بسم اللّٰہ الرحمن الرحیم

دانشگاه کاشان
دانشکده مهندسی
گروه مهندسی کامپیوتر

تحلیل داده جهت رتبه بندی و پیشبینی امتیاز کاربران Zomato براساس ویژگی های رستوران

نگارش شده توسط:

محمد خاکی

استاد راهنما:

دکتر فرشته دهقانی

زمستان ۱۴۰۰

چکیده

رتبه‌بندی رستوران‌ها در ابعاد مختلف برای کمک به کاربران برخط، امری ضروری جهت راهنمایی مناسب آنها است و افزایش میزان رضایتمندی آنان را به همراه دارد به علت اهمیت این موضوع و همچنین ایجاد معیاری هوشمند برای تعیین سطح یک رستوران تازه وارد شده به سیستم سفارش آنلاین، برآن آمدم تا مدلی برای طبقه‌بندی رستوران‌ها ارائه کنم.

در این پروژه به تحلیل داده‌های رستوران‌های زوماتو پرداخته خواهد شد و بر اساس ویژگی‌های هر رستوران نسبت به پیشبینی امتیازی که کاربران برای مجموعه متصور خواهند بود اقدام می‌کنیم تا رتبه بندی صحیحی از رستوران ها داشته باشیم. تحلیل ها بر روی هشت ویژگی هر رستوران که به دو دسته کلی محیط رستوران و خدمات تحویل غذا تقسیم شده انجام می‌شود و در نهایت با استفاده از داده‌های رستوران‌های قبلی به روش دسته‌بندی با روش ماشین بردار پشتیبان نسبت به رتبه‌بندی رستوران‌ها پرداخته می‌شود.

از دست آوردهای این پروژه می‌توان به این اشاره کرد که رستوران‌هایی که به تازگی شروع به کار کرده‌اند و امتیاز زیادی از کاربران ندارند برحسب پیشبینی، رتبه بندی‌ای برای آنان صورت می‌گیرد.

مدل طراحی شده در این پروژه قادر به امتیازدهی با دقت ۶۰ درصد به رستوران‌ها می‌باشد با عنایت به این موضوع که برای رستوران‌هایی که قابلیت سفارش آنلاین را دارند و اتفاقا دغدغه اصلی این پروژه هم می‌باشد، دقت عملکرد تا ۹۰ درصد کارایی دارد.

کلمات کلیدی: رتبه بندی - رستوران - پیشبینی - ماشین بردار پشتیبان - خوشه بندی - یادگیری‌ماشین

فهرست

فهرست اشکال

فهرست جداول

فصل ۱: مقدمه

با افزایش ابزارهای الکترونیک نظیر گوشی‌های تلفن همراه و کامپیوترها در دنیا امروز، سفارش کالاهای مورد نیاز بر بستر اینترنت گسترش زیادی پیدا کرده است. این نحوه خرید، کمک بسیاری به افراد برای صرفه جویی در زمان و هزینه‌های رفت و آمد آنان به صورت مستقیم می‌کند و حتی در سلامتی آنان نقش بسزایی به صورت غیرمستقیم خواهد داشت.

اما تغییر در نحوه خرید افراد نباید در کیفیت خرید آنها موثر باشد و سیستم‌های ارائه دهنده این خدمات باید تلاش خود را بکنند تا تجربه خرید اینترنتی افراد، به خرید حضوری نزدیک و نزدیکتر شود و مشتریان تفاوتی در کیفیت دریافتی در این نوع خرید را حس نکنند بلکه در دسترس‌تر بودن این روش برایشان حس بهتری نیز ایجاد کند.

یکی از این خریدهای پرکاربرد در عصر حاضر سفارش برخط غذا از رستوران‌ها می‌باشد. طبق توضیحاتی که داده شد این نیاز حس می‌شود که باید اطلاعات کافی و مفیدی جهت تصمیم‌گیری برای انتخاب رستوران و پس از آن انتخاب نوع غذا در اختیارمان قرار گیرد تا به عنوان یک مخاطب با توجه به آنکه دیدگاه یا تجربه قبلی‌ای از همه رستوران‌ها نداریم بتوانیم بهترین تصمیم را بگیریم.

علاوه بر ویژگی‌های مشخص شده هر رستوران مثل قیمت غذاها و مکان رستوران یا امکانات موجود در هر یک از آنها، قطعا یکی از متغیرهایی که می‌تواند کمک شایانی به یک سفارش برخط بکند متغیر تجربه می‌باشد. از آنجا که یک نفر نمی‌تواند همه رستوران‌ها را امتحان کند پس می‌توانیم تجربه سایرین را در اختیار دیگران قرار دهیم. مورد با اهمیت دیگر آن است که مخاطبان دوست دارند هرچه سریعتر این انتقال تجربه صورت بگیرد پس چه بهتر که جمیع جهات را با یک عدد به مخاطب ارائه کنیم.

با توجه به مطالب بالا محوریت این پروژه در دو مورد زیر خلاصه می‌شود:

۱) در این پروژه سعی بر آن شده است که با استفاده از سه دسته کلی ویژگی‌های مرتبط به محیط رستوران و کیفیت سرویس خدمات ارسال و نظر کلی مخاطبین گذشته، نسبت به تعیین معیاری برای باکلاسی یا

بی‌کلاسی مجموعه و تعیین اینکه یک رستوران از بخش سفارشات آنلاین خوبی برخوردار است یا خیر اقدام شود. این کار با خوشه‌بندی داده های ثابت رستوران‌ها انجام شده است.

۲) یک مورد دیگر که در این تحقیقات مورد بررسی قرار گرفته نحوه تعیین امتیاز برای رستوران هایی است که به تازگی شروع به فعالیت کرده‌اند چراکه اینگونه مجموعه ها هیچگونه پیشینه‌ای برای رتبه‌بندی ندارند. این بخش از تحقیقات با استفاده از ماشین‌های بردار پشتیبان با استفاده از نتایج داده های مرحله قبل و مدل‌سازی با امتیاز سایر رستوران‌ها صورت پذیرفته است.

در این تحقیق از داده‌های مجموعه رستوران‌های وبسایت زوماتو برای تحلیل مطالب فوق استفاده خواهیم کرد. در این مجموعه داده‌ها، رستوران‌های مختلفی از کشورهای مختلف و با تنوع غذایی زیاد موجود می‌باشد.

در این پروژه تحقیقاتی تمرکز بر روی ویژگی های رستورانی بوده و سیستم بگونه ای طراحی شده است تا بتواند با توجه به نظرات کاربران گذشته در انتخاب مشتریان کمک کند تا تجربه بهتری را لمس بکنند

نتایج این پروژه از آنجایی مورد اهمیت قرار می‌گیرد که این قابلیت را فراهم می‌سازد تا از رستوران های نوپا، تخمینی از عملکرد و رضایتمندی کاربرانشان بر اساس ویژگی‌های مجموعه در آینده داشته باشیم و در صورت داشتن پارامترهای کیفی مناسب، نسبت به استفاده این دسته از رستوران‌ها برای سیستم های توصیه‌گر[۱] اقدام کنیم.

[۱] recommender system

فصل ۲: پیشینه

طبق تحقیقاتی که در [۱] شده است یکی از چالش های رتبه بندی مجموعه ها بر اساس نظر کاربران وجود جریانات ساختگی در امتیازات و برخی ناعدالتی ها در امتیاز دهی توسط مردم است. این تحقیق که در سال ۲۰۱۸ انجام شد راهکارهایی نظیر بررسی عادلانه بودن، سنجش کیفیت، قابل اطمینان بودن و بررسی رفتار کاربران در گذشته پیشنهاد شد. در این پژوهش علاوه بر بررسی کارت های شناسایی، برای افرادی که با ارزهای دیجیتال مثل بیتکوین دست به خرید می‌زنند و اعتبارسنجی آنان قابل ردیابی نیست، الگویی طراحی شد تا اخطاری به مدیران مجموعه برای بررسی دقیق تر این افراد اعلام شود. این الگو بر اساس نسبت دادن ویژگی های ذاتی افراد و ویژگی های درونی و کیفی محصول به هم و مقایسه با کاربرانی که هویت آنان معلوم بود ساخته شده است که نتیجه آن یافتن افراد تقلبی با دقتی در حدود ۸۴ درصد است.

قطعا یکی دیگر از عواملی که می‌تواند بررسی بیشتر نظرات تقلبی را به همراه داشته باشد داشتن یک پیشبینی از آینده است که درصورت عدم تطابق نسبی با واقعیت به عنوان هشدار تلقی گردد. همانطور که در مقاله ای در سال ۲۰۲۱ [۲] بیان شد که محیط یک رستوران و امکان ارسال غذا دو فاکتور مهم در جهت طبقه بندی رستوران ها میتواند باشد. در این مقاله از ویژگی ها یک رستوران دو خروجی جهت تعیین سطح باکلاسی مجموعه و سرویس تحویل غذا آن ارائه می‌دهد و آن را در کنار امتیاز کاربران به عنوان پارامتر های دسته بندی رستوران ها در نظر می‌گیرد گرچه روش دسته بندی در این مقاله با کار صورت گرفته در این پژوهش متفاوت است اما این ایده، پایه ای است جهت ادامه راهبرد این پروژه و افزودن برخی قابلیت های جدید به نتایج آن.

از دیگر فعالیت هایی که بر روی این دیتاست صورت گرفته می‌توان به پروژه [۳] که در ۲۰۱۹ پیاده سازی شده اشاره کرد که از زاویه تحلیل نموداری برای تفکیک تنوع غذایی و کشور های آن بررسی شده و باعث ایجاد پارامتری برای سنجش رستوران بر اساس سلایق و عادات محلی هر منطقه شده است. که طی آن با استخراج اطلاعات منطقه ای رستوران ها مثل کشور هایی که بیشترین رستوران را دارند یا کشور هایی که بیشترین تنوع غذایی را ارائه میدهند یا از زاویه دیگر اثر امتیازی افراد متاهل و مجرد را مقایسه می‌کند. و نشان می‌دهد که نوع مجموعه مثلا رستوران یا کافه یا سوپ فروشی بدون توجه به کیفیت آنها در میانگین امتیازات اثر گذار است. و همچنین فست فود ها از غذا های محلی بیشتر مورد اقبال است. و افراد متاهل عموما

امتیاز کمتری به رستوران ها می‌دهند که نشانگر بعد دیگری از امتیاز دهی به رستوران ها را بیان می‌کند بدین معنی که نتایج پروژه ما و این پروژه و سایر المان ها در کنار هم هستند که میتواند معیار سنجش کامل تری باشد.

فصل ۳: ادبیات موضوع

۳-۱ مقدمه

پیش از ارائه راهکارهای صورت گرفته در این پروژه بهتر است با مفاهیم علمی بکار گرفته شده در این پروژه و اصطلاحات آن آشنا شویم. محوریت این پروژه بر امتیازدهی به مجموعه‌ها می‌باشد تا با استفاده از مفاهیم یادگیری ماشین به سمت هوشمند سازی امتیازدهی حرکت کنیم.

۳-۲ رتبه بندی

محوریت این پروژه در راستای نظر کاربران می‌باشد. بدین صورت که برای سنجش عملکرد رستوران ها از امتیاز کاربران گذشته که از خدمات آن مجموعه استفاده کرده اند بهره می‌گیریم.

امتیاز دهی صرفا مرتبط با پروژه‌ی حاضر نیست بلکه یکی از موثر ترین روش های سنجش عملکردی یک مجموعه است. البته که فرایند امتیازدهی و جمع‌آوری اطلاعات بر بستر آنلاین ساده تر است اما در سرویس های غیر برخط نیز اهمیت ویژه ای برای تصمیم گیری مدیران می‌تواند باشد.

امتیاز کاربران از آن جهت پراهمیت است که نشان دهنده رضایت استفاده از خدمات در یک تجربه واقعی و غیر انتزاعی بوده است.

۳-۳ پایتون

پروژه پیش‌رو با استفاده از زبان برنامه نویسی پایتون و در محیط Jupyter Notebook پیاده سازی شده است. علت استفاده از این محیط رابط گرافیکی بلاک بندی شده است که اجرای قطعه‌ای کد برنامه را راحت می‌کند همچنین پایتون با در اختیار قرار دادن کتابخانه های کاربردی نظیر پانداس[۲] و Scikit Learn مسیر را برای اجرا الگوریتم های یادگیری ماشین فراهم کرده است.

[۲] Pandas

۳-٤ ماشین بردار پشتیبان

الگوریتم ماشین بردار پشتیبان توسط ولادیمیر وپنیک ابداع شد و جزو الگوریتم های تشخیص الگو دسته بندی می‌شود. یکی از مزایای این الگوریتم آن است که برخلاف شبکه های عصبی در مسائل بهینه سازی در کمترین و بیشترین محلی گیر نخواهد افتاد.

یکی از روش های دسته‌بندی اطلاعات حجیم استفاده از ماشین بردار پشتیبان[۳] (SVM) است. در این الگوریتم که یکی از روش های طبقه بندی با نظارت است، بجای ایجاد یک جداکننده همانند خط برای تفکیک داده ها ابتدا یک بُعد به داده ها اضافه کرده و تفکیک سازی و تصمیم گیری را در بعد بالاتر انجام می‌دهد و حائل میان داده‌ای را از بعد اصلی داده ها انتخاب می‌کند تا بتواند در جداسازی داده ها با پیچیدگی بالا عملکرد بهتری داشته باشد. برای ایجاد خط جداساز، الگوریتم خطوط فرضی بسیاری را در نظر می‌گیرد سپس فاصله هر نقطه در فضا را با خط محاسبه نموده و به نسبت فاصله با خط جداساز داده‌ها را در دسته‌های متفاوت قرار می‌دهد. همچنین حداقل فاصله لازم جهت دسته بندی صحیح با مقدار گاما قابل تعیین است.

به طور مثال اگر داده های ما به صورت دو بعدی باشند یعنی داری یک مختصات افقی و عمودی (x,y) باشد یک بعد سومی را به تمامی داده ها اضافه کرده و جداکننده را از همان بعد دوم که صفحه می‌باشد انتخاب می‌کند تا بتواند داده ها را در دسته های مختلف دسته بندی کند. یا در مثالی دیگر داده‌ی یک بعدی را دو بعدی کرده و از خط، برای جداسازی استفاده می‌کند.

۳-٤-۱ انواع هسته‌های ماشین بردار پشتیبان

به جهت بردن داده ها به یک بعد بالاتر از فرم دوگانی آنها استفاده می‌شود که این فرایند با انتقال داده ها توسط یک بردار صورت می‌پذیرد. به انواع این بردار ها هسته[۴] ماشین بردار پشتیبان می‌گویند.

یکی از هسته‌های ماشین بردار پشتیبان هسته‌ی خطی[۵] می‌باشد که بردار مورد نظر برابر است با مجموع حاصل ضرب هر جفت مقادیر ورودی، که در فرمول زیر نیز قابل رویت است

---

[۳] Support Vector Machine

[۴] kernel

[۵] linear

$$k(x, x_i) = sum(x * x_i)$$

یکی دیگر از این هسته‌ها، کرنل چند جمله‌ای می‌باشد. در این هسته تمایل به جداسازی داده‌های غیرخطی و منحنی را دارد. d مقدار درجه چند جمله ای می‌باشد که می‌بایست در طول الگوریتم به صورت دستی تنظیم شود. ضمنا r نیز بیانگر میزان شعاع دایره انحنا جداکننده می‌باشد که عموما آن را برابر ۱ در نظر می‌گیرند

$$k(x, x_i) = sum(x * x_i)^d + r$$

و اما یکی از مهمترین هسته‌ها، کرنل تابع پایه شعاعی یا محور گاوسی[۶] می‌باشد. این کرنل فضای ورودی را در ابعادی نامشخص ترسیم میکند. بدین صورت که میزان فاصله با محور جداکننده را تحت عنوان گاما تعیین می‌کنیم و مجموع فاصله هر جفت مقدار را در گاما ضرب کرده و قرینه آن را به عنوان توان تابع نمایی در نظر می‌گیریم تا بردار مورد نظر بدست آید.

$$k(x, x_i) = exp(-gamma * sum(x - x_i)^2)$$

اینکه از کدام یک از روش ها استفاده کنیم بستگی به ابعاد و نوع داده های ورودی دارد که عموما برای داده های چند برچسبی سعی به استفاده از RBF است.

۳-۵ خوشه بندی K-MEANS

یکی از روش های آنالیز خوشه‌بندی در داده کاوی اطلاعات روش k-means می‌باشد. k-means خوشه بندی با هدف تجزیه n مشاهدات به k خوشه است که در آن هریک از مشاهدات متعلق به خوشه هایی با نزدیکترین میانگین آن است. روش‌ها و الگوریتم‌های متعددی برای تبدیل اشیاء به گروه‌های هم‌شکل یا مشابه وجود دارد اما k-means به دو جهت اهمیت فوق العاده ای نزد تحلیلگران داده دارد اول آنکه قابلیت تفکیک سازی با تعریف یک الگو ساده را فراهم کرده و قابلیت اجرا بر روی داده های چند بعدی را دارد مورد دوم هم به این اشاره دارد که در این الگوریتم قابلیت تعیین تعداد خوشه نهایی وجود دارد.

بدین صورت که تعداد دسته های مورد نظر(K) در ابتدا مشخص می‌شود سپس k نقطه در داده ها را به صورت شانسی برمی‌گزینند و به عنوان مرکز دسته انتخاب می‌کند سپس فاصله اقلیدسی هر یک از نقاط

[۶] RBF

را با k مرکز دسته قبلی محاسبه کرده و هر دسته ای که مرکز دسته آن فاصله کمتری با نقطه مورد نظر داشته باشد را به آن نقطه منتصب می‌کند سپس میانگین نقاط هر دسته به عنوان مرکز دسته بروزرسانی شده در نظر گرفته می‌شود و این فعالیت مجددا ادامه می‌یابد و تا جایی ادامه پیدا می‌کند که دیگر تغییر محسوسی در خوشه‌بندی‌ها رخ ندهد.

درجه پیچیدگی محاسباتی آن برابر با $O(n^{dk+1})$ است، که در آن d بعد ویژگی‌ها و k تعداد خوشه‌ها هستند. همچنین پیچیدگی زمانی برای این الگوریتم برابر با O(nkdi) است که منظور از i تعداد تکرارهای الگوریتم برای رسیدن به جواب بهینه است. در نتیجه با افزایش تعداد پارامترهای مسئله و تعداد دسته‌ها، محاسبات این الگوریتم به صورت نمایی رشد خواهد کرد.

۳-۶ خوشه‌بندی سلسله‌مراتبی

از دیگر روش های خوشه‌بندی که به روش سلسه‌مراتبی نیز معروف است بدین صورت عمل می‌کند که داده ها به صورت درختی طبقه بندی شده و برگ های هر درخت نشانگر یک خوشه می‌باشد. ضمنا این روش بیشتر برای داده هایی صورت می‌گیرد که تنوع مقداری زیادی نداشته و قابلیت تفکیک داشته باشند یعنی به عبارتی شامل مقادیر گسسته باشند.

برای اجرای این روش دو حالت تجمعی و تجزیه ای موجود است. تمجمعی به این معنی که رویکرد این دسته پایین به بالا بوده و از جزئی ترین پارامترها شروع کرده و با خوشه های همسان خود تشکیل یک خوشه جدید را می‌دهد تا در نهایت به یک ریشه واحد برسند. اما در تجزیه‌ای، رویکرد بالا به پایین است و همچون یک درخت تجزیه از کلی‌ترین حالت شروع کرده و براساس فاکتورهای مختلف نسبت به تقسیم و خوشه بندی داده ها می‌پردازد.

اینکه داده ها بر چه اساس تقسیم بشوند و یا در یک گروه قرار بگیرند به تابع اندازه‌گیری تعریف شده بستگی دارد که به صورت عمومی از فاصله های اقلیدسی یا منهتن استفاده می‌شود.

روش پیوند[۷] هر خوشه با خوشه دیگر به جهت ایجاد درخت سلسه‌مراتبی حالات مختلفی دارد. یکی از آنها پیوند کامل[۸] است به این معنی که بیشترین فاصله موجود میان یک نقطه در هرخوشه با خوشه دیگر را پیدا می‌کند و در نهایت هر کدام از حالات که مقدار کمتری داشته باشد به یکدیگر پیوند خورده و خوشه جدید را تولید می‌کند(مطابق روش پایین به بالا). در پیوند میانگین نیز همین اتفاق میافتد با این تفاوت که فاصله مرکز هر دسته معیار است.

پیچیدگی زمانی این الگوریتم $O(n^3)$ بوده و پیچیدگی حافظه‌ای آن $O(n^2)$ است. بنابراین با افزایش حجم داده ها سرعت و حافظه برای اجرای عملیات به شدت افزایش پیدا می‌کند پس روش مناسبی برای کلان داده‌ها[۹] نخواهد بود.

۳-۷ ضریب همبستگی

این معیار نمایانگر میزان شباهت یک مجموعه متغیر به متغیر دیگری می‌باشد و ارتباط خطی بین آن دو را اندازه گیری می‌کند و تحت عنوان ضریب همبستگی[۱۰] شناخته می‌شود. این معیار به این علت بوجود آمد چراکه کوواریانس داده ها داری واحد اندازه‌گیری بوده و معیاری مناسب جهت مقایسه دو گروه نبوده است در نتیجه ضریب همبستگی که مقداری بدون واحد است بوجود آمد. برای بدست آوردن این پارامتر از امید ریاضی آنان استفاده می‌کنیم که به طور خلاصه تر از از تقسیم کوواریانس بر جذر ضرب واریانس هر دو پارامتر بدست می‌آید.

$$Corr(X,Y) = \frac{cov(x,y)}{\sqrt{|var(x) \times var(y)|}}$$

۳-۸ کتابخانه های مورد استفاده

در این پروژه از برخی توابع و کتابخانه های از پیش آماده زبان برنامه نویسی پایتون استفاده شده است که در این بخش به توضیح مختصری از مهمترین آنها خواهیم پرداخت:

---

[۷] linkage

[۸] complete

[۹] Big data

[۱۰] Correlation coefficient

Numpy: یکی از مهمترین کتابخانه‌های مورد استفاده که برای کار با داده ها به صورت آرایه ای و عملیات بر روی آنان استفاده می‌شود.

Pandas: یکی دیگر از مهمترین کتابخانه‌های استفاده شده در پروژه جهت تغییر داده ها به صورت دیتافریم های سطری و ستونی

Matplotlib: کتابخانه‌ای جهت ترسیم نمودار

Sklearn: کتابخانه ای حاوی توابع مفیدی در جهت اجرای الگوریتم های یادگیری ماشین نظیر ماشین بردار پشتیبان و کا-میانگین و محاسبه معیار کاپا و ...

Itertools: برای ساخت جایگشتی از آرایه ها

Scipy: از دیگر کتابخانه‌های یادگیری ماشین که در این پروژه برای اجرا خوشه‌بندی سلسله‌مراتبی مورد استفاده قرار گرفته است.

۳-۹ فاز آموزش و تست

در این نوع دسته بندی انتظار می‌رود که چنانچه داده جدیدی وارد سیستم شود بر اساس اطلاعات قبلی موجود، نسبت به پیشبینی پارامتری از داده جدید اقدام شود. بدین منظور بخشی از داده‌های چند بعدی به عنوان ورودی (X) و یک داده به عنوان خروجی (Y) در نظر گرفته می‌شود که بیانگر آن است که به ازای هر داده در بین چند هزار داده با مقادیر X به مقدار Y دست پیدا کرده‌ایم تا الگویی باشد برای آنکه اگر ورودی جدیدی (X_{new}) وارد سیستم شد بتوانیم خروجی جدید (Y_{new}) را پیشبینی کنیم. به عبارت ساده‌تر $X_{معلوم}$ منجر به $Y_{معلوم}$ می‌شود و به عنوان الگوی پیشبینی در نظر گرفته می‌شود حال چنانچه $X_{معلوم}$ جدیدی وارد سیستم شود که Y آن مجهول است نسبت به کشف Y جدید براساس الگوی تعریف شده اقدام می‌شود.

به جهت سنجش درستی عملکرد این پیشبینی، داده های موجود را به دو دسته آموزش[۱۱] و آزمون[۱۲] تقسیم می‌کنیم بدین صورت که به صورت رندوم ۳۰ درصد از داده‌ها در بخش تست قرار می‌گیرد. در این شرایط ورودی‌های ما به x_train و y_train و x_test و y_test تقسیم می‌شود که x_train و y_train به عنوان

[۱۱] train

[۱۲] test

معلومات مسئله الگوی پیشبینی را می‌سازند سپس با توجه به x_test اقدام به پیشبینی y_test کرده و اسم آن را y_hat قرار می‌دهیم.

حال هرچه تطابق Y_hat که حاصل پیشبینی ما است و Y_test که نتیجه واقعی داده های ما که از پیش معلوم بوده است، بیشتر باشد الگوریتم پیشبینی ما بهتر عمل کرده است و در داده های مجهول آینده نیز بهتر پیشبینی خواهد کرد. این میزان تطابق بر اساس ماتریس اغتشاش قابل اندازه گیری است.

۳-۱۰ معیار های ارزیابی

ارزیابی این پروژه می‌تواند اینگونه تعریف شود که پیشبینی نزدیک تر به واقعیت، اعتبار الگوریتم را نشان دهد به عبارتی پیشبینی هرچه دقیق‌تر و صحیح‌تر پس عملکرد الگوریتم بهتر بدین ترتیب برای سنجش این نزدیکی می‌توان از ماتریس اغتشاش بهره برد.

۳-۱۰-۱ ماتریس اغتشاش

این ماتریس دو بعدی بوده و در سطر آن برچسبی از کلاس داده‌های موجود و در بخش ستونی پیشبینی صورت گرفته خواهد بود. به طور مثال اگر موردی در داده های معلوم مسئله مقداری منتصب به کلاس دوم دارد اما در پیشبینی کلاس سوم تشخیص داده شده است. به سطر دوم ستون سوم از ماتریس، یک واحد اضافه می‌شود و این اتفاق برای تمامی داده‌های بخش تست انجام می‌شود تا ماتریس مذکور تکمیل گردد. ذکر این نکته نیز ضروری است که پر کردن این ماتریس بر اساس فاز آموزش و تستی است که در عناوین بالاتر به شرح آن پرداخته شد.

به جهت تحلیل این ماتریس نیاز به استخراج چهار پارامتر برای استفاده در فرمول های معیار هایمان نیاز داریم. برای داده هایی با یک برچسب به جهت آنکه ماتریس حالت دو بعدی خود را حفظ کند این برچسب را در دو حالت بلی یا خیر تقسیم می‌کنیم برای مثال اگر پارامتر بیماری مورد مطالعه باشد بیمار باشد یا نباشد دو حالت مد نظر خوهد بود.

همین قضیه اما برای داده های چند برچسبی برقرار است ولی دیگر نیاز به تعریف بله یا خیر نیست چرا که سطر و ستون های ماتریس حاصله همان برچسب ها خواهند بود که قطر اصلی ماتریس همان پیشبینی

دقیق خواهد بود. البته که برای هر برچسب می‌توان جداگانه ماتریس ایجاد کرد و دقت کل برابر دقت میانگین همه ماتریس ها است.

جدول ۳-۱: نمای ماتریس اغتشاش

	پیشبینی توسط الگوریتم		
		بلی	خیر
برچسب واقعی	بلی	TP	FN
	خیر	FP	TN

۳-۱۰-۲ معیار های سنجش

برای سنجش عملکرد الگوریتم و ماتریس آن معیاری تحت عنوان دقت[۱۳] وجود دارد که از رابطه زیر بدست می‌آید:

True positive: حالتی که افراد بیمار باشند و الگوریتم نیز بگوید بیمار هستند.

False negative: حالتی که اشخاص بیمار باشند اما الگوریتم بگوید بیمار نیستند.

True negative: حالتی که افراد بیمار نباشند و الگوریتم نیز بگوید بیمار نیستند.

False positive: حالتی که افراد بیمار نباشند و الگوریتم بگوید بیمار هستند.

$$\text{دقت} = \frac{TP + TN}{TP + FN + TN + FP}$$

همانطور که مشاهده می‌کنید در وهله اول TP مهمترین پارامتر است که اهمیت دارد در وهله دوم افزایش مقدار قطر ماتریس پر اهمیت است چراکه پیشبینی دقیق و درست را بیان می‌کند.

[۱۳] accuracy

اما پارامتر دقت یک اشکال بزرگ دارد و تفاوتی میان FN و FP قائل نیست. FN نیز بسیار با اهمیت است به این علت که زیاد بودن این مقدار ضربه مهلکی به الگوریتم در مثال واقعی می‌زند هرچند که در قطر اصلی عملکرد عالی داشته باشد. برای مثال به همان نمونه بیمار اشاره میکنم که اگر فردی بیمار باشد و بیمار تشخیص ندهد عملا الگوریتم هیچکارایی ندارد چراکه به یک بیمار گفته است بیماری نداری.

به همین منظور پارامتر دیگری به اسم پوشش[۱۴] یا حساسیت را خواهیم داشت که بیانگر پوشش برچسب بلی در الگوریتم است. و در کنار دقت معنا پیدا می‌کند یعنی اگر پوشش به میزان قابل توجهی بود می‌توان به دقت اعتماد کرد.

$$\text{پوشش} = \frac{TP}{TP + FN}$$

اما در داده های چند برچسبی ضمن محاسبه میانگین دقت‌ها معیاری که دقیق تر بتواند عملکرد الگوریتم چند برچسبی را بیان کند وجود دارد. این معیار تحت عنوان معیار کاپا معروف است. این معیار علاوه بر بیان دقت در داده‌های چندبرچسبی، میزان شانس دخیل در نتیجه پیشبینی را نیز می‌سنجد.

$$\text{معیار کاپا} = \frac{\text{توافق} - \text{شانس توافق}}{1 - \text{شانس توافق}}$$

توافق[۱۵] احتمال پیشبینی دقیق و درست را بیان می‌کند به عبارتی همان احتمال قطر اصلی است. شانس توافق[۱۶] درصد شانسی بودن پیشبینی درست را می‌سنجد و حاصل مجموع احتمال هر برچسب است. ضمنا احتمال هر برچسب از ضرب احتمال واقعی هر برچسب در احتمال پیشبینی همان برچسب بدست می‌آید.

به جهت تحلیل آماری عموما از محاسبه دقت و امتیاز-f1 [۱۷] استفاده می‌شود اما با توجه به آنکه داده های ما به صورت چند برچسبی[۱۸] هستند از معیار کاپا[۱۹] به جهت سنجش بهره می‌بریم.

---

[۱۴] recall
[۱۵] agreement
[۱۶] chance agree
[۱۷] f۱-score
[۱۸] Multi Labels
[۱۹] Kappa score

مطابق استاندارد، عملکرد الگوریتم بر اساس بازه قرارگیری معیار کاپا سنجیده می‌شود:

معیار کاپا کمتر از ۰: "خیلی بد"

معیار کاپا بین ۰ تا ۲۰: "بد"

معیار کاپا بین ۲۰ تا ۶۰: "متوسط"

معیار کاپا بین ۶۰ تا ۸۰: "خوب"

معیار کاپا بین ۸۰ تا ۱۰۰: "عالی"

۳-۱۱ نتیجه‌گیری

امتیازدهی امری ساده نبوده و تحلیل آن نیز سهل و آسان نخواهد بود. برای یک نتیجه مناسب حتما باید از ابزارهای متنوعی استفاده کرد تا ارزیابی درستی داشته باشیم همانطور که در بخش های قبل دیدیم خوشه‌بندی اطلاعات به تنهایی کافی نیست و باید نتایج آن را در مرحله دسته بندی و به موجب آن در فاز آموزش و تست استفاده کنیم تا نتایج در کنار یکدیگر قابل اعتماد شوند. نمونه ی دیگری که بر این امر صحه می‌گذارد بی ارزش بودن معیار دقت بدون حضور معیار پوشش بود یا معیار دقت به تنهایی قابلیت نمایش نتیجه را در مسائل چند برچسبی به تنهایی ندارد.

یک مسئله می‌تواند از زوایای متعددی مورد بررسی قرار بگیرد همانطور که در پژوهشی به اثر تنوع غذایی با توجه به موقعیت جغرافیایی پرداخته و ما نیز در این پروژه به امتیازدهی بدون توجه به موقعیت جغرافیایی می‌پردازیم و از پارامتر های دیگری استفاده می‌کنیم. پس برای یک نتیجه‌گیری درست علاوه بر استفاده از ابزار های متنوع باید از بررسی در موضوعات متنوع و مقایسه آنان نیز استفاده کرد.

فصل ۴: روش پیشنهادی

٤-١ مقدمه

در این فصل از گزارش به بررسی فعالیت های صورت گرفته در طول پروژه خواهیم پرداخت.

پس از استخراج اطلاعات و تبدیل آنان به قالب‌های کتابخانه‌هایی نظیر پانداس و نامپای، چهار ویژگی مربوط به بخش ارسال و سفارش آنلاین را در یک گروه و چهار ویژگی دیگر را در گروه دیگر قرار دادیم و بوسیله هر یک از این گروه ها اقدام به خوشه بندی اطلاعات نمودم. نتیجه حاصله که هر یک امتیازی از یک تا پنج به این دو بخش را شامل می‌شد به عنوان ورودی دسته‌بندی به روش ماشین بردار پشتیبان در نظر گرفته تا با داده‌های در دسترس، سعی بر ایجاد الگویی برای پیشبینی امتیاز کاربران برای رستوران های آتی شود. شرح بیشتر موارد فوق در ذیل این فصل مورد بررسی خواهد بود.

٤-٢ داده ها

٤-٢-١ منبع اطلاعات

این پروژه از داده های مجموعه رستوران‌های وبسایت زوماتو[۲۰] استفاده کرده است و در وبسایت Kaggle.com [۴] نیز موجود است که شامل اطلاعات مختلفی از رستوران ها نظیر امکان تحویل غذا، تنوع غذایی، میانگین قیمتی غذا، تعداد آرا کاربران، تصاویر منو، لینک مستقیم سفارش، آدرس، کشور، شهر، مختصات و .... می‌باشد.

٤-٢-٢ نوع داده ها

داده ها در پنج فایل JSON که هر یک شامل حدودا ۷۵ رستوران می‌شود قرار دارد و یک فایل CSV که داری اطلاعات کلی رستوران مثل آدرس آن می‌باشد. رابط میان این دو جدول id هر یک از رستوران هاست.

[۲۰] ZOMATO

٤-۲-۳ استخراج داده ها

بجز داده تنوع غذایی[۲۱] و نوع ارز[۲۲] سایر مقادیر مورد استفاده به صورت عددی حاضر بوده است که با استفاده از یک حلقه و ترکیب‌سازی هر پنج فایل در یک فایل، اطلاعات لازم استخراج شد. اما برای متغیر تنوع غذایی تعداد انواع را کشف و به جدول داده ها افزوده شد.

استخراج داده‌ها به این علت انجام پذیرفت چراکه اولا تمامی اطلاعات درون دیتاست مورد استفاده پروژه نبوده و صرفا ۱۱ پارامتر جهت بررسی استخراج شده است ضمنا داده ها به صورت json های تو در تو در دسترس بوده‌اند که به علت بهره‌گیری از کتابخانه پانداس جهت تحلیل بر داده ها باید تبدیلی به فایل های CSV که به صورت جداول سطری و ستونی اند صورت میگرفت. داده های استفاده شده به شرح ذیل می‌باشد:

جدول ۴-۱ : متغیر های استفاده شده از مجموعه داده ها

نام ویژگی	شرح	نوع
Id	شماره اختصاصی هر رستوران	عدد
Has online delivery	سفارش آنلاین	صفر و یک[۲۳]
Has table booking	جدول رزرو آنلاین	صفر و یک
Is delivering now	قابلیت ارسال غذا	صفر و یک
Switch to order menu	موجود بودن منو غذایی	صفر و یک
Cuisines	تنوع غذایی	رشته[۲۴]
Average cost for two	هزینه متوسط برای دو نفر	عدد
Price range	سطح قیمتی	عدد
Votes	تعداد نظرات	عدد
Aggregate rating	میانگین رای کاربران	عدد اعشاری
Currency	نوع ارز	رشته

[۲۱] cuisines
[۲۲] currency
[۲۳] boolean
[۲۴] string

٤-۲-٤ خلاصه وضعیت

داده های استخراج شده شامل اطلاعات ۵۷۳۳ رستوران بوده که میانگین امتیاز کاربران در آن به مقدار ۳.۴۴ و کمترین امتیاز متعلق به رستورانی با امتیاز ۰ و بیشترین نیز ۴.۹ بوده است ضمنا بیشترین تنوع غذایی با هشت نوع غذا و کمترین هم تنها با یک نوع است همچنین باید خاطر نشان کرد که بیشترین تعداد رای متعلق به رستورانی با حدود یازده هزار رای می‌باشد.

نکته قابل توجه این است که حدود ۴۰۲۴ رستوران فاقد بخش تحویل غذا درب منزل[۲۵] می‌باشد و عملا ارسال غذا و سفارش آنلاین در این رستوران ها خدمت‌رسانی نمی‌شود که این تعداد کثیر میتواند برای نتیجه‌گیری پروژه چالش برانگیز باشد.

٤-۲-٥ تغییرات در داده ها

یکسری از داده های استخراجی به صورت اولیه به دلیل پراکندگی و غیر استاندارد بودن قابل استفاده نبوده است.

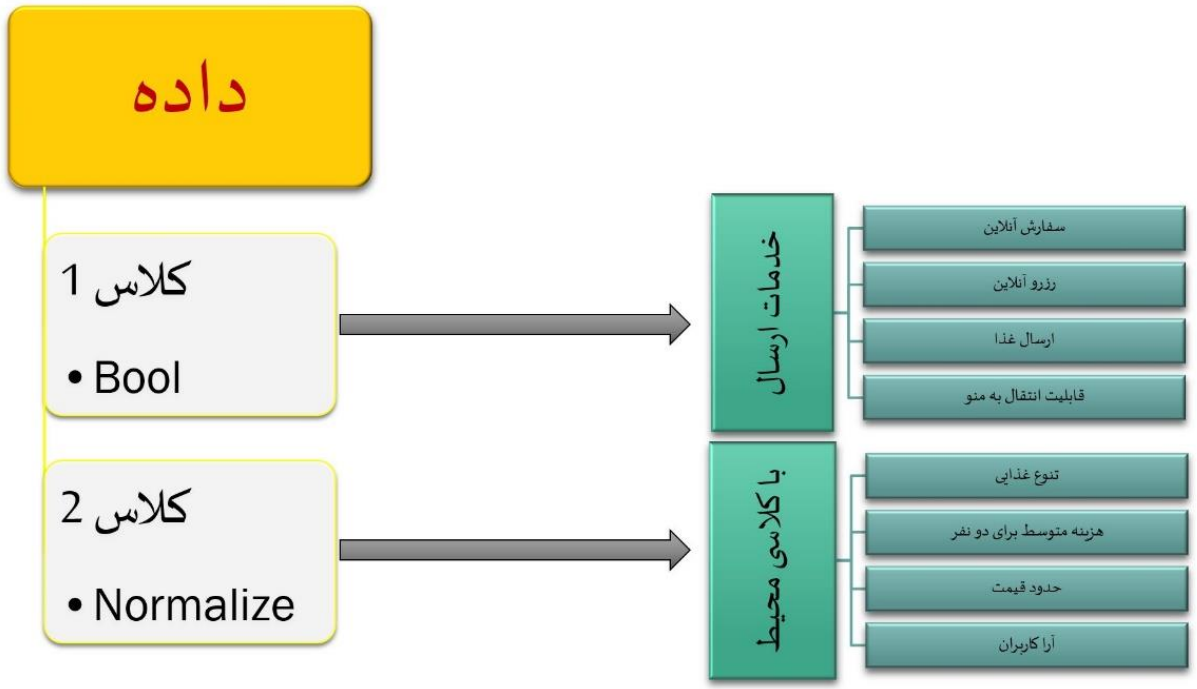
اولین تغییر در داده ها متعلق به متغیر votes می‌باشد که به علت پراکندگی زیاد و بزرگ بودن مقدار، تعداد آرا هریک را تقسیم بر بزرگترین تعداد رای موجود می‌کنیم

دومین مورد مربوط به متغیر average_cost_for_two است چراکه قیمت اعلامی در داده ها با واحد ارزی کشور محل رستوران بیان شده و به جهت قیاس قابل مقایسه نیستند. پس این ستون از داده ها براساس متغیر currency به نرخ دلار تبدیل گشته است.

٤-۲-٦ تقسیم بندی داده ها

همانطور که پیشتر نیز گفته بودیم داده ها در دو دسته "محیط رستوران" و "خدمات تحویل غذا" تقسیم می‌شود.

[۲۵] delivery



شکل ۴-۱: تقسیم بندی داده ها

۴-۳ نرمالسازی

داده های دسته اول که مرتبط با خدمات تحویل غذا است باتوجه به اینکه داری مقادیر صفر و یک می‌باشد نیاز به تغییر ندارد اما داده‌های دسته دوم که مرتبط با محیط رستوران است در مقیاس‌های گوناگونی اند.

برای هم مقیاس سازی داده های دسته دوم باید آنها را نرمالسازی کرد که داده ها با توجه به فرمول زیر و تحت توزیع نرمال مقادیری بین صفر و یک به خود بگیرند.

$$x_{std} = \frac{x\ - x.min}{x.\max - x.min} * (\max - \min) + min$$

٤-۳-۱ گرد کردن میانگین رای کاربران

به علت استفاده از مقدار عبارت aggregate_rating به عنوان ناظر دسته بندی و متغیری که نتیجه گیری و تصمیمات بر اساس آن گرفته می‌شود و همچنین تعریف این مقدار بر اساس استاندارد پنج ستاره[۲۶] که در آن امتیاز کاربران به صورت ستاره ای از یک تا پنج تعریف می‌شود مقدار متغیر aggregate_rating که داری مقداری اعشاری است باید تبدیل به عدد صحیحی بین ۱ تا ۵ گردد.

بدین ترتیب تنها راه تبدیل این پارامتر گرد کردن این اعداد به نزدیک ترین همسایه خود می‌باشد ضمن اینکه مقدار این متغیر در دیتا ما نیز در بین یک تا پنج می‌باشد. اما باتوجه به آنکه این نمرات امتیاز کاربران به رستوران‌ها می‌باشد و مقادیر بسیاری بین ۳.۵ و ۴.۵ خواهد بود اما در عمل تفاوت کیفیت نیز در همین مقادیر رقم می‌خورد و در مصداق واقعی که مربوط به رستوران‌ها می‌باشد تفاوت نمره در این بازه وجود دارد. گردسازی به روش نزدیک‌ترین همسایه باعث گرد شدن بسیاری از مقادیر به عدد ۴ می‌شود و عملا دسته بندی داده ها معنا پیدا نمی‌کند پس طبق تست و بررسی با مقیاس یک‌دهم که مطابق شرایط واقعی و عقلانی پیرامون امتیاز به یک رستوران است مطابق بازه های زیر دست به گردسازی این پارامتر زده شد تا در ادامه کار نتیجه بهتر و واقعی تر نصیبمان شود.

اگر امتیاز کاربران ≤ ۴.۵ آنگاه "۵"

اگر ۴.۵ ≤ امتیاز کاربران ≤ ۳.۸ آنگاه "۴"

اگر ۳.۸ ≤ امتیاز کاربران ≤ ۲.۸ آنگاه "۳"

اگر ۲.۸ ≤ امتیاز کاربران ≤ ۱.۸ آنگاه "۲"

اگر ۱.۸ ≤ امتیاز کاربران آنگاه "۱"

[۲۶] yelp

در شکل زیر گوشه‌ای از چارچوب داده های مورد استفاده در پروژه پس از استاندارد سازی داده‌ها قابل رویت می‌باشد.

has_online_delivery	has_table_booking	is_delivering_now	switch_to_order_menu	aggregate_rating	cuisines	average_cost_for_two	price_range	votes
1.0	1.0	0.0	0.0	4	0.428571	0.002000	0.666667	0.725352
0.0	1.0	0.0	0.0	5	0.714286	0.001875	0.666667	0.071154
1.0	0.0	0.0	0.0	4	0.428571	0.001063	0.333333	0.140571
0.0	1.0	0.0	0.0	4	0.142857	0.002313	0.666667	0.166728
0.0	1.0	0.0	0.0	4	0.428571	0.002000	0.666667	0.076825

شکل ۴-۲: گوشه ای از دیتافریم داده ها

٤-٤ خوشه بندی[۲۷]

٤-٤-١ خوشه اول به روش سلسله‌مراتبی

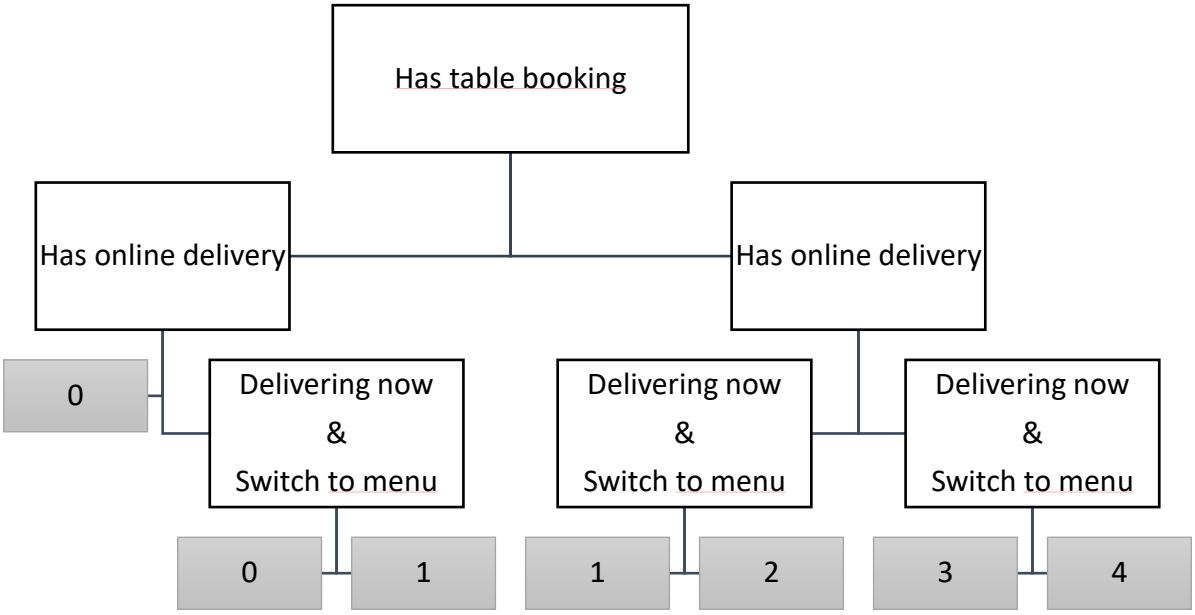
از آنجا که مقادیر خوشه اول به صورت صفر و یک می‌باشد برای تفکیک و درک بهتر خوشه‌بندی از روش سلسله‌مراتبی با شمای درختی بهره خواهیم گرفت.

پس از بررسی ضریب همبستگی هر یک از چهار پارامتر این دسته با مقدار aggregate_rating متوجه می‌شویم بیشترین ارتباط برای متغیر has_table_booking و سپس با has_online_delivery بوده و با دو متغیر دیگر ارتباط بسیار کمی وجود دارد پس ریشه درخت را در دوحالت وجود داشتن یا نداشتن متغیر has_table_booking در نظر می‌گیریم.

به جهت خروجی منطقی‌تر شرایطی را جهت نمره‌دهی در نظر می‌گیریم بدین صورت که وزن وجود یا عدم وجود متغیرهای has_table_booking و has_online_delivery بیشترین تاثیر را داشته باشند و دو متغیر دیگر یعنی is_delivering_now و switch_to_order_menu را به صورت ترکیبی و باهم در نظر می‌گیریم بدین صورت که اگر هر دو مقدار یک را داشتند یک و اگر یکی از آنها صفر باشد صفر را منظور خواهیم‌کرد.

[۲۷] Clustering

سلسله مراتب در نظر گرفته شده در شکل ۴-۳ نمایانگر نحوه عملکرد درخت می‌باشد با این توضیح که سمت چپ عدم وجود و سمت راست یک بودن پارامتر را بیان می‌کند همچنین پس از آن مقادیر بعلاوه یک خواهندشد تا در مقیاس yelp معنا داشته باشد.



شکل ۴-۳: نحوه خوشه‌بندی سلسله‌مراتبی کلاس یک

یکی از پارامترهای موثر بر پیشبینی ما بخش تحویل غذاست اما مقادیر این بخش همبستگی کمی با امتیاز کاربران دارد به همین علت برای بهبود این همبستگی از آرایه‌ای ترکیبی از نتایج بدست آمده در مرحله قبل و متغیر امتیاز کاربران ساخته و با این داده دو بعدی جدید الگوریتم سلسله‌مراتبی[۲۸] را با پیوند[۲۹] میانگین و تعداد دسته ۵ اجرا می‌کنیم. پیوند میانگین بدین معنی است که در هر دفعه از اجرای الگوریتم، فاصله هر نقطه با میانگین نقاط خوشه در هر نوبت مقایسه شود. ضمنا درخت حاصله داری تعداد زیادی شاخه می‌باشد و به جهت نیاز پروژه حاضر به تنها پنج کلاس، از تابع Fcluster به جهت قطع نمودن درخت از ارتفاعی مشخص به نحوی که تعداد ۵ برگ از درخت سلسه‌مراتبی بوجودآید اقدام می‌شود.

[۲۸] hierarchy

[۲۹] Linkage

اجرای سلسله‌مراتبی شاید در بهبود عملکرد خوشه‌بندی کمک کرده باشد اما در بخشی از این خوشه بندی اتفاقی غیر منطقی رخ می‌دهد و آن اینکه احتمال نسبت دادن عددی جز یک به مجموعه‌های فاقد تحویل غذا درب منزل، وجود دارد پس مجددا مقدار برچسب این رستوران ها که طی اجرا الگوریتم به اشتباه برچسب گذاری شده را یک می‌کنیم. نتیجه حاصل شده برای بخش تحویل غذا، ضریب همبستگی را تا پنج درصد بهبود داده است.

نکته جالب توجه اینجاست که این نرخ همبستگی در حدود بیست درصد می‌باشد در حالی که با حذف مجموعه های فاقد تحویل غذا درب منزل با نرخ همبستگی عالی‌ای به نزدیکی ۹۰٪ دست پیدا می‌کنیم که بیانگر آن است که نبود بخش تحویل غذا در امتیاز کاربران چندان موثر نیست

٤-٤-٢ خوشه دوم به روش K-MEANS

در خوشه دوم اما از روش k-means که پیشتر توضیح دادیم استفاده میکنیم. همانطور که میدانیم داده های مورد استفاده در مراحل قبل نرمالسازی شده بودند. در این الگوریتم از حالت k-means++ به جهت انتخاب اولیه از بین نقاط موجود و نه رندوم بهره گرفته شده و تعداد دسته ها نیز ۵ عدد در نظر گرفته‌شده است همچنین تعداد دفعات تکرار الگوریتم بر روی مقدار بیست بار تنظیم خواهد شد.

٤-٤-٣ تصحیح برچسب گذاری

از آنجایی که احتما دارد برچسب گذاری ما بر روی هر دسته به صورت مناسب انجام نشده باشد و همچنین در الگوریتم k-means توجهی به ارزش هر دسته نمی‌شود و صرفا دسته ها را از صفر تا چهار دسته می‌کند. به طور مثال ارزش دسته ها به صورت [۵,۴,۱,۳,۲] است اما به صورت [۵,۴,۳,۲,۱] نماد گذاری شده است، تمام حالت های موجود را با میزان همبستگی با متغیر امتیاز کاربران بررسی کرده و بهترین حالت برچسب گذاری را برای هر خوشه برمی‌گزینیم که باتوجه به ۵ دسته بودن مقادیر، جایگشت آن تعداد ۱۲۰ حالت را برای بررسی در نظر می‌گیرد. تا بدین ترتیب رستوران با محیط بهتر و سرویس خدمات ارسال و سفارش آنلاین بهتر برچسب بالاتری را به خود اختصاص دهد.

٤-٤-٤ تعیین ویژگی رستوران ها

بر اساس خوشه بندی صورت گرفته در مرحله قبل چنانچه خوشه‌ای بیش از نیمی از نمره را یعنی عدد ۳ را کسب کند داری آن ویژگی خواهد بود و مطابق موارد زیر یک ستون به دیتافریم خود میافزاییم.

اگر (کلاس اول ≤ ۳) و (کلاس دوم ≤ ۳) آنگاه **"تحویل غذا درب منزل و محیط باکلاس"**

اگر (کلاس اول ≤ ۳) و (کلاس دوم < ۳) آنگاه **"تحویل غذا درب منزل "**

اگر (کلاس اول < ۳) و (کلاس دوم < ۳) آنگاه **" محیط باکلاس"**

اگر (کلاس اول < ۳) و (کلاس دوم ≤ ۳) آنگاه **" بی‌کلاس و بدون تحویل غذا درب منزل "**

٤-٥ ترکیب دیتاست ها

در ابتدا گفته شد که دیتاست دیگری شامل مشخصات رستوران‌ها نیز وجود دارد حالا با ترکیب آن دسته از اطلاعات با ویژگی جدید افزوده شده به داده‌های ما می‌توانیم نمایشی داشته‌باشیم که تعیین می‌کند هر رستوران ویژگی باکلاسی یا بی‌کلاسی و یا خدمات سفارش آنلاین و ارسال غذا یا عدم آن را داری هستند یا خیر.

٤-٦ افزایش نمونه برداری[۳۰]

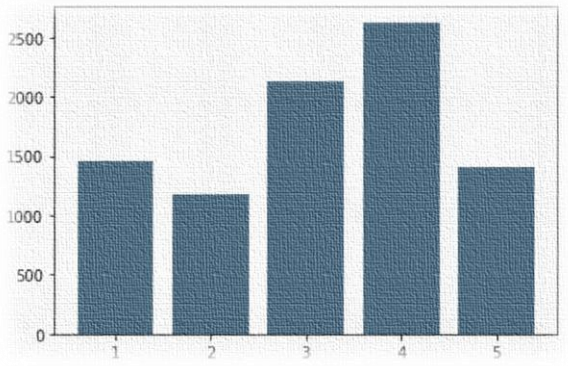
با توجه به آنکه بیشتر داده ها توزیع شدیدی روی مقدار ۴ از پنج را داری هستند تصمیم‌گیری را دچار خطا و اشتباه می‌کند. پس به این علت بر آن آمدم تا داده‌هایی با سایر مقادیر را در جدول داده ها افزایش داده تا نمونه‌های بیشتری از آنها داشته باشیم

Over Sampling در این پروژه با استفاده از تابع SMOTE صورت گرفته و صرفا داده ها را بر اساس استراتژی تعریف شده برای آن تولید مجدد کرده است. این استراتژی را به این صورت تعریف کرده‌ایم که تعداد پارامتر بیشترین(۴) ثابت بماند و تعداد کمترین پارامتر(۲) به میزان ۴۵ درصد بیشترین باشد. سپس تعداد

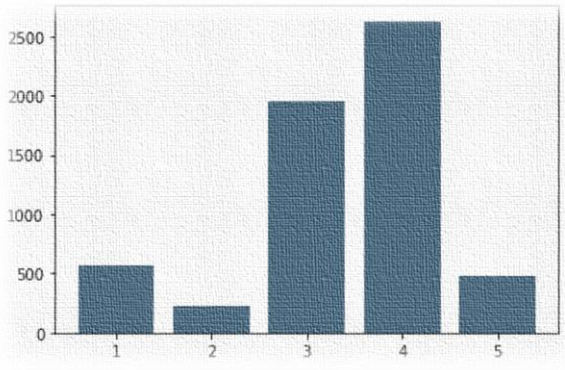
[۳۰] Over Sampling

سایر پارامتر ها(۱و۳و۵) بر اساس نسبت خود به کمترین مقدار تعیین می‌شود تا همچنان تعداد داده‌هایی که از دیگری بیشتر بوده بیشتر بماند و تا حدودی نسبت آن حفظ گردد برای مثال اگر پارامتر چهار ۱۰۰۰ تا و دو ۵۰ تا و یک ۱۰۰ تا آنگاه تعداد پارامتر دو به تعداد چهل‌وپنج درصد بیشترین یعنی ۴۵۰ تا می‌رسد. سپس چون ۵۰ نصف ۱۰۰ است به میزان نصف فاصله دو و چهار که نصف ۵۵۰ یعنی ۲۷۵ تا بیشتر از دو تعداد پارامتر یک می‌شود که برابر است با ۷۲۵تا و به همین ترتیب برای برچسب سه و پنج نمونه‌ها را افزایش می‌دهیم.

شکل زیر نمایانگر تعداد داده‌ها براساس Aggregate Rating است که تفاوت تعداد قبل و بعد از افزایش نمونه برداری را نمایش می‌دهد



شکل ۴-۴: پراکندگی داده ها بعد از افزایش نمونه برداری بر اساس پنج گروه امتیاز کاربران



شکل ۴-۵: پراکندگی داده ها قبل از افزایش نمونه برداری بر اساس پنج گروه امتیاز کاربران

٤-۷ پیشبینی با ماشین بردار پشتیبان

٤-۷-۱ *نحوه اجرا SVM*

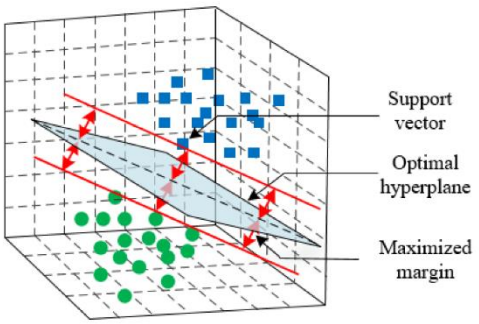
در پیشتر نیز گفتیم که دسته بندی به روش جداسازی برداری به اینگونه است که یک بُعد به داده ها می‌افزاید یعنی چنانچه داده‌ها n بعدی باشند داده‌ها را در n+1 بعد قرار داده و با حائلی از بعد nام نسبت به جداسازی اقدام می‌کند.

این متد تحت کرنل‌های متنوعی که کمی در اصول و فرمول های خود متفاوت‌اند قابل پیاده‌سازی است. از جمله این هسته‌های دسته‌بندی می‌توان به خطی، چند جمله‌ای، بر پایه محور گاوسی[۳۱] (RBF) و سیگموئید اشاره کرد. در بخشی از اجرای پروژه تحت یک حقله سه نوع از کرنل ها بر روی داده‌ها اجرا شدند تا هسته ای برای ادامه مسیر انتخاب شود که نتیجه بهتری را دربر دارد. که مطابق این نتیجه از روش هسته RBF استفاده خواهد شد. همچنین بیشترین فاصله مجاز با خط حائل به جهت جداسازی با پارامتری تحت عنوان گاما[۳۲] با مقدار ۱۰ تعیین شده است. در زیر تفاوت عملکردی هسته های مختلف دسته بندی آورده شده است:

جدول ۴-۲: جدول مقایسه هسته های ماشین بردار پشتیبان

خطی[۳۳]:	۰.۵۴
چند جمله ای[۳۴]:	۰.۵۸
محور گاوسی:	۰.۷۰

برای اجرا دسته‌بندی به روش ماشین بردار پشتیبان از تابعی به نام SVC که از مشتقات کتابخانه SVM است کمک می‌گیرد. داده هایی که باید بر روی این تابع به عنوان ورودی مناسب‌سازی[۳۵] بشوند تحت عنوان train و test شناخته می‌شوند.



شکل ۴-۶ : شِمای چگونگی دسته بندی روش ماشین بردار پشتیبان

[۳۱] RBF
[۳۲] gamma
[۳۳] Linear
[۳۴] Polynomial
[۳۵] fit

٤-۷-۲ پیشبینی امتیاز رستوران

از آن جهت که در بالاتر به آن پی بردیم، رستوران‌هایی که فاقد تحویل غذا درب منزل هستند اثری بر روی امتیاز کاربران ندارد برآن شدیم که پیشبینی رستوران تازه وارد شده به سیستم را از دو راه استخراج کنیم.

۱) استفاده از ماشین بردار پشتیبان برای رستوران های دارای تحویل غذا درب منزل; که نتایج خوشه بندی که در مراحل قبل انجام شد، در بخش خدمات ارسال و سفارش آنلاین به عنوان کلاس اول[۳۶] و محیط رستوران به عنوان کلاس دوم[۳۷] در نظر گرفته شده و تشکیل آرایه دو بعدی را می‌دهند. که این آرایه به عنوان ورودی سیستم، و پارامتر Aggregate Rating به عنوان خروجی سیستم، تابع SVM را قابل اجرا می‌کنند و منجر به ایجاد ماتریس اغتشاش[۳۸] نیز می‌شود.

۲) برای رستوران های فاقد تحویل غذا درب منزل صرفا از classifier.2 برای پیشبینی امتیاز کاربران بهره می‌گیریم

و با توجه به تعداد داده‌های هر روش درستی اجرای الگوریتم را بر اساس میانگین وزن دار آنها محاسبه خواهیم نمود.

٤-۷-۳ پیشبینی رستوران جدید

با توجه به پیاده‌سازی مراحل بالا جهت پیشبینی امتیاز کاربران برای یک رستوران تازه وارد شده می‌توان هشت پارامتر مورد استفاده در پروژه را به تابع Predict New Restaurant ارسال کرد تا فرایند خوشه بندی کلاس اول که بیانگر خدمات تحویل و سفارش آنلاین است و کلاس دوم که سنجش محیط رستوران را بیان می‌کند انجام شود و سپس با توجه به الگوریتم پیشبینی تعریف شده در مرحله قبل نسبت به پیشبینی امتیاز کاربران برای چنین رستورانی اقدام کند.

---

[۳۶] classifier.۱

[۳۷] classifier.۲

[۳۸] confusion matrix

باید این نکته را نیز ذکر کرد که از هشت پارامتر ورودی، متغیر تعداد آرا، براساس میانگین تعداد آرا رستوران های موجود در سیستم به صورت پیشفرض تنظیم شده است چرا که قطعا رستوران تازه وارد تعداد رای خاصی جهت ارسال به تابع پیشبینی ندارد و امکان اختلال در روند پیشبینی وجود دارد پس بدین ترتیب صرفا هفت ویژگی هر رستوران نوپا جهت پیشبینی قابل تعریف شدن است.

٤-٨ نتیجه گیری

یکی از کارهای مفید در این پژوهش استفاده از خوشه بندی پیش از فاز آموزش و تست بود چراکه اولا تعداد پارامترها از ۱۱ تا به دو دسته کاهش یافت دوما گستره پارامتر ها محدود شده تا عملیات پیشبینی با دقت بهتری انجام شود. ضمنا این خوشه‌بندی، باعث امتیازدهی موضوعی به رستوران‌ها هم شد تا ضمن آماده‌سازی داده‌های الگوریتم پیشبینی، یک طبقه‌بندی موضعی پیرامون محیط و تحویل غذا درب منزل مجموعه‌ها داشته باشیم.

یکی از چالش های این مجموعه داده‌ها، توزیع شدید برروی یک مقدار بود که با گردسازی اصولی مقادیر، مطابق شرایط واقعی و پس از آن افزایش نمونه‌های کمتر و در نهایت حذف داده‌های فاقد سرویس تحویل غذا درب منزل، نتیجه بدست آمده بشدت رو به بهبودی قرار گرفت. که این سه اقدام بیشتر جنبه حذف داده های مزاحم و پرت را به همراه داشت تا بتوانیم داده ها را در یک طیف مناسب در اختیار داشته باشیم.

داده ها صرفا با یک مدل قابل تفکیک سازی نیستند به عنوان مثال داده های صفر و یک کلاس اول به صورت سلسله‌مراتبی خوشه‌بندی بهتری را ایجاد کردند و داده های پیوسته مرتبط به محیط رستوران با روش k-means نتیجه بهتری به همراه داشت.

هدف نهایی پیشبینی امتیاز یک رستوران نوپا بوده است اما این فرایند صرفا با اجرای الگوی دسته بندی ماشین بردار پشتیبان قابل دستیابی نبود. و پیش آماده‌سازی داده ها نظیر نرمالسازی و خوشه‌بندی آنها تاثیر بسزایی در نتیجه نهایی داشته است.

فصل ۵: نتیجه گیری

۵-۱ بررسی ماتریس اغتشاش

در شکل زیر confusion matrix حاصل از پیشبینی رستوران های داری ویژگی ارسال غذا و سفارش آنلاین را مشاهده می‌کنید.

جدول ۵-۱: ماتریس اغتشاش رستوران های داری خدمات ارسال و سفارش آنلاین

برچسب واقعی		پیشبینی				
		کلاس ۱	کلاس ۲	کلاس ۳	کلاس ۴	کلاس ۵
	کلاس ۱	۵۲	۰	۰	۰	۰
	کلاس ۲	۳	۱۸۷	۰	۵	۱۱
	کلاس ۳	۰	۱۳	۱۰۲	۷۰	۶
	کلاس ۴	۰	۰	۵۲	۲۲۴	۰
	کلاس ۵	۱	۱۳	۰	۷	۶۲

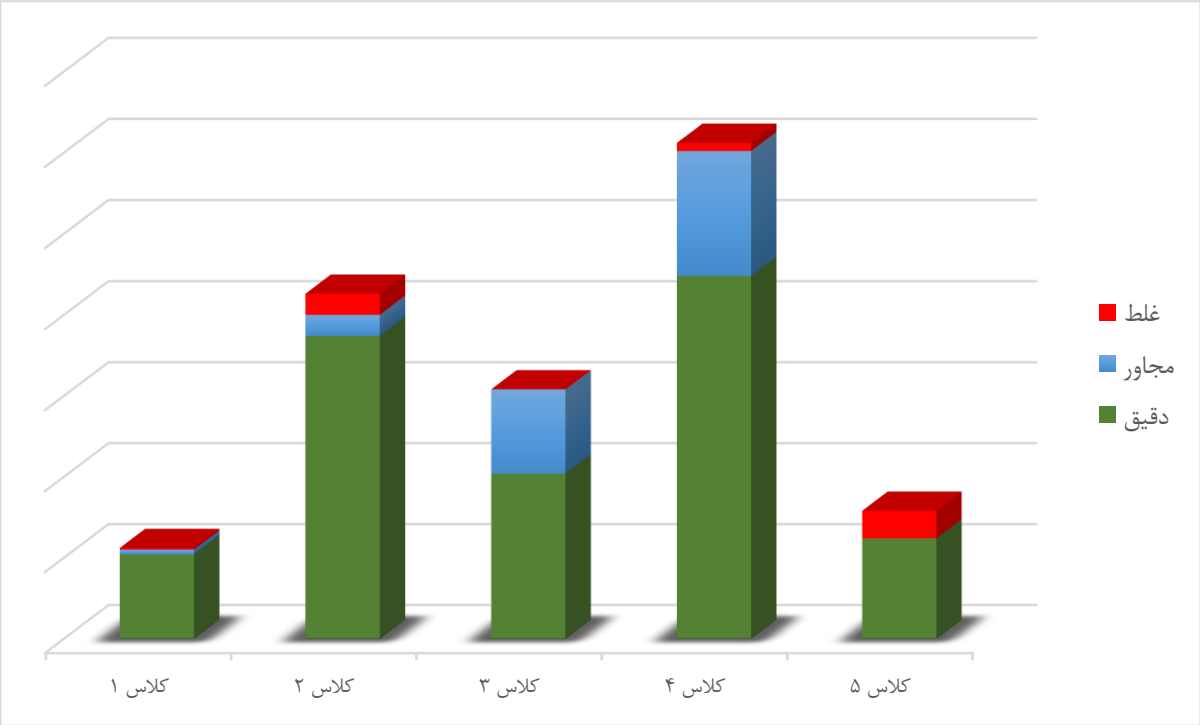
accuracy	۰.۷۸
Kappa score	۷۱٪

بدین ترتیب عملکرد الگوریتم در شرایط خوبی قرار دارد ضمنا برای رستوران های فاقد ارسال و سفارش آنلاین از classifier.2 استفاده می‌شود که ضریب همبستگی این پارامتر با امتیاز کاربران چیزی در حدود ۴۵ درصد می‌باشد.

- پس برای یک نتیجه گیری مطلوب میانگین وزنی این دو حالت را نیز محاسبه می‌کنیم که با توجه به تعداد چهارهزارتایی عدم امکان تحویل غذا درب منزل در داده‌های پیشرو از پنج هزار داده، این میانگین مقداری در حدود شصت درصد درستی پیشبینی را تضمین می‌کند.

- گرچه **۶۰ درصد** دقت عملکرد نسبتا خوبی را رقم می‌زند اما باید توجه داشت که در مثال واقعی رستوران ها، تخمین کلاس مجاور نیز چندان پرت نیست به عنوان مثال رستورانی که امتیاز کاربرانش مقدار ۲ را اخذ کرده است تخمین مقدار ۱ یا ۳ نیز میتواند به صورت نسبی بیانگر عملکرد مجموعه باشد پس در واقعیت میزان قدرت الگوریتم فراتر از میزان شصت خواهد بود.

جدول ۵-۲ : میزان درستی پیشبینی با برچسب های مجاور



- یکی از پیشرفت‌های این پروژه نسبت به تحقیقات [۲] اینگونه بوده است که اولا امکانی جهت پیشبینی امتیاز آتی کاربران به رستوران‌های تازه وارد شده به سیستم فراهم کرده و صرفا به تعیین باکلاسی یا بی‌کلاسی محیط یا تعیین نمره تحویل غذا درب منزل یک مجوعه بسنده نکرده است دوما تحلیل دقیقی پیرامون امتیاز کاربران در رستوران های فاقد تحویل غذا درب منزل و اثر آن در پیشبینی الگوریتم ارائه شده است.

۵-۲ تحلیل رستوران‌های فاقد خدمات ارسال و سفارش برخط

- یکی از نتایج قابل توجه در این پروژه به این نکته اشاره دارد که در دیتاهای مجموعه رستوران‌های زوماتو افراد به کیفیت دریافت شده امتیاز می‌دهند و وجود یا عدم وجود یک سرویس به طور مثال تحویل غذا درب منزل اثر چندانی بر امتیاز آنان نخواهد داشت البته چنانچه این سرویس نیز ارائه شود میزان کیفیت ارائه خدمات قطعا موثر است.

	id	Restaurant Name	Country	classifier.1	classifier.2	aggregate_rating	output label
1161	5704255	Famous Dave's Barbecue	UAE	5	4	5	Delivery & Classy ambiance
1162	5701978	Pizza Di Rocco	UAE	4	3	4	Delivery & Classy ambiance
1167	5700052	Cho Gao - Crowne Plaza Abu Dhabi	UAE	4	4	4	Delivery & Classy ambiance
1168	5702418	Gazebo	UAE	4	5	4	Delivery & Classy ambiance
1169	5700386	Sangeetha Vegetarian Restaurant	UAE	3	2	3	Delivery
...	...	...	...	...	...	...	...
5627	6103868	Nobu	United Kingdom	3	4	4	Delivery & Classy ambiance
5628	6104220	Roti Chai	United Kingdom	5	3	5	Delivery & Classy ambiance
5639	6800569	Chaophraya	United Kingdom	3	4	4	Delivery & Classy ambiance
5651	6801873	Mr Cooper's House & Garden - The Midland	United Kingdom	3	3	4	Delivery & Classy ambiance
5667	18295472	Gymkhana	Qatar	5	4	5	Delivery & Classy ambiance

1703 rows × 7 columns

شکل ۵-۱: اطلاعات برخی از رستوران ها خوشه بندی شده

	id	Restaurant Name	Country	classifier.1	classifier.2	aggregate_rating	output label
0	6600681	Chez Michou	Brazil	1	2	3	Not any
1	6601005	Cafí© Daniel Briand	Brazil	1	1	4	Not any
2	6600292	Casa do Biscoito Mineiro	Brazil	1	2	3	Not any
3	6600441	Maori	Brazil	1	3	4	Classy ambiance
4	6600970	Pizza íæ Bessa	Brazil	1	2	3	Not any
...	...	...	...	...	...	...	...
5722	5915730	NamlÛ± Gurme	Turkey	1	3	4	Classy ambiance
5723	5908749	Ceviz AÛôacÛ±	Turkey	1	3	4	Classy ambiance
5724	5915807	Huqqa	Turkey	1	4	3	Classy ambiance
5725	5916112	A□ô□ôk Kahve	Turkey	1	4	4	Classy ambiance
5726	5927402	Walter's Coffee Roastery	Turkey	1	2	4	Not any

5727 rows × 7 columns

شکل ۵-۲: لیست رستوران ها خوشه بندی شده با تحویل غذا درب منزل

این مورد را می‌توان در امتیاز حاصل از Classifier.۱ و Classifier.۲ در مقایسه شکل ۵-۱ و شکل ۵-۲ نیز به طور واضح مشاهده کرد. بدین صورت که در شکل اول تناسب میان نمره کاربران با خوشه بندی‌های صورت گرفته موجود است اما در شکل دوم به وفور می‌بینیم که برای مثال با مقدار کلاس اول ۱ نمره نهایی کاربر مقدار ۴ یا ۳ و یا حتی ۵ می‌باشد.

فصل ۶: جمع بندی

همانطور که در ذیل این پروژه به آن دست یافتیم طبقه‌بندی رستوران‌ها برای استفاده در سیستم‌های توصیه‌گر امری الزامی است چراکه متغیر های متنوعی علاوه بر امتیاز مستقیم کاربران بر عملکرد مجموعه موثر است و تناسب میان این دو باید سنجیده شود.

امتیازدهی به عملکرد یک مجموعه و سیستم صرفا مختص رستوران‌ها نیست و در هر سیستمی امکان بکارگیری از این ظرفیت وجود دارد تا تجربه بهتری برای مخاطب رقم خورده و به تبع آن درآمد مجموعه نیز روندی صعودی پیدا کند. البته مقیاس های امتیاز گذاری می‌تواند متنوع باشد همانطور که ما در این پروژه از امتیاز دهی پنج ستاره‌ای بهره بردیم در سایر روش‌ها می‌تواند به صورت کیفی این طبقه بندی را انجام دهند گرچه امتیاز دهی به صورت عددی و در بازه های مختلف امکان تحلیل داده را ساده‌تر خواهد نمود.

دیدیم که در امتیاز دهی به یک مجموعه رستوران، وجود یا عدم وجود تحویل غذا درب منزل نقشی در امتیاز کاربران نداشت گرچه کیفیت سرویس با اهمیت بود اما در ادامه راه باید ویژگی های دیگری را نیز بجز امتیاز کاربران در نظر گرفت تا ترکیب پارامتر دیگری با این امتیاز، جلوه‌گر بهتری از توصیف آن مجموعه باشد تا عدم ارائه سرویس خدمات ارسال و سفارش آنلاین در نتیجه نهایی تاثیر منفی خود را نمایان کند.

در این پروژه از یازده ویژگی که هشتای آنها به طور مستقیم بود در پیشبینی عملکرد یک رستوران استفاده شد، اما پر واضح است که برای افزایش بهره‌وری الگوریتم هوشمند سیستم، نیاز به پارامترهای بیشتری برای بررسی هست، همانطورکه در تحقیق انجام شده توسط آقای لیو و خانم تسه [۵] عواملی چون سرعت خدمات، حجم غذا، نورپردازی و جو موجود در محیط تاثیر بسزایی در تحلیل عملکرد رستوران ها دارد.

منابع

[۱] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, and V. S. Subrahmanian, “REV2: Fraudulent User Prediction in Rating Platforms,” in Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, Marina Del Rey, CA, USA ,۲۰۱۸ ,pp. 333–341.

[۲] N. F. AL-Bakri, A. F. Al-zubidi, A. B. Alnajjar, and E. Qahtan, “Multi label restaurant classification using support vector machine,” *Periodicals of Engineering and Natural Sciences (PEN)*, vol. 9, no. 2, pp. 774-783, 2021.

[۳] R. SALUJA. "Food&Restaurants," https://www.kaggle.com/ritesaluja/food-restaurants.

[٤] S. Mehta. "Zomato Restaurants Data," https://www.kaggle.com/shrutimehta/zomato-restaurants-data.

[٥] P. Liu, and E. C.-Y. Tse, “Exploring factors on customers’ restaurant choice: an analysis of restaurant attributes,” *British Food Journal*, vol. 120, no. 10, pp. 2289-2303, 2018.