دومین کنفرانس ملی کامپیوتر ، فناوری اطلاعات و کاربردهای هوش مصنوعی

1 **اسفند ماه** 1397 - **دانشگاه شهید چمران اهواز**

یافتن مشتریان احتمالی خرید بیمه با تکنیک‌های داده‌کاوی

امیر اولیایی، دانشجوی کارشناسی مهندسی کامپیوتر - نرم‌افزار، stu.amir.oliaei@kashmar.ac.ir [۱]

علی حاضری، دانشجوی کارشناسی مهندسی کامپیوتر - نرم‌افزار، stu.ali.hazeri@kashmar.ac.ir [۲]

محمد جمالی، دانشجوی کارشناسی مهندسی کامپیوتر - نرم‌افزار، stu.mohammad.jamali@kashmar.ac.ir [۳]

عاطفه خزاعی، دکتری مهندسی کامپیوتر - نرم‌افزار، atefeh.khazaei@kashmar.ac.ir [۴]

چکیده

در سال‌های اخیر صنعت بیمه رشدی چشمگیر داشته است و شرکت‌های مختلف با خدمات گوناگون پا به عرصه گذاشته‌اند. بازاریابی موفق یکی از اهداف اصلی شرکت‌های بیمه است؛ پیدا کردن افرادی که احتمال می‌رود از خدمات بیمه استفاده کنند، بسیار مهم است و منجر به مدیریت هرچه بهتر سرمایه و هزینه‌ها می‌شود. در این پژوهش با استفاده از یک مجموعه داده، حاوی اطلاعات مختلف ازجمله مشخصات مشتری، اطلاعات فروش، پوشش‌های خاص، موقعیت جغرافیایی و وضعیت اموال، به ارائه مدلی می‌پردازیم که بتواند احتمال اینکه که فرد از خدمات بیمه استفاده می‌کند یا خیر، را پیش‌بینی کند. برای ارائه این مدل تکنیک‌ها و الگوریتم‌های مختلف داده‌کاوی را بکار گرفته‌ایم. پس از پیش‌پردازش‌های لازم بر روی داده‌ها، از الگوریتم‌های جنگل تصادفی، درخت تصمیم و بیزین ساده برای ساخت مدل استفاده کردیم. الگوریتم جنگل تصادفی با بیش از 99 درصد صحت بهترین عملکرد را نشان می‌دهد. نتایج حاصل از این پژوهش نشان می‌دهد که به‌کارگیری تکنیک‌های داده‌کاوی برای پیش‌بینی مشتریان احتمالی می‌تواند به شرکت‌های بیمه کمک کند تا مشتریان احتمالی خود را پیدا و برای فروش خدمات خود بر روی آن افراد سرمایه‌گذاری کنند. این امر باعث کاهش هزینه‌ها ازجمله کاهش هزینه تبلیغات می‌شود.

واژه‌های کلیدی:

مشتری بیمه، داده‌کاوی، جنگل تصادفی، درخت تصمیم، بیز ساده.

[۱۲۳۴] مرکز آموزش عالی کاشمر - کاشمر، بلوار سید مرتضی، مرکز آموزش عالی کاشمر -

1- مقدمه

در جهانی که پر است از مخاطرات و حوادث غیرقابل‌پیش‌بینی، سرمایه‌های مادی افراد دائماً در معرض تهدیدات جدی قرار دارند و این امر ضرورت استفاده از بیمه را دوچندان می‌کند. همین امر منجر به رشد روزافزون شرکت‌های بیمه و ارائه خدمات متنوع شده است. بازاریابی و یافتن مشتری یکی از فعالیت‌های اصلی شرکت‌های بیمه می‌باشد [1].

کسب‌وکار در بستر وب مجموعه داده‌های غول‌پیکری را ایجاد می‌کند؛ ازجمله‌ی آن‌ها می‌توان به فروش، معاملات، سهام سوابق تجاری، توضیحات محصول، تبلیغات فروش، مشخصات شرکت‌هایی مانند بیمه، عملکرد و بازخورد مشتریان آن‌ها اشاره کرد. مدیران کسب‌وکارهای مختلف، می‌دانند که جمع‌آوری اطلاعات و داده‌ها از مشتریان و مخاطبان یک شرکت، یکی از عوامل موردنیاز برای رشد و توسعه این شرکت‌ها است؛ اما این فقط بخش اول راهی است که باید طی شود. اگر داده‌ها فقط جمع شوند و بلااستفاده بمانند، عملاً هدف اصلی از جمع‌آوری این اطلاعات، برآورده نشده است. گام مهمی که پس از جمع‌آوری داده‌ها می‌بایست برداشته شود، استخراج دانش از این داده‌ها است و این مهم با استفاده از تکنیک‌های داده‌کاوی حاصل خواهد شد [2].

در این مقاله قصد داریم با به‌کارگیری تکنیک‌های داده‌کاوی بر داده‌های جمع‌آوری‌شده از یک شرکت بیمه بزرگ به دانشی دست‌یابیم که برای شرکت‌های بیمه بسیار حائز اهمیت است. هدف اصلی ما یافتن مشتریان احتمالی که قصد استفاده از خدمات بیمه را دارند، می‌باشد [3]. در ادامه، ابتدا به پیشینه‌ی پژوهش‌های انجام شده در حوزه‌ی بیمه با استفاده از تکنیک‌های داده‌کاوی، پرداخته خواهد شد. پس از آن، داده‌های مورد استفاده در این پژوهش معرفی خواهند شد. در بخش 4 ابعاد مختلف روش مورد استفاده در این پژوهش (پیش‌پردازش‌ها، مدل‌سازی و ارزیابی، و نتایج حاصل‌شده) ارائه می‌شود. بخش پایانی مقاله، بحث و نتیجه‌گیری این پژوهش را در بردارد.

2- پیشینه پژوهش

در بازار بیمه نیز همانند بسیاری از حوزه‌های دیگر، داده‌کاوی نفوذ کرده است. پژوهشگران بسیاری با به‌کارگیری تکنیک‌های داده‌کاوی و ارائه‌ی نتایج به مسئولین امر در حوزه‌ی بیمه به رشد روزافزون بیمه‌ها کمک قابل‌توجهی کرده‌اند. در این بخش به‌طور خلاصه به برخی از برجسته‌ترین پژوهش‌های انجام شده این حوزه اشاره می‌کنیم.

کلداگ مله[۲] و کنیتو اسیوتی[۳] در پژوهشی که نتایج آن را در سال 2012 منتشر کردند به بررسی داده‌های بیمه درمانی پرداختند. هدف این پژوهش ارائه‌ی روشی برای تشخیص تقلب با استفاده از داده‌کاوی در بیمه‌های درمانی بوده است. برای این هدف روش‌های تشخیص آنومالی و ماشین بردار پشتیبان[4] را بکار گرفتند. تمرکز این پژوهش بر شناسایی افرادی که با فریب عمدی یا بدرفتاری برای به دست آوردن برخی از مزایا، در قالب هزینه‌های بهداشتی تقلب می‌کنند، بوده است. با نتایج حاصل از این مدل، بازرسان بیمه می‌توانند تحقیقات بیشتری را برای افرادی که توسط این مدل‌ها شناسایی شده‌اند، انجام دهند [4].

لینگ[5] و همکارانش در پژوهشی (سال 1998) به بررسی داده‌های بانک، بیمه و شرکت بونس[6] پرداختند. هدف این پژوهش بازاریابی غیرمستقیم و دادن راه‌حل برای حل مشکلات بازاریابی با استفاده از تکنیک‌های داده‌کاوی بوده است. برای رسیدن به این هدف روش‌های بیزین ساده[7]، نزدیک‌ترین همسایه[8] و شبکه عصبی[9] را بکار گرفتند. مدل ارائه شده در این پژوهش دارای صحت بیش از 70 درصد می‌باشد [5].

---

[2] Kirlidog Melih
[3] Cuneyt Asuk
[4] Support Vector Machine(SVM)
[5] Ling
[6] Bonus
[7] Naive Bayes
[8] Nearest neighbour
[9] Neural network

اسمیت کیت[10] و همکارانش در پژوهشی که آن را در سال 2000 منتشر کردند به بررسی داده‌های بیمه پرداختند. هدف این پژوهش تجزیه و تحلیل الگوهای حفظ بیمه با استفاده از تکنیک‌های داده‌کاوی بوده است. در این پژوهش برای نیل به این هدف، پروسه کشف دانش را در یک چارچوب جامع با استفاده از آزمون فرضیه‌ها[11]، خوشه‌بندی، درخت تصمیم[12] و شبکه عصبی انجام دادند. مدل ارائه شده در این پژوهش راهکاری را به شرکت‌های بیمه ارائه می‌کند تا بتوانند مشتریان خود را حفظ کنند [6].

چو[13] و نای[14] (سال 2003) به بررسی داده‌های بیمه پرداختند. هدف این پژوهش ارائه روشی برای انتخاب عوامل مؤثر در فروش بیمه بوده است. برای رسیدن به این هدف روش‌های تجزیه و تحلیل محرمانه، درخت تصمیم و شبکه عصبی را بکار گرفتند. آن‌ها با پیش‌بینی طول مدت خدمات، حق فروش بیمه و شاخص‌های پایداری عوامل بیمه به مدیران کمک می‌کنند تا عوامل اصلی مؤثر در فروش بیمه را شناسایی کنند [7].

3- داده‌ها

داده‌های مورد استفاده در این پژوهش متعلق به شرکت بیمه هوم‌سایت[15] می‌باشد. این شرکت خدمات مختلف بیمه همانند بیمه سرقت، آتش‌سوزی منزل و بیمه درمانی را برای مشتریان خود ارائه می‌کند و برای بازاریابی هرچه بهتر و یافتن مشتریان بیشتر اقدام به جمع‌آوری داده‌های مربوط به افراد مختلف کرده است. این مجموعه داده در سایت کگل[16] موجود است و دارای 130376 نمونه و برای هر نمونه 299 ویژگی می‌باشد. ویژگی‌های این داده‌ها از نوع عددی[17] و اسمی[18] می‌باشند [3].

برچسب مربوط به هر یک از نمونه‌های این مجموعه داده نشان‌دهنده‌ی این است که آیا فرد اقدام به خرید بیمه می‌کند یا خیر. تعداد افرادی که اقدام به خرید بیمه کرده‌اند 24089 نفر و تعداد افرادی اقدام به خرید بیمه نمی‌کنند 106287 نفر می‌باشد. در هنگام مدل‌سازی از اعتبارسنجی متقابل[19] 10 بخشی[20] استفاده می‌کنیم؛ 0.9 از نمونه‌ها برای آموزش[21] و از 0,1 باقی‌مانده برای آزمایش[22] استفاده می‌شوند و این عمل 10 مرتبه اجرا شده و در هر بار اجرا یکی از 10 قسمت برای آزمایش و باقی نمونه‌ها برای آموزش در نظر گرفته می‌شود (جدول 1) [3].

جدول 1- مشخصات داده‌های شرکت بیمه مورد استفاده در این پژوهش

داده‌ها	نمونه	ویژگی‌ها
کل داده‌ها	130376	299
آموزش	117338	299
آزمایش	13038	298

[10] Smith Kate
[11] Hypothesis testing
[12] PART
[13] Cho
[14] Ngai
[15] Homesite
[16] kaggle
[17] Numeric
[18] Nominal
[19] Cross validation
[20] 10 fold
[21] Train
[22] Test

4- روش انجام کار

4-1- پیش‌پردازش

یکی از مهم‌ترین مراحل انجام یک پژوهش داده‌کاوی، پیش‌پردازش داده‌ها می‌باشد. داده‌های ما داری تعدادی مقادیر ازدست‌رفته[23] هستند. دو ویژگی از مجموع 299 ویژگی این مجموعه داده، دارای مقادیر ازدست‌رفته می‌باشند. ازآنجاکه مقادیر بیش از 90 درصد نمونه‌ها، برای این دو ویژگی یا ازدست‌رفته هستند و یا پرتکرارترین مقدار ثابت را دارند (به‌عبارت‌دیگر تنوع مقادیر در این دو ستون پایین است) از این 2 ویژگی در ساخت مدل‌ها صرف نظر می‌کنیم. پایین بودن تنوع مقادیر در یک ویژگی باعث می‌شود که آن ویژگی نتواند در ساخت مدل مؤثر واقع شود [2]. با در نظر داشتن این نکته تعداد 2 ویژگی که دارای مقدار ثابتی هستند نیز از مجموعه‌ی داده‌ها حذف شدند.

یکی از ویژگی‌های موجود در این داده‌ها، از نوع تاریخ می‌باشد. برای اینکه بتوانیم از این ویژگی در ساخت مدل به شکل مؤثرتری استفاده کنیم به جای ویژگی تاریخ دو ویژگی روز و ماه را از آن استخراج کردیم.

به کمک محاسبه‌ی مقادیر همبستگی[24] می‌توان از چگونگی ارتباط میان ویژگی‌ها مطلع شد. هر چه این مقادیر بیشتر باشند همبستگی میان ویژگی‌ها بیشتر است و به این معنا است که این دو ویژگی حاوی اطلاعات تقریباً یکسانی هستند و رفتاری مشابه دارند [2]. از بین ویژگی‌هایی که دارای همبستگی بالای 0/8 با یکدیگر هستند، آن ویژگی که همبستگی کمتری با ستون هدف دارد را حذف می‌کنیم. علاوه بر این، ویژگی‌هایی که مقدار همبستگی آن‌ها با ستون هدف کمتر از 0/01 بودند، نیز حذف شدند. با انجام این پردازش، 30 ویژگی از مدل‌سازی ما حذف شدند. لازم به ذکر است که این مقادیر آستانه به‌صورت تجربی برای این پژوهش حاصل شده‌اند.

نویزها[25] به دلایل مختلفی، چون بروز اشتباه در فرآیند جمع‌آوری و یا وارد کردن اطلاعات به سیستم بوجود می‌آید؛ شناسایی آن‌ها به ما کمک می‌کند تا با آگاهی و دقت بیشتری به طراحی مدل بپردازیم [2]. پس از یافتن نمونه‌های نویزدار، آن‌ها را حذف کردیم؛ این کار باعث شد که تعداد 20 ویژگی دارای مقادیری تقریباً ثابت شوند، همان‌طور که پیش از این اشاره شده ویژگی‌های دارای مقادیر ثابت در مدل‌سازی مؤثر نیستند در نتیجه از این ویژگی‌ها صرف نظر شد.

لازم به ذکر است که در پایان مرحله‌ی پیش‌پردازش، در مجموع 245 ویژگی برای ساخت مدل باقی مانده است.

4-2- مدل‌سازی و ارزیابی

در این پژوهش برای مدل‌سازی از سه الگوریتم معروف داده‌کاوی استفاده کرده‌ایم. این الگوریتم‌ها عبارت‌اند از:

- درخت تصمیم، یکی از الگوریتم‌های پرکاربرد داده‌کاوی است. در این الگوریتم که از نوع تکرارشونده و حریصانه است، یک درخت تصمیم با یافتن بهترین ویژگی جداکننده در هر گام و سپس جداسازی متوالی داده‌ها به گروه‌های مجزا، ساخته می‌شوند [2].
- بیزین ساده، بر پایه‌ی قضیه بیز برای مدل‌سازی پیش‌گویانه ارائه شده است. قضیه بیز از روشی برای دسته‌بندی پدیده‌ها بر پایه احتمال وقوع یا عدم وقوع یک پدیده استفاده می‌کند و احتمال رخ دادن یک پدیده محاسبه و دسته‌بندی می‌شود [2].
- جنگل تصادفی[26]، یک الگوریتم یادگیری ماشین است که اغلب اوقات نتایج بسیار خوبی را حتی بدون تنظیم فراپارامترهای آن، فراهم می‌کند. این الگوریتم درخت‌های تصمیم زیادی را تولید می‌کند. هر درخت یک طبقه‌بند مجزا است و نتیجه‌ی نهایی در جنگل تصادفی با رأی‌گیری از این درخت‌ها حاصل می‌شود. رأی‌گیری از بین چندین طبقه‌بند باعث می‌شود که مدل نهایی در برابر نویز مقاوم باشد [2].

---

[23] Null
[24] Correlation
[25] Noise
[26] Random forest

برای ارزیابی و مقایسه‌ی مدل‌ها از معیارهای صحت[27]- بازخوانی[28] - دقت[29]- معیار اف[30] استفاده کرده‌ایم:

- صحت: اولین معیار یا سنجه‌ای که به ذهن می‌رسد، معیار صحت یا میزان تشخیص درست مدل است؛ یعنی نسبت تشخیص‌های درست (مثبت واقعی[31]+ منفی واقعی[32]) به کل داده‌ها. این معیار برای ارزیابی مدل‌ها در زمان استفاده از داده‌های نامتوازن (یعنی تفاوت زیادی در تعداد نمونه‌های دسته‌ها وجود دارد) کافی نیست؛ زیرا این عدم توازن باعث می‌شود مدل‌های متمایل به دسته پرتعداد، شناسایی نشوند و به اشتباه یک مدل بد، مدل خوبی معرفی شود.
- بازخوانی: ارزیابی اینکه کل نمونه‌های واقعاً مثبت شامل نمونه‌هایی است که درست، مثبت شناسایی شده‌اند (مثبت واقعی) و نمونه-هایی که مثبت بوده‌اند اما نادرست، منفی شناسایی شده‌اند (منفی اشتباه[33]). در این معیار بر تعداد نمونه‌های مثبت شناسایی شده به کل نمونه‌های مثبت تمرکز می‌شود.
- دقت: در کنار معیار بازخوانی معیار دیگری را به نام دقت، برابر تعداد نمونه‌های تشخیصی مثبت واقعی به کل نمونه‌های مثبت اعلام شده تعریف می‌شود تا میزان مثبت‌های اشتباه[34] هم در نظر گرفته شود.
- معیار اف: میانگین هارمونیک بازخوانی و دقت می‌باشد [2].

4-3- نتایج

در بخش‌های پیشین به مشخصات داده‌های این پژوهش، پیش‌پردازش‌های انجام شده بر روی داده‌ها، الگوریتم‌های مورد استفاده برای ساخت مدل و معیارهای ارزیابی در نظر گرفته شده اشاره شد. پس از پیاده‌سازی و انجام آزمایش‌ها نتایج قابل‌توجهی حاصل شد. خلاصه‌ای از نتایج این آزمایش‌ها در جدول 2 ارائه شده است. همان‌طور که در این جدول مشخص است، الگوریتم جنگل تصادفی در همه‌ی معیارهای ارزیابی توانسته است نتایج بهتری را ارائه کند. جهت ارائه‌ی جزئیات بیشتر، ماتریس سردرگمی[35]، الگوریتم جنگل تصادفی در جدول 3 ارائه شده است.

جدول 2- مقایسه الگوریتم‌های مختلف معیارهای ارزیابی صحت، بازخوانی، دقت و معیار اف

	صحت	بازخوانی	دقت	معیار اف
جنگل تصادفی	0,996172	0,996039	0,999263	0,997648
بیزین ساده	0,869968	0,880663	0,955845	0,916715
درخت تصمیم	0,953703	0,985885	0,958397	0,971946

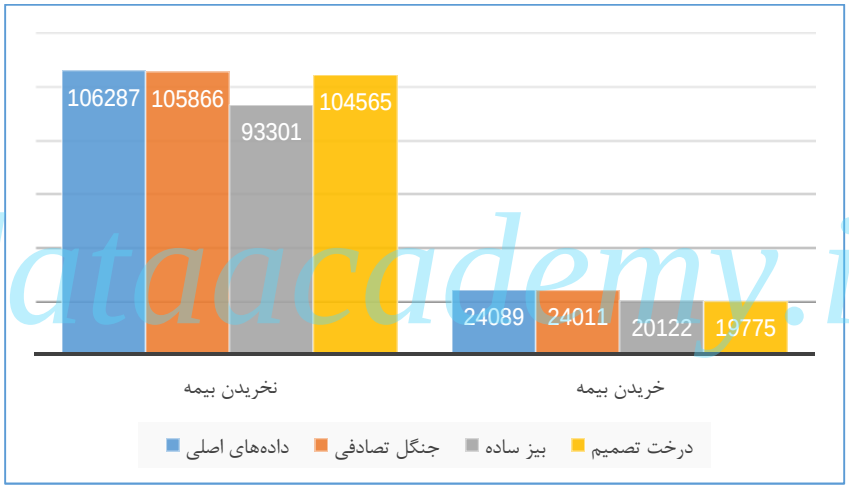
جدول 3- نتیجه مدل‌سازی با الگوریتم جنگل تصادفی

	0	1	جمع
0	105866	421	106287
1	78	24011	24089
جمع	105944	24432	130376

[27] Accuracy
[28] Recall
[29] Precision
[30] F-measure
[31] True Positive (TP)
[32] True Negative (TN)
[33] False Negative (FN)
[34] False Positive (FP)
[35] Confusion matrix

5- بحث و نتیجه‌گیری

هدف پژوهش ما، ارائه‌ی یک مدل پیشگو برای مشتریان احتمالی استفاده از خدمات بیمه می‌باشد. در سال‌های اخیر صنعت بیمه رشدی چشمگیر داشته است و بازاریابی موفق یکی از اهداف اصلی شرکت‌های بیمه است؛ پیدا کردن افرادی که احتمال می‌رود از خدمات بیمه استفاده کنند، بسیار مهم است و منجر به مدیریت هرچه بهتر سرمایه و هزینه‌ها می‌شود. نتایج حاصل از آزمایش‌های ما نشان داد که تکنیک‌های داده‌کاوی می‌توانند رویکردی مناسب برای دست‌یابی به چنین دانش سودمندی باشند. داده‌های این پژوهش بسیار نامتوازن هستند اما ما با پیش‌پردازش‌هایی مناسب و استفاده از الگوریتم‌هایی کارا توانستیم به مدل پیش‌گویی با صحت بیش از 99 درصد دست‌یابیم. نمودار ارائه شده در شکل 1 فروانی مثبت صحیح (افراد خریدار بیمه) و منفی صحیح (افرادی که بیمه نمی‌خرند) را برای الگوریتم‌های مختلف و داده‌های اصلی با هم مقایسه می‌کند. همان‌طور که در بخش پیشین اشاره شد، الگوریتم جنگل تصادفی توانسته است بهترین نتایج را ارائه دهد. با استفاده از نتایج این پژوهش مسئولین بیمه می‌توانند، مشتریان احتمالی خود را پیدا و برای فروش خدمات خود بر روی آن افراد سرمایه‌گذاری کنند.



شکل 1: مقایسه‌ی فروانی دو کلاس خریدن/نخریدن بیمه برای الگوریتم‌های مختلف و داده‌های اصلی

مراجع

[1] J. Hsia, E. Kemper, C. Kiefe, J. Zapka, S. Sofaer, M. Pettinger, D. Bowen, M. Limacher, L. Lillington, E. Mason, and Women's Health Initiative Investigators, "The importance of health insurance as a determinant of cancer screening: evidence from the Women's Health Initiative," *Preventive medicine*, vol. 3, no. 31, pp. 261-270, 2000.

[2] J. Han, J. J. Pei, and M. Kamber,, . Data mining: concepts and techniques, 3nd ed., Morgan Kaufmann, 2011.

[3] Kaggle Inc, "kaggle," 1 April 2001. [Online]. Available: https://www.kaggle.com. [Accessed 24 January 2019].

[4] M. Kirlidog, and C. Asuk, , "A fraud detection approach with data mining in health insurance," *Procedia-Social and Behavioral Sciences*, vol. 62, pp. 989-994, 2012.

[5] C. X. Ling, and C. Li,, "Data mining for direct marketing: Problems and solutions," in *Kdd*, Newyork, 1998.

[6] K.A. Smith, R.J. Willis, and M. Brooks, , "An analysis of customer retention and insurance claim patterns using data mining: A case study," *Journal of the operational research society*, vol. 5, no. 51, pp. 532-541, 2000.

[7] V. Cho, and E.W. Ngai,, "Data mining for selection of insurance sales agents," *Expert systems*, vol. 3, no. 20, pp.

123-132, 2003.

Finding Potential Insurance Customers Using Data-mining Techniques

Amir Oliaei, Undergraduate student, Kashmar Higher Education Institute
Ali Hazeri, Undergraduate student, Kashmar Higher Education Institute
Mohammad Jamali, Undergraduate student, Kashmar Higher Education Institute
Atefeh Khazaei, PhD in Software Engineering, Kashmar Higher Education Institute

Abstract

In recent years, the insurance industry has grown dramatically, with various companies and services. The successful marketing is one of the main goals of insurance companies; finding people who would like to use insurance services is very important and leads to better management of the fund and expense. In this research, using a dataset containing various information such as customer characteristics, sales information, specific coverage, geographic location and status of properties, we present a model that predicts the probability of buying insurance by persons. To provide this model, we have applied various data-mining techniques and algorithms. After data preprocessing, we used Random Forest, Decision Trees, and Naïve Bayes algorithms to construct the model. Random Forest algorithm shows the best performance with more than 99% accuracy. The results of this research show that the data-mining techniques can be useful to predict potential customers and help insurers to find their customers and invest on them to sell their services. This model reduces the costs, such as the advertising cost.

Keywords: Insurance customer, Data-mining, Random Forest, Decision Tree, Naïve Bayes.