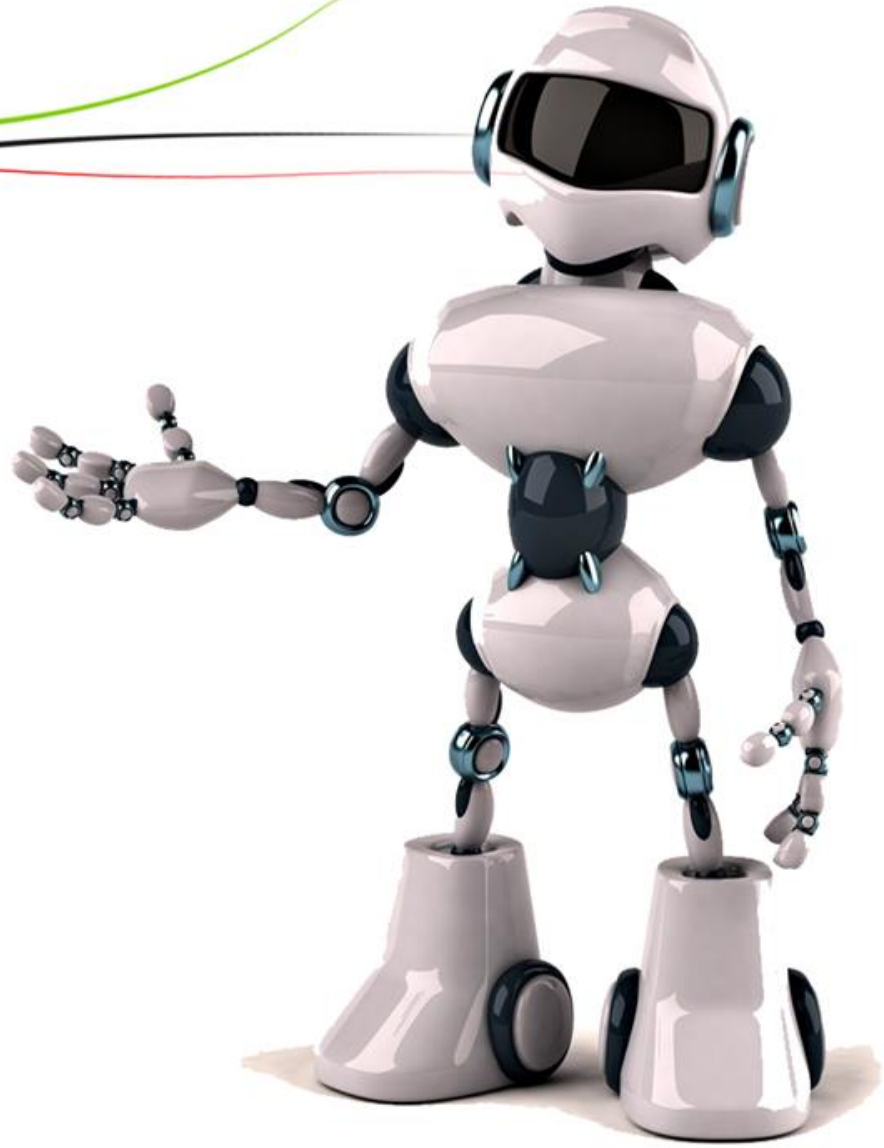
Credit Card Fraud Detection - Machine learning methods



استاد: دکتر یغمایی

استاد حل تمرین: مهندس شکری

درس: یادگیری ماشین

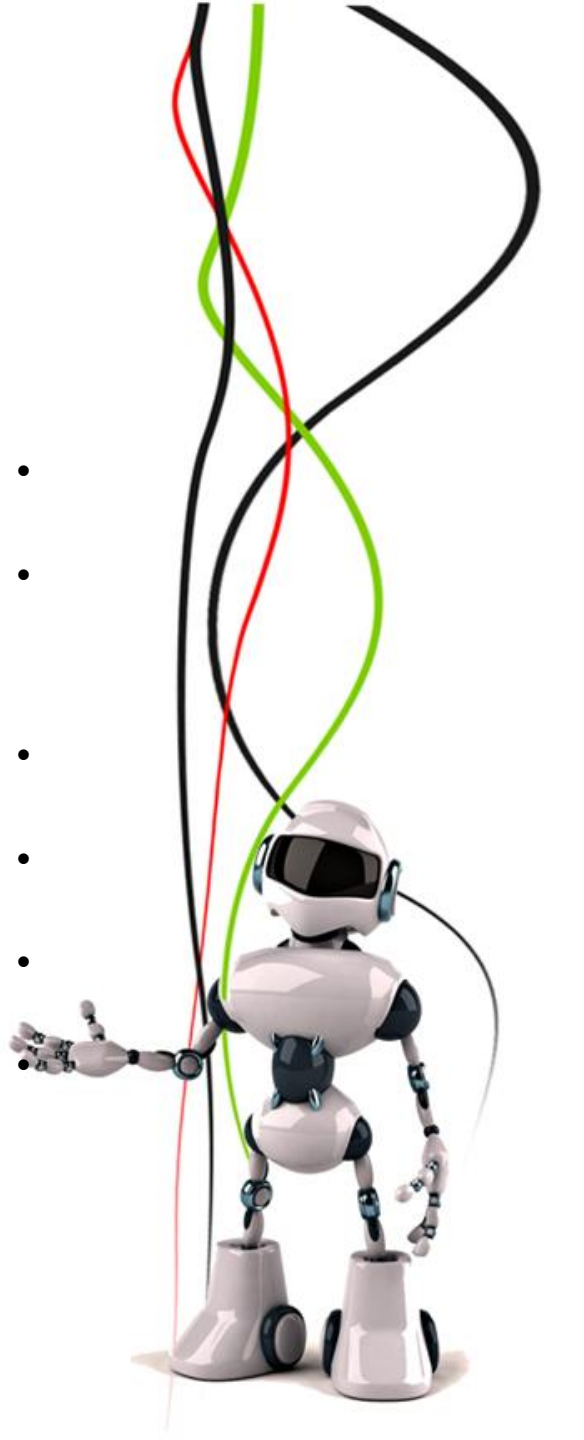
دانشجو: همایون طوسی

پاییز: ۱۴۰۰

چکیده

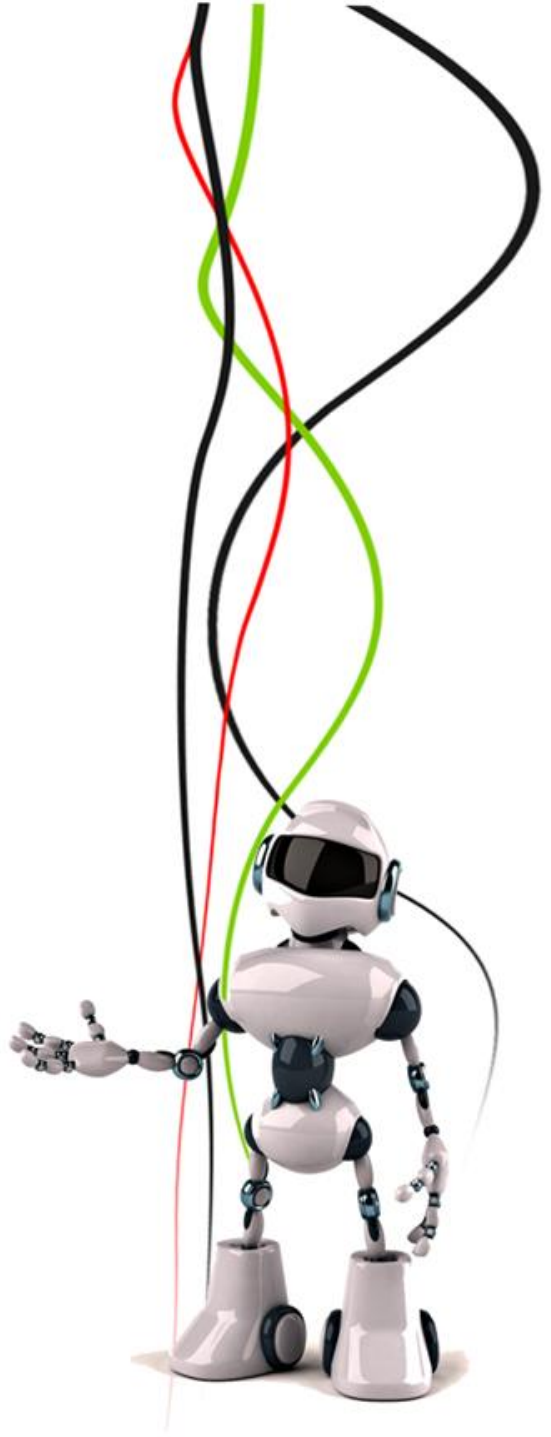
تشخیص تقلب در کارت اعتباری با استفاده از متدهای یادگیری ماشین

- تقلب در کارت اعتباری به از دست دادن فیزیکی کارت اعتباری یا از دست دادن اطلاعات حساس کارت اعتباری اشاره دارد
- این تحقیق چندین الگوریتم را نشان می‌دهد که می‌توانند برای طبقه‌بندی معاملات به عنوان تقلب یا تقلب واقعی مورد استفاده قرار گیرند. مجموعه داده تشخیص تقلب کارت اعتباری در این تحقیق مورد استفاده قرار گرفت
- از آنجا که مجموعه داده‌ها بسیار نامتعادل بودند، تکنیک SMOTE برای نمونه‌گیری بیش از حد مورد استفاده قرار گرفت.
- علاوه بر این، انتخاب ویژگی انجام شد و مجموعه داده‌ها به دو بخش داده‌های آموزشی و داده‌های تست تقسیم شدند.
- الگوریتم های مورد استفاده در این آزمایش عبارتند از رگرسیون لجستیک، جنگل تصادفی، نایو بیز و پرسپترون چند لایه
- نتایج نشان می‌دهد که هر الگوریتم می‌تواند برای تشخیص تقلب کارت اعتباری با دقت بالا مورد استفاده قرار گیرد. مدل پیشنهادی می‌تواند برای تشخیص دیگر بی‌نظمی‌ها مورد استفاده قرار گیرد.



مقدمه

- تعداد رو به رشدی از شرکت‌های جدید در سراسر جهان وجود دارد
- تمام این شرکت‌ها در تلاش برای ارائه بهترین کیفیت خدمات به مشتریان خود هستند.
- به منظور موفقیت در این امر، شرکت‌ها داده‌های زیادی را به صورت روزانه پردازش می‌کنند.
- این داده‌ها از تعداد زیادی از منابع می‌آیند و در فرمت های مختلف هستند. علاوه بر این، این داده‌ها شامل برخی از بخش‌های کلیدی کسب وکار آینده شرکت است
- به همین دلیل، شرکت‌ها باید آن داده‌ها را ذخیره کنند، آن را پردازش کنند و آنچه واقعا مهم است، تا آن را ایمن نگه دارند .
- بدون تامین امنیت داده‌ها، بسیاری از آن‌ها می‌توانند توسط شرکت‌های دیگر و یا حتی بدتر از آن مورد استفاده قرار گیرند، می‌توانند به سرقت برده شوند .
- در بیشتر موارد، اطلاعات مالی به سرقت می‌رود که می‌تواند به کل شرکت یا فرد آسیب برساند.
- میزان کلاهبرداری به طور نسبی یکسان است و یا به دلیل سیستم‌های تشخیص تقلب پیچیده کاهش‌یافته است. با این حال، کلاهبرداران به طور مداوم با روش‌های جدیدی برای سرقت اطلاعات مواجه می‌شوند.



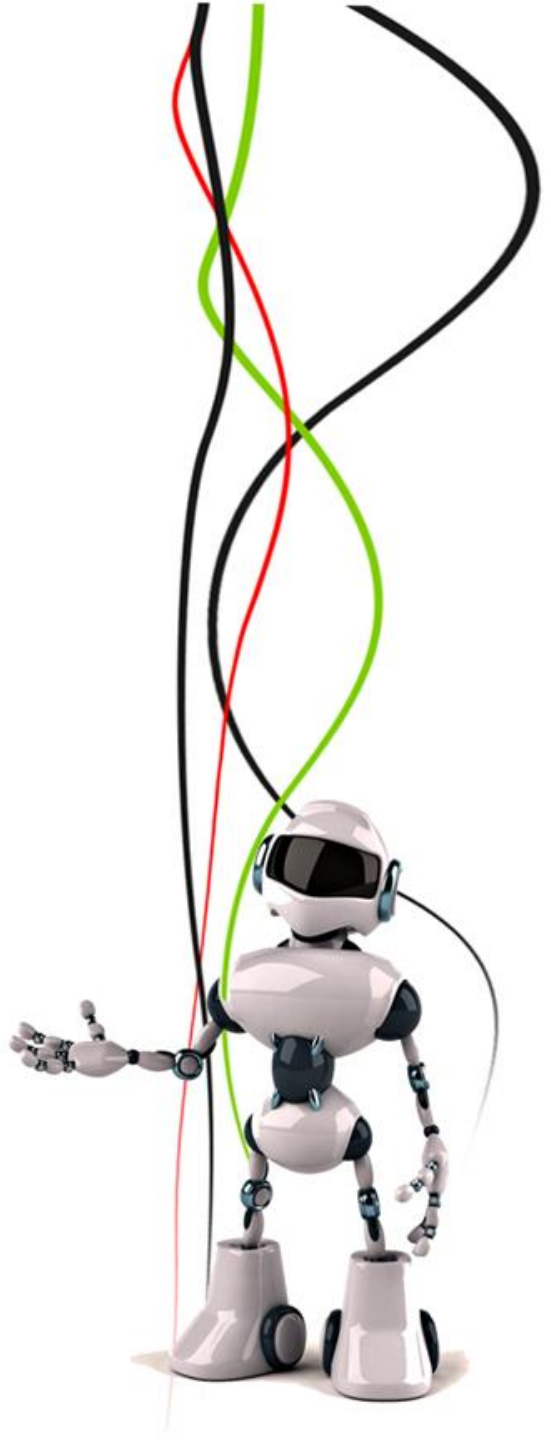
- دو نوع کلاهبرداری کارت اعتباری وجود دارد:
- سرقت کارت فیزیکی است
- سرقت اطلاعات حساس از کارت است، مانند شماره کارت، کد **CVV**، نوع کارت و غیره.

نکته: با سرقت اطلاعات کارت اعتباری، کلاهبردار می‌تواند مقدار زیادی پول را به سرقت برده و یا قبل از این که صاحب کارت متوجه شود، مقدار زیادی خرید انجام دهد. به همین دلیل، شرکت‌ها از روش‌های مختلف یادگیری ماشین برای تشخیص اینکه کدام تراکنشها جعلی هستند و کدام نیستند، استفاده می کنند.

هدف از این مقاله، تجزیه و تحلیل الگوریتم های مختلف یادگیری ماشین، مانند :

- (رگرسیون لجستیک LR است)
- Random forest (RF)
- Naïve bayes (NB)
- (Mlp)

به منظور اینکه کدام الگوریتم مناسب تر برای تشخیص تقلب است



فعالیت‌های کلاهبرداری باعث ضرر و زیان زیادی می‌شوند، که محققان را بر آن داشت تا راه حلی بیابند که بتواند کلاهبرداری‌ها را شناسایی کرده و از آن‌ها جلوگیری کند. تاکنون چندین روش پیشنهاد و آزمایش شده‌اند. برخی از آن‌ها به طور خلاصه در زیر مرور شده‌اند.

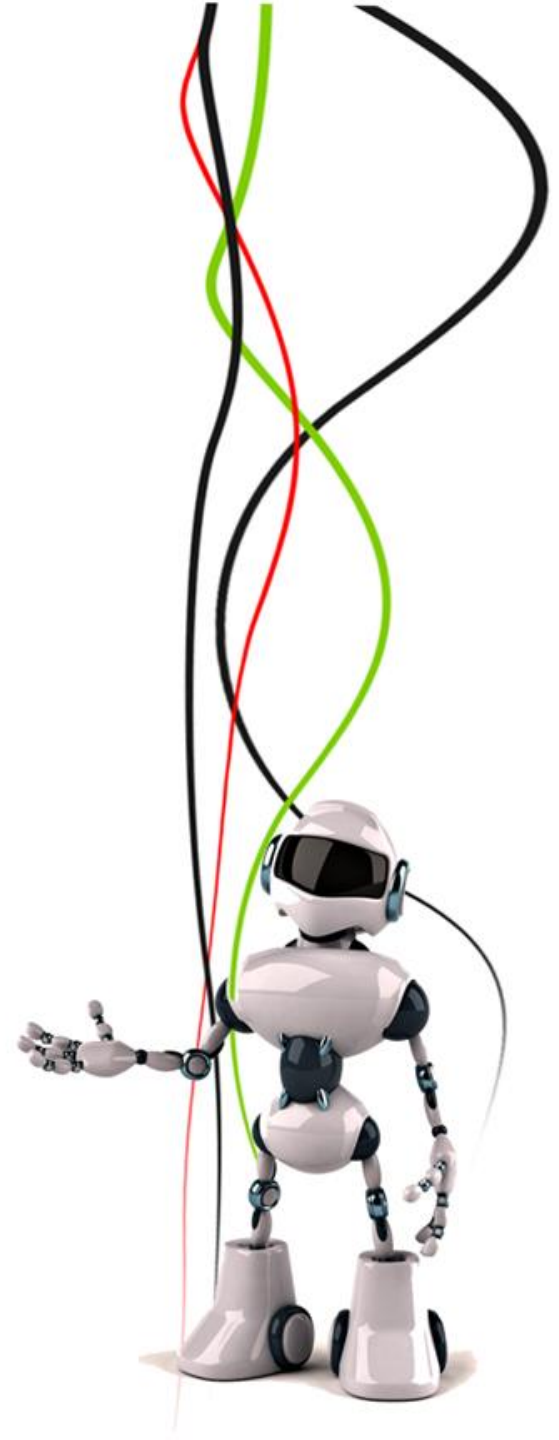
- الگوریتم های کلاسیک مانند بوت کردن گرادیان (GB)
- ماشین بردار پشتیبان (SVM)
- درخت تصمیم
- RF ، LR

مفید بودن آن‌ها را اثبات کرده‌اند

مقایسه‌ای بین مدل‌های مبتنی بردرخت تصمیم ، LR ، RF انجام شد. در میان این سه مدل :

- RF با دقت ۹۵,۵ ٪ بهترین
- درخت تصمیم با دقت ۹۴,۳ ٪
- LR با دقت ۹۰ ٪ بهترین بودند

نزدیک‌ترین همسایه‌ها ((KNNو تکنیک‌های تشخیص داده‌های پرت نیز می‌توانند در تشخیص تقلب موثر باشند. ثابت شده‌است که آن‌ها در به حداقل رساندن نرخ هشدار اشتباه و افزایش نرخ تشخیص تقلب مفید هستند.



هدف اصلی ما این است که نشان دهیم الگوریتم های مختلف یادگیری ماشین می‌توانند نتایج مناسبی را با پیش‌پردازش مناسب ارائه دهند.

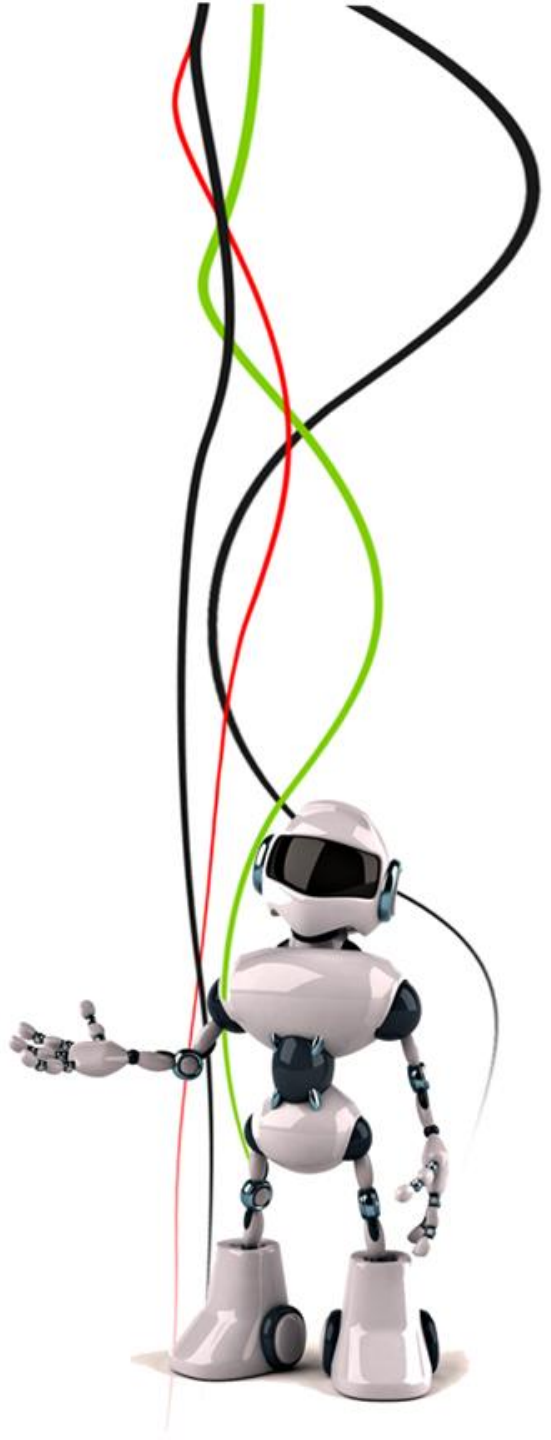
مقاله ذکر شده از تکنیک نمونه‌گیری کم‌تر استفاده کردند، و این انگیزه‌ای برای استفاده از یک روش نمونه‌گیری بیش از حد متفاوت بود.

با توجه به حقایق داده‌شده، نویسندگان این مقاله تصمیم به مقایسه مناسب بودن LR، RF، NB،MLP برای تشخیص تقلب کارت اعتباری گرفتند. به منظور دستیابی به این هدف، آزمایشی انجام شد.

A. دیتاست

در این تحقیق از مجموعه داده تشخیص تقلب کارت اعتباری استفاده شده‌است که می‌تواند از کاگل دانلود شود. این مجموعه داده شامل تراکنش هایی است که در دو روز گذشته توسط دارندگان سهام اروپایی در سپتامبر ۲۰۱۳ انجام شده‌است. از آنجا که برخی از متغیرهای ورودی شامل اطلاعات مالی هستند

به منظور ناشناس نگه داشتن این داده‌ها، تبدیل این متغیرهای ورودی انجام شد. سه مورد از این ویژگی‌ها تبدیل نشدند. ویژگی "time" زمان بین اولین معامله و هر معامله دیگر در مجموعه داده را نشان می‌دهد. ویژگی "Amount" مقدار تراکنش های انجام‌شده توسط کارت اعتباری است. ویژگی class و تنها ۲ ارزش می‌گیرد: ارزش ۱ در مورد معامله کلاهبرداری و درغیراینصورت۰.



B : پیش‌پردازش

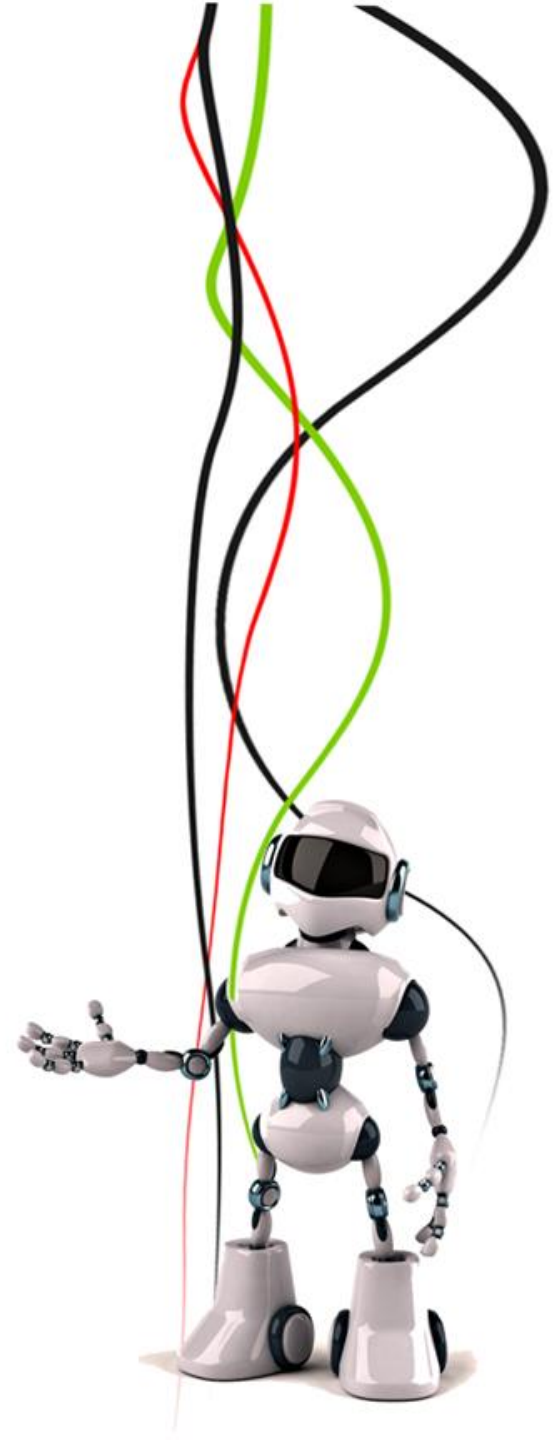
انتخاب ویژگی یک تکنیک اساسی است که متغیرهایی را انتخاب می‌کند که بیش‌ترین ارتباط را با مجموعه داده داده‌شده دارند. با انتخاب دقیق ویژگی‌های مناسب و حذف ویژگی‌های کم‌تر مهم می توان بیش از حد مناسب را کاهش داد، دقت را بهبود بخشید و زمان آموزش را کاهش داد.

الگوریتم های مورد استفاده در این آزمایش

- رگرسیون منطقی
- نایو بیز
- جنگل تصادفی
- پرسپترون چندلایه

نتایج:

- برای تعیین اینکه کدام الگوریتم برای مشکل تشخیص تراکنش های کلاهبرداری مناسب‌تر است، از معیارهای مختلفی برای مقایسه الگوریتم استفاده شده‌است. اکثر معیارهای مورد استفاده برای تعیین نتایج الگوریتم های یادگیری ماشین عبارتند از دقت، یادآوری .
- ارزیابی عملکرد یک مدل مطابق با این معیارها انجام شد. مدل‌ها بر روی داده‌های اصلی و بیش از حد نمونه‌برداری شده آزمایش شدند و نتایج نشان داد که نمونه‌گیری بسیار مهم است.



از آنجا که مجموعه تست شامل ۲۰ % کل مجموعه داده است، مجموع کل نمونه‌ها ۵۶۹۶۲ است. از مجموع ۹۸ معامله کلاهبرداری،

مدل (LR)جدول ۱به این نتیجه رسید:

TABLE 1:CONFUSION MATRIX FOR LR

		Predicted	
		0	1
Actual	0	55424	1440
	1	8	90

دقت: ۵۸.۸۲ درصد

به یاد آوری: ۹۱.۸۴ % دقت: ۹۷.۴۶ %.

جدول ۱: ترکیب برای LR

TABLE 2:CONFUSION MATRIX FOR NB

		Predicted	
		0	1
Actual	0	56444	420
	1	17	81

RF model obtained following results (Table 3):

مدل NB نتایج زیر را به دست آورد

(جدول ۲): ۸۲.۶۵test %، دقت: ۹۹.۲۳ %

جدول ۲: ترکیب برای NB

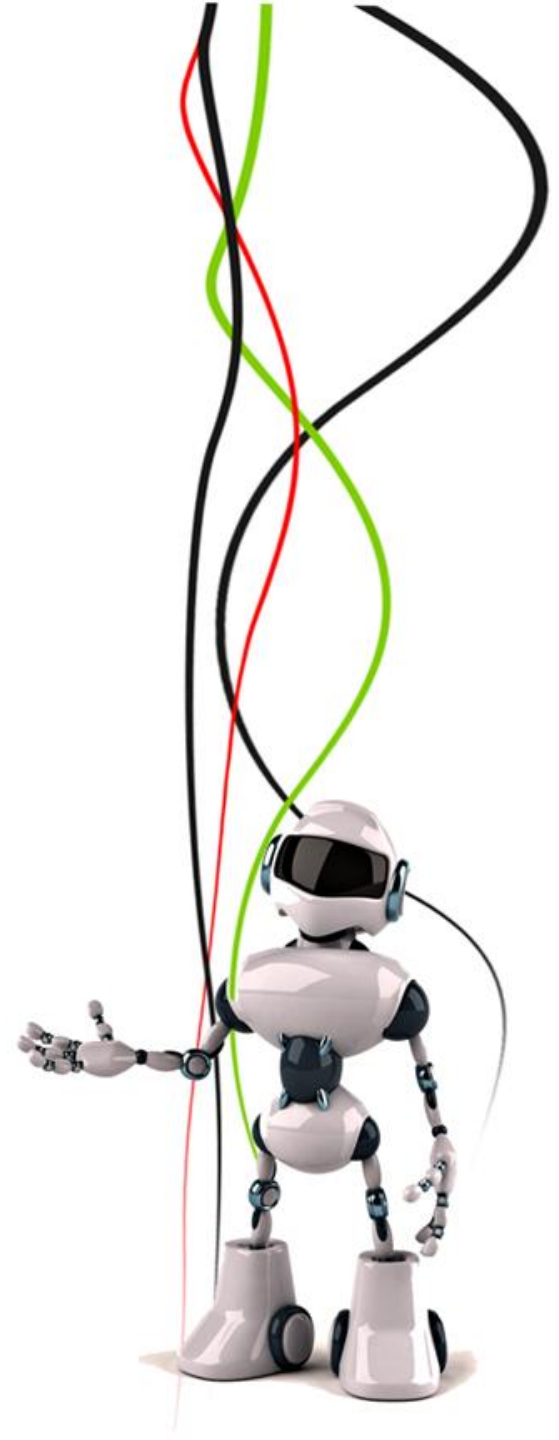


TABLE 3:CONFUSION MATRIX FOR RF

		Predicted	
		0	1
Actual	0	56861	3
	1	18	80

MLP model obtained following results (Table 4):

- precision: 79.21%,
- recall: 81.63%,
- accuracy: 99.93%

TABLE 4:CONFUSION MATRIX FOR MLP

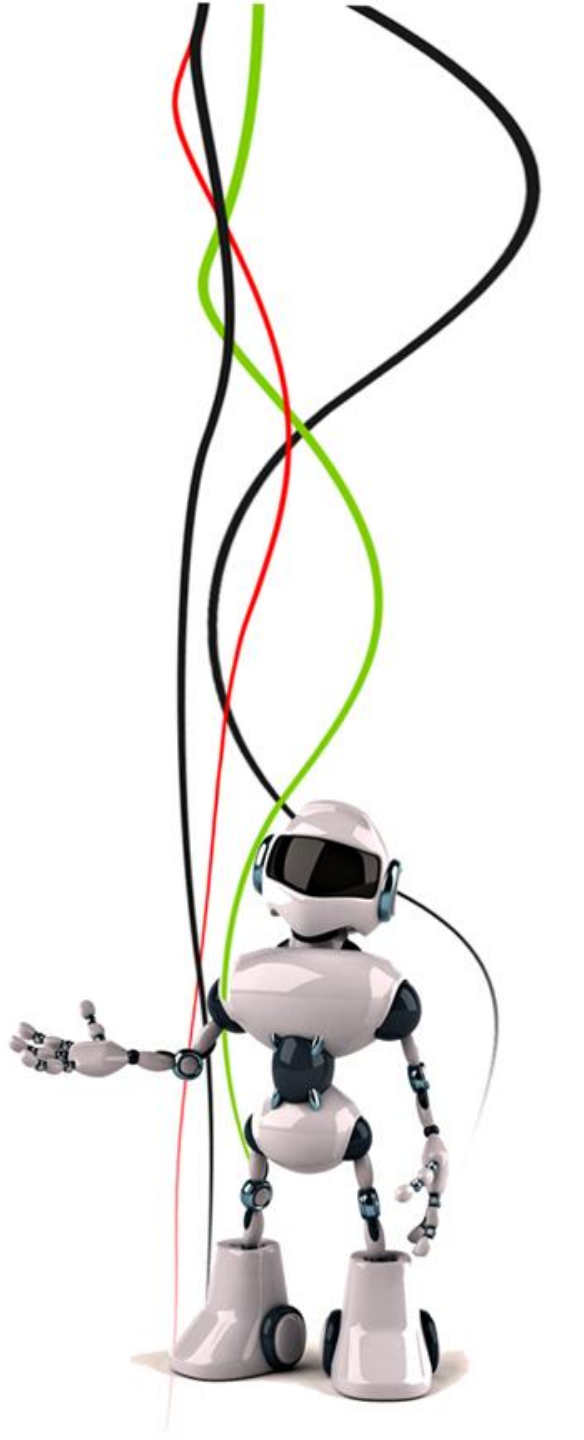
		Predicted	
		0	1
Actual	0	56843	21
	1	18	80

مدل RF نتایج زیر را به دست آورد

- (جدول ۳): دقت: ۹۶.۳۸ %
- یادآوری: ۸۱.۶۳ %
- دقت: ۹۹.۹۶ درصد
- جدول ۳: ترکیب شیمیایی برای RF

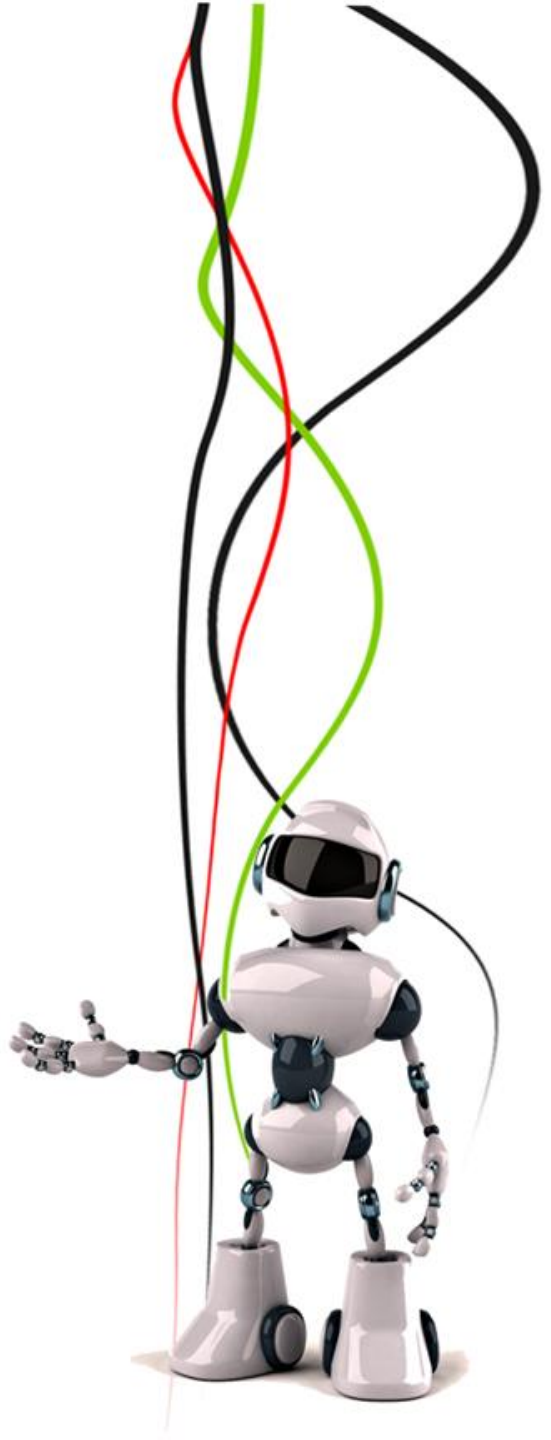
مدل MLP نتایج زیر را به دست آورد

- (جدول ۴): دقت: ۷۹.۲۱ %
- یادآوری: ۸۱/۶۳ %، دقت: ۹۹/۹۳ %



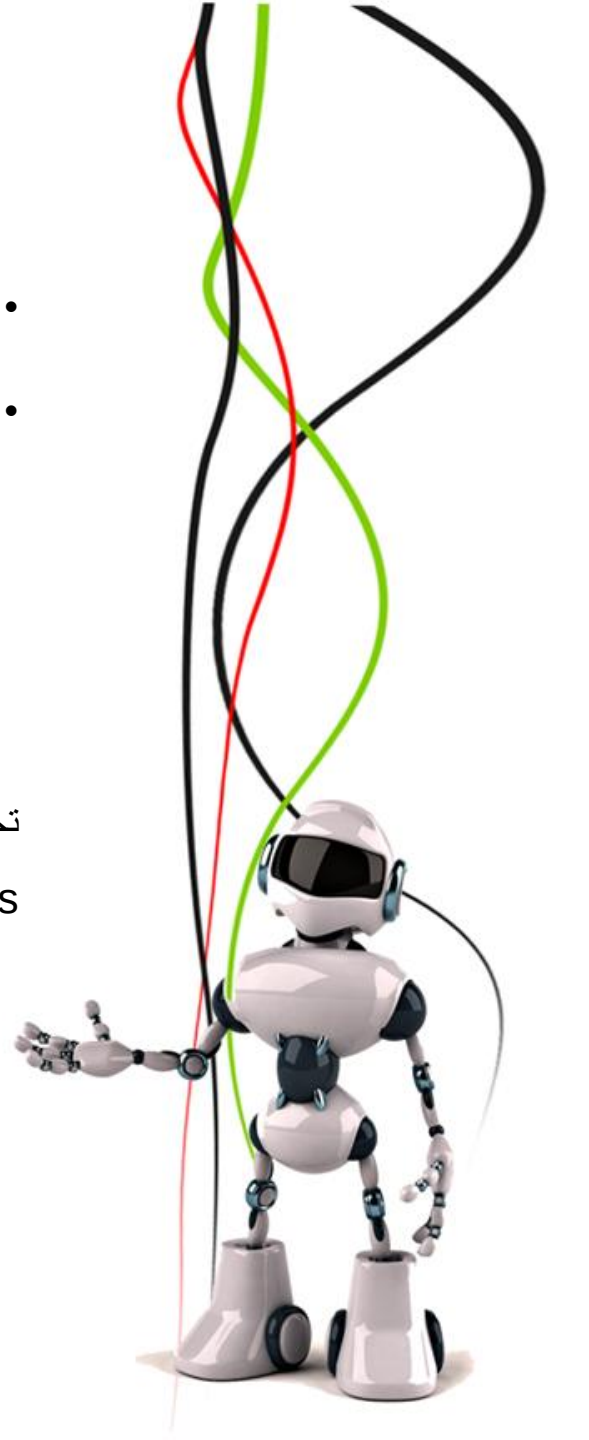
با تجزیه و تحلیل نتایج به‌دست‌آمده، واضح است که دقت بسیار بالا است، اگرچه این بدان معنی نیست که نتایج کامل هستند:

- مقایسه نتایج به‌دست‌آمده با نتایج به‌دست‌آمده در تحقیقات بر روی همان مجموعه داده، با الگوریتم های کلاسیک نشان می‌دهد که نمونه‌گیری بیش از حد از داده‌ها می‌تواند نرخ تشخیص تقلب را بهبود بخشد ثابت شده‌است که الگوریتم های کلاسیک می‌توانند به اندازه الگوریتم های یادگیری عمیق موفق باشند.
- الگوریتم های یادگیری عمیق را به عنوان الگوریتم بهینه برای این نوع از مسائل نشان می‌دهند، اما باید با توجه به شرایطی که از این الگوریتم ها باید استفاده شود، تصمیم‌گیری شود.
- به عنوان مثال، شبکه‌های عمیق با داده‌های بیشتر بهتر کار می‌کنند و می‌توانند راحت‌تر از الگوریتم های کلاسیک با حوزه‌های مختلف سازگار شوند. از سوی دیگر، اگر داده‌های زیادی وجود نداشته باشد، احتمالا بهتر است که با الگوریتم های کلاسیک کار کنیم. تفسیر این الگوریتم ها هم از نظر مالی و هم از نظر محاسباتی آسان‌تر و ارزان‌تر است



- هدف اصلی این مقاله مقایسه برخی از الگوریتم های یادگیری ماشین برای تشخیص تقلب ، معاملات بود.
- از این رو، مقایسه انجام شد و مشخص شد که (Random Forest algorithm) بهترین نتایج را ارائه می‌دهد، به عنوان مثال، بهترین طبقه‌بندی را ارائه می‌دهد که آیا تراکنش‌ها تقلب هستند یا خیر. این امر با استفاده از معیارهای مختلف، مانند یادآوری، دقت ایجاد شد. برای این نوع مشکل، مهم است که به یاد آوری با ارزش بالا داشته باشیم. انتخاب ویژگی و تعادل مجموعه داده نشان داده‌است که در دستیابی به نتایج قابل توجه بسیار مهم است.

تحقیقات بیشتر باید بر روی الگوریتم های مختلف یادگیری ماشین مانند الگوریتم های ژنتیک، و انواع مختلف stacked classifiers، همراه با انتخاب ویژگی‌های گسترده برای به دست آوردن نتایج بهتر تمرکز کند



از حسن توجه شما سپاسگذارم