کاربرد یادگیری ماشین در فرایند جانهی داده‌های آمارگیری‌های خانواری

مجری

لیدا کلهری ندرآبادی

همکاران

مریم میرزایی

عباس مرادی

حسن رنجی

بهنام بهرامی

اشکان شباک

پژوهشکده‌ی آمار

گروه پژوهشی طرح‌های فنی و روش‌های آماری

پاییز ۱۴۰۳

بسم الله الرحمن الرحیم

کد شناسه: RP-0303

عنوان: کاربرد یادگیری ماشین در فرایند جانهی داده‌های آمارگیری‌های خانواری

مجری: لیدا کلهری ندرآبادی

همکاران: مریم میرزایی، عباس مرادی، حسن رنجی، بهنام بهرامی، اشکان شباک

گروه پژوهشی: طرح‌های فنی و روش‌های آماری

تاریخ انتشار: پاییز ۱۴۰۳

نوبت انتشار: اول

صفحه‌آرا: ساناز مهندسی

طراح جلد: ساناز مهندسی

❖ حق مالکیت معنوی این طرح پژوهشی متعلق به پژوهشکده‌ی آمار است و نقل مطالب فقط با ذکر مأخذ مجاز است.

پژوهشکده‌ی آمار

تهران، خیابان دکتر فاطمی، خیابان باباطاهر، خیابان سرتیپ فکوری،
شماره‌ی ۱۴۵
۰۲۱ ۸۸۶۳۰۴۴۰ -۳
www.srtc.ac.ir

به نام خداوند جان و خرد

پیش‌گفتار

داده‌های همراه با خطای اندازه‌گیری و بی‌پاسخی از مسائلی است که همه مراکز ملی آمار تقریباً در هر مجموعه داده‌ای با آن‌ها سروکار داشته‌اند. خطای اندازه‌گیری و بی‌پاسخی تأثیر منفی بر کیفیت جمع آوری داده‌ها و کیفیت نتایج تولید شده از آن دارند. بنابراین نمی‌توانیم به سادگی این موارد را در مجموعه داده‌ها نادیده بگیریم. عدم توجه به این مشکلات منجر به ارائه برآوردهای اریب می‌شود. درک خوب و تجزیه و تحلیل این خطاها، پیشنهاد یک فرایند ویرایش و جانهی را که داده‌های با کیفیت خوبی را تولید می‌کند، تضمین می‌کند.

در دهه‌های اخیر الگوریتم‌های مختلفی برای جانهی مقادیر گم‌شده با یادگیری ماشین توسعه یافته‌اند که می‌توانند کلاس گسترده‌تری از روش‌های تحلیل انعطاف‌پذیر را ارائه دهد که با منابع مدرن داده متناسب‌تر باشد. از این‌رو سازمان‌های آماری مختلف به کاربردهای مختلف از مباحث یادگیری ماشین روی آورده‌اند که یکی از آن‌ها جانهی با استفاده از الگوریتم‌های یادگیری ماشین است.

بر این اساس پژوهشکده‌ی آمار با توجه به رسالت خود در زمینه‌ی اجرای طرح‌های پژوهشی برای بررسی روش‌های نوین آماری و با هدف افزایش کیفیت آمارها، طرح پژوهشی « کاربرد یادگیری ماشین در فرایند جانهی داده‌های آمارگیری‌های خانواری» را در دستور کار خود قرار داد. در این طرح مطالعاتی ضمن مرور روش‌های جانهی با استفاده از یادگیری ماشین، به تجربه‌ی کشورهای مختلف در استفاده از روش‌های یادگیری ماشین در آمار رسمی پرداخته می‌شود. همچنین کاربردی از آن بر داده‌های آمارگیری هزینه و درآمد خانوار مرکز آمار ایران ارائه می‌شود.

این طرح در گروه پژوهشی طرح‌های فنی و روش‌های آماری پژوهشکده‌ی آمار با همکاری سرکار خانم دکتر لیدا کلهری ندرآبادی به‌عنوان مجری طرح سرکار خانم مریم میرزایی، جناب آقای دکتر عباس مرادی، جناب آقای جناب آقای حسن رنجی، بهنام بهرامی و جناب آقای دکتر اشکان شباک به‌عنوان همکاران اصلی طرح پژوهشی به انجام رسیده است، که بدین وسیله از ایشان صمیمانه تشکر و قدردانی می‌شود.

گروه پژوهشی طرح‌های فنی و روش‌های آماری

پژوهشکده‌ی آمار

فهرست مطالب

مرور روش‌های جانهی ۱
۱-۱- روش‌های جانهی ۲
۱-۱-۱- حذف داده‌های گم‌شده ۳
۱-۱-۲- جانهی قیاسی ۳
۱-۱-۳- جانهی میانگینی ۳
۱-۱-۴- جانهی فراوانی ۴
۱-۱-۵- جانهی میانگینی تصادفی ۵
۱-۱-۶- جانهی بی‌درنگی ۵
۱-۱-۷- جانهی بی‌درنگی تصادفی ۵
۱-۱-۸- جانهی بی‌درنگی دنباله‌ای ۶
۱-۱-۹- جانهی بی‌درنگی با استفاده از k نزدیکترین همسایه ۶
۱-۱-۱۰- جانهی چندمتغیره به کمک معادله‌ی زنجیر (MICE) ۸
۱-۱-۱۱- جانهی بادرنگی ۹
۱-۱-۱۲- جانشانی ۱۰
۱-۱-۱۳- جانهی رگرسیونی ۱۰
۱-۱-۱۴- جانهی رگرسیونی تصادفی ۱۱
۱-۱-۱۵- جانهی چندگانه (MI) ۱۲
۱-۱-۱۶- جانهی کسری ۱۲
۱-۱-۱۷- جانهی مقادیر گم‌شده با استفاده از درون‌یابی و برون‌یابی ۱۴
استفاده از روش‌های یادگیری ماشین برای جانهی در سازمان‌های ملی آمار ۲۳
۲-۱- استفاده از روش‌های هوش مصنوعی و یادگیری ماشین در اداره آمار آلمان ۲۴
۲-۱-۱- جانهی وضعیت مادر بودن و تعداد فرزندان ۲۴
۲-۱-۱- روش‌های یادگیری ماشین برای جانهی وضعیت مادر بودن و تعداد فرزندان ۲۷
۲-۲- نیوزلند ۳۰

۲-۲-۱-آمارگیری اقتصادی خانوار ۳۰
۲-۲-۲- کاوش قوانین انجمنی برای ویرایش ۳۱
۲-۲-۳- مطالعه موردی و نتایج ۳۲
۲-۲-۴- کاربرد روش‌های رده‌بندی در جانهی (درخت رگرسیون) ۳۴
۲-۳- کانادا ۳۷
۲-۳-۱-چارچوب ارزیابی ۳۸
۲-۳-۲- پارامتر سازی در CANCEIS ۴۰
۲-۴-فرایند جانهی داده‌های سرشماری ثبتی‌مبنای اتریش ۴۰
۲-۴-۱- فرایند داده‌ها ۴۱
۲-۴-۲- فرایند جانهی ۴۲
۲-۴-۳-ویرایش قطعیتی ۴۳
۲-۴-۴- همبستگی و خوشه‌بندی ۴۴
۲-۴-۵- روش بی‌درنگ ۴۵
۲-۴-۶- رگرسیون لوژستیک ۴۷
۲-۴-۷- میزان مقادیر گم‌شده متغیرها ۴۸
۲-۴-۸- ارزیابی کیفیت ۴۸
۲-۵- جانهی داده‌های گم‌شده در هلند ۴۹
کاربرد جانهی در آمارگیری هزینه و درآمد خانوار ۵۵
۳-۱- آمارگیری هزینه و درآمد خانوار ۵۶
۳-۱-۱-معرفی متغیرها ۵۶
۳-۲- روش‌های انتخاب ویژگی ۵۹
۳-۲-۱-روش پالایه ۵۹
۳-۲-۲-روش بسته‌بند ۶۰
۳-۲-۳-روش نشانیده ۶۱
۳-۲-۴-روش ماشین بردار پشتیبان ۶۱
۳-۲-۵-روش رگرسیون بردار پشتیبان ۶۴
۳-۲-۶-الگوریتم درخت تصمیم ۶۶
۳-۲-۷-الگوریتم جنگل تصادفی ۶۷
۳-۳-تحلیل داده‌ها ۶۸
۳-۴-بحث و نتیجه‌گیری ۸۴
پیوست‌ها ۸۷
مرجع‌ها ۹۷

واژه‌نامه ۱۰۱

فهرست جدول‌ها

جدول ۲-۱- نتایج بدست آمده برای داده‌های تحت آزمون ۲۸
جدول ۲-۲- توزیع وضعیت تعداد فرزندان قبل و بعد از جانهی..... ۲۹
جدول ۲-۳- خلاصه‌ای از اطمینان ، c، برای درآمد با بالاترین مدرک تحصیلی..... ۳۳
جدول ۲-۴- خلاصه ای از اطمینان ، c ، برای درآمد بر اساس سن ۳۴
جدول ۲-۵- وزن متغیرهای جورسازی برای جانهی درآمد از بخش مشاغل..... ۳۶
جدول ۲-۶- تشخیص اریبی مقادیر جانهی شده با مدل رگرسیون ۳۷
جدول ۲-۷- اندازه‌های همبستگی برای آخرین مدرک تحصیلی ۴۵
جدول ۳-۱- متغیرهای تأثیرگذار در استان‌های تحت مطالعه به تفکیک مناطق شهری و روستایی ۷۲
جدول ۳-۲- میانگین مربعات خطا (زمان) حاصل از اجرای سه مدل رگرسیون بردار پشتیبان، جنگل تصادفی و رگرسیون لگ‌نرمال با ۲۵۰ تکرار برای مناطق شهری. ۷۳
جدول ۳-۳- میانگین مربعات خطا (زمان) حاصل از اجرای سه مدل رگرسیون بردار پشتیبان، جنگل تصادفی و رگرسیون لگ‌نرمال با ۲۵۰ تکرار برای مناطق روستایی. ۷۴
جدول ۳-۴- میانگین مربعات خطا (زمان) حاصل از اجرای جانهی میانگین، میانه و مد با ۲۵۰ تکرار برای مناطق شهری..... ۷۵
جدول ۳-۵- میانگین مربعات خطا (زمان) حاصل از اجرای جانهی میانگین، میانه و مد با ۲۵۰ تکرار برای مناطق روستایی..... ۷۶
جدول ۳-۶- میانگین مربعات خطا (زمان) حاصل از اجرای سه مدل رگرسیون بردار پشتیبان، جنگل تصادفی و رگرسیون لگ‌نرمال با ۵۰۰ تکرار برای مناطق شهری. ۷۷
جدول ۳-۷- میانگین مربعات خطا (زمان) حاصل از اجرای سه مدل رگرسیون بردار پشتیبان، جنگل تصادفی و رگرسیون لگ‌نرمال با ۵۰۰ تکرار برای مناطق روستایی. ۷۸
جدول ۳-۸- میانگین مربعات خطا (زمان) حاصل از اجرای جانهی میانگین، میانه و مد با ۵۰۰ تکرار برای مناطق شهری..... ۷۹

جدول ۳-۹- میانگین مربعات خطا (زمان) حاصل از اجرای جانهی میانگین، میانه و مد با ۵۰۰ تکرار برای مناطق روستایی..... ۸۰
جدول ۳-۱۰- میانگین مربعات خطا (زمان) حاصل از اجرای سه مدل رگرسیون بردار پشتیبان، جنگل تصادفی و رگرسیون لگ‌نرمال با ۱۰۰۰ تکرار برای مناطق شهری. ۸۱
جدول ۳-۱۱-میانگین مربعات خطا (زمان) حاصل از اجرای سه مدل رگرسیون بردار پشتیبان، جنگل تصادفی و رگرسیون لگ‌نرمال با ۱۰۰۰ تکرار برای مناطق روستایی. ۸۲
جدول ۳-۱۲- میانگین مربعات خطا(زمان) حاصل از اجرای جانهی میانگین، میانه و مد با ۱۰۰۰ تکرار برای مناطق شهری..... ۸۳
جدول ۳-۱۳- میانگین مربعات خطا ح(زمان) اصل از اجرای جانهی میانگین، میانه و مد با ۱۰۰۰ تکرار برای مناطق روستایی..... ۸۴

فهرست شکل‌ها

شکل ۱-۱- استفاده از جانهی میانگینی ۴
شکل ۱-۲- استفاده از جانهی فراوانی ۴
شکل ۱-۳- میانگین ۳۰ مشاهده نزدیک..... ۷
شکل ۱-۴- جانهی چندمتغیره به کمک معادله‌ی زنجیر..... ۹
شکل ۱-۵- انواع جداکننده‌ها..... ۱۶
شکل ۱-۶- در جستجوی بهترین جداساز..... ۱۶
شکل ۱-۷- بهترین جداساز (SVM)..... ۱۶
شکل ۱-۸- مثالی از یک درخت تصمیم ساده..... ۱۹
شکل ۲-۱- اختلافات نسبت تعداد فرزندان به سن زنان قبل و بعد از جانهی بر حسب درصد..... ۲۶
شکل ۲-۲- اختلاف سهم زنان بدون فرزند در سنین مختلف قبل و بعد از جانهی بر حسب درصد برای ۶ روش یادگیری ماشین..... ۳۰
شکل ۲-۳-درخت نمونه برای کمک به تعیین متغیرهای مرتبط برای جانهی درآمد از بخش مشاغل ۳۵
شکل ۲-۴-فرایند داده‌ها و ارزیابی کیفیت برای آمارهای ثبتی‌مبنا ۴۱
شکل ۲-۵-جریان کار ساختاریافته فرایند جانهی..... ۴۳
شکل ۳-۱-نمای هندسی ماشین بردار پشتیبان..... ۶۲
شکل ۳-۲- نمای هندسی مرز تصمیم‌گیری بهینه برای روش ماشین بردار پشتیبان با حاشیه سخت ۶۳
شکل ۳-۳- نمای هندسی مرز تصمیم‌گیری بهینه برای روش رگرسیون بردار پشتیبان..... ۶۵
شکل ۳-۴- نمای هندسی مرز تصمیم‌گیری بهینه برای روش رگرسیون بردار پشتیبان در حضور متغیر کمکی.... ۶۶
شکل ۳-۵- توزیع فراوانی درآمد خانوار. در مناطق شهری و روستایی استان تهران ۶۹
شکل ۳-۶- توزیع فراوانی لگاریتم درآمد خانوار در مناطق شهری و روستایی استان تهران..... ۶۹
شکل ۳-۷- شناسایی نقاط دور افتاده برای مناطق شهری و روستایی استان تهران..... ۷۰

۱

مرور روش‌های جانهی

هدف از آمارگیری‌ها تولید شاخص‌هایی از طریق اندازه‌گیری صفت مورد نظر در جامعه است. حال اگر نتوان موارد تعیین‌شده را به‌طور کامل اندازه‌گیری کرد و نرخ بی‌پاسخی بالا باشد، برآوردها تنها بر اساس موارد با پاسخ به دست می‌آیند. در این شرایط امکان دارد کیفیت آماره‌های آمارگیری با تهدید مواجه شوند. به بیان دیگر، بی‌پاسخی می‌تواند کیفیت آماره‌های آمارگیری را تحت تأثیر قرار دهد و آن‌ها را دچار انحراف کند. اگر این انحراف نظام‌مند باشد، یعنی اجراهای یک طرح آمارگیری به سمت خاصی تمایل دارد و در نتیجه، «اریبی بی‌پاسخی» به‌وجود می‌آید. با توجه به آن‌چه گفته شد، آمارشناسان همواره در تلاش هستند تا در حد امکان نرخ بی‌پاسخی را کاهش دهند. استفاده از روش‌هایی چون مراجعه‌ی مجدد، اختصاص مدت زمان طولانی برای گردآوری داده‌ها، بارکار کم‌تر برای پرسش‌گر و پاسخ‌گو با طراحی پرسش‌نامه‌های تا حد امکان کوتاه، نامه‌های پیش‌آگاهی بخش، استفاده از پاسخ‌گویان نماینده، تنظیم رفتار پرسش‌گر با توجه به نگرانی‌های واحد نمونه‌ای، جور کردن پرسش‌گر با مشخصات واحد نمونه‌ای، نامه‌های تشویقی برای امتناع‌های اولیه، جا به جایی پرسش‌گر و تغییر شیوه گردآوری داده‌ها برای پاسخ‌گویان بی‌علاقه و غیره از جمله مواردی است که به بالا بردن نرخ پاسخ در آمارگیری‌ها کمک می‌کند. تقریباً تمام این روش‌ها نیازمند صرف زمان بیش‌تر، تلاش یا ارتباط بیش‌تر با واحدهای نمونه‌ای هستند و به‌طور کلی موجب افزایش هزینه‌های آمارگیری می‌شوند. بنا بر این در بسیاری از موارد، سازمان‌های آماری ملی با به‌کار بردن روش‌های تعدیل بی‌پاسخی درصدد کم کردن اثرهای نامطلوب بی‌پاسخی هستند. «جانهی»، «تعدیل وزن‌دهی» و «مراجعه‌ی مجدد» سه رویکرد کلّی برای رویارویی با بی‌پاسخی هستند

در این پژوهش استفاده از روش‌های یادگیری ماشین در جانهی متغیر درآمد به عنوان یکی از مهم‌ترین متغیرها در آمارگیری هزینه و درآمد خانوار، مورد توجه است. از این‌رو در ادامه این فصل برخی روش‌های جانهی به اختصار مرور می‌شوند. در فصل دوم مطالعه تطبیقی استفاده از روش‌های یادگیری ماشین برای جانهی در سازمان‌های آماری ارائه شده است. در فصل سوم ضمن تشریح روش رگرسیون بردار پشتیبان و جنگل تصادفی، به‌کارگیری