. بنابراین با استفاده از روش یادگیری ماشین می‌توان به طبقه‌بندی مکان گردشگری براساس زمینهٔ کاربر پرداخت. به عبارت دیگر، از اطلاعات زمینهٔ کاربر می‌توان برای استخراج دانش و ارائهٔ پیشنهاد مکان گردشگری متناسب با زمینه‌های کاربر استفاده کرد. بدین‌منظور، روش یادگیری ماشین با هدف تخمینِ مناسب‌ترین مکان برای کاربر با توجه به زمینه‌های کاربر بررسی می‌شود تا بتوان به‌صورت هوشمند پیش‌بینی کرد که مردم به چه مکانی بیشتر علاقه دارند. در این زمینه به مقایسهٔ نتایج دو الگوریتم SVM و درخت تصمیم‌گیری پرداخته می‌شود.

همان‌طور که در شکل ۴ مشخص است، در این روش امتیاز گردشگر به هر مکان، جنسیت، سن و

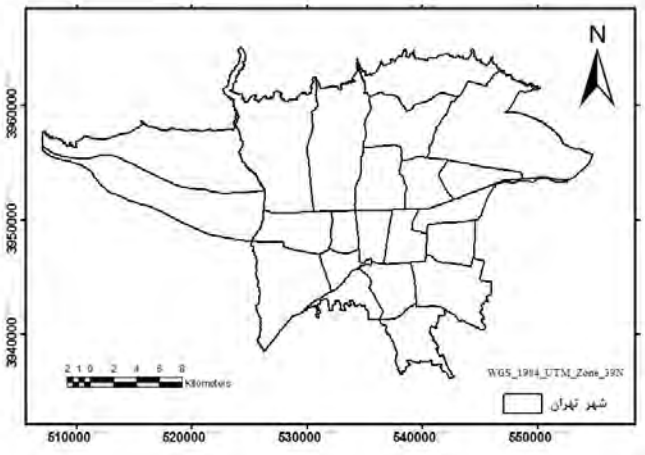
میزان تحصیلات کاربر درحکم ورودی الگوریتم‌های یادگیری ماشین و نوع مکان گردشگری درحکم خروجی الگوریتم در نظر گرفته شده است. برای این منظور، هشت نوع جاذبهٔ گردشگری ـ پارک، رستوران، موزه، کافه، مرکز خرید، سینما و تماشاخانه ـ در نظر گرفته شده است. نخست دو الگوریتم SVM و درخت تصمیم‌گیری با استفاده از داده‌های جمع‌آوری‌شده از کاربران برای طبقه‌بندی جاذبه‌های گردشگری آموزش داده می‌شوند و سپس مدل آموزش‌دیده‌شده می‌تواند برای تعیین نوع مکان مورد علاقهٔ کاربر استفاده شود. با این هدف، اطلاعات سن، جنسیت و میزان تحصیلات هر کاربر به‌منزلهٔ اطلاعات زمینه و همچنین نظر و امتیاز هر کاربر به مکان گردشگری و همچنین موقعیت مکان گردشگری جمع‌آوری می‌شود.



شکل ۴: روش‌شناسی

منطقهٔ مورد مطالعه

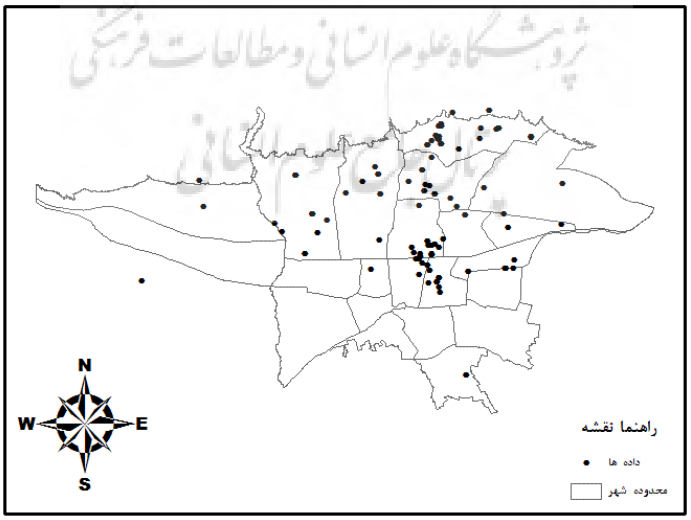
منطقهٔ مورد مطالعه در این پژوهش شهر تهران است. این شهر در حدود عرض جغرافیایی ۳۵/۶۹۷۷۴۲ درجه و طول جغرافیایی ۵۱/۴۲۱۵۲۸ درجه قرار گرفته است. شکل ۵ نمای کلی از منطقهٔ مورد مطالعه را نشان می‌دهد. تهران یکی از مهم‌ترین مراکز گردشگری ایران به‌شمار می‌آید. این شهر دربرگیرندهٔ مجموعه‌ای از جاذبه‌های گردشگری، شامل کاخ‌ها و موزه‌هاست. میدان و برج آزادی، برج میلاد، پل طبیعت و کاخ گلستان از جاذبه‌های گردشگری مهم شهر تهران به‌شمار می‌روند. براساس سایت مرکز آمار ایران در سال ۲۰۱۶ میلادی، شهر تهران یکی از مهم‌ترین شهرهای خاورمیانه در زمینهٔ گردشگری بوده است. همچنین تهران پس از شهرهای دبی، ژوهانسبورگ، ریاض و ابوظبی، در رتبهٔ پنجم از دیدگاه شمار ورود گردشگران خارجی در سال ۲۰۱۶ در مناطق آفریقا و خاورمیانه قرار داشته است و گردشگران خارجیِ تهران در این سال، نیم میلیارد دلار هزینه‌کرد داشته‌اند.



شکل ۵: منطقهٔ مورد مطالعه

جمع‌آوری داده

برای برآورد نوع مکان گردشگری مورد علاقهٔ کاربر نیاز است که از ویژگی‌های کاربر و مکان گردشگری مطلع بود. برای این منظور، اطلاعات و مختصات مربوط به مکان‌های گردشگری، شامل نوع مکان گردشگری (موزه‌ها، پارک‌ها، کافه‌ها، رستوران‌ها و...)، موقعیت مکانی (به‌صورت آدرس یا مختصات GPS)، امتیاز اختصاص‌داده‌شده به هر مکان (از ۱ تا ۱۰، ۱ بیانگر کمترین امتیاز مکان گردشگری و ۱۰ بیانگر بیشترین امتیاز مکان گردشگری است) و نام مکان و همچنین اطلاعات زمینهٔ کاربر یعنی سن، جنسیت و مدرک تحصیلی در این مطالعه جمع‌آوری شده است. اطلاعات به سه روش کاغذی، اپلیکیشن[1] تحت سیستم‌عامل اندروید و فرم اینترنتی گوگل جمع‌آوری شده است. پراکندگی مکانی داده‌های جمع‌آوری‌شده در شکل ۶ مشخص شده است.



شکل ۶: پراکندگی مکانی داده‌ها

1. Application

روش کاغذی

در این روش فرمی کاغذی تهیه و به‌صورت حضوری از مردم برای پرکردن اطلاعات آن کمک گرفته شد (شکل ۷). در این روش باید نشانی واردشده به فرمت مکانی تبدیل شود.

روش اپلیکیشن

در این روش با استفاده از Android Studio اپلیکیشنی در محیط اندروید ساخته شد (شکل ۸-الف). این نرم‌افزار برای گوشی‌های با سیستم‌عامل اندروید بالاتر از ۴ مناسب است و برای دیتابیس آن از سرویس Firebase گوگل بهره برده است. در این نرم‌افزار نیز کاربر اطلاعات خود و مکان را وارد می‌کند و موقعیت کاربر را نیز حسگر موقعیت‌یاب جهانی گوشی[1] همراه ثبت می‌کند.


فرم جمع‌آوری اطلاعات گردشگری

اطلاعات کاربر

سن: ..... جنسیت: □ زن □ مرد

تحصیلات:

□دانش آموز □دانشجو لیسانس □دانشجو فوق لیسانس □دانشجو دکتری □ لیسانس □فوق لیسانس □دکتر

اطلاعات مکان گردشگری

نوع مکان گردشگری:

□ رستوران □کافه □پارک و بوستان □سینما □میراث فرهنگی □میراث فرهنگی □تماشاخانه □موزه □مرکز خرید

نام مکان گردشگری: .....

امتیاز (بین ۱ تا ۱۰): .....

نظر و انتقاد درمورد مکان بازدید شده:

.....

.....

.....

نشانی:.....

.....

.....

.....

شکل ۷: فرم کاغذی

روش فرم اینترنتی گوگل

این روش نیز مانند روش کاغذی است با این تفاوت که در بستر اینترنت منتشر می‌شود و از سرویس

1. Global Positioning System (GPS)

گوگل استفاده می‌کند. در این روش نیز همانند روش کاغذی باید نشانی‌های واردشده به‌دست کاربران، به فرمت مکانی تبدیل شوند (شکل ۸–ب). درنهایت ۲۲۰ رکورد اطلاعات از هر سه روش جمع‌آوری شد که نمونه‌ای از آن‌ها در شکل ۹ آمده است.

پیش‌پردازش داده‌ها

قبل از به‌کارگیری روش درخت تصمیم‌گیری، پیش‌پردازشی بر روی داده‌ها صورت می‌گیرد. در این پژوهش براساس امتیازی که کاربر به مکان گردشگری داده است این مکان‌ها به سه دسته طبقه‌بندی شده‌اند. امتیاز ۹ و ۱۰ مکان‌های مورد علاقهٔ کاربر، امتیازهای ۵، ۶، ۷ و ۸ مکان‌هایی با علاقهٔ کمتر و امتیازهای زیر ۵ مکان‌هایی که مورد علاقه کاربر نبوده‌اند در نظر گرفته شده‌اند. برای این پژوهش، داده‌های دستهٔ اول برای ورودی درخت تصمیم‌گیری انتخاب شده‌اند. از بین مکان‌هایی که به‌منزلهٔ مکان گردشگری مشخص شده بود چهار مکان پارک، رستوران، موزه و کافه انتخاب شدند و مابقی مکان‌ها به دلیل حجم کم دادهٔ جمع‌آوری‌شده وارد مرحلهٔ پردازش نشدند. همچنین تحصیلات کاربران نیز برای راحتی در اجرای درخت به‌صورت A تا I کدگذاری شدند که توضیح آن در جدول ۱ آمده است.

ب الف

شکل ۸ – الف: محیط اپلیکیشن ب: فرم اینترنتی گوگل

جدول ۱: کدگذاری میزان تحصیلات

A	دانش‌آموز	F	دکتر
B	دیپلم	G	دانشجو لیسانس
C	فوق‌دیپلم	H	دانشجو فوق‌لیسانس
D	لیسانس	I	دانشجو دکتری
E	فوق‌لیسانس		

پیاده‌سازی و ارزیابی

روند پیاده‌سازی در شکل ۱۰ نشان داده شده است. پس از جمع‌آوری داده‌ها از سه طریق فرم اینترنتی، فرم کاغذی و اپلیکیشن، بر روی آن‌ها پیش‌پردازش صورت گرفته است. در این مطالعه ۹۰ درصد از داده‌ها به‌منزلهٔ آموزش و ۱۰ درصد به‌منزلهٔ تست در نظر گرفته شده است. درنهایت پنج دادهٔ سن، جنسیت، میزان تحصیلات، امتیاز و نوع مکان گردشگری پس از پیش‌پردازش وارد مرحلهٔ پردازش شدند و درنهایت پس از ارزیابی نتایج هر روش، روش بهتر مشخص شد. به‌منظور پیاده‌سازی درخت تصمیم‌گیری و SVM، پنج ستون از داده‌های جمع‌آوری‌شده شامل نوع کاربری، سن، جنسیت، تحصیلات و امتیاز کاربر به مکان‌های گردشگری وارد نرم‌افزار Weka شد. سپس به‌وسیلهٔ ابزار داده‌کاوی J48 (شکل ۱۱) و روش SVM، بر روی این داده‌ها پردازش صورت گرفت. برای بررسی نتایج به‌دست‌آمده، از سه شاخص آماری موارد صحیح طبقه‌بندی‌شده،[1] جذر میانگین مربعات خطا[2] (رابطهٔ (۱)) و میانگین مطلق خطا[3] (رابطهٔ (۲)) استفاده شد. نتایج حاصل از طبقه‌بندی در جدول ۲ آمده است.

TOURIST INFORMATIONS

Type	Age	Sex	Education	Name	Rate	Comment	Location	Ave Rating	Category Rating Ave
مرکز خرید	27	m	دانشجو لیسانس	پالادیم	10	خیلی عالی و شیک ولی خیلی گرون	35.79810739 51.41270892	8.8	7.4222222
	23	F	دانشجو لیسانس		10	گرونه ولی جای خوبی برای گذران وقت			
	21	M	دانشجو لیسانس		8	نوع معماری و کیفیت عالی			
	20	M	دانشجو لیسانس		8	-			
	23	M	دانشجو لیسانس		8				
	34	زن	فوق لیسانس	بازار تهران	10		خیابان بانزده خرداد	10	
	36	زن	لیسانس	پاساژ سانا	8	گران بودن اجناس، کافه و رستوران خوب و کیفیت خوب اجناس	نیاوران-بلوار اندرزگو-مرکز خرید سانا	8	
	24	مرد	دانشجو فوق لیسانس	مرکز خرید دنیای نور	4		بزرگراه رسالت غرب. مرکز خرید دنیای نور	4	
	26	زن	دانشجو فوق لیسانس	پاساژ کسا	5		پیروزی، جهارراه کوکاکولا	5	
	26	زن	دانشجو فوق لیسانس	اسکان	5		35.7636693 51.4108794	5	
	26	زن	دانشجو فوق لیسانس	بام لند	10		همت، خروجی بلوار کاشان، میدان	10	
	40	f	دیپلم	کوروش	9			9	
	44	f	دیپلم		10				
	35	f	فوق لیسانس		10				
	30	f	دانشجو دکتری		7	هزینه پارکینگ بسیار بالاست	ستاری ، مرکز خرید کوروش	7	

شکل ۹: نمونهٔ اطلاعات جمع‌آوری‌شده

1. Correctly Classified Instances
2. Root Mean Square Error
3. Mean Absolute Error

$$RMSE = \sqrt{\frac{1}{n}\sum\left[y_i - x_i\right]^2} \quad (۱)$$

$$MAE = \frac{\sum_{i=1}^{n}|x_i - y_i|}{n} \quad (۲)$$

جدول ۲: نتایج درخت تصمیم‌گیری و SVM

	جذر میانگین مربعات خطا	درصد موارد صحیح طبقه‌بندی‌شده	میانگین مطلق خطا
SVM	361/0	73/75	0/285
J48	0/287	77/5	164/0

جدول ۲ نشان می‌دهد که روش درخت تصمیم‌گیری در هر سه عامل تعیین‌شده عملکرد بهتری از خود نشان داده است. یکی از روش‌های دیگر ارزیابی، استفاده از مساحت زیر-منحنی Recall-Precision است. هرچه مساحت زیر-نمودار PR بیشتر باشد، نشان‌دهندهٔ مقادیر بالا برای Precision و Recall است. مقدار بالای دقت نشان‌دهندهٔ مقدار پایین مثبت کاذب[1] و مقدار بالای Recall نشان‌دهندهٔ مقدار پایین منفی کاذب[2] است؛ بنابراین نموداری که بالای نمودار دیگر قرار می‌گیرد نشان می‌دهد که به‌نسبت روش دیگر، عملکرد بهتری داشته است. شکل ۱۲ نشان‌دهندهٔ نمودار recall-precision برای دو روش SVM و درخت تصمیم‌گیری است. با توجه به نمودار، درخت تصمیم‌گیری در تمامی قسمت‌ها بالاتر از نمودار SVM است.

شکل ۱۰: روند پیاده‌سازی

شاخص دیگر استفاده‌شده در این مطالعه برای بررسی عملکرد دو روش درخت تصمیم‌گیری و SVM، ماتریس خطای[3] طبقه‌بندی است (جدول ۳). همان‌طور که در جدول ۳ مشخص است، در این شاخص نیز درخت تصمیم‌گیری عملکرد بهتری دارد و در سه نوع مکان، پارک و رستوران به‌نسبت SVM عملکرد بهتری داشته و فقط در طبقهٔ کافه به‌نسبت SVM عملکرد ضعیف‌تری داشته است.

1. False Positive
2. False Negative
3. Confusion Matrix

برای بررسی دقیق‌تر جدول ۳ می‌توان از شاخص دقت کلی استفاده کرد. دقت کلی از جمع اعداد قطر اصلی ماتریس خطا تقسیم بر مجموع کل اعداد ماتریس به‌دست می‌آید (رابطه (۳)).

$$\text{دقت کلی} = \frac{1}{n}\sum p_{ii} \qquad (۳)$$

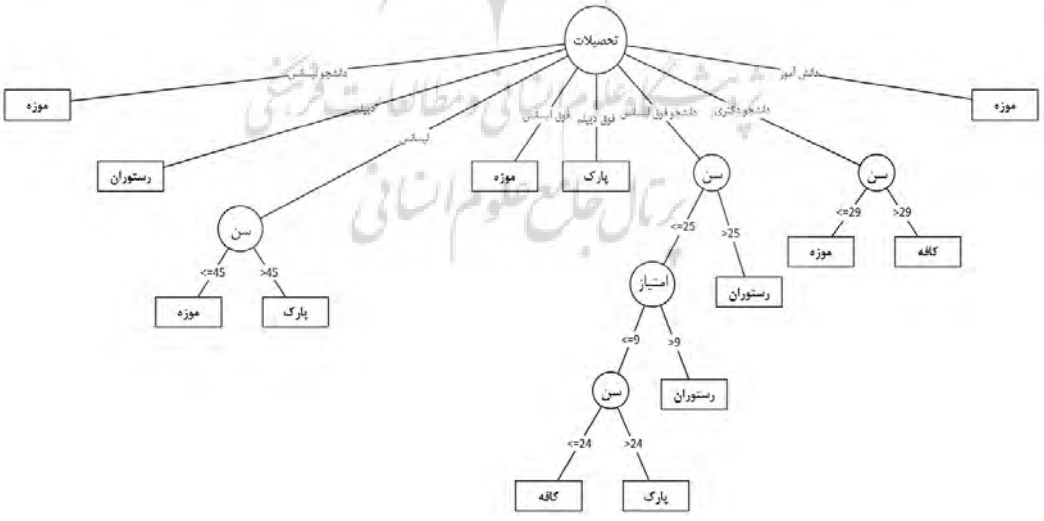
دقت کلی محاسبه‌شده برای این درخت تصمیم‌گیری برابر با مقدار ۷۷ درصد و برای روش SVM برابر ۷۳ درصد است.

جدول ۳: ماتریس خطای طبقه‌بندی

	موزه		کافه		رستوران		پارک	
	DT	SVM	DT	SVM	DT	SVM	DT	SVM
پارک	۶	۶	۱	۳	۱	۰	۶	۵
رستوران	۱	۱	۰	۳	۱۴	۱۱	۰	۰
کافه	۳	۳	۱۰	۱۱	۲	۱	۰	۰
موزه	۳۲	۳۲	۳	۳	۱	۱	۰	۰

با توجه به ایراداتی که به روش دقت کلی گرفته می‌شود، از شاخص کاپا[1] نیز برای حصول اطمینان استفاده شده است. شاخص کاپا، که براساس رابطه (۴) محاسبه می‌شود، در این پژوهش برای درخت تصمیم‌گیری مقدار ۰/۶۶۷ و برای SVM برابر ۰/۶۱۱ است.

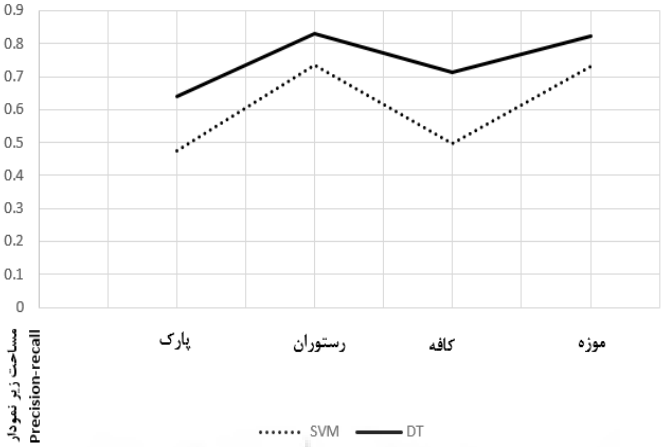
$$kappa = \frac{p_o - p_c}{1 - p_c} * 100 \qquad (۴)$$



شکل ۱۱: نتیجهٔ درخت تصمیم‌گیری

1. Kappa Coefficient

در رابطهٔ بالا، p_o پیکسل‌های درست طبقه‌بندی‌شده مشاهداتی و p_c توافق مورد انتظار درمورد پیکسل‌هاست.



شکل ۱۲: نمودار Recall-Precision

نتیجه‌گیری

با افزایش چشمگیر تعداد انتخاب‌ها در بسته‌های سفر، هتل‌ها، جاذبه‌های گردشگری، پیداکردن آنچه گردشگر بدان نیاز دارد بسیار دشوار شده است. به همین علت به ابزارهای داده‌کاوی برای تحلیل داده‌ها توجه شده است. داده‌کاوی درواقع جست‌وجوی خودکار منابع داده‌ای بزرگ برای یافتن الگوها و وابستگی‌هایی است که تحلیل‌های ساده و معمول آماری قادر به آن کار نیستند. یکی از زمینه‌های استفاده از این ابزار، تحلیل داده‌های وسیع و مدل‌سازی پیش‌گویانه با روش‌های محاسباتی جدید گردشگری است. هدف این مطالعه تجزیه و تحلیل داده‌های گردشگری با دو روش درخت تصمیم‌گیری و SVM است تا دقت این دو روش برای داده‌های استفاده‌شده بررسی شود و روش بهتر استخراج شود. برای این منظور، با استفاده از نرم‌افزار Weka پنج ستون از داده‌های جمع‌آوری‌شده، شامل سن، جنسیت، میزان تحصیلات، نوع مکان گردشگری و امتیازی که هر کاربر به مکان گردشگری داده است وارد نرم‌افزار شد، سپس دو روش درخت تصمیم‌گیری و ماشین بردار پشتیبان بر روی داده‌ها اعمال شد. این دو روش با معیارهای گوناگون ارزیابی شدند. درمورد معیار موارد صحیح طبقه‌بندی‌شده، درخت تصمیم‌گیری ۷۷/۵ درصد و SVM ۷۳/۷۵ درصد است. معیار جـذر میانگین مربعات خطا برای درخت تصمیم‌گیری ۰/۲۸۷ و برای SVM ۰/۳۶۱ محاسبه شد. میانگین مطلق خطا برای درخت تصمیم‌گیری ۰/۱۶۴ و برای SVM ۰/۲۸۵ است. همچنین دقت کلی، که از طریق ماتریس خطا محاسبه می‌شود، نیز برای هریک از روش‌ها به‌دست آمد که این معیار برای درخت تصمیم‌گیری مقدار ۷۷ درصد و برای SVM ۷۳ درصد محاسبه شد. برای حصول اطمینان از نتایج به‌دست‌آمده، از معیار کاپا نیز استفاده شد که مقدار آن برای درخت تصمیم‌گیری ۰/۶۶۷ و برای SVM ۰/۶۱۱ به‌دست آمد. در این مطالعه از نمودار recall-precision نیز استفاده شد که درخت

تصمیم‌گیری در تمام نمودار بالاتر از SVM قرار گرفت که حاکی از دقت بالاتر این روش است. گفتنی است که این دقت با شبکهٔ آموزش‌داده‌شده با ۱۹۸ رکورد، درحکم دادهٔ ورودی، به‌دست آمده است؛ درصورتی که اگر تعداد داده‌های ورودی بیشتری استفاده شود، دقت بیشتری به‌دست خواهد آمد.

این مطالعه نشان داد که روش درخت تصمیم‌گیری در بیشتر شاخص‌های آماری عملکرد بهتری دارد که یکی از دلایل آن می‌تواند حجم و نوع دادهٔ استفاده‌شده باشد. با توجه به اهمیت روزافزون گردشگری و رقابت روزافزون سازمان‌های گردشگری، مطالعه بر روی داده‌های گردشگری بسیار ارزشمند است تا ازطریق آن، خدمات رضایت‌بخش‌تری به کاربر ـ با توجه به علایق او ـ داده شود. با استفاده از این روش و اطلاعات زمینهٔ کاربر می‌توان سیستم‌های توصیه‌گر گردشگری طراحی کرد، سیستم‌هایی که بتواند مکان‌های گردشگری که بیشترین شباهت را با علایق و ترجیحات گردشگر دارند، به گردشگر توصیه کند. طراحی سیستم توصیه‌گرِ همراه براساس اطلاعات زمینهٔ کاربر از فعالیت‌های آیندهٔ این تحقیق است.

منابع

انصاری، آذرنوش و اسدی، علی (۱۳۹۵). «گردشگری، ارزیابی وفاداری گردشگر به مقصد با رویکرد داده‌کاوی گردشگران داخلی شهر اصفهان». فصل‌نامهٔ *مطالعات مدیریت گردشگری*، سال یازدهم، شمارهٔ ۳۵، ص ۸۵-۱۰۶.

برزمینی، س. (۱۳۹۶). «داده‌کاوی و کاربرد آن در صنعت گردشگری». کنفرانس ملی علوم مهندسی.

صفدری و همکاران (۲۰۱۸). «مقایسهٔ الگوریتم‌های مختلف طبقه‌بندی داده‌ها برای تعیین نوع زردی در نوزادان». مجلهٔ *پیاورد سلامت*، سال یازدهم، شمارهٔ ۵، ص ۵۴۱-۵۴۸.

مسلمی نجار کلایی، فردین و همکاران (۱۳۹۴). «پیش‌بینی زمان سفر در مسیرهای برون‌شهری با استفاده از تکنیک‌های داده‌کاوی مکانی، مطالعهٔ موردی: مسیر قائمشهر به بابل و ساری به قائمشهر». *مجلهٔ رایانش نرم و فناوری اطلاعات*، سال چهارم، شمارهٔ ۳، ص ۳-۱۵.

محمدی، شهریار و پیرمحمدیانی، راجیار (۱۳۹۴). «امتیازبندی رفتاری مشتریان بانک با استفاده از رویکرد داده‌کاوی و فرآیند تحلیل سلسله‌مراتبی». *رایانش نرم و فناوری اطلاعات*، سال چهارم، شمارهٔ ۳، ص ۵۲-۶۵.

Batet, M. et al. (2012). "Turist@: Agent-based personalised recommendation of tourist activities" in *Expert Systems with Applications*. 39(8), 7319-7329.

Buhalis, D., & Amaranggana, A. (2015). "Smart tourism destinations enhancing tourism experience through personalisation of services" in *Information and communication technologies in tourism* Springer, 377-389.

Chen, K. Y., & Wang C. H. (2007). "Support vector regression with genetic algorithms in forecasting tourism demand". *Tourism Management*, 28(1), 215-226.

Cho, Y.H., Kim, J. K., & Kim, S.H.J.E.s.w.A. (2002). "A personalized recommender system based on web usage mining and decision tree induction". 23(3), 329-342.

Dey, A. (2001). "Understanding and using context*". personal ubiquitous computing*, 5(1), 4-7.

Golbandi, N., Koren, Y., & Lempel, R. (2011). "Adaptive bootstrapping of recommender systems using decision trees". in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM.

Gunn, S. R. (1998). "Support vector machines for classification and regression". *ISIS technical report*, 14(1), 5-6.

Jadhav, S. D., & Channe, Hemlata (2016). "Comparative study of K-NN, naive Bayes and decision tree classification techniques". *International Journal of Science and Research (IJSR)*, 5(1), 1842-1845.

Kim, S.S., Timothy, D.J. and Hwang, J.J.T.M. (2011). "Understanding Japanese tourists' shopping preferences using the Decision Tree Analysis method". *Tourism Management*, 32(3), 544-554.

Lim, K. H. (2015). "Recommending tours and places-of-interest based on user interests from geo-tagged photos". in *Proceedings of the ACM SIGMOD on PhD Symposium*. ACM.

Liu, H. H., Chang, L. C., Li, C. W., & Yang, C. H. (2018). "Particle Swarm Optimization-Based Support Vector Regression for Tourist Arrivals Forecasting . *Computational Intelligence and Neuroscience*, 2018.

Nikam, S. S. (2015). "A comparative study of classification techniques in data mining algorithms. *Oriental journal of computer science & technology*, 8(1), 13-19.

Perera, C. et al. (2014). "Context aware computing for the internet of things: A survey". *IEEE commun surv.tutor*, 16(1), 414-454.

Rafidah, A. et al. (2017). "A Wavelet Support Vector Machine Combination Model for Singapore Tourist Arrival to Malaysia" in *IOP Conference Series: Materials Science and Engineering*. IOP Publishing.

Ramya, R. et al. (2018). A Review Of Different Classification Techniques In Machine Learning Using Weka For Plant Disease Detection.

Stein, G. et al. (2005).“Decision tree classifier for network intrusion detection with GA-based feature selection” in *Proceedings of the 43rd annual Southeast regional conference-2*, 136-141, ACM.

Swain, P. H., & Hauska, H. (1977). “The decision tree classifier: Design and potential”. *IEEE Transactions on Geoscience Electronics*, 15(3), 142-147.

Xu, X., Law, R., & Wu, T. (2007). “Classification of business travelers using SVMs combined with kernel principal component analysis”. in *International Conference on Advanced Data Mining and Applications*.

Yoneyama, Y. et al. (2002). “Increased plasma adenosine concentrations and the severity of preeclampsia”. 100(6), 1266-1270.