
بسم الله الرحمن الرحيم

چکیده

مطالعات نشان می‌دهد که سوانح هوایی به دلیل یک عامل منفرد نیست، بلکه به علت زنجیره‌ای از خطاها در مراحل مختلف به وجود می‌آیند. مطالعات متعددی به بررسی عوامل انسانی در سوانح هوایی پرداخته‌اند و با روش‌های آماری مختلف این عوامل را طبقه‌بندی کرده‌اند. اما روش‌های مبتنی بر داده‌کاوی و مدلسازی عوامل با استفاده از یادگیری ماشین و پیش‌بینی متغیرهای مختلف یک حادثه، همچنان نیاز به توسعه دارند.

مجموعه‌داده بررسی شده در این مطالعه، شامل ۴۷۹ حادثه هوایی است که در بازه زمانی بین سال‌های ۲۰۰۶ تا ۲۰۱۵ توسط هیئت ایمنی حمل و نقل[۱] بررسی و گزارش شده است. این مجموعه‌داده طی مطالعه‌ای که در سال ۲۰۱۸ تحت عنوان «ارزیابی پیش‌شرط‌های مؤثر بر خطای انسانی علامت‌دار در سوانح هوانوردی عمومی و شرکت‌های هواپیمایی[۲]» مورد بررسی قرار گرفته است و مدل سیستم تجزیه‌وتحلیل و طبقه‌بندی عوامل انسانی روی آن پیاده‌سازی شده است. هدف این پژوهش، بررسی هر کدام از عوامل و میزان تاثیر آن‌ها بر حادثه، مدلسازی عوامل تاثیرگذار بر سوانح هوایی به کمک تکنیک‌های یادگیری ماشین و پیش‌بینی نوع حادثه با توجه به ویژگی‌های مختلف می‌باشد.

نتایج پژوهش نشان می‌دهد که خطاهای مبتنی بر عملکرد و خطاهای مبتنی بر اعمال اشتباه از خطاهای موثر در سوانح هوایی محسوب می‌شوند. همچنین الگوریتم‌های یادگیری ماشین بکارگرفته شده در این مطالعه از عملکرد نسبتا خوبی برخوردار هستند. الگوریتم‌های دسته‌بندی مبتنی بر رای‌گیری، دسته‌بندی جنگل تصادفی، دسته‌بندی XGboost و دسته‌بندی درخت تصمیم، ۴ الگوریتم برتر در این مطالعه هستند که دقت آن‌ها روی مجموعه‌داده دیده نشده توسط الگوریتم‌ها بیشتر از ۸۰ درصد می‌باشد.

واژگان کلیدی: سوانح هوایی، داده‌کاوی، یادگیری ماشین، چارچوب HFACS، خطای انسانی

[۱] NTSB: National Transportation Safety Board

[۲] Anthony J. Erjavac , Ronald Iammartino , John M. Fossaceca , Evaluation of preconditions affecting symptomatic human error in general aviation and air carrier aviation accidents, Reliability Engineering and System Safety (۲۰۱۸), doi: ۱۰,۱۰۱٦/j.ress.۲۰۱۸,۰٥,۰۲۱

فهرست مطالب

فهرست اشکال

فهرست جداول

فهرست نمودارها

۱- مقدمه

۱-۱- مقدمه

خطای انسانی[۳]، به عنوان عامل مهمی در بروز حوادث بزرگ در بسیاری از صنایع شناخته می‌شود. با اینکه تخمین‌ها و نرخ‌های گزارش شده در حوادث بسته به محدوده صنعت متفاوت است، مطالعات نشان می‌دهد که ۶۰ درصد تا ۸۰ درصد از خرابی‌های سیستم تا حدی به عملکرد انسان بستگی دارد [۱].

مطالعات نشان می‌دهد که سوانح هوایی به دلیل یک عامل منفرد نیست، بلکه به علت زنجیره‌ای از خطاها در مراحل مختلف به وجود می‌آیند. برای کاهش خطاهای انسانی در تصادفات، لازم است به جای تمرکز بر کسانی که آن خطا را مرتکب شده‌اند، روی علل آن تمرکز کرد. بسیاری از عوامل مانند استرس، خستگی و آموزش ناکافی می‌توانند سبب بروز خطاهای انسانی در طول پرواز گردند. چالش اصلی، دشوار بودن ردیابی عواملی است که منجر به خطاهای انسانی می‌شوند. مطالعات، این عوامل را به عنوان عوامل آشکار و پنهان نام برده‌اند. در ادبیات، مدلی تحت عنوان مدل پنیر سوییسی[۴] تعریف شده است که نشان می‌دهد حوادث در نتیجه ترکیب بیش از یک خطا رخ می‌دهند. همچنین در ادبیات آمده است که در سال ۲۰۰۱، شاپل[۵] و ویگمن[۶] سیستم تجزیه‌وتحلیل و دسته‌بندی عوامل انسانی[۷] را براساس مدل پنیر سوییسی توسعه داده‌اند. این مدل عوامل انسانی را در چهار سطح طبقه‌بندی می‌کند و علاوه بر هوانوردی در بسیاری از زمینه‌های مختلف مورد استفاده قرار گرفته است و ابزار ارزشمندی برای تشخیص عوامل انسانی در حوادث می‌باشد. تشخیص و طبقه‌بندی عوامل انسانی در سوانح هوایی برای اتخاذ اقدامات احتیاطی موثر بسیار حائز اهمیت است [۲].

۱-۲- ضرورت انجام تحقیق

از آنجایی‌که که انسان موجودی پیچیده است و عملکرد آن به عوامل مختلفی بستگی دارد، بنابراین مدلسازی خطاهای انسانی همچنان امری چالش برانگیز است. مطالعات متعددی به بررسی عوامل انسانی در سوانح هوایی پرداخته‌اند و با روش‌های آماری مختلف این عوامل را طبقه‌بندی کرده‌اند. اما روش‌های مبتنی بر داده‌کاوی و مدلسازی عوامل با استفاده از یادگیری ماشین و پیش‌بینی متغیرهای مختلف یک حادثه، همچنان نیاز به توسعه دارند.

۱-۳- اهداف تحقیق

مجموعه داده شامل ۴۷۹ حادثه هوایی است که در بازه زمانی بین سال‌های ۲۰۰۶ تا ۲۰۱۵ توسط هیئت ایمنی حمل و نقل[۸] بررسی و گزارش شده است. این مجموعه‌داده طی مطالعه‌ای که در سال ۲۰۱۸ تحت عنوان «ارزیابی پیش‌شرط‌های مؤثر بر خطای انسانی علامت‌دار در سوانح هوانوردی عمومی

---

[۳] Human Error
[۴] Swiss Cheese Model
[۵] Shappell
[۶] Wiegmann
[۷] HFACS: Human Factor Analysis and Classification System
[۸] NTSB: National Transportation Safety Board

و شرکت‌های هواپیمایی[۹]» مورد بررسی قرار گرفته است و مدل سیستم تجزیه‌وتحلیل و دسته‌بندی عوامل انسانی روی آن پیاده‌سازی شده است. هدف این پژوهش، بررسی هر کدام از عوامل و میزان تاثیر آن‌ها بر حادثه، مدلسازی عوامل تاثیرگذار بر حوادث هوایی به کمک تکنیک‌های یادگیری ماشین و پیش‌بینی نوع حادثه با توجه به ویژگی‌های مختلف می‌باشد.

۱-۴- طرح مساله

آنچه که به عنوان مساله و سوال برای این پژوهش در نظر گرفته شده است، به شرح زیر است:

۱. کدام یک از عوامل انسانی بیشترین تاثیر را بر یک حادثه هوانوردی دارند؟

۲. مدل‌های یادگیری ماشین تا چه میزان در مدلسازی و پیش‌بینی حوادث و آسیب‌های ناشی از آن در صنعت هوانوردی کاربرد دارند؟

۲- پیشینه تحقیق

۲-۱- مرور ادبیات

۲-۱-۱- یک مدل ترکیبی HFACS-BN برای تجزیه و تحلیل آگاهی متخصصان هوانوردی مغولستان از عوامل انسانی مرتبط با ایمنی هوانوردی

برای روشن شدن بهتر تاثیرات عوامل انسانی بر خطرات سوانح هوایی، این مطالعه با استفاده از یک مدل ترکیبی HFACS تحت عنوان HFACS-BN (سیستم تجزیه و تحلیل و طبقه‌بندی عوامل انسانی ؛ شبکه بیزی) انجام شده است. پژوهشگران پرسشنامه‌ای بر اساس چارچوب ۴ سطحی HFACS طراحی و اجرا کردند و داده‌های معتبری را از ۱۸۰ نفر از ۶۴۹ نفر متخصص هوانوردی که در فرودگاه بین‌المللی اولان‌باتور، مغولستان در سال ۲۰۱۷ کار می‌کردند جمع‌آوری کردند. این مدل ۳۵ عامل اصلی انسانی را از ۱۲۹ عامل شناسایی کرد. بررسی‌ها و نتایج این مقاله بیانگر این است که سطح اعمال ناایمن بیشترین تاثیر را بر خطرات در بین چهار سطح موجود دارد، در حالی که سطح نظارت ناایمن کمترین سهم را دارد. همچنین مشخص شده است که افزایش آگاهی متخصصان هوانوردی از عوامل انسانی باید از اثرات زنجیره‌ای علی در میان عوامل انسانی استفاده کامل کند [۳].

۲-۱-۲- تجزیه و تحلیل عوامل انسانی در پنجاه پرواز کنترل شده به سوانح هوایی زمینی

CFIT به عنوان یک برخورد غیر عمدی با زمین (زمین، یک کوه، یک بدنه آبی یا یک مانع) در حالی که یک هواپیما تحت کنترل مثبت است، تعریف می شود. هدف این مقاله شناسایی عوامل انسانی درگیر با سوانح هوایی است که منجر به CFIT شده است. در این مطالعه از HFACS برای تعیین عوامل دخیل در ۵۰ حادثه CFIT از ۲۴ شهرستان در یک دوره ۱۰ ساله، یعنی ۲۰۰۷-۲۰۱۷ استفاده شده است. از مصاحبه با پنج کارشناس ارشد ایمنی هوانوردی برای ارائه درک بهتری از عوامل انسانی مؤثر بر ایمنی پرواز استفاده شد. در این مطالعه ۱۲۸۹ مورد عامل انسانی و عاملی فردی با اقدامات ناایمن و پیش‌شرط اقدامات ناایمن به عنوان زیرمجموعه‌های اصلی تصادفات شناسایی شد. این مطالعه نشان داد که CFIT

[۹] Anthony J. Erjavac , Ronald Iammartino , John M. Fossaceca , Evaluation of preconditions affecting symptomatic human error in general aviation and air carrier aviation accidents, Reliability Engineering and System Safety (۲۰۱۸), doi: ۱۰،۱۰۱٦/j.ress.۲۰۱۸،۰٥،۰۲۱

در طیف وسیعی از تجربیات خلبانان رخ می دهد و ۴۴ درصد از تصادفات در پروازهای کروز رخ می‌دهد. حواس‌پرتی، نارضایتی و خستگی همه عناصری هستند که خدمه پرواز ممکن است به عنوان مشارکت کنندگان در CFIT در طول سفر هوایی تجربه کنند [۴].

۲-۱-۳- کاربرد HFACS در حوادث و سوانح هوایی شبانه

سوانح و حوادث هوانوردی تجاری در ساعات مشخصی از روز شیوع بیشتری دارد. مشکلات عملیاتی (مانند دید در شب، کوری فلاش، توهم سیاهچاله و انعکاس) که خلبانان هنگام انجام پروازهای شبانه با آن مواجه می شوند، امنیت پرواز را تهدید می‌کند. هدف مقاله حاضر بررسی عوامل مؤثر در بروز سوانح هوایی تجاری در شب است. در این مقاله، گزارش‌های سوانح مربوط به ۳۰ سقوط هواپیمای تجاری که طی پنج سال گذشته رخ داده است، مورد تجزیه و تحلیل قرار گرفته است. عوامل مؤثر در این حوادث با استفاده از چارچوب HFACS مورد بررسی قرار گرفت. بررسی‌ها بیانگر این است که محیط فیزیکی مهم ترین عامل سببی است. خطاهای مبتنی بر مهارت دومین عامل موثر و خطاهای ادراکی و خطاهای تصمیم‌گیری به عنوان سومین عامل موثر در رتبه بندی قرار گرفتند. همچنین نتایج این بررسی نشان می‌دهد که چندین عامل سببی در ایجاد سوانح ناو هواپیمابر تجاری در شب وجود دارد و صرفاً به دلیل خطاهای ادراکی رخ نداده‌اند [۵].

۲-۱-۴- سیستم تجزیه و تحلیل و طبقه بندی عوامل انسانی

در این مطالعه، چارچوب HFACS برای شناسایی عوامل انسانی که در سانحه پرواز ۲۱۴ خطوط هوایی آسیانا در ۶ جولای ۲۰۱۳ رخ داد، استفاده شده است. نتایج این مطالعه نشان می‌دهد که آموزش ناکافی خلبان، عدم نظارت توسط سطوح بالا، و انحراف مکرر از رویه های عملیاتی استاندارد تا حد زیادی در این حادثه نقش داشته‌اند. این یافته‌ها بر سطوح مختلف سازمانی که در سوانح هوانوردی نقش دارند، تأکید می‌کنند و اهمیت بالای رویکردی فعالانه برای ایمنی و کاهش خطرات در سطوح بالای سازمان را قبل از اینکه منجر به فاجعه در خط مقدم شوند، برجسته می‌کنند. در مورد پرواز ۲۱۴، بیشترین موارد قابل اجرا در این مورد طراحی پیچیده هواپیما و آموزش ناکافی خلبانان بود. بوئینگ ER۷۷۷-۲۰۰ یک هواپیمای بسیار پیچیده است که شامل تعداد زیادی از فناوری‌های پیچیده منحصر به فرد هواپیماهای بوئینگ است. متأسفانه، روش‌های آموزشی خطوط هوایی آسیانا کافی نبود و خلبانان آن را از عملکرد داخلی سیستم‌های مختلف بی‌اطلاع می‌کرد [۶].

۲-۱-۵- پردازش زبان طبیعی برای شناسایی عوامل انسانی در سوانح هوایی: روش SHEL

حوادث در هوانوردی اتفاق نادری است. از این رو، سیستم‌های مدیریت ایمنی هوانوردی با انجام تجزیه و تحلیل علل ریشه‌ای سوانح، به دنبال ریشه‌یابی این علل هستند. از آنجایی که استاندارد فعلی بر طبقه‌بندی دستی انجام شده توسط کارکنان آموزش دیده متکی است، هیچ استاندارد فنی از قبل برای شناسایی خودکار عوامل انسانی تعریف نشده است. این مقاله این موضوع را بررسی می‌کند و تکنیک‌های یادگیری ماشین را با استفاده از فناوری‌های پیشرفته پردازش زبان طبیعی پیشنهاد می‌کند. سپس این تکنیک‌ها با استفاده از مدل SHEL تطبیق داده می‌شوند و روی مجموعه‌ای از حوادث واقعی آزمایش می‌شوند. نتایج محاسباتی دقت و اثربخشی روش پیشنهادی را نشان می‌دهد. علاوه بر این،

استفاده از این روش برای اسناد واقعی بررسی شده توسط کارشناسان کاهش زمان مورد نیاز را برای حداقل ۳۰ درصد در مقایسه با روش‌های استاندارد شناسایی عوامل انسانی تخمین می‌زند [۷].

۲-۱-۶- ادغام عوامل خلبان در تجزیه و تحلیل ریسک سوانح هوانوردی غیرنظامی از سال ۲۰۰۸ تا ۲۰۲۰: یک رویکرد شبکه بیزی مبتنی بر داده

این مطالعه یک رویکرد شبکه بیزی مبتنی بر داده را برای بررسی اثرات علی مشترک خلبان و سایر عوامل بر ایمنی هوانوردی غیرنظامی معرفی می‌کند. تعداد کل ۱۶۳ تصادف مربوط به خلبان فردی در پایگاه داده سوانح هوایی هیئت ملی ایمنی حمل و نقل از سال ۲۰۰۸ تا ۲۰۲۰ تجزیه و تحلیل شده است، با تمرکز بر استخراج اثرات علی عوامل خطر بالقوه مختلف، از جمله عوامل خلبان، بر سوانح هوانوردی غیرنظامی. مدل‌سازی وابستگی متقابل بین عوامل مؤثر بر خطر و اثر کمک‌کننده علّی آن‌ها بر پیامد حادثه توسط یک شبکه تقویت‌شده درختی ساختار یافته و با تحلیل حساسیت تأیید می‌شود. تازگی این مطالعه ترکیب عوامل خلبان به دست آمده از پایگاه داده سوانح هوانوردی غیرنظامی در تجزیه و تحلیل ریسک، همراه با سایر عوامل خارجی است. نتایج نشان می‌دهد که شرایط آب و هوایی و مراحل پرواز با انواع تلفات سوانح هوانوردی غیرنظامی نسبت به اقدام و تصمیم خلبان ارتباط بیشتری دارد و سه عامل خلبان دیگر تنها در صدمات کشنده در سوانح هواپیمایی کشوری نقش دارند [۸].

۲-۱-۷- عوامل انسانی هوانوردی: بررسی اجمالی مفهومی

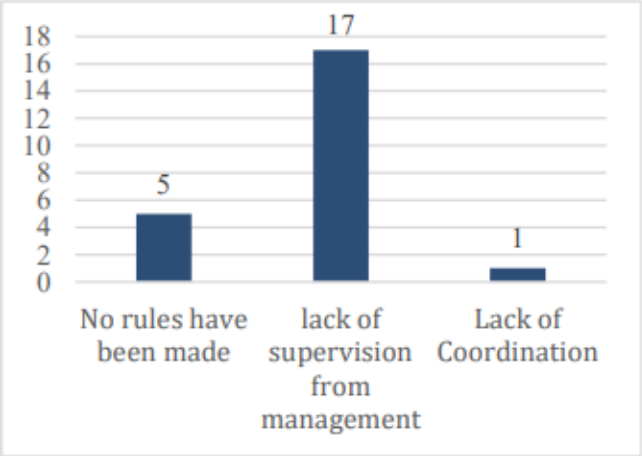
اهمیت عوامل انسانی در افزایش ایمنی غیرقابل‌اجتناب است، به همین دلیل از بیش از ۲۰ سال پیش آموزش عوامل انسانی در هوانوردی اجرا می‌شود. ایجاد سیستم‌های گزارش‌دهی و یادگیری کامل یکی از بزرگترین موفقیت‌های هوانوردی بوده است. از حدود ده سال پیش تا به امروز پیشرفت‌های چشمگیری در بحث عوامل انسانی بوجود آمده است. منصفانه است که بگوییم در طی این ۱۰ سال، تئوری و روش‌ها تکامل یافته‌اند و خوب است که ببینیم فاکتورهای انسانی برای حل مسائل مختلف استفاده شده است. این برنامه‌های کاربردی تاکنون چندین اکتشاف جذاب و نتایج خوبی را به همراه داشته‌اند. ۶۰ سال اول تحقیقات عوامل انسانی در صنعت هوانوردی به طور قابل توجهی ایمنی را بهبود بخشیده است. برای ۶۰ سال آینده، وظیفه ادامه افزایش استانداردهای ایمنی و در عین حال تلاش برای به کارگیری اصول ارگونومیک برای بهبود اثربخشی و عملکرد سازمانی خواهد بود [۹].

۲-۱-۸- یادگیری ماشین و پردازش زبان طبیعی برای پیش‌بینی عوامل انسانی در گزارش‌های حوادث هوانوردی

این مطالعه روشی را برای شناسایی و طبقه‌بندی دسته‌های عوامل انسانی از گزارش‌های حوادث هوانوردی ارائه می‌کند و چارچوبی تحت عنوان سیستم طبقه‌بندی و تحلیل عوامل انسانی مبتنی بر یادگیری ماشین را ارائه داده است. برای شناسایی و استخراج ویژگی‌ها، ترکیبی از روش‌های پیش پردازش متن و پردازش زبان طبیعی توسعه داده شده است. برای مدلسازی داده‌ها، تکنیک‌های LS نیمه نظارت یافته و SVM در نظر گرفته شده‌اند. برای بهینه‌سازی، بهبود مدل و تجزیه و تحلیل فراپارامترها از روش بهینه‌سازی بیزی استفاده شده است. نتایج این مطالعه نشان می‌دهد که برای طبقه‌بندی عوامل انسانی بر اساس داده‌های متنی می‌توان از روش ارائه شده استفاده کرد [۱۰].

۲-۱-۹- تحلیلی بر حوادث و سوانح جدی هوانوردی با استفاده از مدل شِل[۱۰]

این تحقیق از نوع توصیفی-تحلیلی با روش کمی است. داده‌ها از گزارش‌های نهایی منتشر شده توسط کمیته ملی ایمنی حمل و نقل اندونزی از سال ۲۰۱۵ تا ۲۰۱۹ جمع آوری شده است. علل اصلی با استفاده از تحلیل مدل شِل طبقه‌بندی شده‌اند. نتایج نشان می‌دهد که شایع‌ترین علت حوادث هوایی در اندونزی، عدم هماهنگی بین نرم‌افزارهای زنده است که درصد آن به ۶۴ درصد رسیده است. شکل زیر نمودار فراوانی علل حوادث هوایی در اندونزی طی سال‌های ۲۰۱۵ تا ۲۰۱۹ را نشان می‌دهد. در ۱۷ رویداد ناشی از عدم نظارت مدیریت، ۳ مرحله وجود داشت که رویدادها رخ داده است که شامل مرحله فرود شامل ۵۹٪ یا ۱۰ رویداد، مرحله پرواز شامل ۱۸٪ یا ۳ رویداد و مرحله فرود آمدن شامل ۲۴٪ یا ۴ رویداد است. همچنین بررسی فراوانی داده‌ها نشان می‌دهد که در بازه زمانی بررسی شده تعداد سوانح جدی هوایی روند کاهشی داشته است. در صورت عدم رفع علل پنهانی مانند عدم نظارت مدیریت، در دسترس نبودن قوانین و عدم هماهنگی، ممکن است حوادث جدی رخ دهد [۱۱].



نمودار (۲-۱): نمودار فراوانی علل حوادث هوایی در اندونزی طی سال‌های ۲۰۱۵ تا ۲۰۱۹

۲-۱-۱۰- تجزیه و تحلیل سوانح هوایی غیر نظامی ترکیه بین سال‌های ۲۰۰۳ تا ۲۰۱۷

این مطالعه، با هدف بررسی عوامل موثر بر سوانح هوانوردی غیرنظامی در ترکیه و افزایش ایمنی هوانوردی از طریق افزایش آگاهی نسبت به عوامل مؤثر در سوانح انجام شده است. گزارش‌ها به صورت تاریخچه‌نگر با استفاده از چارچوب HFACS مورد تجزیه و تحلیل قرار گرفته‌اند. ۵۹ سانحه هوایی در این مطالعه وارد شده‌اند. مدیریت منابع خدمه ٪۴۱.۴، از دست دادن آگاهی موقعیت ٪۳۹.۰ و هواشناسی ٪۲۹.۲ بیشترین عوامل مؤثر در ۴۱ حادثه هواپیما، هلیکوپتر، گلایدر بوده‌اند، در حالی که هواشناسی ٪۷۷.۷ و مدیریت منابع خدمه ٪۶۱.۱ بیشترین عوامل مؤثر در ۱۸ تصادف بالن بودند. این یافته‌ها نشان می‌دهد که عوامل انسانی هنوز هم از عوامل اصلی در سوانح هوایی هستند. ادغام HFACS در سیستم مدیریت ایمنی هوانوردی ممکن است به کاهش نرخ سوانح هوایی کمک کند [۲].

[۱۰] Shell Model

۲-۱-۱۱- کاربرد چارچوب HFACS-HFIX در یافته‌های NTSC و توصیه‌ها با مطالعه موردی سوانح هوایی فرودگاه Wamena

فرودگاه وامنا در سال های ۲۰۰۲، ۲۰۰۸، ۲۰۰۹، ۲۰۱۳، ۲۰۱۵ و ۲۰۱۶ حوادثی را تجربه کرده است. ساختاربندی تحقیقات NTSC، تحت چارچوب HFACS برای درک نوع خرابی‌های عامل انسانی و استراتژی HFIX برای بستن خرابی‌ها با اعمال توصیه‌ها، باید در بررسی سانحه هوایی انجام شود. در این مطالعه یازده کارشناس و متخصص هوانوردی برای اعتبار بخشیدن به چارچوب مورد مصاحبه قرار گرفتند. در تصادفات ۲۰۰۸، ۲۰۱۳ و ۲۰۱۶ لایه‌هایی بدون هیچ گونه خرابی وجود داشت. حوادث در سال های ۲۰۰۲، ۲۰۰۹، ۲۰۱۳ و ۲۰۱۵ دارای شکست در لایه‌ای هستند که با دو یا چند توصیه مداخله می‌کند. شکست، خطا یا نقض مکرر حوادث تکراری در سال‌های ۲۰۰۲، ۲۰۰۹، ۲۰۱۳، ۲۰۱۵ و ۲۰۱۶ رویکردی تثبیت نشده است و با مداخلات موثر مسدود نشده است. HFACS و HFIX برای چارچوب بررسی تصادف مفید هستند و از وقوع حادثه مشابه در آینده جلوگیری می‌کنند [۱۲].

۲-۱-۱۲- شناسایی عوامل ایجاد کننده سوانح پروازی بر اساس مدل SHELLO و روش همبستگی خاکستری آنتروپی بهبود یافته

به منظور شناسایی موثر عوامل ایجاد کننده اصلی در سوانح پروازی هوانوردی غیرنظامی، بررسی قوانین ایجاد سوانح پرواز، و ایجاد یک مکانیسم پیشگیری آینده‌نگر و موثر برای سوانح پروازی، این مقاله ابتدا براساس مدل SHELLO، طبقه‌بندی عوامل موثر بر سوانح پرواز را بر اساس انسان، سخت افزار، نرم افزار، محیط و سازمان و تعامل با نیازهای تحلیل به وجود آمدن سوانح پروازی ایجاد می‌کند. سپس، با توجه به ویژگی‌های تصادفی و عدم قطعیت فعالیت‌های پروازی هوانوردی غیرنظامی و ویژگی‌های خاکستری عوامل حادثه‌آفرین پرواز، ریتم الگوی همبستگی خاکستری آنتروپی بهبود یافته همراه با ویژگی‌های نمونه داده‌ها برای طبقه‌بندی عوامل موثر ایجاد می‌شود. در نهایت، الگوریتم برای شناسایی و اولویت‌بندی علل سوانح پرواز استفاده می‌شود. عوامل اصلی که باعث ایجاد پروازهای ناایمن در هوانوردی غیرنظامی چین (۲۰۱۵-۲۰۱۹) می‌شوند به ترتیب عبارتند از: خطاهای ادراکی، خطاهای مبتنی بر مهارت، اشتباهات تصمیم گیری، تخلفات، انحراف اجرای SOP و غیره [۱۳].

۲-۲- جدول مرور ادبیات

جدول (۲-۱): جدول مرور ادبیات

ردیف	عنوان مقاله	سال چاپ	نام مجله	نویسندگان	روش مورد استفاده	روش تحقیق مقاله	نتایج و دستاوردهای مقاله
۱	A Hybrid HFACS-BN Model for Analysis of Mongolian Aviation Professionals' Awareness of Human Factors Related to Aviation Safety	۲۰۱۸	Sustainability	Tuqiang Zhou et al.	HFACS-BN	پژوهشگران پرسشنامه‌ای بر اساس چارچوب ۴ سطحی HFACS طراحی و اجرا کردند و داده‌های معتبری را از ۱۸۰ نفر از ۶۴۹ نفر متخصص هوانوردی که در فرودگاه بین‌المللی اولان‌باتور، مغولستان در سال ۲۰۱۷ کار می‌کردند جمع‌آوری کردند. این مدل ۳۵ عامل اصلی انسانی را از ۱۲۹ عامل شناسایی کرد.	بررسی‌ها و نتایج این مقاله بیانگر این است که سطح اعمال ناایمن بیشترین تاثیر را بر خطرات در بین چهار سطح موجود دارد، در حالی که سطح نظارت ناایمن کمترین سهم را دارد. همچنین مشخص شده است که افزایش آگاهی متخصصان هوانوردی از عوامل انسانی باید از اثرات زنجیره‌ای علی در میان عوامل انسانی استفاده کامل کند.
۲	An analysis of human factors in fifty controlled flight into terrain aviation accidents from ۲۰۰۷ □□ ۲۰۱۷	۲۰۱۹	Journal of Safety Research	Damien Kelly, Marina Efthymiou	HFACS	در این مطالعه از چارچوب سیستم تجزیه و تحلیل و طبقه‌بندی عوامل انسانی (HFACS) برای تعیین عوامل دخیل در ۵۰ حادثه CFIT از ۲۴ کشور در یک دوره ۱۰ ساله، یعنی ۲۰۰۷-۲۰۱۷ استفاده شده است. از مصاحبه با پنج کارشناس ارشد ایمنی هوانوردی برای ارائه درک بهتری از عوامل انسانی مؤثر بر ایمنی پرواز استفاده شد.	در این مطالعه ۱۲۸۹ مورد عامل انسانی و عاملی فردی با اقدامات ناایمن و پیش‌شرط اقدامات ناایمن به عنوان زیرمجموعه‌های اصلی تصادفات شناسایی شد. این مطالعه نشان داد که CFIT در طیف وسیعی از تجربیات خلبانان رخ می دهد و ۴۴ درصد از تصادفات در پروازهای کروز رخ می دهد.

۳	Application of HFACS to the Nighttime Aviation Accidents and Incidents	۲۰۲۰	Journal of Aviation	Bilal KILIÇ, Ercan GÜMÜŞ	HFACS	داده های سوانح هواپیماهای تجاری که در طول دهه گذشته در شب رخ داده اند از پایگاه داده حوادث و حوادث NTSB به دست آمده است.	بررسی‌ها بیانگر این است که محیط فیزیکی مهم ترین عامل سببی است. خطاهای مبتنی بر مهارت دومین عامل موثر و خطاهای ادراکی و خطاهای تصمیم به عنوان سومین عوامل موثر در رتبه بندی قرار گرفتند. همچنین نتایج این بررسی نشان می‌دهد که چندین عامل سببی در ایجاد سوانح ناو هواپیمابر تجاری در شب وجود دارد و صرفاً به دلیل خطاهای ادراکی رخ نداده‌اند.
۴	Human Factors Analysis and Classification System (HFACS) As Applied to Asiana Airlines Flight ۲۱٤	۲۰۲۰	Journal of Purdue Undergraduate Research	Alex Small, Flavio A. C. Mendonca	HFACS	شناسایی عوامل انسانی موثر در سانحه پرواز ۲۱۴ خطوط هوایی آسیانا، در ۶ جولای ۲۰۱۳	یافته های این مطالعه بر سطوح مختلف سازمانی که در سوانح هوایی نقش دارند و به ویژه نقش بالاترین سطوح یک سازمان را در فرآیند ایمنی برجسته می کند، تأکید می کند. با شروع بررسی‌های سوانح در بالای یک سازمان، اجرای HFACS می‌تواند از رسیدن خطرات به سطوح پایین‌تر جلوگیری کند. این مطالعه همچنین ضرورت تمرین یک رویکرد پیشگیرانه برای ایمنی، کاهش خطرات در یک سازمان را قبل از اینکه منجر به فاجعه شود، برجسته می‌کند.

۵	Natural Language Processing for the identification of Human factors in aviation accidents causes An application to the SHEL methodology	۲۰۲۱	ELSEVIER	Guido Perboli et al.	SHEL Model	مجموعه دامنه از کتب و اسناد حوادث هوایی استخراج شد. با این حال، آموزش یک شبکه عصبی بر روی متن خام به‌دست‌آمده، منجر به هر یک از شکل‌های انحرافی یک کلمه می‌شود که توسط یک بردار جاسازی جداگانه نمایش داده می‌شود. این به نوبه خود منجر به اشکالات و ناکارآمدی های بسیاری می‌شود. حفظ یک بردار مجزا برای هر شکل عطف هر کلمه باعث می‌شود مدل افزایش حجم دهد و مصرف حافظه را افزایش دهد.	نتایج محاسباتی دقت و اثربخشی روش پیشنهادی را نشان می‌دهد. علاوه بر این، استفاده از روش برای اسناد واقعی بررسی شده توسط کارشناسان کاهش زمان مورد نیاز را برای حداقل ۳۰٪ در مقایسه با روش های استاندارد شناسایی عوامل انسانی تخمین می زند.
۶	Incorporation of Pilot Factors into Risk Analysis of Civil Aviation Accidents from ۲۰۰۸ to ۲۰۲۰ A Data-Driven Bayesian Network Approach	۲۰۲۲	MDPI	Chenyang Zhang et al.	شبکه بیزی مبتنی بر داده (BN)	تعداد کل ۱۶۳ تصادف مربوط به خلبان فردی در پایگاه داده سوانح هوایی هیئت ملی ایمنی حمل و نقل (NTSB) از سال ۲۰۰۸ تا ۲۰۲۰ تجزیه و تحلیل شده است.	تایج نشان می دهد که شرایط آب و هوایی و مراحل پرواز با انواع تلفات سوانح هوانوردی غیرنظامی نسبت به اقدام و تصمیم خلبان ارتباط بیشتری دارد و سه عامل خلبان دیگر تنها در صدمات کشنده در سوانح هواپیمایی کشوری نقش دارند.
۷	Aviation Human Factors Conceptual Overview	۲۰۲۳	IJERED	Shreya Mane	Conceptual Overview	در این مقاله با مطالعه پژوهش‌های گذشته، به مرور ادبیات در بحث فاکتورهای انسانی پرداخته شده است.	از حدود ده سال پیش تا به امروز پیشرفت‌های چشمگیری در بحث عوامل انسانی بوجود آمده است. در طی این ۱۰ سال، تئوری و روش‌ها تکامل یافته‌اند و نتایج خوبی را به همراه داشته‌اند.

۸	Machine Learning and Natural Language Processing for Prediction of Human Factors in Aviation Incident Reports	۲۰۲۱	MDPI	Tomás Madeira et al.	یادگیری ماشین و پردازش زبان طبیعی	برای شناسایی و استخراج ویژگی‌ها، ترکیبی از روش‌های پیش پردازش متن و پردازش زبان طبیعی توسعه داده شده است. برای مدلسازی داده‌ها، تکنیک‌های LS نیمه نظارت یافته و SVM در نظر گرفته شده‌اند. برای بهینه‌سازی، بهبود مدل و تجزیه و تحلیل فراپارامترها از روش بهینه‌سازی بیزی استفاده شده است.	این مطالعه روشی را برای شناسایی و طبقه‌بندی دسته‌های عوامل انسانی از گزارش‌های حوادث هوانوردی ارائه می‌کند و چارچوبی تحت عنوان سیستم طبقه‌بندی و تحلیل عوامل انسانی مبتنی بر یادگیری ماشین را ارائه داده است. نتایج این مطالعه نشان می‌دهد که برای طبقه‌بندی عوامل انسانی بر اساس داده‌های متنی می‌توان از روش ارائه شده استفاده کرد.
۹	An Analysis on Serious Incidents and Accidents in Aviation Using Shell Model	۲۰۲۲	Journal Perhubungan Udara	Pangsa Rizkina Aswia et al.	مدل Shell	داده‌ها از گزارش‌های نهایی منتشر شده توسط کمیته ملی ایمنی حمل و نقل اندونزی از سال ۲۰۱۵ تا ۲۰۱۹ جمع آوری شده است. علل اصلی با استفاده از تحلیل مدل شِل طبقه‌بندی شده‌اند.	شایع‌ترین علت حوادث هوایی در اندونزی، عدم هماهنگی بین نرم‌افزارهای زنده است که درصد آن به ۶۴ درصد رسیده است. علاوه بر این، این عدم تطابق ناشی از عدم نظارت مدیریت در ۱۷ رویداد یا ۷۴ درصد بوده است. ۵ رویداد یا ۲۲٪ ناشی از در دسترس نبودن قوانین بود و عدم هماهنگی با ۱ رویداد یا ۴٪.

۱۰	Analysis of Turkish Civil Aviation Accidents Between ۲۰۰۳ and ۲۰۱۷	۲۰۲۲	Journal of Aviation	Erdinç E, Ahmet U A.	HFACS	گزارش‌ها به صورت تاریخچه‌نگر با استفاده از تجزیه و تحلیل سیستم طبقه‌بندی و تجزیه و تحلیل عوامل انسانی مورد تجزیه و تحلیل قرار گرفته‌اند. ۵۹ سانحه هوایی در این مطالعه وارد شده‌اند.	یافته‌ها نشان می‌دهد که عوامل انسانی هنوز هم از عوامل اصلی در سوانح هوایی هستند. آموزش‌های آکادمیک مانند مدیریت منابع خدمه، هواشناسی هوانوردی و از دست دادن آگاهی موقعیت باید بیشتر به هوانوردان داده شود تا از حوادث جلوگیری شود. ادغام HFACS در "سیستم مدیریت ایمنی هوانوردی" ممکن است به کاهش نرخ سوانح هوایی کمک کند.
۱۱	Application of HFACS-HFIX framework in NTSC’S findings and recommendations Wamena air accidents’ case study	۲۰۲۲	IJIM	Aloysius Sigit Haryono et al.	HFACS-HFIX	به عنوان مفهوم پنیر سوئیسی، حادثه زمانی رخ می‌دهد که خطاها به لایه‌های دفاعی ایمنی در خط مستقیم نفوذ کرده‌اند. ساختاربندی تحقیقات NTSC، تحت چارچوب HFACS برای درک نوع خرابی‌های عامل انسانی و استراتژی HFIX برای بستن خرابی‌ها با اعمال توصیه ها، باید در بررسی سانحه هوایی انجام شود. در این مطالعه یازده کارشناس و متخصص هوانوردی برای اعتبار بخشیدن به چارچوب مورد مصاحبه قرار گرفتند.	در تصادفات ۲۰۰۸، ۲۰۱۳ و ۲۰۱۶ لایه‌هایی بدون هیچ گونه خرابی وجود داشت. حوادث در سال های ۲۰۰۲، ۲۰۰۹، ۲۰۱۳ و ۲۰۱۵ دارای شکست در لایه‌ای هستند که با دو یا چند توصیه مداخله می‌کند. شکست، خطا یا نقض مکرر حوادث تکراری در سال‌های ۲۰۰۲، ۲۰۰۹، ۲۰۱۳، ۲۰۱۵ و ۲۰۱۶ رویکردی تثبیت نشده است و با مداخلات موثر مسدود نشده است. HFACS و HFIX برای چارچوب بررسی تصادف مفید هستند و از وقوع حادثه مشابه در آینده جلوگیری می‌کنند.

۱۲	Identification of flight accidents causative factors base on SHELLO and improved entropy gray correlation method	۲۰۲۳	Heliyon	Nongtian Chen et al.	روش SHELO و روش همبستگی خاکستری آنتروپی بهبودیافته	به منظور شناسایی موثر عوامل ایجاد کننده اصلی در سوانح پروازی هوانوردی غیرنظامی، بررسی قوانین ایجاد سوانح پرواز، و ایجاد یک مکانیسم پیشگیری آینده‌نگر و موثر برای سوانح پروازی، این مقاله ابتدا براساس مدل SHELLO، طبقه‌بندی عوامل موثر بر سوانح پرواز را بر اساس انسان، سخت افزار، نرم افزار، محیط و سازمان و تعامل با نیازهای تحلیل به وجود آمدن سوانح پروازی ایجاد می‌کند. سپس، با توجه به ویژگی‌های تصادفی و عدم قطعیت فعالیت‌های پروازی هوانوردی غیرنظامی و ویژگی‌های خاکستری عوامل حادثه‌آفرین پرواز، ریتم الگوی همبستگی خاکستری آنتروپی بهبود یافته همراه با ویژگی‌های نمونه داده‌ها برای طبقه‌بندی عوامل موثر ایجاد می‌شود. در نهایت، الگوریتم برای شناسایی و اولویت‌بندی علل سوانح پرواز استفاده می‌شود.	عوامل اصلی که باعث ایجاد پروازهای ناایمن در هوانوردی غیرنظامی چین (۲۰۱۵-۲۰۱۹) می‌شوند به ترتیب عبارتند از: خطاهای ادراکی، خطاهای مبتنی بر مهارت، اشتباهات تصمیم گیری، تخلفات، انحراف اجرای SOP، زمین ناهموار برای صعود و فرود، مکانیسم مدیریت ایمنی ضعیف، شرایط آب و هوایی بد.

۲-۳- جدول شکاف تحقیقاتی

جدول (۲-۲): جدول شکاف تحقیقاتی

ردیف	مرور ادبیات	مطالعه موردی	روش‌های ارزیابی و طبقه‌بندی				یادگیری عمیق	یادگیری ماشین	تکنیک‌های آماری	
			پنیر سوییسی	مبتنی بر چارچوب HFACS	مبتنی بر مدل Shell	شبکه بیزی	NLP		روش تجزیه و تحلیل فراوانی	سایر
۱		❊		❊		❊				
۲		❊	❊	❊						
۳		❊		❊						
۴	❊	❊	❊	❊						
۵		❊			❊		❊	❊		❊
۶		❊				❊	❊	❊ (TAN)		
۷	❊									
۸		❊						❊ (LS , SVM)	❊	❊ (بهینه‌سازی بیزی و fANOVA)
۹		❊			❊				❊	
۱۰		❊		❊						❊ (کای اسکوئر و فیشر)
۱۱		❊	❊	❊						
۱۲		❊			❊					❊
مطالعه ما	❊	❊	❊	❊				❊	❊	**❊ (داده‌کاوی)**

۲-۴- تبیین شکاف تحقیقاتی

همانطور که در جدول شکاف مشاهده می‌شود، با توجه به اهمیت موضوع سوانح هوایی و ارتباط آن با جان افراد، اکثر مقالات از مطالعه موردی برای مطالعه خود استفاده کرده‌اند. روش‌های مبتنی بر چارچوب HFACS در نیمی از مقالات مورد مطالعه، بکار گرفته شده‌اند که نشان از عملکرد مناسب این روش در بررسی حوادث هوایی دارد. مطالعات متعددی به بررسی عوامل انسانی در سوانح هوایی پرداخته‌اند و با روش‌های ارزیابی و طبقه‌بندی مختلف مانند مدل شِل و شبکه بیزی این عوامل را دسته‌بندی کرده‌اند. اما روش‌های مبتنی بر داده‌کاوی و

مدلسازی عوامل با استفاده از یادگیری ماشین و پیش‌بینی متغیرهای مختلف یک حادثه، همچنان نیاز به توسعه دارند. زیرا تکنیک‌های بکار گرفته شده در مقالات، متنوع نبوده و از الگوریتم‌های گسترده‌ای استفاده نشده است.

بنابراین ما در این مطالعه، تلاش کردیم ابتدا به بررسی و داده‌کاوی داده‌های سوانح هوایی که با روش HFACS دسته‌بندی شده‌اند، بپردازیم و در ادامه با استفاده از الگوریتم‌های یادگیری ماشین عوامل انسانی موثر بر سوانح هوایی را مدلسازی کرده و نوع حادثه را پیش‌بینی کنیم.

۳- مطالعه موردی

در این مطالعه، برای تجزیه‌وتحلیل سوانح هوانوردی و پیاده‌سازی الگوریتم‌های یادگیری ماشین، ما از مجموعه‌داده اطلاعات سوانح هوانوردی که توسط هیئت ملی ایمنی حمل و نقل، منتشر شده است، استفاده کرده‌ایم. هیئت ملی ایمنی حمل و نقل یک آژانس تحقیقاتی مستقل دولت ایالات متحده است که مسئول بررسی حوادث حمل‌ونقل غیرنظامی می‌باشد. این مجموعه‌داده در ابتدا شامل ۲۸ ستون (ویژگی[۱۱]) و ۴۷۹ سطر (نمونه[۱۲]) بوده است که پس از انتخاب ویژگی‌ها[۱۳]، پاکسازی داده[۱۴] و کاهش ابعاد[۱۵] مجموعه داده، تعداد ستون‌ها به ۱۳ ستون و سطرها به ۴۷۷ سطر تقلیل یافت.

در ادامه، در جدول (۵-۱) ویژگی‌های مجموعه داده پاکسازی شده، به صورت مختصر توضیح داده شده است.

جدول (۵-۱): تعریف ویژگی‌های مجموعه داده پاکسازی شده

نام ویژگی	تعریف
Performance-Based Errors	خطاهای مبتنی بر عملکرد
Judgment & Decision-Making Errors	خطاهای مبتنی بر قضاوت و تصمیم‌گیری
Violations	خطاهای ناشی از رخ دادن تخلفات
Physical Environment	خطاهای مبتنی بر محیط فیزیکی
Inadequate Supervision	خطاهای مبتنی بر نظارت ناکافی
Technology Failure	خطاهای مبتنی بر شکست فناوری
Acts	خطاهای مبتنی بر اعمال اشتباه
Preconditions	خطاهای مبتنی بر پیش‌بینی
Supervision	خطاهای مبتنی بر نظارت
Organization	خطاهای مبتنی بر سازمان
Fatal or Serious	خطای کشنده یا جدی
Flight Segment ۱=Taxi	بخش پرواز
۹۱=۱/۱۲۱=۰	۱۲۱ پروازهای تجاری و ۹۱ پروازهای غیرتجاری

[۱۱] Feature
[۱۲] Instance (Record)
[۱۳] Feature Selection
[۱۴] Data Cleaning
[۱۵] Dimension Reduction

۴- مبانی نظری

خطای انسانی عامل اصلی حوادث در صنعت هوانوردی است، زیرا قابلیت اطمینان فناوری و ایمنی سیستم دستخوش پیشرفت‌های قابل توجهی شده است. برای مدل‌سازی رویدادهایی که منجر به حوادث خطای انسانی می‌شوند، روش‌های بهبود یافته و بهینه مورد نیاز است. سیستم تجزیه‌وتحلیل و دسته‌بندی عوامل انسانی برای تعریف چارچوبی به کار می‌رود که برای شناسایی مناطق کانونی برای جامعه ایمنی به منظور کاهش خرابی‌های مشابه در آینده در نظر گرفته شده است. این مدل در طبقه بندی عوامل و خطاهای انسانی که منجر به سوانح‌هوایی شده‌اند بسیار کارآمد و نتایج آن بسیار حائز اهمیت است [۱].

۴-۱- مدل پنیر سوئیسی[۱۶]

مدل پنیر سوئیسی یکی از مدل‌های مدیریت و تحلیل خطر از جمله ایمنی حمل و نقل هوایی، مهندسی، بهداشت و درمان، سازمان خدمات اورژانس و به عنوان یکی از اصول ماورای لایه‌های امنیتی مثلا در امنیت کامپیوتر می‌باشد. بر اساس این مدل، خطاها و رویدادها اغلب چندعاملی هستند و مستلزم آن است که به‌طور همزمان یک سری از لایه‌های محافظتی با شکست روبرو شوند و لذا می‌توان هرگونه اشتباه یا خطا را از طریق اقدامات محافظتی سیستم، پرسنل درون سیستم یا هر دو کاهش داد [۱۴].

۴-۲- مدل سیستم دسته‌بندی و تجزیه‌وتحلیل عوامل انسانی[۱۷]

با وجود اینکه روش‌های ارزیابی ریسک و پیشگیری از حوادث در صنایع مختلف به‌کارگرفته شده‌است، اما هنوز حوادث بی‌شماری در صنایع قابل‌مشاهده است. از این رو، اجرای یک روش تجزیه‌وتحلیل حادثه می‌تواند علل ریشه‌ای و سببی حوادث را شناسایی کند. روش سیستم دسته‌بندی و تجزیه‌وتحلیل عوامل انسانی یا همان HFACS با تجزیه‌وتحلیل حوادث گذشته می‌تواند خطاهای انسانی را در صنعت مورد بررسی قراردهد. هدف این پژوهش، شناسایی و تجزیه‌وتحلیل خطاهای انسانی در صنعت هوانوردی و سوانح آن با استفاده از روش HFACS می باشد.

۴-۳- داده‌کاوی

داده‌کاوی، دانش را در یک مجموعه‌داده بزرگ بررسی می‌کند و آن را به یک ساختار قابل درک تبدیل می‌کند. رویکردهای مختلفی بر اساس هدفی که باید به آن دست یافت، وجود دارد. کشف گروه‌هایی از داده‌ها (به عنوان مثال، تجزیه‌وتحلیل خوشه‌ای)، داده‌های غیرمعمول (مانند تشخیص ناهنجاری) و روابط بین متغیرها (به عنوان مثال، قوانین تلازمی) مثال‌هایی از کاربرد داده‌کاوی می‌باشند [۱۵].

۴-۴- یادگیری ماشین[۱۸]

یادگیری ماشین شاخه‌ای از هوش مصنوعی و علوم کامپیوتر است که بر استفاده از داده‌ها و الگوریتم‌ها برای تقلید از روشی که انسان‌ها یاد می‌گیرند، تمرکز دارد و به تدریج دقت آن را بهبود می‌بخشد. یادگیری ماشین جزء

---

[۱۶] Swiss Cheese Model
[۱۷] HFACS: Human Factor Analysis & Classification System
[۱۸] Machine Learning

مهمی از حوزه رو به رشد علم داده است که از طریق استفاده از روش‌های آماری، الگوریتم‌ها، برای دسته‌بندی یا پیش‌بینی و کشف بینش‌های کلیدی در پروژه‌های داده‌کاوی آموزش داده می‌شوند. این بینش‌ها متعاقباً تصمیم‌گیری را در برنامه‌ها و کسب‌وکارها هدایت می‌کنند و به طور ایده‌آل بر معیارهای رشد کلیدی تأثیر می‌گذارند. الگوریتم‌های یادگیری ماشین از داده‌های ساختاریافته و برچسب‌گذاری‌شده برای پیش‌بینی استفاده می‌کنند. به این معنی که ویژگی‌های خاصی از داده‌های ورودی برای مدل تعریف شده و در جداول سازمان‌دهی می‌شوند. این لزوما به این معنی نیست که از داده‌های بدون ساختار استفاده نمی‌کنند. این فقط به این معنی است که معمولا برای سازماندهی داده‌ها در قالبی ساختاریافته، داده‌ها فرآیند پیش‌پردازش را طی می‌کنند [۱۶].

۴-۴-۱- انواع یادگیری ماشین

- یادگیری بانظارت[۱۹]
- یادگیری بدون نظارت[۲۰]
- یادگیری تقویتی[۲۱]

۴-۴-۱-۱- الگوریتم‌های یادگیری ماشین بانظارت

این نوع از یادگیری، یک نوع یادگیری ماشین است که در آن الگوریتم از داده‌های برچسب‌دار یاد می‌گیرد. داده برچسب‌گذاری شده، به معنای مجموعه‌داده‌ای است که متغیر هدف مربوطه آن از قبل مشخص است. یادگیری بانظارت دو نوع دارد.

- **دسته‌بندی:** در این نوع از الگوریتم‌ها، کلاس مجموعه‌داده بر اساس متغیر ورودی مستقل پیش‌بینی می‌شود. کلاس مقادیر مقوله‌ای گسسته است. مثلا تصویر حیوان گربه یا سگ است.
- **رگرسیون:** در این نوع از الگوریتم‌ها متغیرهای خروجی پیوسته بر اساس متغیر ورودی مستقل پیش‌بینی می‌شود. برای مثال پیش‌بینی قیمت مسکن بر اساس پارامترهای مختلف مانند سن خانه، فاصله از جاده اصلی، موقعیت مکانی، مساحت و غیره.

۴-۴-۱-۲- الگوریتم‌های یادگیری ماشین بدون‌نظارت

در یادگیری بدون نظارت، الگوریتم باید خود به تنهایی به‌دنبال ساختارهای جالب موجود در داده‌ها باشد. به بیان ریاضی، یادگیری بدون نظارت مربوط به زمانی است که در مجموعه‌داده فقط متغیرهای ورودی وجود داشته باشند و هیچ متغیر داده خروجی موجود نباشد. به این نوع یادگیری، بدون نظارت گفته می‌شود. زیرا برخلاف یادگیری بانظارت، هیچ پاسخ صحیح داده شده‌ای وجود ندارد و ماشین خود باید به دنبال پاسخ باشد.

به بیان دیگر، هنگامی که الگوریتم برای کار کردن از مجموعه داده‌ای بهره گیرد که فاقد داده‌های برچسب‌دار (متغیرهای خروجی) است، از مکانیزم دیگری برای یادگیری و تصمیم‌گیری استفاده می‌کند. به چنین نوع یادگیری، بدون نظارت گفته می‌شود. یادگیری بدون نظارت قابل تقسیم به مسائل خوشه‌بندی و انجمنی است.

---

[۱۹] Supervised Learning
[۲۰] Unsupervised Learning
[۲۱] Reinforcement Learning

- **قوانین انجمنی:** یک مساله یادگیری هنگامی قوانین انجمنی محسوب می‌شود که هدف کشف کردن قواعدی باشد که بخش بزرگی از داده‌ها را توصیف می‌کنند. مثلا شخصی که کالای الف را خریداری کند، تمایل به خرید کالای ب نیز دارد.
- **خوشه‌بندی:** یک مساله هنگامی خوشه‌بندی محسوب می‌شود که قصد کشف گروه‌های ذاتی (داده‌هایی که ذاتا در یک گروه خاص می‌گنجند) در داده‌ها وجود داشته باشد. مثلا، بخش‌بندی مشتریان بر اساس رفتار خرید آن‌ها.

۴-۴-۱-۳- یادگیری تقویتی

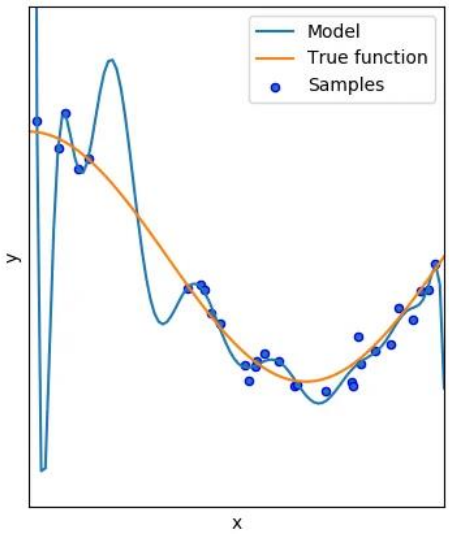
یک برنامه رایانه‌ای که با محیط پویا در تعامل است باید به هدف خاصی دست‌یابد (مانند بازی کردن با یک رقیب یا راندن خودرو). این برنامه بازخوردهایی را با عنوان پاداش‌ها و تنبیه‌ها فراهم و فضای مساله خود را بر همین اساس هدایت می‌کند. با استفاده از یادگیری تقویتی، ماشین می‌آموزد که تصمیمات مشخصی را در محیطی که دائم در معرض آزمون و خطا است، اتخاذ کند [۱۷].

۴-۴-۲- بیش برازش[۲۲]، کم برازش[۲۳] و برازش مناسب

مدل بیش‌برازش، مدلی بسیار پیچیده برای داده‌ها است. به این معنی که در تحلیل رگرسیونی، مدلی با بیشترین پارامترها ایجاد می‌شود. در چنین حالتی، مدل با تغییرات جهشی سعی در پوشش داده‌های حاصل از نمونه و حتی مقدارهای نویز می‌کند. در حالیکه چنین مدلی باید منعکس‌کننده رفتار جامعه باشد. در این گونه موارد، اگر مدل رگرسیون بدست آمده، برای پیش‌بینی نمونه دیگری به کار رود، مقدارهای پیش‌بینی شده اصلا مناسب به نظر نخواهند رسید.

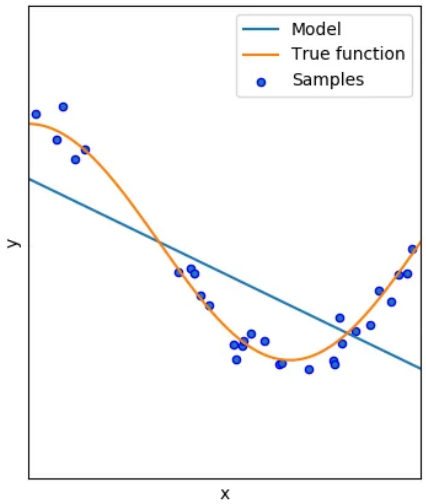
در تصویر زیر، نمودار حاصل از بیش‌برازش روی داده‌های حاصل از نمونه دیده می‌شود. خط آبی، نشان دهنده منحنی برازش شده روی داده‌ها است و خط نارنجی تابعی است که مدل واقعی جامعه آماری را نشان می‌دهد. نقاط آبی رنگ نیز نمونه‌های تصادفی از جامعه آماری را نشان می‌دهند. در مدل بیش‌برازش، نقطه‌های حاصل از نمونه بهترین برازش را دارند و خط آبی تقریباً از همه آن‌ها عبور کرده است.

[۲۲] Overfitting
[۲۳] Underfitting



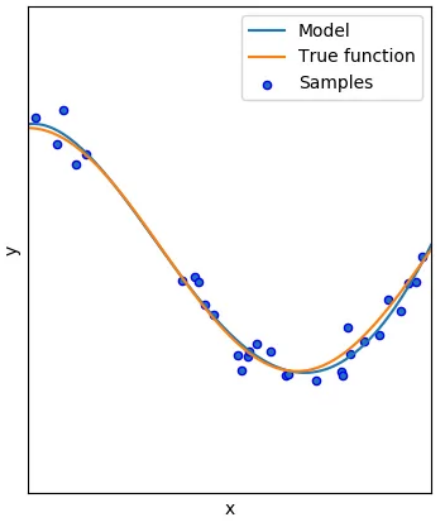
نمودار (۴-۱): منحنی بیش‌برازش بر اساس چندجمله‌ای مرتبه ۱۵

همچنین در زمانی که پارامترهای مدل رگرسیونی به صورت کم برازش برآورد می‌شوند، جانب احتیاط حفظ شده و مدل سعی می‌کند با کمترین پارامترها، عمل برازش را انجام دهد. در نتیجه خطای حاصل از این مدل حتی براساس نمونه‌های به کار رفته نیز بسیار زیاد است. در تصویر زیر، یک نمونه از مدل رگرسیونی کم‌برازش دیده می‌شود. درجه منحنی به کار رفته در این حالت ۱ است که معادله خط محسوب می‌شود.



نمودار (۴-۲): منحنی کم‌برازش براساس چندجمله‌ای مرتبه ۱

انتظار ما از یک تحلیل رگرسیون مناسب، ایجاد مدلی است که نه تنها بتواند برای داده‌های مربوط به نمونه، برازش مناسب را انجام دهد، بلکه برای داده‌هایی جدید نیز امکان برآورد مناسب وجود داشته باشد. همانطور که در تصویر زیر دیده می‌شود، مدل مناسب دارای خطای کوچکی است و قابلیت پیش‌بینی برای داده‌های جدید را دارد [۱۷].



نمودار (۴-۳): منحنی برازش مناسب براساس چندجمله‌ای مرتبه ۴

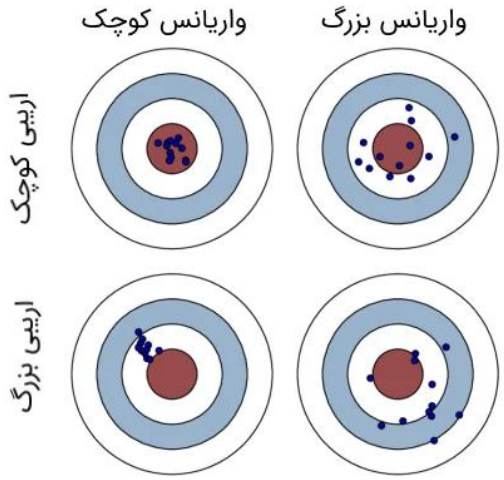
۴-۴-۳- موازنه واریانس و بایاس[۲۴]

خطای بایاس: بایاس در واقع میزان اختلاف نقاط پیش‌بینی شده از متغیر هدف واقعی است. وجود فرضیه‌های مختلف روی مدل و الگوریتم یادگیری منجر به ایجاد خطای اریبی می‌شود. بزرگ بودن اریبی می‌تواند الگوریتم یا مدل آماری را از کشف روابط بین ویژگی‌ها و متغیر پاسخ باز دارد. اغلب بزرگ بودن خطای اریبی، منجر به کم‌برازش می‌شود.

خطای واریانس: واریانس میزان پراکندگی نقاط را نشان می‌دهد. هر چه واریانس بیشتر باشد، پراکندگی داده‌ها بیشتر است. حساسیت زیاد مدل با تغییرات کوچک روی داده‌های آموزشی، نشانگر وجود واریانس زیاد است. این امر نشانگر آن است که اگر مدل آموزش داده شده را روی داده‌های آزمایشی به کارگیریم، نتایج حاصل با داده‌های واقعی فاصله زیادی خواهند داشت. متاسفانه افزایش واریانس در این حالت منجر به مدل‌بندی مقادیر نویز[۲۵] شده و به جای پیش‌بینی صحیح، دچار پیچیدگی و مشکل بیش‌برازش می‌شود [۱۷].

شکل زیر مصورسازی مفهوم موازنه واریانس و بایاس را نشان می‌دهد.

[۲۴] Bias-Variance Tradeoff
[۲۵] Noise



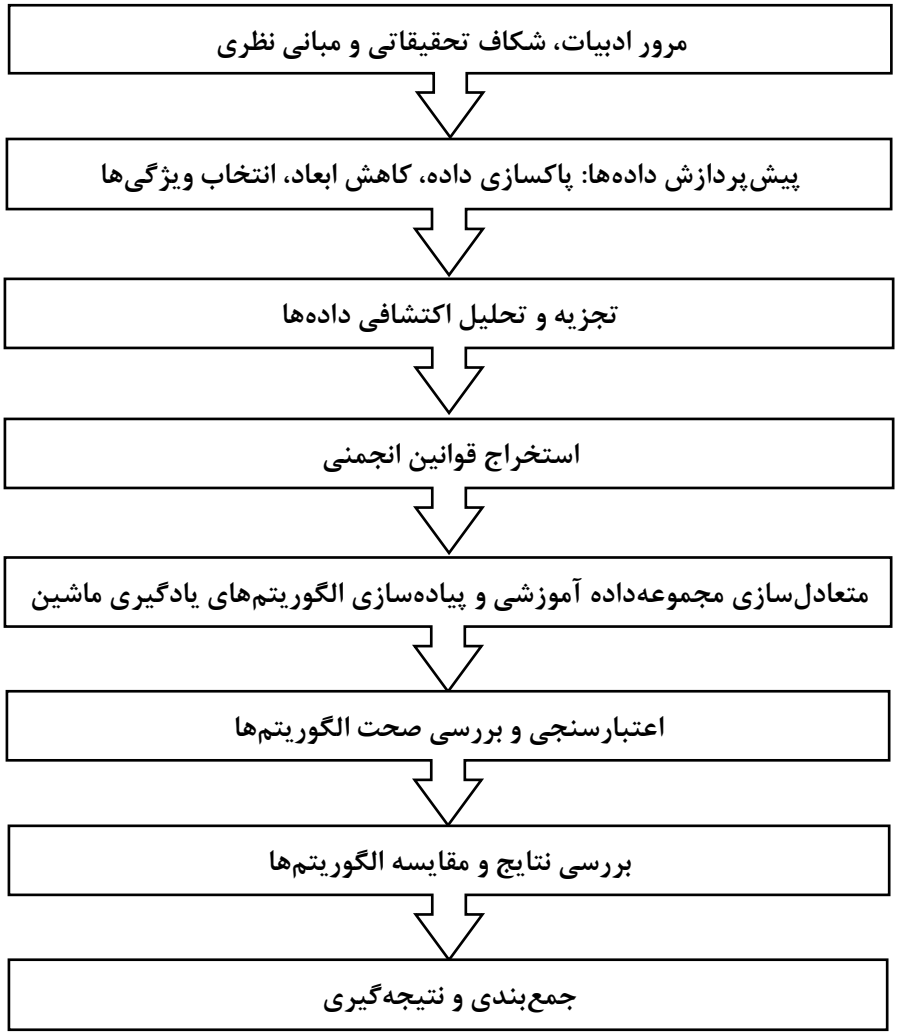
شکل (۴-۱): مصورسازی موازنه واریانس و بایاس برای برآوردگر

۴-۵- کاربرد یادگیری ماشین در سوانح هوایی

در بخش هوانوردی، عوامل انسانی عامل اصلی حوادث ایمنی هستند. سیستم‌های پیش‌بینی هوشمند، که قادر به ارزیابی وضعیت انسانی و مدیریت ریسک هستند، در طول سال‌ها برای شناسایی و پیشگیری از عوامل انسانی توسعه یافته‌اند. یادگیری ماشین اغلب در شرایطی که مشکل دامنه بر اساس ترکیبی از عوامل است، بهتر از دیگر روش‌های موجود عمل می‌کند. از دیگر مواردی که باعث حائز اهمیت بودن این حوزه می‌شود می‌توان به این موضوع اشاره کرد که خطاهای انسانی در این حوزه می‌تواند تبعات جبران ناپذیری داشته باشد که اهم آن‌ها، از دست رفتن جان مسافران و سرنشینان هواپیما است. این مورد باعث می‌شود تا ضرورت تحلیل و پیشبینی‌هایی با دقت بالا در این حوزه حس شود که یادگیری ماشین می‌تواند در این زمینه پاسخگو باشد [۱۸].

۵- روش تحقیق

روش تحقیق این مطالعه، در شکل زیر آمده است.



شکل (۵-۱): روش تحقیق مطالعه

۵-۱- پیش‌پردازش داده‌ها: پاکسازی داده، کاهش ابعاد، انتخاب ویژگی‌ها

مجموعه‌داده توسط هیئت ملی ایمنی حمل‌ونقل ایالات متحده منتشر شده است. دو سطر از داده‌ها که تکراری بوده‌اند، حذف گردید. سطرهایی که داده‌های خالی داشتند، حذف گردید. تعدادی از ستون‌ها به علت اینکه توزیع داده‌های ۰ و ۱ آن‌ها به صورت متعادل و بالانس نبود، حذف گردید تا در قسمت پیش‌بینی بر روی مجموعه‌داده مدل دچار گمراهی و اشتباه نشود.

۵-۲- تجزیه و تحلیل اکتشافی داده‌ها

در این بخش، تجزیه تحلیل فراوانی روی ستون‌های (ویژگی‌های) موجود در مجموعه‌داده صورت گرفته است و با استخراج دانش، پیشنهاداتی برای بهبود ارائه گردیده است. همچنین همبستگی و ارتباط بین ستون‌ها نیز در این بخش مورد بررسی قرار گرفته است.

۵-۳- استخراج قوانین انجمنی

استخراج قوانین انجمنی، تکنیکی است که برای کشف روابط پنهان بین متغیرها در مجموعه‌داده‌های بزرگ استفاده می‌شود. این یک روش محبوب در داده‌کاوی و یادگیری ماشین است و کاربردهای گسترده‌ای در زمینه‌های مختلف مانند تجزیه‌وتحلیل سبد بازار، تقسیم‌بندی مشتریان و کشف تقلب دارد. هدف از استخراج قوانین انجمنی کشف قوانینی است که روابط بین متغیرهای مختلف در مجموعه‌داده را توصیف می‌کند.

برای مثال، مجموعه داده‌ای از سوانح هوایی را در نظر بگیرید. استخراج قوانین انجمنی می‌تواند برای شناسایی روابط بین عواملی که سبب بروز حادثه می‌شود، استفاده شود. برای مثال، قانون «اگر خطای مبتنی بر اعمال نادرست اتفاق بیوفتد، احتمالا خطای مبتنی بر عملکرد نیز اتفاق می‌افتد.» یک قانون انجمنی است که می‌تواند از این مجموعه‌داده استخراج شود. ما می‌توانیم از چنین قوانینی برای اطلاع از تصمیم‌گیری‌ها در مورد آموزش، محل‌قرارگیری تجهیزات و مواردی از این دست استفاده کنیم.

الگوریتم‌های مختلفی برای استخراج قوانین انجمنی وجود دارند. در ادامه به پرکاربردترین آن‌ها اشاره می‌شود.

- **الگوریتم Apriori**: الگوریتم Apriori یکی از پرکاربردترین الگوریتم‌ها برای استخراج قوانین انجمنی است. این الگوریتم، ابتدا مجموعه موارد پرتکرار در مجموعه‌داده را شناسایی می‌کند (مجموعه‌هایی که در تعداد معینی از رکوردها ظاهر می‌شوند). سپس از این مجموعه موارد پرتکرار، برای تولید قوانین انجمنی استفاده می‌کند. الگوریتم Apriori از یک رویکرد پایین به بالا استفاده می‌کند، که از موارد جداگانه شروع می‌شود و به تدریج به مجموعه‌های موارد پیچیده‌تر می‌رسد.
- **الگوریتم FP-Growth** [۲۶]: الگوریتم رشد الگوی پرتکرار، یکی دیگر از الگوریتم‌های محبوب برای استخراج قوانین انجمنی است که با ساختن یک ساختار درخت مانند به نام FP-tree کار می‌کند که مجموعه موارد پرتکرار در مجموعه‌داده را رمزگذاری می‌کند. سپس از FP-tree برای ایجاد قوانین انجمنی به روشی مشابه الگوریتم Apriori استفاده می‌شود. الگوریتم رشد الگوری پرتکرار، به طور کلی سریعتر از الگوریتم Apriori است.
- **الگوریتم ECLAT** [۲۷]: الگوریتم خوشه‌بندی کلاس هم‌ارز و پیمایش شبکه از پایین به بالا، نوعی از الگوریتم Apriori است که از رویکرد بالا به پایین به جای رویکرد از پایین به بالا استفاده می‌کند. با تقسیم موارد به کلاس‌های معادل بر اساس پشتیبانی آن‌ها (تعداد رکوردهایی که در آن‌ها ظاهر می‌شوند) کار می‌کند. سپس قوانین انجمنی با ترکیب این کلاس‌های هم‌ارزی در یک ساختار شبکه مانند ایجاد می‌شود. این یک نسخه کارآمدتر و مقیاس پذیرتر از الگوریتم Apriori است.

در این مطالعه ما از الگوریتم Apriori برای استخراج قوانین انجمنی استفاده کرده‌ایم. در ادامه نحوه عملکرد این الگوریتم را بیان می‌کنیم.

[۲۶] Frequent Pattern Growth
[۲۷] Equivalence Class Clustering and bottom-up Lattice Traversal

۵-۳-۱- الگوریتم Apriori

این الگوریتم، با تنظیم حداقل آستانه پشتیبانی[۲۸] شروع می‌شود. این عدد حداقل تعداد دفعاتی است که یک مورد باید در پایگاه داده تکرار شود تا بتوان آن را به عنوان مجموعه موارد پرتکرار در نظر گرفت. سپس الگوریتم هر مجموعه مواردی را که حداقل آستانه پشتیبانی را برآورده نمی‌کنند، فیلتر می‌کند.

سپس الگوریتم لیستی از تمام ترکیبات ممکن از مجموعه موارد پرتکرار ایجاد می‌کند و تعداد دفعاتی که هر ترکیب در پایگاه داده ظاهر می‌شود را می‌شمارد. سپس الگوریتم فهرستی از قوانین مرتبط را بر اساس ترکیبات پرتکرار مجموعه موارد تولید می‌کند.

قدرت[۲۹] قوانین انجمنی با استفاده از معیار اطمینان[۳۰] اندازه‌گیری می‌شود، که احتمال وجود مورد ب با توجه به وجود مورد الف است. سپس الگوریتم، قوانین انجمنی را که حداقل آستانه اطمینان را برآورده نمی‌کند، فیلتر می‌کند. از این قوانین به عنوان قوانین انجمنی قوی یاد می‌شود. در نهایت، الگوریتم لیستی از قوانین مرتبط قوی را به عنوان خروجی برمی‌گرداند. در ادامه به معیارهای ارزیابی و تحلیل قوانین انجمنی می‌پردازیم.

۵-۳-۲- معیارهای ارزیابی قوانین انجمنی

در استخراج قوانین انجمنی، معمولا از چندین معیار برای ارزیابی کیفیت و اهمیت قوانین کشف شده استفاده می‌شود. این معیارها را می‌توان برای ارزیابی کیفیت و اهمیت قوانین مرتبط و انتخاب مناسب‌ترین قوانین برای یک کاربرد خاص مورد استفاده قرار داد.

تفسیر نتایج معیارهای استخراج قوانین انجمنی مستلزم درک معنا و مفاهیم هر معیار و همچنین نحوه استفاده از آن‌ها برای ارزیابی کیفیت و اهمیت قوانین مرتبط کشف شده است. در اینجا چند دستورالعمل برای تفسیر نتایج معیارهای استخراج قانون انجمنی اصلی آورده شده است.

- **Support**: پشتیبانی معیاری است که نشان می‌دهد یک مورد یا مجموعه موارد به دفعات در مجموعه‌داده ظاهر می‌شود و به شکل زیر محاسبه می‌گردد. پشتیبانی زیاد نشان می‌دهد که یک مورد یا مجموعه موارد در مجموعه‌داده پرتکرار است، در حالی که پشتیبانی کم نشان‌دهنده نادر بودن آن است.

$$Support(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ of\ transactions}$$

- **Confidence**: معیاری برای سنجش قدرت ارتباط بین دو مورد است. به عنوان تعداد رکوردهای حاوی هر دو مورد تقسیم بر تعداد رکوردهای حاوی اولین مورد محاسبه می‌شود. اطمینان بالا نشان می‌دهد که وجود مورد اول یک پیش‌بینی‌کننده قوی برای حضور مورد دوم است.

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

[۲۸] Minimum Support Threshold
[۲۹] Strength
[۳۰] Confidence

- **Lift:** این معیار اندازه‌گیری قدرت ارتباط بین دو مورد با در نظرگرفتن فراوانی هر دو مورد در مجموعه‌داده است. به عنوان معیار اطمینان تقسیم بر معیار پشتیبانی مورد دوم محاسبه می‌شود. این معیار، برای مقایسه قدرت ارتباط بین دو مورد با قدرت مورد انتظار انجمن در صورتی که موارد مستقل باشند، استفاده می‌شود. مقدار بیشتر از ۱ نشان می‌دهد که ارتباط بین دو مورد قوی‌تر از حد انتظار بر اساس فراوانی اقلام است. این نشان می‌دهد که این ارتباط ممکن است معنی‌دار باشد و ارزش بررسی بیشتر را داشته باشد. مقدار کمتر از ۱ نشان می‌دهد که ارتباط ضعیف‌تر از حد انتظار است و احتمالا کمتر قابل توجه باشد.

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}{Fraction\ of\ transactions\ containing\ Y}$$

۵-۴- پیاده‌سازی الگوریتم‌های یادگیری ماشین

۵-۴-۱- آماده‌سازی مجموعه‌داده برای پیاده‌سازی الگوریتم‌های یادگیری ماشین

از آن‌جایی که ستون‌هایی که واریانس کمی دارند در مدل‌های پیش‌بینی تاثیر چندانی ندارند، برای کاهش زمان پردازش مدل، از مجموعه داده حذف شدند.

۵-۴-۲- تقسیم مجموعه داده به سه قسمت آموزشی[۳۱]، اعتبارسنجی[۳۲] و تست[۳۳]

پس از آماده‌سازی مجموعه‌داده، برای پیاده‌سازی الگوریتم‌های یادگیری ماشین، مجموعه‌داده را به سه قسمت آموزشی، اعتبارسنجی و تست تبدیل کرده‌ایم. ۷۰ درصد مجموعه‌داده، به‌عنوان مجموعه‌داده آموزشی، ۱۰ درصد برای اعتبارسنجی اولیه هر مدل و ۲۰ درصد از مجموعه‌داده به‌عنوان مجموعه‌داده تست در نظر گرفته شده‌است که این مجموعه‌داده برای اندازه‌گیری صحت و دقت مدل‌ها کنار گذاشته می‌شود و مدل در نهایت روی آن آزمایش می‌گردد.

۵-۴-۳- متعادل‌سازی مجموعه‌داده آموزشی

با توجه به اینکه داده‌ها به صورت متعادل در متغیر هدف توزیع نشده‌اند، ما روی مجموعه‌داده آموزشی فرآیند متعادل‌سازی را انجام دادیم و در نهایت با ارزشیابی با داده‌های واقعی مدل‌ها را ارزیابی کردیم.

یک مجموعه‌داده نامتعادل را می‌توان به عنوان «یک مشکل مدل‌سازی پیش‌بینی دسته‌بندی که در آن توزیع مثال‌ها در بین کلاس‌ها برابر نیست» تعریف کرد. یعنی توزیع کلاس مساوی یا نزدیک به مساوی نیست و در عوض نامتعادل یا منحرف است. این موضوع می‌تواند سبب انحراف در پیش‌بینی الگوریتم‌های یادگیری ماشین استاندارد شود.

این مشکل در سناریوهایی که تشخیص ناهنجاری بسیار مهم است مانند شناسایی بیماری‌های نادر، تراکنش‌های متقلبانه در بانک‌ها، شناسایی نرخ ریزش مشتری (یعنی چه کسری از مشتریان به استفاده از خدمات ادامه می‌دهند) و غیره کاربرد دارد.

---

[۳۱] Train Dataset
[۳۲] Validation Dataset
[۳۳] Test Dataset

در این مطالعه، ما از روش نمونه‌برداری مجدد برای متعادل‌سازی مجموعه‌داده آموزشی استفاده کرده‌ایم که به صورت مختصر به آن می‌پردازیم.

۵-۴-۳-۱- نمونه‌برداری مجدد از مجموعه‌داده[۳۴]

در این استراتژی، قبل از ارائه داده‌ها به عنوان ورودی به الگوریتم یادگیری ماشین، بر تعادل کلاس‌ها در داده‌های آموزشی تمرکز می‌کنیم. هدف اصلی از متعادل کردن کلاس‌ها افزایش فراوانی دسته اقلیت[۳۵] یا کاهش فراوانی دسته اکثریت[۳۶] است. این کار برای به‌دست آوردن تقریبا همان تعداد نمونه برای هر دو کلاس انجام می‌شود. دو رویکرد نمونه‌برداری برای ایجاد یک مجموعه‌داده متعادل از یک مجموعه نامتعادل وجود دارد که در ادامه به آن‌ها می‌پردازیم [۱۷].

- **نمونه‌برداری کم**[۳۷]: نمونه‌گیری کم، مجموعه‌داده را با کاهش اندازه کلاس بیشتر متعادل می‌کند. این روش زمانی استفاده می‌شود که مقدار داده کافی باشد. با نگه‌داشتن تمام نمونه‌ها در کلاس نادر و انتخاب تصادفی تعداد مساوی از نمونه‌ها در کلاس بیشتر، می‌توان یک مجموعه‌داده جدید متعادل برای مدل‌سازی بازیابی کرد.
- **نمونه‌برداری بیش از حد**[۳۸]: در مقابل، نمونه‌برداری بیش از حد زمانی استفاده می‌شود که کمیت داده‌ها کافی نباشد. این روش، سعی می‌کند با افزایش اندازه نمونه‌های کمیاب، مجموعه‌داده را متعادل کند.

در این مطالعه، ما از روش نمونه‌برداری بیش از حد برای نمونه‌برداری مجدد از مجموعه‌داده و متعادل‌سازی مجموعه‌داده آموزشی استفاده کرده‌ایم. شکل زیر تفاوت دو رویکرد نمونه‌برداری مجدد را نشان می‌دهد.

شکل (۵-۲): تفاوت دو ریکرد نمونه‌برداری مجدد برای متعادل‌سازی داده‌ها

۵-۴-۴- پیاده‌سازی الگوریتم‌های یادگیری ماشین

برای پیش‌بینی کشنده یا جدی‌بودن حوادث از الگوریتم‌های دسته‌بندی بانظارت زیر استفاده شده است. عدد ۱ به معنی کشنده‌بودن حادثه و عدد ۰ به معنی غیرکشنده‌بودن حادثه می‌باشند.

---

[۳۴] Resampling The Dataset
[۳۵] Minority Class
[۳۶] Majority Class
[۳۷] Under-sampling
[۳۸] Over-sampling

۵-۴-۴-۱- رگرسیون لجستیک[۳۹]

در این مدل، احتمالاتی که نتایج احتمالی یک رویداد را توصیف می‌کنند، با استفاده از یک تابع لجستیک مدل‌سازی می‌شوند. رگرسیون لجستیک، یک مدل آماری رگرسیون برای متغیرهای وابسته دوسویی مانند بیماری یا سلامت و مرگ یا زندگی است. این مدل را می‌توان به عنوان مدل خطی تعمیم‌یافته‌ای که از تابع لوجیت[۴۰] به عنوان تابع پیوند استفاده می‌کند و خطایش از توزیع چندجمله‌ای پیروی می‌کند، به‌حساب‌آورد. منظور از دوسویی بودن، رخ داد یک واقعه تصادفی در دو موقعیت ممکنه است. در این مطالعه متغیر هدف یعنی کشنده یا غیرکشنده‌بودن حادثه نیز یک متغیر وابسته دوسویی است. رابطه‌ی زیر، رابطه تابع لوجیت را نشان می‌دهد.

$$\operatorname{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i},$$

همچنین p در رابطه فوق برابر است با:

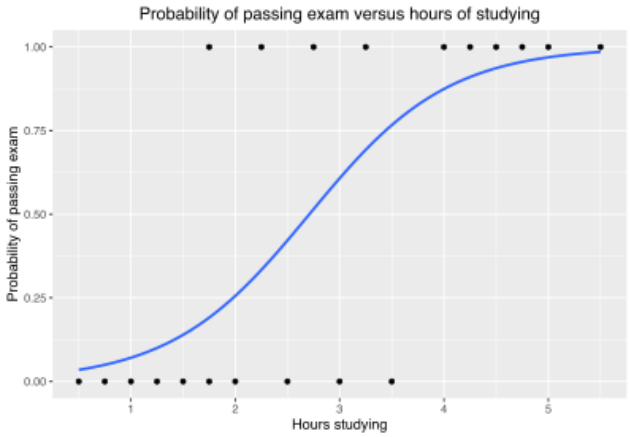
$$p = \Pr(y_i = 1).$$

$$p = \Pr(y_i = 1|\overrightarrow{x_i}; \vec{\beta}) = \frac{e^{\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}}}{1 + e^{\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i})}}.$$

در نهایت رگرسیون لجستیک را می‌توان به شکل زیر بازنویسی کرد.

$$\Pr(y_i = 1|\overrightarrow{x_i}; \vec{\beta}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i})}} = \sigma\left(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}\right)$$

نمودار زیر، یک منحنی رگرسیون لجستیک را نشان می‌دهد که احتمال قبولی در امتحان را در مقابل ساعات مطالعه مورد بررسی قرار داده است [۱۹].



نمودار (۵-۱): مثالی از منحنی رگرسیون لجستیک

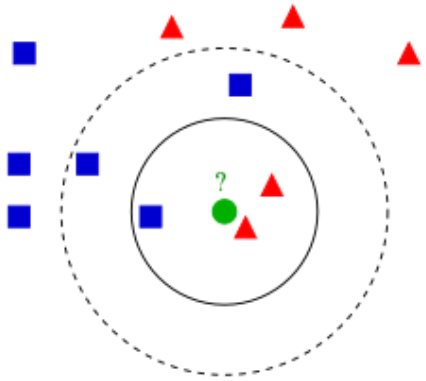
۵-۴-۴-۲- K نزدیک‌ترین همسایه[۴۱]

روش K نزدیک‌ترین همسایه، که روش آماری مبتنی بر فاصله است که برای دسته‌بندی آماری و رگرسیون استفاده می‌شود. در هر دو حالت، K شامل نزدیک‌ترین مثال آموزشی در فضای داده‌ای می‌باشد و خروجی آن بسته به نوع مورد استفاده در دسته‌بندی و رگرسیون متغیر است.

---

[۳۹] Logistic Regression
[۴۰] logit
[۴۱] K-Nearest Neighbors

در حالت طبقه‌بندی با توجه به مقدار مشخص شده برای K، به محاسبه فاصله نقطه‌ای که می‌خواهیم برچسب آن را مشخص کنیم با نزدیک‌ترین نقاط می‌پردازد و با توجه به تعداد رای حداکثری این نقاط همسایه، در رابطه با برچسب نقطه مورد نظر تصمیم‌گیری صورت می‌گیرد. برای محاسبه این فاصله می‌توان از روش‌های مختلفی استفاده کرد که یکی از مطرح‌ترین این روش‌ها، فاصله اقلیدسی است. در حالت رگرسیون نیز میانگین مقادیر به‌دست آمده از K خروجی آن می‌باشد. از آن‌جا که محاسبات این الگوریتم بر اساس فاصله است نرمال‌سازی داده‌ها می‌تواند به بهبود عملکرد آن کمک کند. شکل زیر، مثالی از پیاده‌سازی این الگوریتم را نشان می‌دهد.



شکل (۵-۳): مثالی از پیاده‌سازی الگوریتم K نزدیک همسایه برای طبقه‌بندی داده‌ها

در شکل فوق، نقطه سبز رنگ نمونه تست می‌باشد که باید به مربع‌های آبی یا قرمز دسته‌بندی شود. اگر K=۳ در نظر گرفته شود، این داده به مثلث‌های قرمز نسبت داده می‌شود. اما اگر K=۵ باشد، به مربع‌های آبی اختصاص داده می‌شود [۱۹].

۵-۴-۴-۳- درخت تصمیم‌گیری[۴۲]

درخت تصمیم، یک مدل سلسله مراتبی پشتیبانی تصمیم است که از یک مدل درخت مانند از تصمیمات و پیامدهای احتمالی آن‌ها، از جمله نتایج رویدادهای شانسی، هزینه‌های منابع و مطلوبیت استفاده می‌کند. درخت‌های تصمیم معمولا در تحقیقات عملیاتی، به‌ویژه در تجزیه‌وتحلیل تصمیم‌گیری برای کمک به شناسایی استراتژی که به احتمال زیاد به یک هدف می‌رسد، استفاده می‌شوند.

درخت تصمیم دارای اجزای زیر است:

- **گره اصلی**[۴۳]: ویژگی کلیدی در مجموعه‌داده
- **گره داخلی**[۴۴]: گره‌هایی که یک یال ورودی و دو یا چند یال خروجی دارند.
- **گره برگ**[۴۵]: گره پایانی بدون یال خروجی

درخت تصمیم از یک گره اصلی شروع می‌شود و با بررسی شرایط مختلف و اختصاص آن به سایر گره‌ها ادامه می‌یابد. درخت تصمیم زمانی کامل می‌شود که تمام شرایط به یک گره برگ منتهی شوند. گره برگ حاوی برچسب دسته‌بندی می‌باشد.

---

[۴۲] Decision Tree
[۴۳] Root Node
[۴۴] Internal Node
[۴۵] Leaf Node

برای تقسیم بهینه ویژگی‌ها دو روش وجود دارد:

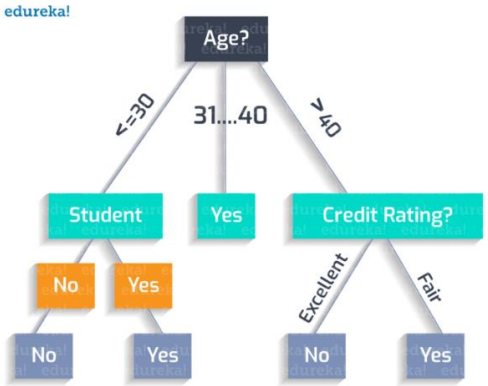
- **روش شاخص جینی**[۴۶]: ناخالصی جینی تعداد برچسب‌گذاری اشتباه هر عنصر مجموعه‌داده را هنگامی که به طور تصادفی برچسب‌گذاری می‌شود، اندازه‌گیری می‌کند. در شکل زیر فرمول شاخص جینی مشاهده می‌شود که در آن p_j احتمال کلاس j است. حداقل مقدار شاخص جینی ۰ است که زمانی اتفاق می‌افتد که گره خالص باشد، به این معنی که تمام عناصر موجود در گره از یک کلاس منحصربه‌فرد هستند. بنابراین، این گره دوباره تقسیم نخواهد شد. تقسیم بهینه توسط ویژگی‌هایی با شاخص جینی کمتر انتخاب می‌شود. علاوه بر این، زمانی که احتمال دو کلاس یکسان باشد، شاخص جینی حداکثر مقدار (۰.۵) را دریافت می‌کند.

$$GiniIndex = 1 - \sum_j p_j^2$$

- **روش آنتروپی**[۴۷]: آنتروپی معیاری از اطلاعات است که نشان‌دهنده بی نظمی ویژگی‌ها با متغیر هدف است. مشابه شاخص جینی، تقسیم بهینه توسط ویژگی با آنتروپی کمتر انتخاب می‌شود. مقدار آنتروپی زمانی حداکثر مقدار خود (۱) را به‌دست می‌آورد که احتمال دو کلاس یکسان باشد و هنگامی که یک گره خالص باشد، مقدار آنتروپی حداقل مقدار خود یعنی ۰ است. فرمول محاسبه آنتروپی به شکل زیر است که در آن p_j احتمال کلاس j است.

$$Entropy = -\sum_j p_j \cdot log_2 \cdot p_j$$

شکل زیر نمونه‌ای از یک درخت تصمیم با تقسیم ویژگی‌ها با روش شاخص جینی را نشان می‌دهد که هدف آن پیش‌بینی خرید لپ‌تاپ توسط کاربر می‌باشد. همانطور که مشاهده می‌شود، ویژگی سن[۴۸] به عنوان گره اصلی انتخاب شده است و سایر ویژگی‌ها در گره‌های داخلی قرار دارند و با بررسی شرایط مختلف، گره‌های برگ مشخص شده‌اند [۱۹].

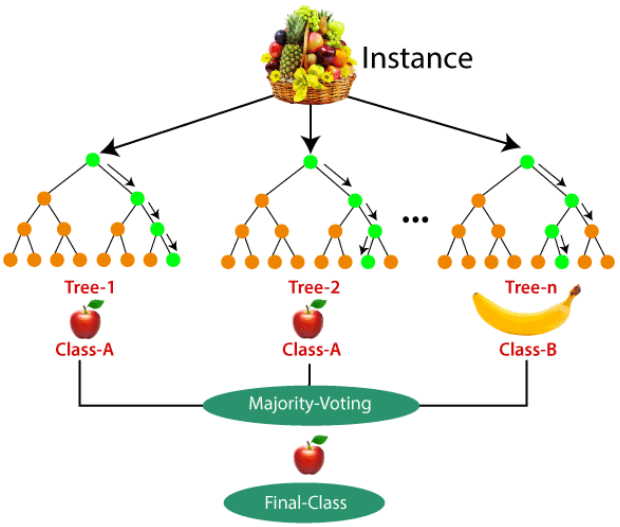


شکل (۵-۴): مثالی از پیاده‌سازی الگوریتم درخت تصمیم برای طبقه‌بندی داده‌ها

[۴۶] Gini Index
[۴۷] Entropy
[۴۸] Age

۵-۴-۴-۴- جنگل تصادفی[۴۹]

جنگل تصادفی، یک روش یادگیری ترکیبی برای دسته‌بندی و رگرسیون می‌باشد، که بر اساس ساختاری متشکل از شمار بسیاری درخت تصمیم، در زمان آموزش عمل می‌کند. عملکرد الگوریتم جنگل تصادفی معمولا بهتر از الگوریتم درخت تصمیم است، اما این بهبود عملکرد تا حدی به نوع داده هم بستگی دارد. برای کاربرد دسته‌بندی، خروجی جنگل تصادفی کلاسی است که توسط اکثر درختان انتخاب شده است. شکل زیر مثالی ساده از پیاده‌سازی الگوریتم جنگل تصادفی را بر روی نمونه‌ای از میوه‌ها نشان می‌دهد [۱۹].



شکل (۵-۵): مثالی از پیاده‌سازی الگوریتم جنگل تصادفی برای دسته‌بندی داده‌ها

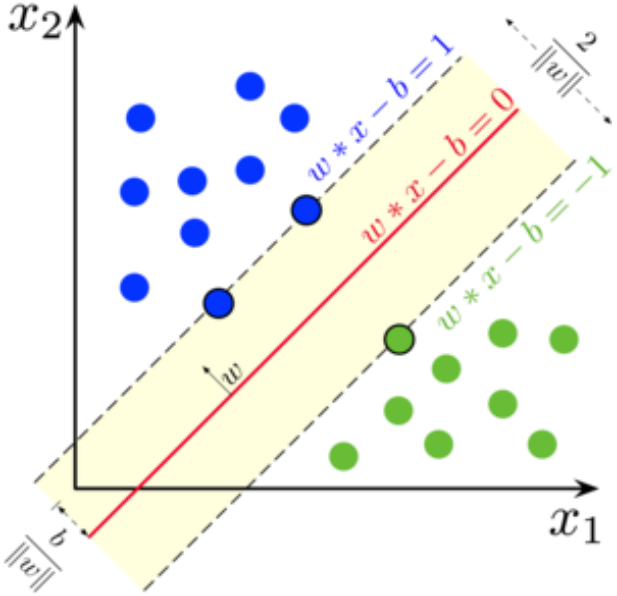
۵-۴-۴-۵- ماشین بردار پشتیبان[۵۰]

ماشین بردار پشتیبان، یکی از روش‌های یادگیری بانظارت است که از آن برای طبقه‌بندی و رگرسیون استفاده می‌شود. مبنای کاری دسته‌بندی این الگوریتم، دسته‌بندی خطی داده‌ها است و در تقسیم خطی داده‌ها، سعی می‌کند ابرصفحه‌ای را انتخاب کند که حاشیه اطمینان بیشتری داشته باشد. حل معادله پیدا کردن خط بهینه برای داده‌ها به وسیله روش‌های برنامه‌ریزی غیرخطی که روش‌های شناخته شده‌ای در حل مسائل محدودیت‌دار هستند، صورت می‌گیرد. قبل از تقسیمِ خطی برای اینکه ماشین بتواند داده‌های با پیچیدگی بالا را دسته‌بندی کند، داده‌ها به وسیله‌ی تابعِ فی[۵۱] به فضای با ابعاد خیلی بالاتر برده می‌شود. برای اینکه بتوان مسئله ابعاد خیلی بالا را با استفاده از این روش‌ها حل کنیم، از قضیه دوگانی لاگرانژ[۵۲] برای تبدیلِ مسئله‌ی مینیمم‌سازی مورد نظر به فرم دوگانی آن که در آن به جای تابع پیچیده‌ی فی که ما را به فضایی با ابعاد بالا می‌برد، تابعِ ساده‌تری به نامِ تابع هسته[۵۳] که ضرب برداری تابع فی است ظاهر می‌شود، استفاده می‌کنیم. از توابع هسته مختلفی از جمله

---

[۴۹] Random Forest
[۵۰] Support Vector Machine
[۵۱] Phi Function
[۵۲] Lagrange Duality Theorems
[۵۳] Kernel

هسته‌های نمایی، چندجمله‌ای و سیگموید[۵۴] می‌توان بدین منظور استفاده نمود. شکل زیر، مثالی از عملکرد الگوریتم ماشین بردار پشتیبان را نشان می‌دهد. به ابرصفحه‌های حاشیه، بردارهای پشتیبان گفته می‌شود [۱۹].



شکل (۵-۶): ابرصفحه‌ای با حداکثر حاشیه برای یک ماشین بردار پشتیبان که با نمونه داده‌هایی از دو دسته یادگرفته شده‌است.

۵-۴-۴-۶- بیز ساده برنولی[۵۵]

بیز ساده برنولی، یکی از الگوریتم‌های بیز ساده است که اساس توزیع برنولی می‌باشد و فقط مقادیر دودویی، یعنی ۰ یا ۱ را می‌پذیرد. اگر ویژگی‌های مجموعه‌داده دودویی[۵۶] باشند، می‌توان از این الگوریتم استفاده کرد. فرمول بیز ساده برنولی به شکل زیر است [۱۹].

$$P(x_i \mid y) = P(x_i = ۱ \mid y)x_i + (۱ - P(x_i = ۱ \mid y))(۱ - x_i)$$

۵-۴-۴-۷- دسته‌بندی کیسه‌ای[۵۷]

الگوریتم دسته‌بندی کیسه‌ای یک فرابرآوردگر[۵۸] ترکیبی[۵۹] است که هر کدام از طبقه‌بندی‌کننده‌های پایه را بر روی زیرمجموعه‌های تصادفی مجموعه‌داده اصلی قرار می‌دهد و سپس پیش‌بینی‌های فردی آن‌ها (چه با رأی‌گیری[۶۰] یا با میانگین‌گیری[۶۱]) را جمع‌آوری می‌کند تا یک پیش‌بینی نهایی را تشکیل دهد. این فرابرآوردگر

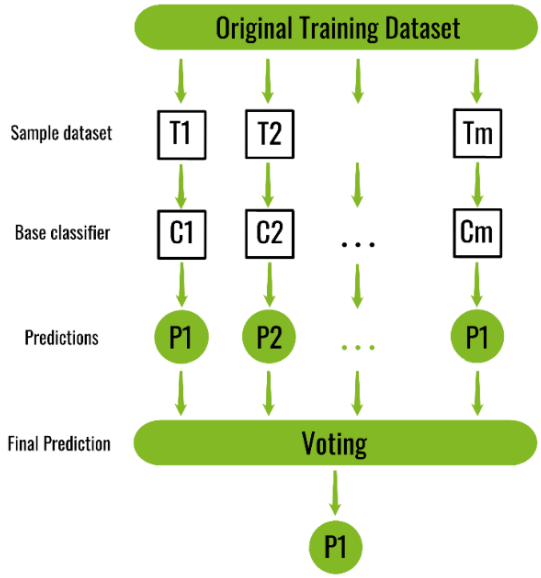
---

[۵۴] Sigmoid Function

تابع سیگموئید تابعی حقیقی، یکنوا، کران‌دار و مشتق‌پذیر است که به ازای کلیه مقادیر حقیقی قابل تعریف بوده دارای مشتق نامنفی است که دارای یک نقطه‌ی عطف است. این تابع به لحاظ گرافیکی شکلی شبیه حرف S انگلیسی و سیگما در یونانی دارد. دامنه توابع سیگموئید شامل تمامی اعداد حقیقی بوده و مقدار بازگشتی این تابع نیز به طور یکنواخت از ۰ تا ۱ یا باتوجه به نوع تابع از ۱ تا ۱- تغییر می‌کند.

[۵۵] Bernoulli Naive Bayes
[۵۶] Binary
[۵۷] Bagging Classifier
[۵۸] Meta-Estimator
[۵۹] Ensemble
[۶۰] Voting
[۶۱] Averaging

معمولا می‌تواند به عنوان راهی برای کاهش واریانس تخمین‌گر جعبه سیاه[۶۲] (به عنوان مثال، درخت تصمیم)، با ورود تصادفی به مراحل ایجاد آن و سپس ساختن مجموعه‌ای از آن استفاده شود [۱۹].



شکل (۵-۷): نحوه عملکرد الگوریتم دسته‌بندی کیسه‌ای

۵-۴-۴-۸- دسته‌بندی تقویتی گرادیان[۶۳]

طبقه‌بندی‌کننده تقویتی گرادیان یک الگوریتم یادگیری ماشین است که بسیاری از مدل‌های یادگیری ضعیف را با هم ترکیب می‌کند تا یک مدل پیش‌بینی قوی ایجاد کند. معمولا هنگام انجام الگوریتم دسته‌بندی تقویتی گرادیان از درختان تصمیم استفاده می‌شود. مدل تقویتی گرادیان به دلیل اثربخشی در طبقه‌بندی مجموعه‌داده‌های پیچیده، محبوب شده‌است.

مدل تقویتی گرادیان ترکیبی خطی از یک سری مدل‌های ضعیف است که به صورت تناوبی برای ایجاد یک مدل نهایی قوی ساخته شده‌است. این روش به خانواده الگوریتم‌های یادگیری گروهی تعلق دارد و عملکرد آن همواره از الگوریتم‌های اساسی یا ضعیف (مثلا درخت تصمیم) یا روش‌های براساس کیسه‌گذاری (مانند جنگل تصادفی) بهتر است اما این موضوع تا حدی از مشخصات داده‌های ورودی تأثیر می‌پذیرد.

روش این الگوریتم بدین ترتیب است که تابع هزینه[۶۴] را به کمینه‌ترین مقدار خود برساند. در علم آمار، معمولا تابع هزینه برای اینکه مشخص شود تخمین پارامترمان تا چه حد موفق بوده، استفاده می‌شود. تابع هزینه، تابعی است که برای سنجش میزان موفقیت تخمینگر از تخمین پارامتر نسبت به مقادیر واقعی از آن استفاده می‌شود. در مسائل دسته‌بندی، تابع هزینه در اصل به نوعی تعداد دسته‌بندی‌های اشتباه توسط تخمینگر را نمایان می‌کند.

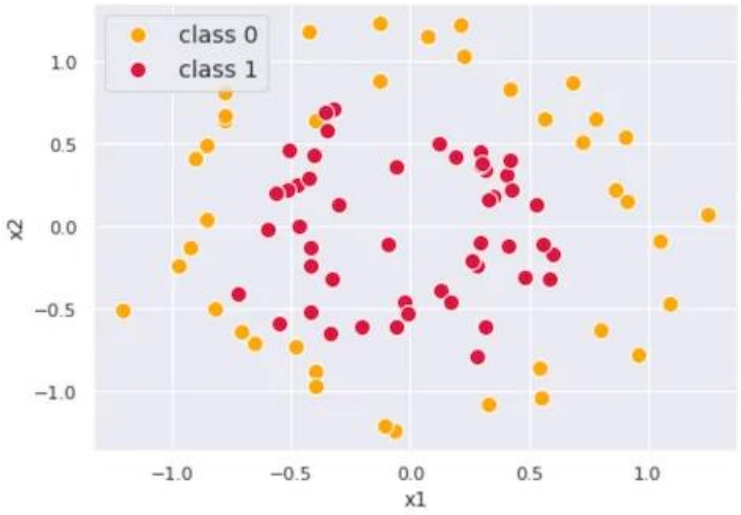
الگوریتم یادگیری تقویتی یک الگوریتم تقویتی قدرتمند است که چندین یادگیرنده ضعیف را با یادگیرندگان قوی ترکیب می‌کند، که در آن هر مدل جدید برای به حداقل رساندن تابع هزینه مانند میانگین مربعات خطا یا آنتروپی متقابل مدل قبلی، با استفاده از گرادیان نزول آموزش داده می‌شود. در هر تکرار، الگوریتم گرادیان تابع

---

[۶۲] Black-Box Estimator
[۶۳] Gradient Boosting Classifier Algorithm
[۶۴] Loss Function

هزینه را با توجه به پیش‌بینی‌های مجموعه فعلی محاسبه می‌کند و سپس یک مدل ضعیف جدید را برای به حداقل رساندن این گرادیان آموزش می‌دهد. سپس پیش‌بینی‌های مدل جدید به مجموعه اضافه می‌شود و این فرآیند تا زمانی که یک معیار توقف برآورده شود، تکرار می‌شود. در این الگوریتم، وزن نمونه‌های آموزشی بهینه‌سازی نمی‌شود. در عوض، هر پیش‌بینی‌کننده با استفاده از خطاهای باقی‌مانده قبلی به عنوان برچسب، آموزش داده می‌شود [۱۹].

در ادامه روند پیاده‌سازی مدل تقویتی گرادیان در قالب یک مثال آموزشی توضیح داده خواهد شد. شکل نمایش داده‌های دسته‌بندی در شکل زیر نشان داده شده است.

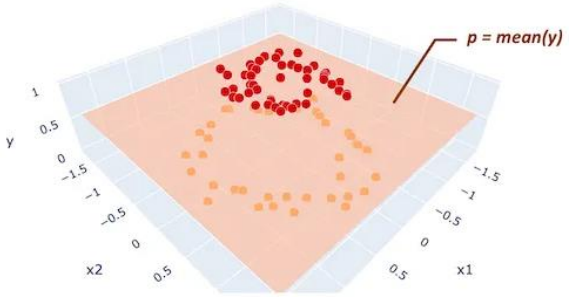


شکل (۵-۸): مثالی از مسئله دسته‌بندی دو کلاسه

هدف ساخت یک مدل تقویتی گرادیان است که داده‌ها را به دو دسته دسته‌بندی کند. اولین گام، ایجاد یک پیش‌بینی یکنواخت بر روی احتمال کلاس ۱ (ما آن را p می‌نامیم) برای تمام نقاط داده است که در واقع همان میانگین کلاس می‌باشد.

$$p = P(y = 1) = \bar{y}$$

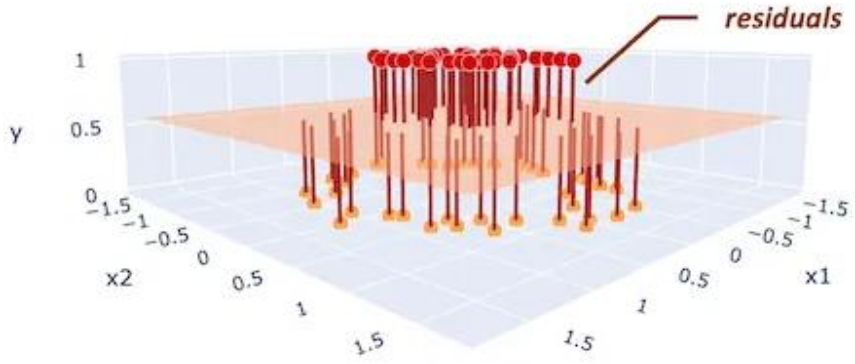
در اینجا یک نمایش سه بعدی از داده‌ها و پیش بینی اولیه آمده است. در این لحظه، پیش‌بینی فقط صفحه‌ای است که همیشه مقدار یکنواخت p = mean(y) را در محور y دارد.



شکل (۵-۹): نمایش صفحه پیش‌بینی به شکل سه‌بعدی

در این مثال، میانگین y، ۰.۵۶ است. از آن‌جایی که بزرگتر از ۰.۵ است، همه چیز با این پیش‌بینی اولیه در کلاس ۱ دسته‌بندی می‌شود. ممکن است به نظر برسد که این پیش‌بینی ارزش یکسان، منطقی نیست، لازم به ذکر است که با اضافه کردن مدل‌های ضعیف بیشتر به آن، پیش‌بینی بهبود می‌یابد.

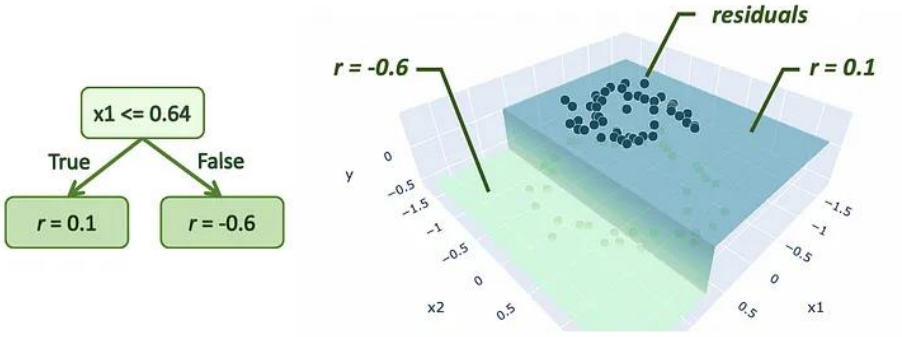
برای بهبود کیفیت پیش‌بینی، ممکن است روی باقی‌مانده‌ها (یعنی خطای پیش‌بینی) از پیش‌بینی اولیه تمرکز کنیم، زیرا این همان چیزی است که باید به حداقل برسد. باقی‌مانده‌ها به صورت $r_i = y_i - p$ تعریف می‌شوند (i نشان دهنده شاخص هر نقطه داده است). در شکل زیر باقیمانده‌ها به صورت خطوط قهوه‌ای نشان داده شده‌اند که خطوط عمود از هر نقطه داده به صفحه پیش‌بینی هستند.



شکل (۵-۱۰): شکل نمایش باقی‌مانده‌ها

برای به حداقل رساندن این باقی‌مانده‌ها، یک مدل درخت رگرسیون با x_1 و x_2 به‌عنوان ویژگی‌های آن و باقی‌مانده r به‌عنوان هدف آن باید ساخته شود. اگر بتوان درختی ساخت که الگوهایی را بین x و r پیدا کند، می‌توان با استفاده از آن الگوهای یافت شده، باقی‌مانده‌های حاصل از پیش‌بینی اولیه p را کاهش داد.

برای ساده‌کردن نمایش، درختان بسیار ساده‌ای که هر کدام فقط دارای یک تقسیم و دو گره برگ هستند، ساخته شده‌اند که به آن «استامپ[۶۵]» می‌گویند. لازم به ذکر است که درخت‌های تقویت‌کننده گرادیان معمولا درختان کمی عمیق‌تر مانند درخت‌هایی با ۸ تا ۳۲ گره برگ دارند. در اینجا ما اولین درخت ایجاد شده باقی‌مانده‌ها را با دو مقدار مختلف (۰،۱,-۰،۶) = r پیش‌بینی می‌کند.



شکل (۵-۱۱): درخت ساخته شده برای متغیرهای x و باقی‌مانده r

[۶۵] Stump

در ادامه گاما طبق فرمول زیر محاسبه می‌شود. مقادیر گاما را به پیش‌بینی اولیه خود اضافه می‌کنیم تا باقی‌مانده‌ها را کاهش دهیم.

$$\gamma_j = \frac{\sum_{x_i \in R_j} (y_i - p)}{\sum_{x_i \in R_j} p(1-p)}$$

γ is computed for each terminal node j

Aggregating for all the data points x_i that belongs to terminal node j

مقادیر گاما ۱ و گاما ۲ بدین ترتیب محاسبه می‌شوند.

$$\gamma_1 = \frac{\sum_{x_i \in R_1} (y_i - 0.56)}{\sum_{x_i \in R_1} 0.56 \cdot (1 - 0.56)} = 0.3$$

$$\gamma_2 = \frac{\sum_{x_i \in R_2} (y_i - 0.56)}{\sum_{x_i \in R_2} 0.56 \cdot (1 - 0.56)} = -2.2$$

برای اینکه گاما را به مقدار p اضافه شود، به شکل زیر عمل می‌کنیم. ابتدا مقدار log(odds) را از p بدست می‌آوریم (به آن F(x) گفته می‌شود). سپس گاما را به آن اضافه می‌کنیم.

$$log(odds) = log\left(\frac{p}{1-p}\right)$$

برای اینکه مدل بیش از حد از آموزش[۶۶] نبیند و خطای آن کاهش نیابد می‌توان مقدار گاما را در یک وزنی (بین ۰ تا ۱) که به آن نرخ یادگیری[۶۷] ν گفته می‌شود، ضرب کرد و سپس به مقدار log(odds) یا همان F(x) اضافه نمود تا پیش‌بینی بروز شود.

$$F_1(x) = F_0(x) + \nu \cdot \gamma$$

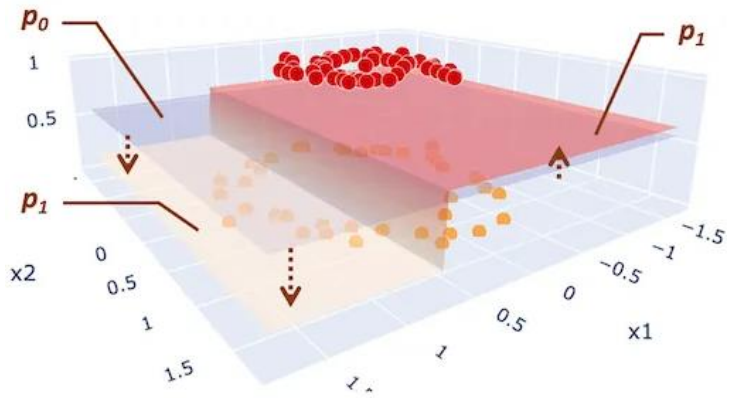
Updated prediction — Initial prediction — Learning rate

در این مثال، ما از نرخ یادگیری نسبتا بزرگ ν = ۰،۹ استفاده می‌کنیم تا فرآیند بهینه‌سازی را آسان‌تر درک کنیم، اما معمولا قرار است مقادیر بسیار کوچک‌تری مانند ۰.۱ در نظر گرفته شود. با جایگزینی مقادیر واقعی برای متغیرهای سمت راست معادله بالا، پیش‌بینی به‌روز $F_1(x)$ را بدست می‌آید.

$$F_1(x) = \begin{cases} log\left(\frac{0.56}{1-0.56}\right) + 0.9 \cdot 0.3 = 0.5 & if\ x_1 \le 0.64 \\ log\left(\frac{0.56}{1-0.56}\right) - 0.9 \cdot 2.2 = -1.7 & otherwise \end{cases}$$

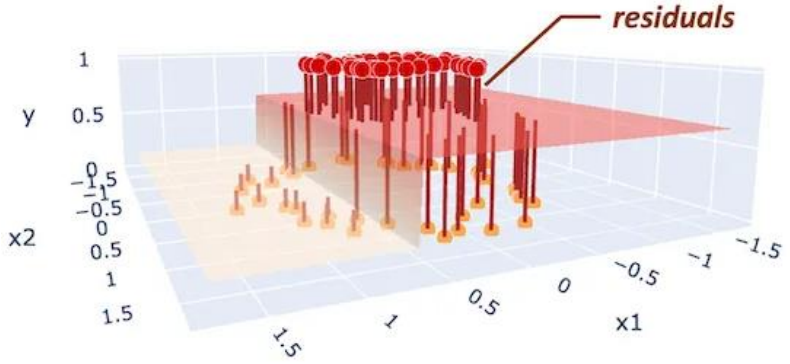
اگر log(odds) را دوباره به p تبدیل کنیم. شکلی پله‌مانند از داده‌ها مانند شکل زیر بدست می‌آید.

[۶۶] Overfit
[۶۷] Learning Rate



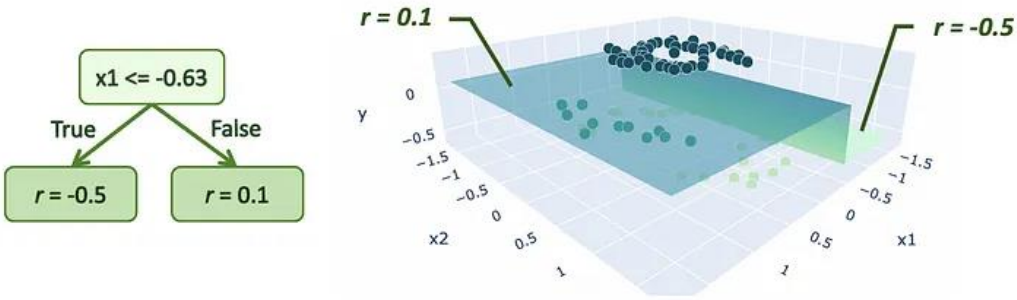
شکل (۵-۱۲): نمایش پیش‌بینی بروز شده

اکنون، باقی‌مانده‌های بروز شده r به شکل زیر است.



شکل (۵-۱۳): نمایش باقی‌مانده‌های بروز شده

مجددا یک درخت رگرسیون با استفاده از همان x_1 و x_2 به عنوان ویژگی‌های ورودی برای باقی‌مانده‌های بروز شده ایجاد می‌کنیم.



شکل (۵-۱۴): درخت ساخته شده برای متغیرهای x و باقی‌مانده r بروز شده

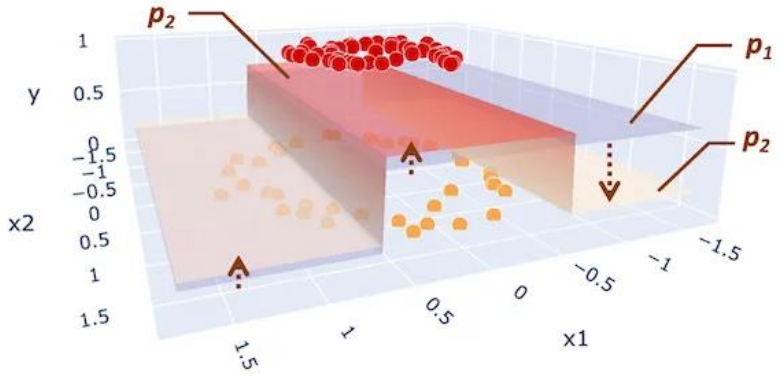
حال مجددا گاما به همان روش قبل محاسبه کرده و $F_2(x)$ را بدست می‌آوریم.

$$F_2(x) = \begin{cases} F_1(x) - \nu \cdot 2.3 = 0.5 - 0.9 \cdot 2.3 = -1.6 & if\ x_1 \le -0.63 \\ F_1(x) + \nu \cdot 0.4 = 0.5 + 0.9 \cdot 0.4 = 0.9 & else\ if\ -0.63 < x_1 \le 0.64 \\ F_1(x) + \nu \cdot 0.4 = -1.7 + 0.9 \cdot 0.4 = -1.3 & otherwise \end{cases}$$

These are γ computed with this formula:

$$\gamma_j = \frac{\sum_{x_i \in R_j} (y_i - p)}{\sum_{x_i \in R_j} p(1-p)}$$

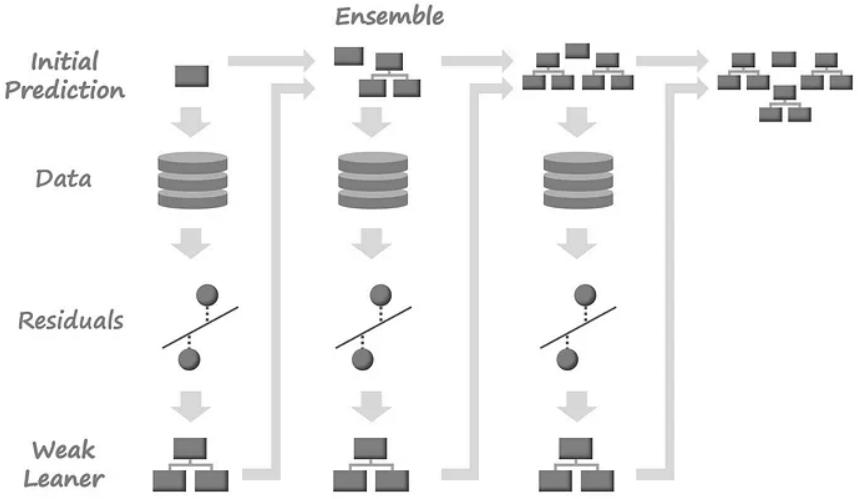
سپس $F_2(x)$ را به $p_2(x)$ تبدیل می‌کنیم و به شکل زیر می‌رسیم.



شکل (۵-۱۵): نمایش پیش‌بینی بروز شده

سپس، این مراحل را تکرار می‌کنیم تا زمانی که پیش‌بینی مدل متوقف شود. در نهایت، می‌توان دید که پیش‌بینی ترکیبی $p(x)$ به هدف ما نزدیک‌تر می‌شود. زیرا درخت‌های بیشتری را به مدل ترکیبی اضافه می‌کنیم. این روشی است که الگوریتم تقویتی گرادیان برای پیش‌بینی اهداف پیچیده با ترکیب چندین مدل ضعیف انجام می‌دهد.

تصویر زیر به طور خلاصه کل فرآیند این الگوریتم را نشان می‌دهد.



شکل (۵-۱۶): فرآیند الگوریتم تقویتی گرادیان

۵-۴-۴-۹- دسته‌بندی XGboost [۶۸]

XGBoost یک الگوریتم یادگیری ماشین مبتنی بر درخت تصمیم است که از یک چارچوب تقویت گرادیان استفاده می‌کند. شکل زیر فرآیند بهینه‌سازی الگوریتم ماشین تقویت گرادیان[۶۹] را توسط XGboost نشان می‌دهد.

شکل (۵-۱۷): نحوه بهینه‌سازی در الگوریتم XGboost

XGBoost مخفف واژه تقویت گرادیان شدید[۷۰] است و به دلیل توانایی آن در مدیریت مجموعه‌داده‌های بزرگ و توانایی آن برای دستیابی به عملکرد پیشرفته در بسیاری از وظایف یادگیری ماشین، به یکی از محبوب‌ترین و پرکاربردترین الگوریتم‌های یادگیری ماشین تبدیل شده است.

در این الگوریتم درخت‌های تصمیم به صورت متوالی ایجاد می‌شوند. وزن‌ها نقش مهمی در XGBoost دارند. وزن‌ها به همه متغیرهای مستقل اختصاص داده می‌شوند که سپس به درخت تصمیم که نتایج را پیش‌بینی می‌کند، وارد می‌شوند. وزن متغیرهای پیش‌بینی‌شده اشتباه توسط درخت، افزایش می‌یابد و این متغیرها سپس به درخت تصمیم دوم تغذیه می‌شوند. سپس این دسته‌بندی‌کننده‌ها یا پیش‌بینی‌کننده‌های منفرد برای ارائه یک مدل قوی و دقیق‌تر جمع می‌شوند.

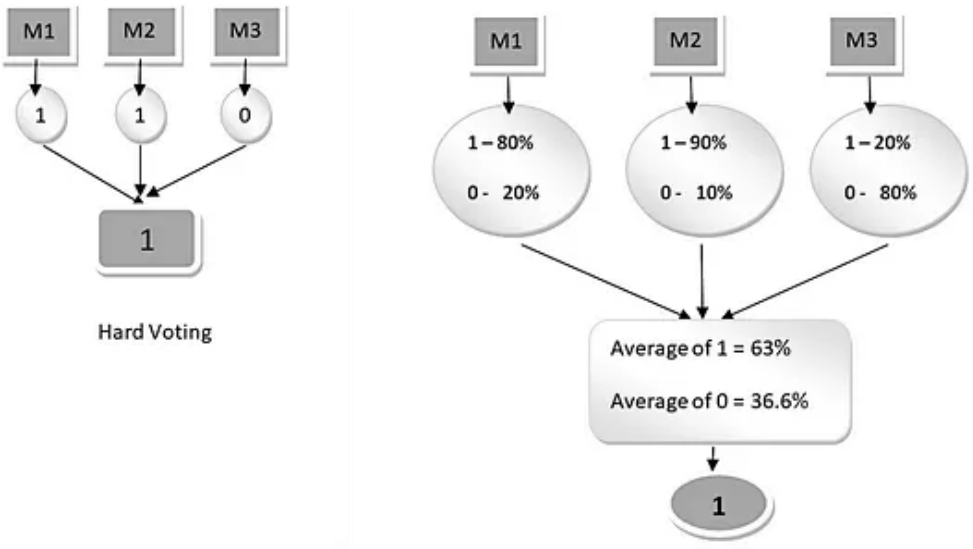
۵-۴-۴-۱۰- دسته‌بندی براساس رای‌گیری[۷۱]

این الگوریتم، یک تخمین‌گر یادگیری ماشین است که مدل‌های پایه یا برآوردگرهای مختلفی را آموزش می‌دهد و بر اساس جمع‌آوری یافته‌های هر تخمین‌گر پایه، پیش‌بینی می‌کند. معیارهای تجمیع می‌تواند دو نوع باشد:

- **سخت:** دسته‌بندی بر اساس کلاس خروجی پیش‌بینی شده محاسبه می‌شود.
- **نرم:** دسته‌بندی بر اساس احتمال پیش‌بینی شده کلاس خروجی محاسبه می‌شود.

---

[۶۸] XGboost Classifier
[۶۹] Gradient Boosting Machines (GBMs)
[۷۰] Extreme Gradient Boosting
[۷۱] Voting Classifier



شکل (۵-۱۸): نحوه تجمیع مدل‌ها و پیش‌بینی الگوریتم دسته‌بندی براساس رای‌گیری

۵-۵- اعتبارسنجی و بررسی صحت الگوریتم‌ها

پس از پیاده‌سازی و آموزش مدل‌های روی داده‌های آموزشی، اعتبارسنجی الگوریتم‌ها با داده‌های اعتبارسنجی انجام شد. در ادامه به معیارهای اعتبارسنجی الگوریتم‌ها می‌پردازیم.

۵-۵-۱- ماتریس اغتشاش[۷۲]

برای اینکه بتوانیم نتایج دسته‌بندی الگوریتم را با داده‌های واقعی مقایسه کنیم، از ماتریس اغتشاش استفاده می‌کنیم. جدول زیر، ماتریس اغتشاش را نشان می‌دهد.

جدول (۵-۲): ماتریس اغتشاش

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

سطرها مقادیر واقعی و ستون‌ها مقادیر پیش‌بینی شده را نشان می‌دهند. سلول‌های این ماتریس مفاهیم زیر را ارائه می‌دهند.

- **مثبت-صحیح**[۷۳]: نشان می‌دهد که مدل یک نتیجه مثبت را پیش‌بینی کرده است و مشاهده واقعی درست بوده است.

[۷۲] Confusion Matrices
[۷۳] True Positive

- **مثبت-کاذب**[۷۴]: نشان می‌دهد که مدل یک نتیجه مثبت را پیش‌بینی کرده است، اما مشاهده واقعی نادرست بوده است.
- **منفی-کاذب**[۷۵]: نشان می‌دهد که مدل یک نتیجه منفی را پیش بینی کرده است، در حالی که مشاهده واقعی درست بوده است.
- **منفی-صحیح**[۷۶]: نشان می‌دهد مدل یک نتیجه منفی را پیش‌بینی کرده است، و نتیجه واقعی نیز نادرست بوده است.

۵-۵-۲- دقت[۷۷]

دقت معمولا برای قضاوت در مورد عملکرد مدل استفاده می‌شود، فرمول محاسبه دقت به شکل زیر است.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

در واقع دقت، میزان پیش‌بینی درست مدل را بر کل محاسبه می‌کند.

۵-۵-۳- صحت[۷۸]

این معیار، اندازه‌گیری مثبت‌های واقعی نسبت به تعداد کل مثبت‌های پیش‌بینی شده توسط مدل را محاسبه می‌کند. در واقع این معیار، میزان مثبت‌بودن پیش‌بینی‌های مثبت مدل را اندازه‌گیری می‌کند.

$$Precision = \frac{TP}{TP + FP}$$

۵-۵-۴- پوشش[۷۹]

معیار پوشش قادر به سنجش مثبت پیش‌بینی شده مدل نسبت به تعداد پیامدهای مثبت واقعی است. با استفاده از این معیار، می‌توان ارزیابی کرد که مدل چقدر قادر به شناسایی نتایج واقعی است.

$$Recall = \frac{TP}{TP + FN}$$

۵-۵-۵- امتیاز F۱[۸۰]

این معیار، میانگین هارمونیک بین دقت و پوشش است. فرمول آن به شکل زیر است.

$$F1\ Score = \frac{۲ * (Precision * Recall)}{Precision + Recall}$$

این معیار هنگامی که داده‌ها به صورت نامتوازن پخش شده‌اند، دید بهتری از عملکرد مدل ارائه می‌دهد.

۵-۵-۶- اعتبارسنجی متقابل[۸۱]

اعتبارسنجی متقابل، یک روش آماری است که برای تخمین عملکرد مدل‌های یادگیری ماشین استفاده می‌شود. این روش برای ارزیابی چگونگی تعمیم نتایج یک تحلیل آماری به یک مجموعه‌داده دیده‌نشده است. این روش،

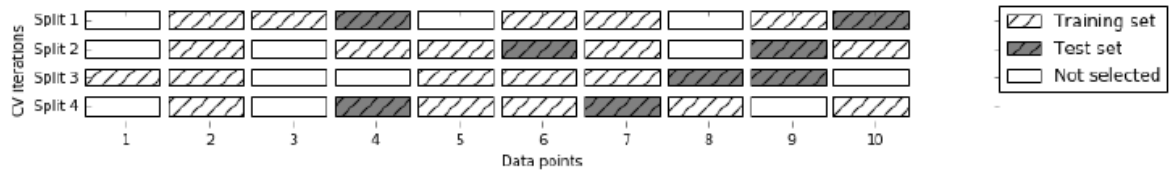
---

[۷۴] False Positive
[۷۵] False Negative
[۷۶] True Negative
[۷۷] Accuracy
[۷۸] Precision
[۷۹] Recall
[۸۰] F۱ Score
[۸۱] Cross Validation

آموزش بیش برازش مدل را شناسایی می‌کند و با بررسی داده‌های دیده نشده، نتیجه دقیق‌تری راجع به عملکرد مدل، ارائه می‌دهد.

در این پژوهش ما از روش اعتبارسنجی متقابل مونت کارلو[۸۲] برای ارزیابی نهایی عملکرد الگوریتم‌ها روی داده‌های دیده نشده، استفاده کرده‌ایم. در ادامه به نحوه عملکرد این معیار اعتبارسنجی می‌پردازیم.

این معیار، یک استراتژی بسیار انعطاف‌پذیر برای اعتبارسنجی متقابل است. در این تکنیک، مجموعه‌داده‌ها به طور تصادفی به مجموعه‌های آموزشی و اعتبارسنجی تقسیم می‌شوند. درصدی از مجموعه داده‌ای را که قرار است به‌عنوان مجموعه آموزشی استفاده شود و درصدی که به عنوان مجموعه اعتبارسنجی استفاده می‌شود، را مشخص می‌کنیم. اگر مجموع درصدها به ۱۰۰ نرسد، از مجموعه‌داده باقی‌مانده استفاده نمی‌شود. سپس این تقسیم‌بندی به تعداد دفعاتی که مشخص می‌کنیم، تکرار می‌شود و دقت هر تکرار محاسبه می‌گردد. می‌توان میانگین نهایی دقت دفعات تکرار را به‌عنوان میزان دقت نهایی مدل روی داده‌های دیده نشده در نظر گرفت.



شکل (۵-۱۹): نحوه عملکرد اعتبارسنجی متقابل مونت کارلو

۵-۵-۷- منحنی AUC – ROC [۸۳]

این معیار ارزیابی روی داده‌های دیده‌نشده برای مقایسه نهایی مدل‌ها، پیاده‌سازی شده است. برای درک بهتر این معیار، مفاهیم زیر مطرح می‌گردد.

- **نرخ مثبت صحیح**[۸۴]: یک معیار ارزیابی عملکرد می‌باشد. همان مفهوم معیار پوشش می‌باشد و مشخص می‌کند که به چه نسبتی پیش‌بینی صحیح صورت گرفته است. فرمول آن به شکل زیر است.

$$\mathrm{TPR} = \frac{TP}{P} = \frac{TP}{TP + FN}$$

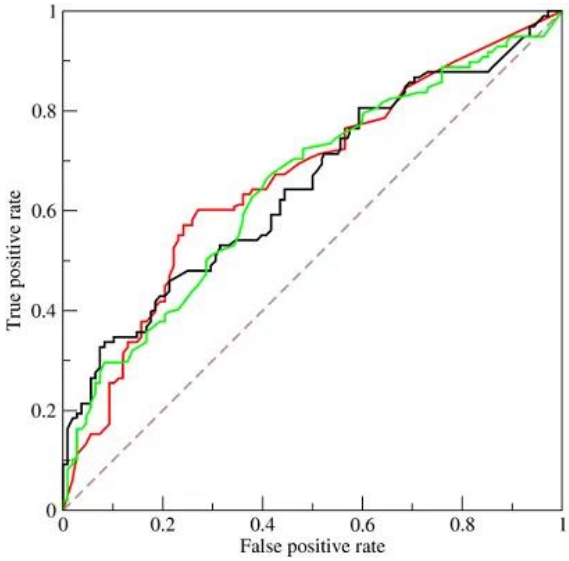
- **نرخ مثبت کاذب**[۸۵]: یک معیار ارزیابی عملکرد می‌باشد و نشانگر تعداد شناسایی‌های مثبت از میان مشاهدات منفی است. فرمول آن به شکل زیر است.

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

- **منحنی مشخصه عملکرد**[۸۶]: یک منحنی مشخصه عملکرد، یک نمودار برای نمایش توانایی ارزیابی یک سیستم دسته‌بندی دودویی محسوب می‌شود که آستانه تشخیص آن نیز متغیر است. این منحنی توسط ترسیم

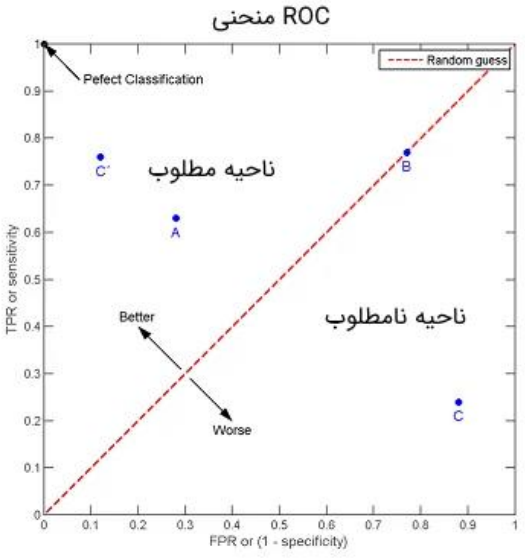
[۸۲] Monte Carlo Cross-Validation (Shuffle Split)
[۸۳] AUC (Area Under the Curve) - ROC (Receiver Operating Characteristics) curve
[۸۴] True Positive Rate (TPR)
[۸۵] False Positive Rate (FPR)
[۸۶] ROC: Receiver Operating Characteristics curve

نرخ مثبت صحیح برحسب نرخ مثبت کاذب، ایجاد می‌شود. نمودار زیر منحنی مشخصه عملکرد را برای سه مدل دسته‌بندی مختلف نشان می‌دهد.



نمودار (۵-۲): منحنی مشخصه عملکرد برای سه روش مختلف دسته‌بندی

با توجه به نمودار زیر، بهترین عملکرد دسته‌بندی در این نمودار در نقطه‌ای با مختصات (۰,۱) رخ خواهد داد که در آن کمترین نرخ اشتباه و بیشترین نرخ بازیابی یا حساسیت را داریم. این نقطه بیانگر «بهترین دسته‌بندی[۸۷]» است.



نمودار (۵-۳): نواحی مطلوب و نامطلوب در منحنی ROC

همچنین در نمودار فوق، خط منقطعی که از میان نمودار عبور کرده و نقطه (۰,۰) را به (۱,۱) پیوند می‌دهند، حدس تصادفی است که به صورت ناحیه ۵۰٪-۵۰٪ نیز شناخته می‌شود. اگر نقطه‌ای روی این خط منقطع قرار گرفته باشد، تشخیص درستی نسبت به قرارگیری در هر گروه، برایش وجود ندارد. در حقیقت در نیمی از موارد

[۸۷] Perfect Classification

می‌تواند در یک دسته و در نیمی از موارد نیز در دسته دیگری، طبقه‌بندی شود و نقشی در تعیین خطا نخواهد داشت. یکی از نمونه‌های معروف برای دسته‌بندی به صورت تصادفی، تصمیم تعلق نقطه به هر یک از دو گروه بوسیله پرتاب سکه است. هر چه تعداد نمونه‌ها در دسته‌بندی تصادفی بیشتر شود، این خط به قطر نواحی ROC نزدیکتر خواهد شد.

AUC: مساحت زیر منحنی مشخصه عملکرد را AUC می‌گویند که نشان‌دهنده درجه یا معیار تفکیک‌پذیری است. این معیار نشان می‌دهد که مدل چقدر می‌تواند بین کلاس‌ها تمایز قائل شود. یک مدل عالی دارای AUC برابر با ۱ است.

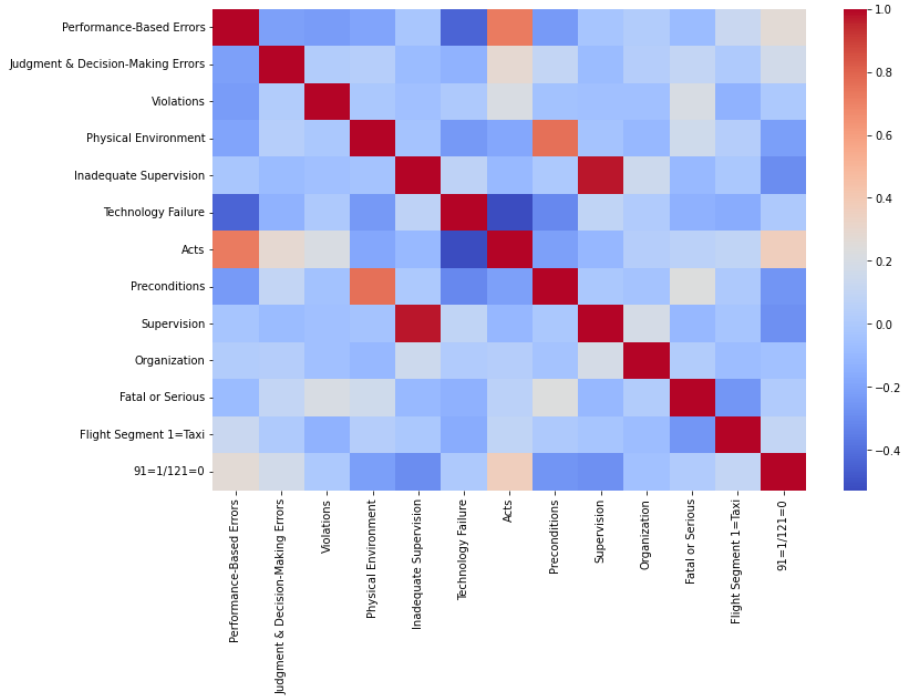
۶-۵- بررسی نتایج و مقایسه الگوریتم‌ها

در این بخش، مقایسه الگوریتم‌ها براساس معیارهای اعتبارسنجی، صورت گرفته است و نتایج در قالب جدول ارائه گردیده است.

۶- یافته‌های تحقیق

۶-۱- تجزیه و تحلیل اکتشافی

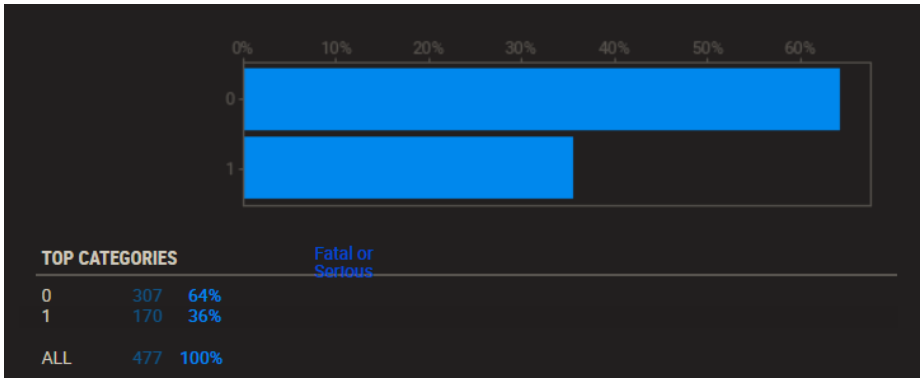
ابتدا همبستگی میان ویژگی‌ها بررسی شده است. شکل زیر، همبستگی بین ویژگی‌ها را نشان می‌دهد. همبستگی یک معیار آماری است که میزان ارتباط خطی دو متغیر را بیان می‌کند (به این معنی که آن‌ها با هم با یک نرخ ثابت تغییر می‌کنند). این یک ابزار رایج برای توصیف روابط ساده بدون اظهار نظر در مورد علت و معلول است. همبستگی عددی بین ۱- و ۱+ است. همبستگی مثبت، نشان‌دهنده میزان افزایش یا کاهش آن متغیرها به صورت موازی است و همبستگی منفی نشان‌دهنده میزان افزایش یک متغیر با کاهش متغیر دیگر است.



شکل (۶-۱): همبستگی ویژگی‌ها در مجموعه داده

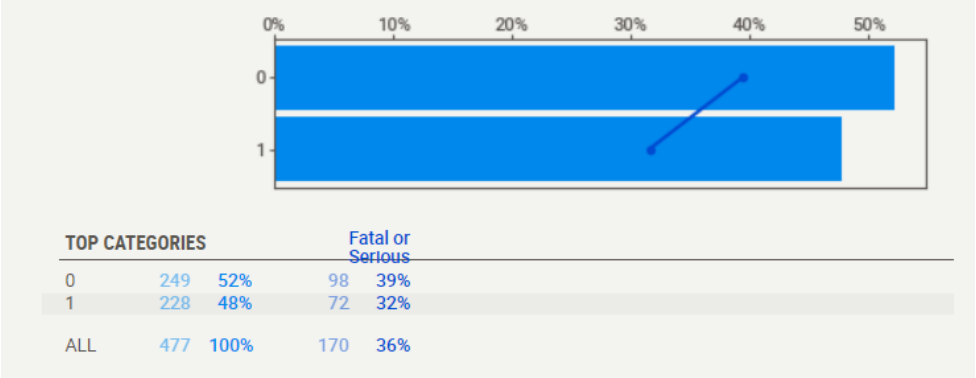
همانطور که از شکل فوق مشاهده می‌گردد، ویژگی‌های خطاهای مبتنی برنظارت و خطاهای مبتنی بر نظارت ناکافی همبستگی بسیار بالایی (حدود ۱) بایکدیگر دارند. همچنین ویژگی خطای مبتنی بر پیش‌بینی با خطاهای مبتنی بر محیط فیزیکی همبستگی نسبتا بالایی دارند (بیشتر از ۰.۵). بین ویژگی خطاهای مبتنی بر عملکرد و خطاهای مبتنی بر اعمال اشتباه هم همبستگی نسبتا بالایی (بیشتر از ۰.۵) وجود دارد. بین برخی از ویژگی‌های دیگر نیز همبستگی کمی مشاهده می‌گردد.

متغیر هدف در این مطالعه، کشنده یا جدی‌بودن یک حادثه می‌باشد. همانطور که در شکل زیر مشاهده می‌شود، در مجموعه داده مورد بررسی، ۶۴ درصد حوادث غیرکشنده (معادل ۳۰۷ پرواز از ۴۷۷ پرواز) و ۳۶ درصد حوادث کشنده (معادل ۱۷۰ پرواز از ۴۷۷ پرواز) بوده‌اند. این موضوع نشان می‌دهد که مجموعه‌داده مورد بررسی نامتعادل می‌باشد. برای رفع این مشکل و جلوگیری از انحراف پیش‌بینی مدل‌ها، از روش نمونه‌برداری بیش از حد برای نمونه‌برداری مجدد و متعادل‌سازی داده‌ها استفاده شده است.



نمودار (۶-۱): متغیر هدف مطالعه

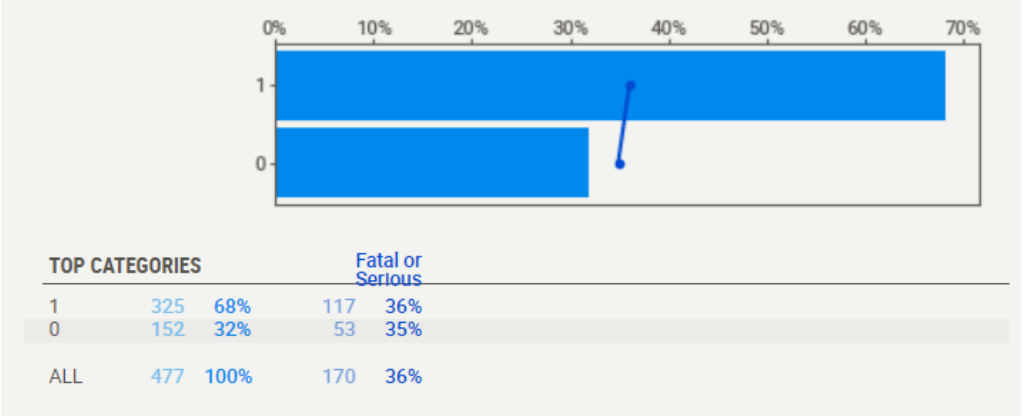
در ادامه به بررسی خطاهای مبتنی بر عملکرد می‌پردازیم.



نمودار (۶-۲): نمودار خطاهای مبتنی بر عملکرد به تفکیک کشنده و غیرکشنده بودن حادثه

همانطور که مشاهده می‌شود، ۴۸ درصد از حوادث دارای ویژگی خطای مبتنی بر عملکرد هستند که این موضوع بیانگر این است که خطای مبتنی بر عملکرد از خطاهای موثر در سوانح هوایی محسوب می‌شود. همچنین ۳۲ درصد از خطاهای مبتنی بر عملکرد باعث بروز حادثه کشنده شده‌اند.

شکل زیر، نمودار پروازهای تجاری و غیرتجاری را نشان می‌دهد.



نمودار (۶-۳): نمودار پروازهای تجاری و غیرتجاری به تفکیک کشنده و غیرکشنده بودن حادثه

همانطور که در نمودار فوق مشاهده می‌گردد، ۳۲ درصد از پروازها، پروازهای تجاری و ۶۸ درصد از کل پروازها را، پروازهای غیرتجاری تشکیل می‌دهد. نرخ کشنده‌بودن حادثه در هر دو دسته از پروازها حدود ۳۵ درصد است که این نرخ در پروازهای تجاری معادل ۵۳ پرواز و در پروازهای غیرتجاری معادل ۱۱۷ پرواز است.

۶-۲- ارزیابی و مقایسه عملکرد الگوریتم‌های یادگیری ماشین

در جدول زیر، مقایسه ارزیابی الگوریتم‌های اجرا شده روی داده‌های دیده نشده قبل از متعادل‌سازی مجموعه‌داده آموزشی، توسط مدل‌ها صورت گرفته است. همانطور که مشاهده می‌شود، الگوریتم XGboost در میان تمام الگوریتم‌های پیاده‌سازی شده، بهترین عملکرد را داشته است.

معیارهای ارزیابی این الگوریتم نشان می‌دهد که دقت مدل روی داده‌های آموزشی ۸۴.۷۸ درصد، دقت مدل روی داده‌های تست ۸۱.۲۵ درصد، معیار صحت ۶۳.۳۳ درصد و معیار پوشش ۷۳.۰۳ درصد شده‌اند. معیار ارزیابی امتیاز F۱ نیز ۶۷.۸۶ درصد بدست آمده است. با توجه به اعداد دقت مدل‌های روی داده‌های آموزشی و تست و کاهش قابل توجه اعداد در معیارهای دیگر، می‌توان دید که عملکرد مدل در داده‌های نامتعادل دچار افت می‌شود.

جدول (۶-۱): مقایسه الگوریتم‌های یادگیری ماشین روی داده‌های نامتعادل

	MLA used	Train Accuracy	Test Accuracy	Precission	Recall	AUC	F1-Score
2	XGBClassifier	0.8478	0.8125	0.633333	0.730769	0.786813	0.678571
3	GradientBoostingClassifier	0.8031	0.8125	0.642857	0.692308	0.774725	0.666667
8	VotingClassifier	0.8504	0.8125	0.633333	0.730769	0.786813	0.678571
1	RandomForestClassifier	0.8530	0.8021	0.600000	0.807692	0.803846	0.688525
6	DecisionTreeClassifier	0.8530	0.7917	0.593750	0.730769	0.772527	0.655172
4	BaggingClassifier	0.8425	0.7812	0.571429	0.769231	0.777473	0.655738
7	BernoulliNB	0.6772	0.7708	0.583333	0.538462	0.697802	0.560000
0	LogisticRegression	0.6772	0.7500	0.555556	0.384615	0.635165	0.454545
5	SVC	0.7533	0.7396	0.517241	0.576923	0.688462	0.545455

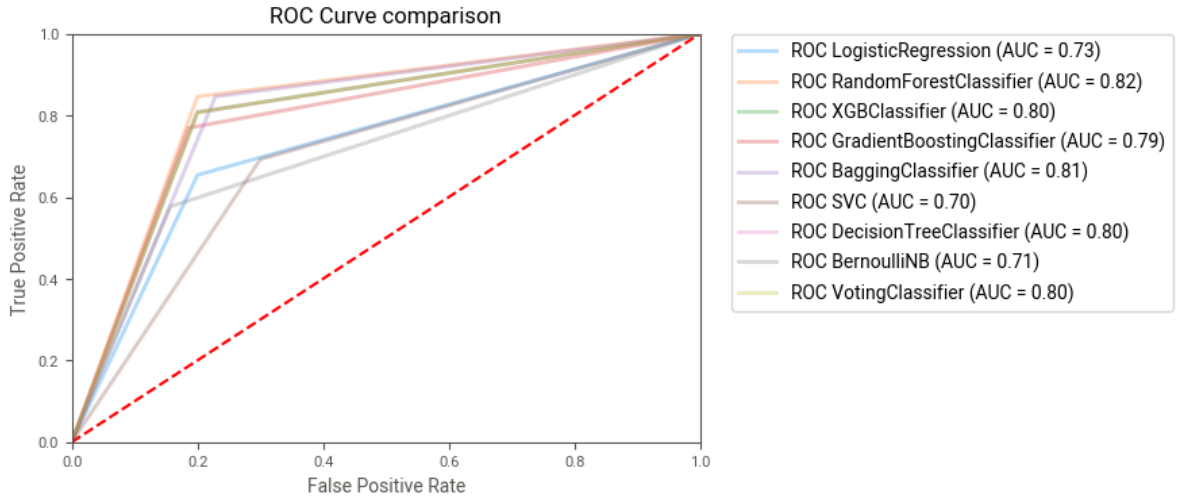
در جدول زیر، مقایسه ارزیابی الگوریتم‌های اجرا شده روی داده‌های دیده نشده بعد از متعادل‌سازی مجموعه‌داده آموزشی، توسط مدل‌ها صورت گرفته است. همانطور که مشاهده می‌شود، الگوریتم جنگل تصادفی در میان تمام الگوریتم‌های پیاده‌سازی شده، بهترین عملکرد را داشته است.

معیارهای ارزیابی این الگوریتم نشان می‌دهد که دقت مدل روی داده‌های آموزشی ۸۵ درصد، دقت مدل روی داده‌های تست ۸۱.۲۵ درصد، معیار صحت ۶۱.۳۳ درصد و معیار پوشش ۸۴.۶۲ درصد شده‌اند. معیار ارزیابی امتیاز F^1 نیز ۷۰.۹۷ درصد بدست آمده است. با مقایسه این جدول با جدول قبل، می‌توان دید که پس از متعادل‌سازی مجموعه‌داده آموزشی، عملکرد مدل‌های یادگیری ماشین کمی بهبود یافته است.

جدول (۶-۲): مقایسه الگوریتم‌های یادگیری ماشین روی داده‌های نامتعادل

	MLA used	Train Accuracy	Test Accuracy	Precission	Recall	AUC	F1-Score
1	RandomForestClassifier	0.8500	0.8125	0.611111	0.846154	0.823077	0.709677
2	XGBClassifier	0.8457	0.8021	0.600000	0.807692	0.803846	0.688525
3	GradientBoostingClassifier	0.8174	0.8021	0.606061	0.769231	0.791758	0.677966
4	BaggingClassifier	0.8457	0.8021	0.594595	0.846154	0.815934	0.698413
6	DecisionTreeClassifier	0.8500	0.8021	0.600000	0.807692	0.803846	0.688525
8	VotingClassifier	0.8478	0.8021	0.600000	0.807692	0.803846	0.688525
7	BernoulliNB	0.6304	0.7708	0.576923	0.576923	0.709890	0.576923
0	LogisticRegression	0.6870	0.7604	0.548387	0.653846	0.726923	0.596491
5	SVC	0.7717	0.6979	0.461538	0.692308	0.696154	0.553846

همچنین نمودار زیر، منحنی ROC الگوریتم‌های بکارگرفته شده را نشان می‌دهد. مساحت زیر محنی مربوط به مدل دسته‌بندی جنگل تصادفی بیشترین مقدار را در بین الگوریتم‌های پیاده‌سازی شده روی داده‌های متعادل در اختیار دارد.



نمودار (۶-۴): منحنی ROC برای الگوریتم‌های بکارگرفته شده

۶-۳- استخراج قوانین انجمنی

الگوریتم Apriori برای استخراج قوانین انجمنی در این مطالعه استفاده شده است. حداقل حد آستانه پشتیبانی، ۰.۲ در نظر گرفته شده است. جدول زیر تعدادی از قوانین را نشان می‌دهد که بیشترین تکرار را در مجموعه‌داده داشته‌اند. در واقع بیشترین میزان پشتیبانی را دارند.

جدول (۶-۳): ترتیب قوانین انجمنی استخراج شده براساس معیار پشتیبانی از زیاد به کم

	support	itemsets
5	0.681342	(91=1/121=0)
2	0.637317	(Acts)
10	0.515723	(Acts, 91=1/121=0)
0	0.477987	(Performance-Based Errors)
6	0.477987	(Acts, Performance-Based Errors)
7	0.387841	(91=1/121=0, Performance-Based Errors)
12	0.387841	(Acts, 91=1/121=0, Performance-Based Errors)
4	0.356394	(Fatal or Serious)
1	0.320755	(Technology Failure)
3	0.253669	(Preconditions)
11	0.245283	(91=1/121=0, Fatal or Serious)
9	0.241090	(Acts, Fatal or Serious)
8	0.218029	(91=1/121=0, Technology Failure)
13	0.205451	(Acts, 91=1/121=0, Fatal or Serious)

همانطور که مشاهده می‌شود، خطاهای مبتنی بر اعمال اشتباه و خطاهای مبتنی بر عملکرد، تجاری یا غیرتجاری بودن پرواز و خطاهای مبتنی بر عملکرد، در میان موارد دوتایی بیشترین تعداد را داشته‌اند. جدول زیر قوانین انجمنی استخراج شده از جدول فوق را نشان می‌دهد که معیارهای Confidence و Lift در آن محاسبه شده است. ترتیب این قوانین براساس بیشترین میزان معیار Lift می‌باشد [۲۰].

جدول (۶-۴): قوانین انجمنی استخراج شده از مجموعه‌داده، ترتیب براساس معیار Lift از زیاد به کم

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
17	(Performance-Based Errors)	(Acts, 91=1/121=0)	0.477987	0.515723	0.387841	0.811404	1.573331	0.141331	2.567793	0.698079
12	(Acts, 91=1/121=0)	(Performance-Based Errors)	0.515723	0.477987	0.387841	0.752033	1.573331	0.141331	2.105165	0.752475
0	(Acts)	(Performance-Based Errors)	0.637317	0.477987	0.477987	0.750000	1.569079	0.173358	2.088050	1.000000
1	(Performance-Based Errors)	(Acts)	0.477987	0.637317	0.477987	1.000000	1.569079	0.173358	inf	0.694779
15	(Acts)	(91=1/121=0, Performance-Based Errors)	0.637317	0.387841	0.387841	0.608553	1.569079	0.140663	1.563836	1.000000
14	(91=1/121=0, Performance-Based Errors)	(Acts)	0.387841	0.637317	0.387841	1.000000	1.569079	0.140663	inf	0.592466
21	(Acts)	(91=1/121=0, Fatal or Serious)	0.637317	0.245283	0.205451	0.322368	1.314271	0.049128	1.113757	0.659313
20	(91=1/121=0, Fatal or Serious)	(Acts)	0.245283	0.637317	0.205451	0.837607	1.314271	0.049128	2.233366	0.316837
22	(91=1/121=0)	(Acts, Fatal or Serious)	0.681342	0.241090	0.205451	0.301538	1.250729	0.041186	1.086545	0.629095
19	(Acts, Fatal or Serious)	(91=1/121=0)	0.241090	0.681342	0.205451	0.852174	1.250729	0.041186	2.155630	0.264150
16	(91=1/121=0)	(Acts, Performance-Based Errors)	0.681342	0.477987	0.387841	0.569231	1.190891	0.062168	1.211815	0.503023
13	(Acts, Performance-Based Errors)	(91=1/121=0)	0.477987	0.681342	0.387841	0.811404	1.190891	0.062168	1.689630	0.307066
3	(Performance-Based Errors)	(91=1/121=0)	0.477987	0.681342	0.387841	0.811404	1.190891	0.062168	1.689630	0.307066
2	(91=1/121=0)	(Performance-Based Errors)	0.681342	0.477987	0.387841	0.569231	1.190891	0.062168	1.211815	0.503023
9	(91=1/121=0)	(Acts)	0.681342	0.637317	0.515723	0.756923	1.187672	0.081493	1.492052	0.495881
8	(Acts)	(91=1/121=0)	0.637317	0.681342	0.515723	0.809211	1.187672	0.081493	1.670209	0.435688
18	(Acts, 91=1/121=0)	(Fatal or Serious)	0.515723	0.356394	0.205451	0.398374	1.117791	0.021650	1.069777	0.217599
23	(Fatal or Serious)	(Acts, 91=1/121=0)	0.356394	0.515723	0.205451	0.576471	1.117791	0.021650	1.143431	0.163731
7	(Fatal or Serious)	(Acts)	0.356394	0.637317	0.241090	0.676471	1.061436	0.013954	1.121022	0.089931
6	(Acts)	(Fatal or Serious)	0.637317	0.356394	0.241090	0.378289	1.061436	0.013954	1.035218	0.159588

در سطر اول جدول مشاهده می‌گردد، خطاهای مبتنی بر عملکرد و تجاری یا غیرتجاری بودن پرواز به هم وابسته هستند و در کنار یکدیگر می‌آیند. سایر موارد و قوانین استخراج شده نیز در فایل پیوست کدها قابل مشاهده است.

قانون «اگر خطای مبتنی بر عملکرد اتفاق بیوفتد، خطای مبتنی بر اعمال اشتباه در پروازهای غیرتجاری اتفاق می‌افتد.» را می‌توان از سطر اول جدول فوق استنباط کرد. مشاهده می‌شود که این قانون در ۳۸.۷۸ درصد کل مجموعه‌داده مشاهده شده است. احتمال وقوع خطای مبتنی بر اعمال اشتباه در پروازهای غیرتجاری به شرط وقوع خطای مبتنی بر عملکرد ۰.۸۱ می‌باشد. معیار Lift نیز عدد ۱.۵۷ بدست آمده است که این مقدار نشان از معنادار بودن این قانون دارد.

۷- نتیجه‌گیری و پیشنهاد

در این مطالعه، برای تجزیه و تحلیل سوانح هوایی و پیاده‌سازی الگوریتم‌های یادگیری ماشین، ما از مجموعه‌داده شامل ۴۷۹ حادثه هوایی است که در بازه زمانی بین سال‌های ۲۰۰۶ تا ۲۰۱۵ توسط هیئت ایمنی حمل و نقل[۸۸] بررسی و گزارش شده است، استفاده کرده‌ایم. این مجموعه‌داده طی مطالعه‌ای که در سال ۲۰۱۸ تحت عنوان «ارزیابی پیش‌شرط‌های مؤثر بر خطای انسانی علامت‌دار در سوانح هوانوردی عمومی و شرکت‌های هواپیمایی[۸۹]» مورد بررسی قرار گرفته است و مدل سیستم تجزیه و تحلیل و طبقه‌بندی عوامل انسانی روی آن پیاده‌سازی شده است. تجزیه و تحلیل اکتشافی داده‌ها و نتایج این مطالعه نشان می‌دهد:

[۸۸] NTSB: National Transportation Safety Board

[۸۹] Anthony J. Erjavac , Ronald Iammartino , John M. Fossaceca , Evaluation of preconditions affecting symptomatic human error in general aviation and air carrier aviation accidents, Reliability Engineering and System Safety (۲۰۱۸), doi: ۱۰،۱۰۱۶/j.ress.۲۰۱۸،۰۵،۰۲۱

- ۴۸ درصد از حوادث دارای خطای مبتنی بر عملکرد هستند که این خطاها با خطاهای مبتی بر اعمال اشتباه همبستگی بالایی دارند. همچنین ۳۲ درصد از خطاهای مبتنی بر عملکرد باعث بروز حادثه کشنده شده‌اند. این موضوع بیانگر این است که خطای مبتنی بر عملکرد و خطاهای مبتنی بر اعمال اشتباه از خطاهای موثر در سوانح هوایی محسوب می‌شوند.
- نرخ کشنده‌بودن حادثه در هر دو دسته از پروازه‌های تجاری و غیرتجاری حدود ۳۵ درصد است که این نرخ در پروازهای تجاری معادل ۵۳ پرواز و در پروازهای غیرتجاری معادل ۱۱۷ پرواز است.
- الگوریتم‌های یادگیری ماشین بکارگرفته شده در این مطالعه از عملکرد نسبتا خوبی برخوردار هستند. الگوریتم‌های دسته‌بندی مبتنی بر رای‌گیری، دسته‌بندی جنگل تصادفی، دسته‌بندی XGboost و دسته‌بندی درخت تصمیم، ۴ الگوریتم برتر در این مطالعه هستند که دقت آن‌ها روی مجموعه‌داده دیده نشده، توسط الگوریتم‌ها بیشتر از ۸۰ درصد می‌باشد.
- نبود مجموعه‌های بزرگ از سوانح هوایی، از محدودیت‌های این حوزه می‌باشد. در مطالعات آینده، پیشنهاد می‌گردد با جمع‌آوری داده‌های بزرگ‌تر، عملکرد الگوریتم‌های یادگیری ماشین مورد بررسی قرار گیرد.
- سوانح هوایی از آن‌جایی‌که مستقیما با جان انسان‌ها در ارتباط است، حوزه‌ای بسیار مهم می‌باشد و لازم است مطالعات بیشتری برای کاهش و جلوگیری از این سوانح صورت بگیرد.
- پیشنهاد دیگر این است که از الگوریتم‌های یادگیری ماشینی که به‌طور ویژه برای داده‌های نامتعادل ساخته‌شده‌اند، برای پیش‌بینی متغیر هدف در این مجموعه‌داده استفاده گردد و عملکرد آن‌ها با الگوریتم‌های استفاده شده در این مطالعه، مقایسه گردد.

۸- منابع

[۱] A. J. Erjavac, R. Iammartino, and J. M. Fossaceca, "Evaluation of preconditions affecting symptomatic human error in general aviation and air carrier aviation accidents," *Reliability Engineering & System Safety*, vol. ۱۷۸, pp. ۱۵٦-۱٦۳, ۲۰۱۸.

[۲] E. Ercan and A. U. AVCI, "Analysis of Turkish Civil Aviation Accidents Between ۲۰۰۳ and ۲۰۱۷," *Journal of Aviation*, vol. ٦, no. ۲, pp. ٥٦-٦۲, ۲۰۲۲.

[۳] T. Zhou, J. Zhang, and D. Baasansuren, "A hybrid HFACS-BN model for analysis of Mongolian aviation professionals' awareness of human factors related to aviation safety," *Sustainability*, vol. ۱۰, no. ۱۲, p. ٤٥۲۲, ۲۰۱۸.

[٤] D. Kelly and M. Efthymiou, "An analysis of human factors in fifty controlled flight into terrain aviation accidents from ۲۰۰۷ to ۲۰۱۷," *Journal of safety research*, vol. ٦۹, pp. ۱٥٥-۱٦٥, ۲۰۱۹.

[٥] B. KILIC and E. GÜMÜŞ, "Application of HFACS to the nighttime aviation accidents and incidents," *Journal of Aviation*, vol. ٤, no. ۲, pp. ۱۰-۱٦, ۲۰۲۰.

[٦] A. Small, "Human factors analysis and classification system (HFACS): as applied to Asiana airlines flight ۲۱٤," *The Journal of Purdue Undergraduate Research*, vol. ۱۰, no. ۱, p. ۱۸, ۲۰۲۰.

[۷] G. Perboli, M. Gajetti, S. Fedorov, and S. L. Giudice, "Natural Language Processing for the identification of Human factors in aviation accidents causes: An application to the SHEL methodology," *Expert Systems with Applications*, vol. ۱۸٦, p. ۱۱٥٦۹٤, ۲۰۲۱.

[۸] C. Zhang *et al.*, "Incorporation of Pilot Factors into Risk Analysis of Civil Aviation Accidents from ۲۰۰۸ to ۲۰۲۰: A Data-Driven Bayesian Network Approach," *Aerospace*, vol. ۱۰, no. ۱, p. ۹, ۲۰۲۳.

[۹] S. Mane, "Aviation Human Factors: Conceptual Overview".

[۱۰] T. Madeira, R. Melício, D. Valério, and L. Santos, "Machine learning and natural language processing for prediction of human factors in aviation incident reports," *Aerospace*, vol. ۸, no. ۲, p. ٤۷, ۲۰۲۱.

[۱۱] P. R. Aswia, D. Lestary, F. Masykur, and G. T. Putra, "An Analysis on Serious Incidents and Accidents in Aviation Using Shell Model," *WARTA ARDHIA* ,vol. ٤۸, no. ۱, pp. ۳٥-٤۲, ۲۰۲۲.

[۱۲] A. Haryono, T. D. Sofianti, and D. Hendriana, "APPLICATION OF HFACS-HFIX FRAMEWORK IN NTSC'S FINDINGS AND RECOMMENDATIONS: WAMENA AIR ACCIDENTS'CASE STUDY," *International Journal of Industrial Management*, vol. ۱۳, no ,۱ .pp. ٤۷۱-٤۷۸, ۲۰۲۲.

[۱۳] N. Chen, Y. Sun, Z. Wang, and C. Peng, "Identification of Flight Accidents Causative Factors Base on SHELLO and Improved Entropy Gray Correlation Method," *Available at SSRN 4236059*.

[۱٤] J. Reason, E. Hollnagel, and J. Paries" ,Revisiting the Swiss cheese model of accidents," *Journal of Clinical Engineering*, vol. ۲۷, no. ٤, pp. ۱۱۰-۱۱٥, ۲۰۰٦.

[۱٥] A. T. Cabello, M. Martínez-Rojas, J. A. Carrillo-Castrillo, and J. C. Rubio-Romero, "Occupational accident analysis according to professionals of different construction phases using association rules," *Safety science*, vol. ۱٤٤, p. ۱۰٥٤٥۷, ۲۰۲۱.

[۱٦] J. Hurwitz and D. Kirsch, "Machine learning for dummies," *IBM Limited Edition*, vol. ۷٥, ۲۰۱۸.

[١٧] https://medium.com/analytics-vidhya/what-is-an-imbalanced-data-how-to-handle-imbalanced-data-in-python-e٦٠٦٧٧٩٢٩٥٠f (accessed.
[١٨] B. Cankaya, K. Topuz, and A. M. Glassman, "Business Inferences and Risk Modeling with Machine Learning; The Case of Aviation Incidents," *Business Inferences and Risk Modeling with Machine Learning; The Case of Aviation Incidents*, vol. ١١, p. ١٢٣٨, ٢٠٢٣.
[١٩] [Online]. Available: https://en.wikipedia.org/wiki/Machine_Learning_Models.

[٢٠] https://www.datacamp.com/tutorial/association-rule-mining-python