



دانشگاه پیام نور

دانشکده فنی مهندسی

دانشگاه پیام نور مرکز تهران

پایان نامه

برای دریافت درجه کارشناسی ارشد

رشته مدیریت فناوری اطلاعات

گروه مهندسی کامپیوتر و فناوری اطلاعات

پیش بینی میزان ریسک مشتریان هدف در صنعت بیمه با استفاده از روش های داده کاوی (مورد کاوی

بیمه بدنه شرکت بیمه ایران)

سامرند خالهء

استاد راهنما:

دکتر نسترن حاجی حیدری

استاد مشاور:

دکتر احمد فراهی

اردیبهشت ۹۱

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

تقدیم ہے:

پدرو مادر کے تقدرو مہربانم کہ

ہمیشہ شمع راہ من بودہ اندو

خواہر دوست دانتیم

کہ حامی ہمیشگی من است.

تشکر و قدردانی:

قبل از هرچیز پروردگار بزرگ را شاکرم که به من نیرو داد تا در این راه قدم گذارم. لازم می‌دانم از زحمات استاد عزیز و گرانقدر سرکار خانم دکتر نسترن حاجی حیدری نهایت تشکر و قدردانی را به عمل آورم. بی‌شک رهنمودهای ایشان، مهم‌ترین عامل موفقیت این پایان‌نامه بوده است. همچنین از حمایت‌ها و مشاوره استاد گرامی جناب آقای دکتر احمد فراهی در تمام مراحل این تحقیق کمال تشکر را دارم.

جا دارد از زحمات استاد ارجمند و دوست عزیز، دکتر مهدی فسقوری که نخستین الهام بخش من در زمینه فعالیت در شاخه داده‌کاوی بودند کمال تشکر را داشته باشم.

چکیده:

پایه گذاری سیستمی که ریسک مشتریان را کنترل می‌کند، یک بخش مهم در مدیریت علمی یک شرکت بیمه تلقی می‌گردد. با توجه به اهمیت ریسک بیمه های خودرو در این تحقیق، تکنیک داده‌کاوی برای تحلیل ریسک مشتریان در یک شرکت بیمه ایرانی مورد استفاده قرار می‌گیرد. هدف آنست که نهایتاً مدل تصمیم‌گیری ارائه گردد تا مشتریان قبل از بیمه نمودن شناسایی شوند. روش تحقیق بر اساس یک فرایند استاندارد داده‌کاوی می‌باشد بدین صورت که داده‌های مشتریان سابق جمع‌آوری و پالایش شده و متغیرهای موثر بر رفتار ریسک مشتریان شناسایی گردیده‌اند. سپس بر پایه اینکه مشتریان ریسکی داشته‌اند یا خیر در دو طبقه خوب و بد قرار داده شده‌اند و برای هر مشتری یک برچسب کلاس در نظر گرفته شده است. بعد از پیش پردازش و تغییر روی رکوردها (مشتریان) و فیلدها (ویژگی‌های مشتریان)، تکنیک C5 بر روی داده‌های نهایی اجرا گردید و الگوهای پنهان به صورت یک درخت در قوانین داده‌کاوی ارائه گردید. داده‌های نهایی همچنین با چند الگوریتم دیگر مانند درخت‌های تصمیم سنتی، شبکه‌های عصبی، رگرسیون لجستیک، شبکه‌های بیزین، ماشین بردار پشتیبان نیز مدل‌سازی گردید. نتایج نشان داده الگوریتم نتایج بهتری را برای طبقه بندی مشتریان نسبت به سایر الگوریتم‌ها داشته است. همچنین مدل ترکیبی طبقه مشتریان را بهتر از سایر تکنیک‌ها پیش‌بینی نموده است.

کلمات کلیدی:

بیمه بدنه اتومبیل، طبقه‌بندی ریسک مشتریان، طبقه بندی، داده‌کاوی، الگوریتم C5

فهرست مطالب

صفحه	عنوان
۲	۱- مقدمه.....
۲	۱-۱- مقدمه.....
۲	۲-۱- تعریف مسئله و سؤالات اصلی تحقیق.....
۵	۳-۱- اهداف تحقیق.....
۶	۴-۱- روش تحقیق.....
۶	۵-۱- مراحل انجام تحقیق.....
۸	۶-۱- جامعه آماری، روش نمونه‌گیری و حجم نمونهها.....
۸	۷-۱- ساختار پایان‌نامه.....
۱۰	۲- مبانی نظری و ادبیات تحقیق.....
۱۰	۲-۱- مقدمه.....
۱۰	۲-۲- کاربردهای داده‌کاوی در صنعت بیمه.....
۱۱	۱-۲-۲- اصول و مفاهیم ریسک در بیمه.....
۱۴	۲-۳- داده‌کاوی و کشف دانش از پایگاه‌داده.....
۱۵	۱-۳-۲- مفهوم داده‌کاوی.....
۱۶	۲-۳-۲- فرایند کلی داده‌کاوی.....
۱۸	۳-۳-۲- روش‌ها و تکنیک‌های داده‌کاوی.....
۲۲	۴-۲- طبقه‌بندی (دسته‌بندی).....
۲۲	۱-۴-۲- الگوریتم‌های مختلف طبقه‌بندی.....
۲۴	۵-۲- درخت تصمیم.....

۲۶	۱-۵-۲- استخراج قاعده از یک درخت تصمیم
۲۶	۲-۵-۲- درخت تصمیم سنتی و انواع الگوریتم‌های آن
۲۸	۳-۵-۲- نقاط قوت و ضعف الگوریتم‌های درخت تصمیم
۲۹	۶-۲- مروری بر الگوریتم‌های مورد استفاده در این پژوهش
۲۹	۱-۶-۲- ساختار الگوریتمی درخت تصمیم
۳۴	۲-۶-۲- CART
۳۵	۲-۶-۳- CHAID
۳۵	۴-۶-۲- QUEST
۳۶	۵-۶-۲- شبکه‌های بیزین
۳۷	۶-۶-۲- شبکه‌های عصبی
۴۱	۲-۶-۷- ماشین بردار پشتیبان
۴۳	۸-۶-۲- رگرسیون لجستیک
۴۴	۹-۶-۲- تحلیل تمایزی
۴۶	۷-۲- مروری بر تحقیقات پیشین
۴۶	۱-۷-۲- مطالعات داخلی
۴۸	۲-۷-۲- مطالعات خارجی
۵۰	۸-۲- نتیجه‌گیری
۵۳	۳- روش‌شناسی تحقیق
۵۳	۱-۳- مقدمه
۵۳	۲-۳- فرایند داده‌کاوی
۵۴	۱-۳-۲- مدل فرایند داده‌کاوی بر اساس استاندارد CRISP-DM
۵۸	۳-۳- داده‌های تحقیق
۵۹	۱-۳-۳- شیوه و ابزار جمع‌آوری اطلاعات
۵۹	۲-۳-۳- نوع داده‌ها و مقیاس آن‌ها

- ۳-۳-۳- جامعه آماری و نمونه‌گیری ۵۹
- ۳-۴-۴- اعتبار و کارایی روش ارائه شده ۶۰
- ۳-۵-۵- ساختار اجرایی تحقیق ۶۱
- ۳-۵-۱- درک مسئله کسب و کار ۶۱
- ۳-۵-۲- درک داده‌ها ۶۳
- ۳-۵-۳- آماده‌سازی داده ۶۴
- ۳-۵-۴- مدل‌سازی ۶۷
- ۳-۵-۵- ارزیابی نتایج ۶۹
- ۳-۵-۶- بکارگیری مدل ۶۹
- ۳-۶- جمع‌بندی ۷۰

۴- تجزیه و تحلیل داده‌ها ۷۲

- ۴-۱- مقدمه ۷۲
- ۴-۲- توصیف داده‌ها ۷۲
- ۴-۲-۱- ویژگی داده‌ها ۷۲
- ۴-۲-۲- نمونه‌های مورد بررسی پایگاه داده ۷۳
- ۴-۲-۳- رکوردهای پایگاه داده ۷۳
- ۴-۲-۴- فیلدهای پایگاه داده ۷۴
- ۴-۲-۵- نوع و مقیاس داده‌ها ۷۶
- ۴-۳- آماده‌سازی داده‌ها برای مدل ۷۷
- ۴-۴- درک مسئله و کسب و کار ۷۷
- ۴-۵- درک داده‌ها ۷۷
- ۴-۶- پالایش داده‌ها ۷۸
- ۴-۷- تغییرات داده ۷۹
- ۴-۷-۱- برچسب‌گذاری طبقات مشتریان ۸۱

۸۲	۸-۴- تحلیل داده‌ها
۸۲	۱-۸-۴- مدلسازی
۸۶	۲-۸-۴- دقت الگوریتم
۸۶	۳-۸-۴- اهمیت متغیرها
۸۷	۹-۴- اعتبار و کارایی مدل
۸۷	۱-۹-۴- ارزیابی اعتبار مدل
۸۸	۲-۹-۴- کارایی مدل در مقایسه با الگوریتم‌های دیگر
۹۸	۳-۹-۴- بررسی صحت الگوریتم‌ها در حالت سه کلاس
۹۹	۴-۹-۴- نتایج ترکیب مدل‌های طبقه بندی
۱۰۰	۱۰-۴- نتیجه‌گیری

۵- نتیجه‌گیری ۱۰۲

۱۰۲	۱-۵- مقدمه
۱۰۲	۲-۵- نتایج حاصل از تحقیق
۱۰۲	۱-۲-۵- پاسخ به سؤالات تحقیق
۱۰۴	۳-۵- نتایج مستقیم حاصل از تحقیق
۱۰۵	۴-۵- مقایسه یافته‌های پژوهش با پژوهش‌های قبلی
۱۰۶	۵-۵- پیشنهادات و کارهای آینده
۱۰۸	۶-۵- محدودیت‌های تحقیق

مراجع ۱۲۲

واژه‌نامه ۱۲۷

فهرست اشکال

-
- شکل ۱-۱ مدل اجرایی تحقیق..... ۷
- شکل ۱-۲. کاربردهای داده‌کاوی در صنعت بیمه..... ۱۱
- شکل ۲-۲. گام‌های اصلی فرایند اکتشاف دانش از پایگاه داده (مروج، ۱۳۸۳)..... ۱۸
- شکل ۳-۲. ساختار کلی یک مدل امتیازدهی (اخباری، ۱۳۸۷)..... ۲۳
- شکل ۴-۲. نمونه ای از شبکه عصبی..... ۳۸
- شکل ۵-۲. ماشین بردار پشتیبان در حالت خطی..... ۴۲
- شکل ۱-۴. توزیع داده‌ها در هر فیلد..... ۷۵
- شکل ۲-۴. توزیع طبقات مشتریان..... ۷۵
- شکل ۳-۴. پالایش اولیه و یکپارچه‌سازی پایگاه داده‌ها..... ۷۸
- شکل ۴-۴. حذف برخی رکوردها..... ۷۹
- شکل ۵-۴. حذف برخی فیلدها..... ۷۹
- شکل ۶-۴. فراخوانی داده‌ها در نرم افزار..... ۸۳
- شکل ۷-۴. استریم مدل‌سازی الگوریتم‌های طبقه بندی..... ۸۳
- شکل ۸-۴. درخت تصمیم C5 ایجاد شده..... ۸۴
- شکل ۹-۴. اهمیت شاخص‌ها..... ۸۷
- شکل ۱۰-۴. اولویت بندی الگوریتم‌ها به لحاظ میزان صحت..... ۹۷

شکل ۴-۱۱. استریم پیش بینی وضعیت مشتریان جدید..... ۹۸

شکل ۴-۱۲. استریم مدل ترکیبی الگوریتم های طبقه بندی..... ۹۹

فهرست جداول

جدول ۱-۲. مطالعات داخلی و خارجی مرتبط با پژوهش پیش رو.....	۴۹
جدول ۱-۳. مراحل مدل فرایندی داده‌کاوی بر اساس استاندارد CRISP-DM.....	۵۵
جدول ۲-۳. وظایف کلی مدل استاندارد CRISP-DM (Chapman, 1999).....	۵۷
جدول ۱-۴. چارچوب کلی پایگاه داده مورد بررسی.....	۷۳
جدول ۲-۴. عناوین فیلدها.....	۷۴
جدول ۳-۴. حالات متغیرهای اسمی.....	۷۶
جدول ۴-۴. مقیاس‌های عددی.....	۷۶
جدول ۵-۴. یک شکل نمودن داده‌ها.....	۸۰
جدول ۶-۴. درخت تصمیم C5.....	۸۶
جدول ۷-۴. شبکه‌های عصبی Neural Network.....	۸۹
جدول ۸-۴. تحلیل تمایزی.....	۹۰
جدول ۹-۴. رگرسیون لجستیک Logistic -R.....	۹۰
جدول ۱۰-۴. ماشین بردار پشتیبان Support Vector Machine.....	۹۱
جدول ۱۱-۴. شبکه‌های بیزین BayesNet.....	۹۲
جدول ۱۲-۴. درخت تصمیم CART.....	۹۳
جدول ۱۳-۴. درخت تصمیم QUEST.....	۹۴

- جدول ۴-۱۴. درخت تصمیم CHAID..... ۹۵
- جدول ۴-۱۵. مقایسه تطبیقی صحت الگوریتم‌ها..... ۹۶
- جدول ۴-۱۶. مقایسه تطبیقی صحت الگوریتم‌ها در حالت سه کلاس..... ۹۸
- جدول ۵-۱. مقایسه یافته‌های این تحقیق با پژوهش مشابه..... ۱۰۵

فصل اول

مقدمه

۱- مقدمه

۱-۱- مقدمه

همانند سایر بخش‌های اقتصادی، صنعت بیمه نیز شاهد تغییرات بسیاری در عرصه فناوری اطلاعات در طی سال‌های گذشته بوده است. پیشرفت در زمینه‌های سخت افزار، نرم افزار و شبکه‌های فناوری اطلاعات فواید بسیاری را از قبیل کاهش هزینه‌ها، کاهش زمان پردازش داده‌ها، افزایش قابلیت‌های سودآوری، همزمان با رو یارویی با چالش‌های جدید در فضایی که هر روز شاهد افزایش میزان رقابت در آن بوده‌ایم به همراه داشته است. ابداعات و نوآوری‌های فنی از قبیل استخراج و نیز نگهداری داده‌ها به طور قابل ملاحظه ای موجبات کاهش هزینه‌های نگهداری، دست یابی و نیز پردازش داده‌ها را فر اهم آورده است. بسیاری از سؤالات تجاری که سابقاً به علت کمبود داده او امکانات پردازشگری غیرممکن، غیرعملی و یا غیرقابل حل می نمودند، در حال حاضر با استفاده از تکنیک داده کاوی قابل بررسی و پاسخ گویی هستند. (حسین زاده، ۱۳۸۶)

۱-۲- تعریف مسئله و سؤالات اصلی تحقیق

پیچیدگی محیطی، شدت رقابت، رواج تکنولوژی‌های نو و پیشرفته، توسعه فناوری اطلاعات و ارتباطات، شیوه‌های نوین عرضه کالاها و خدمات، مسایل زیست محیطی و سمت گیری سازمان‌ها از دارایی‌های مشهود به نامشهود و... از عوامل عمده ای است که موجب شده است سازمان‌ها و بنگاه‌های اقتصادی در دوران حیات خود با ریسک‌های بسیار متعدد و خطرات زیاد و حتی پیش بینی نشده مواجه شوند. به همین جهت به منظور کاهش ریسک و جبران زیان‌های ناشی از آن امروزه

در ادبیات علمی انواع مدیریت ریسک نظیر مدیریت ریسک بنگاه، مدیریت ریسک کسب و کار و مدیریت ریسک استراتژیک مطرح شده و هریک جایگاهی خاص دارند. بدیهی است هر سازمان باتوجه به ماهیت کار خود، ریسک های گوناگونی را تجربه می کند و در شرایط متحول امروز، اساساً موفقیت هر بنگاه به تسلط آن بر ریسک ها و نوع مدیریتی است که بر انواع ریسک ها اعمال می کند. مدیریت ریسک زمانی معنا و مفهوم می یابد که شرایط با احتمال متحمل شدن زیان و عدم اطمینان مواجه شود. این نوع مدیریت شامل حوزه های گسترده ای است که مسایل مالی، عملیاتی، تجاری، استراتژیک و حوزه وسیع تری به نام حوادث خطرآفرین را دربرمی گیرد. درمجموع مدیریت ریسک فرایند سنجش یا ارزیابی ریسک و سپس طرح استراتژی هایی برای اداره ریسک است. بدیهی است باتوجه به شرایط پیچیده و رقابتی کسب و کار در عصر امروز، مدیریت ریسک بیش از گذشته اهمیت خود را بازیافته و مدیران برای بقای بنگاه های خود و کاهش زیان ناگزیرند به آن روی آورده و متعهد به اجرای آن باشند. ریسک دلیل وجود بیمه است و بدون ریسک درواقع بیمه مفهوم خود را از دست می دهد. کار بیمه گری با ریسک و ریسک پذیری و کاهش ریسک و محاسبه ریسک سروکار دارد..

یک از چالش های اصلی شرکت های بیمه کنترل ریسک برای مشتریان به خصوص مشتریان بیمه بدنه می باشد و شناسایی و طبقه بندی مشتریان کم ریسک و پر ریسک می تواند در مدیریت ریسک مشتریان و سیاست های مرتبط مانند روی گردانی و کاهش ریسک و تعدیل نرخ حق بیمه موثر باشد. استفاده از داده کاوی در این زمینه می تواند بسیار راهگشا باشد و این تحقیق به بررسی موضوع طبقه بندی مشتریان به لحاظ ریسک خسارت بیمه بدنه خودرو در یک شرکت بیمه کشور می پردازد.

به دلیل نفوذ زیاد بیمه‌های اتومبیل و نقشی که این نوع بیمه در بین دیگر رشته‌های بیمه در ایران دارد^۱، بیمه بدنه اتومبیل برای این مطالعه انتخاب شد. ولی متأسفانه ابزارهای جدیدی مانند داده‌کاوی در صنعت بیمه ایران تاکنون جایگاه اصلی خود را پیدا نکرده است. اما به علت تغییر محیط به لحاظ قانونی (تبدیل نظام تعرفه‌ای حق بیمه به نظام نرخ‌گذاری آزاد) استفاده از فناوری‌های جدید برای رقابت در یک بازار کاملاً رقابتی از سوی شرکت‌های بیمه به نظر اجتناب‌ناپذیر می‌باشد. در حال حاضر از ابتدا تمامی بیمه‌گذاران حق بیمه واحدی پرداخت می‌کنند که این امر باعث می‌شود بیمه‌گذاران با ریسک پایین‌تر نسبت به بیمه‌گذاران با ریسک بالاتر حق بیمه بیشتری پرداخت نمایند که به نظر منصفانه نمی‌رسد. با دسترسی به ابزارهای تحلیل و طبقه‌بندی ریسک می‌توان این مشکل را برطرف ساخت. با استفاده از داده‌کاوی و تجزیه تحلیل داده‌ها، می‌توان این بیمه‌گذاران را به لحاظ ریسک طبقه‌بندی نمود و با استفاده از این طبقه‌بندی می‌توان طبقه بیمه‌گذاران آتی را پیش‌بینی کرد و در میزان حق دریافتی از بیمه‌گذاران مختلف تعدیل ایجاد نمود و با ایجاد یک سیستم نرخ‌گذاری مبتنی بر ریسک بیمه‌گذاران، میزان رضایت بیمه‌گذاران را افزایش داده و از طرف دیگر بر سودآوری شرکت‌های بیمه تأثیر مثبتی گذاشت.

در این راستا سؤالات زیر مطرح اند:

سؤال اصلی:

^۱ طبق آمار سال ۸۸ بیمه مرکزی ج.ا. ایران سهم بیمه بدنه اتومبیل از نظر میزان حق بیمه ۱۵،۱۱ درصد است که در بین سایر رشته‌های بیمه‌ای جایگاه سوم را داراست.

۱) چگونه می‌توان در یک شرکت بیمه، مشتریان را به لحاظ ریسک خسارتی با استفاده از

الگوریتم‌های داده‌کاوی طبقه بندی نمود؟

سؤالات فرعی:

۱) آیا الگوریتم C5، طبقه بندی صحیح‌تری را نسبت به سایر الگوریتم‌های طبقه بندی داده

کاوی ارائه خواهد کرد؟

۲) آیا شاخص «نوع خودرو» مهم‌ترین شاخص در طبقه بندی مشتریان بیمه خودرو می‌باشد؟

۳) آیا مدل ترکیبی الگوریتم‌های داده کاوی نسبت به سایر الگوریتم‌ها صحت بالاتری دارد؟

۱-۳- اهداف تحقیق

۱) تعیین مهم‌ترین متغیرهای دخیل در پیش بینی رفتار بیمه گذاران بیمه بدنه اتومبیل.

۲) جستجوی الگوهای موجود بین داده های بیمه گذاران سابق و ارائه چارچوب مناسبی برای

طبقه بندی بیمه گذاران بیمه بدنه اتومبیل با استفاده از داده‌کاوی.

۳) مقایسه نتایج پیش بینی الگوریتم‌های مختلف طبقه بندی در داده‌کاوی مشتریان بیمه بدنه

خودرو.

۴) افزایش سودآوری صنعت بیمه در زمینه طبقه بندی بیمه گذاران برای تعیین حق بیمه.

۵) کاهش خطاهای انسانی و جلوگیری از فرایند قضاوتی کارشناسان خبره به وسیله بکارگیری

الگوهای کشف شده این تحقیق در یک سیستم هوشمند (یادگیری ماشین).

۶) - تعیین چارچوب ریسک برای طبقات مختلف بیمه‌گذاران و در نتیجه پرداخت میزان

عادلانه حق بیمه بر مبنای میزان ریسک هر بیمه‌گذار.

۱-۴- روش تحقیق

روش انجام تحقیق از طریق مطالعه و بررسی کتب، مقالات، پایان‌نامه‌های انجام شده داخلی و خارجی، پروژه‌های تحقیقاتی صورت گرفته و اینترنت می‌باشد.

۱-۵- مراحل انجام تحقیق

۱- جمع آوری داده از پایگاه داده بیمه‌گذاران فعلی بانک اطلاعاتی بیمه بدنه خودرو و تقسیم آنها به

دو دسته داده‌های آزمایشی و داده‌های آموزشی؛

۲- تعیین شاخص‌هایی برای تعریف طبقات ریسک بیمه‌گذاران؛

۳- استخراج الگوها با استفاده از داده‌های آزمایشی با تکنیک درخت تصمیم و مقایسه نتایج با

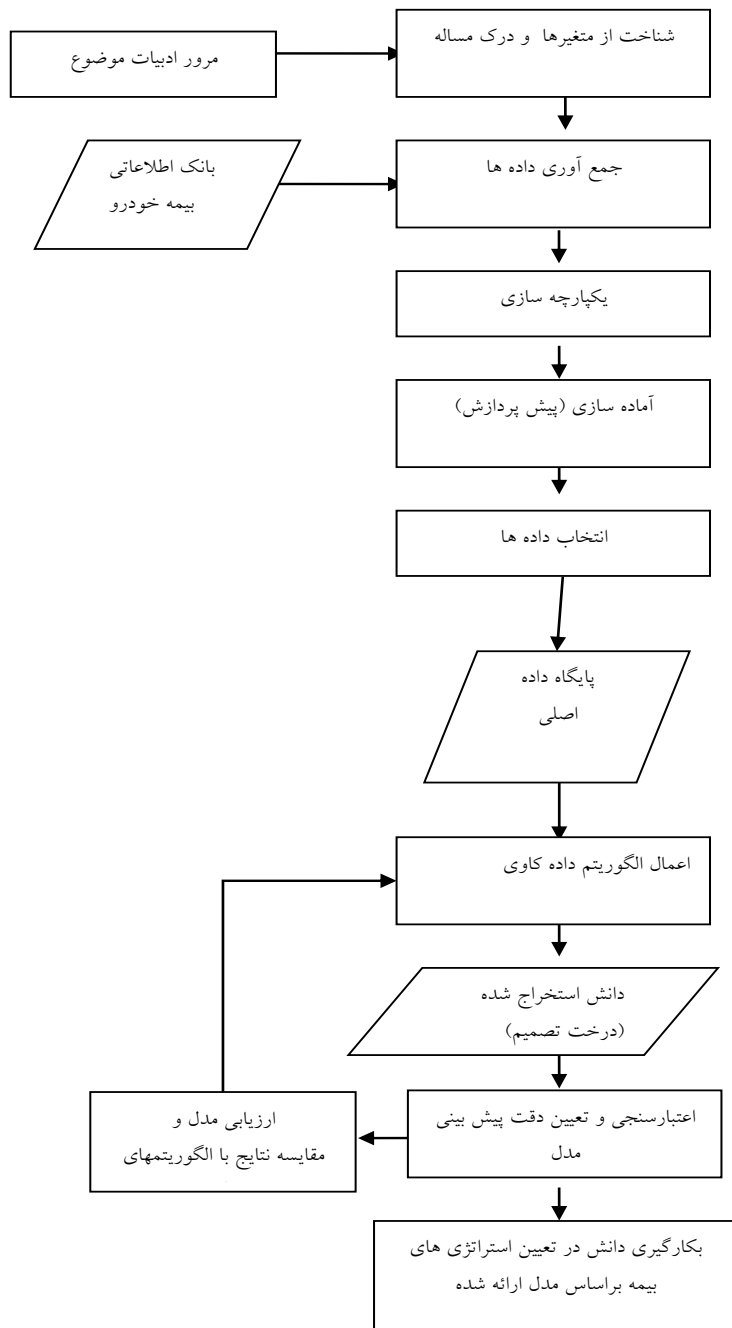
استفاده از روش‌های دیگر؛

۴- ارائه الگوی کشف شده از طبقه بندی بیمه‌گذاران؛

۵- اعتبارسنجی مدل با استفاده از مجموعه داده‌های آزمایشی.

در خصوص مراحل انجام تحقیق و مدل اجرایی در فصل سوم به تفصیل اشاره شده است. در این

فصل مدل اجرایی تحقیق را می‌توان در قالب شکل ۱-۱ نمایش داد.



شکل ۱-۱ مدل اجرایی تحقیق