نشریه علمی (فصلنامه) «انرژی ایران»
دوره ۲۳، شماره ۱، بهار ۱۳۹۹، صفحه ۴۵ – ۲۵

پیش‌بینی عوامل مؤثر در مصرف انرژی خانگی به کمک روش‌های داده‌کاوی

[۱] ریحانه السادات حافظی فرد ، [۲] جمال زارع‌پور احمدآبادی* ، [۳] الهام عباسی هرفته

تاریخ دریافت:
۱۳۹۸ / ۱۱ / ۲۸

تاریخ پذیرش:
۱۳۹۹ / ۶ / ۵

کلمات کلیدی:
انرژی مصرفی خانگی، الگوریتم M5Rules، الگوریتم K–نزدیک‌ترین همسایه، الگوریتم جنگل تصادفی ، ارزیابی همبستگی ویژگی‌ها وسایل روشنایی، دما

چکیده

باتوجه به افزایش جمعیت و اینکه منابع انرژی رو به کاهش است، در این تحقیق به مطالعه انرژی مصرفی خانگی پرداخته شده است. هدف از این پژوهش پیش‌بینی عوامل مؤثر بر انرژی مصرفی خانگی می‌باشد. برای این پیش‌بینی از ۳ الگوریتم M5Rules، K– نزدیک‌ترین همسایه و جنگل تصادفی استفاده شده است که در نرم افزار weka موجود می‌باشد. در این پژوهش از الگوریتم ارزیابی همبستگی ویژگی‌ها برای انتخاب بهترین عوامل نیز استفاده شده است. این الگوریتم مهمترین عوامل مؤثر بر انرژی مصرفی و میزان تأثیر آنها را مشخص می‌کند. نتایج حاصل از این بررسی نشان می‌دهد که چراغ‌ها و وسایل روشنایی، درجه حرارت و دما در اتاق نشیمن، درجه حرارت و دما در خارج از ساختمان، درجه حرارت و دما در خارج از ایستگاه هواشناسی چیورس، سرعت وزیدن باد، رطوبت در منطقه آشپزخانه و درجه حرارت و دما در محل لباسشویی بیشترین تأثیر را در مصرف انرژی خانگی دارد. همچنین از بین الگوریتم‌های آزموده شده، جنگل تصادفی بهترین نتیجه را به‌دست می‌دهد.

۱. کارشناس علوم کامپیوتر، دانشگاه یزد hafezifard.reyhaneh@gmail.com
۲. استادیار علوم کامپیوتر، دانشگاه یزد (نویسنده مسئول) zarepoujamal@yazd.ac.ir
۳. استادیار علوم کامپیوتر، دانشگاه یزد e.abbasi@yazd.ac.ir

۱. مقدمه

پیشرفت‌های اخیر در سیستم‌های کامپیوتری و گستردگی استفاده از آنها بخصوص در جمع‌آوری دادگان، منجر به پیدایش منابع داده‌ای بسیار بزرگی شده است که درپی آن، نیاز به اطلاعات نهفته در این داده‌ها مطرح گشته که به وضوح به کمک روش‌های سنتی قابل دستیابی نیست. حجم بالای داده‌های دائماً در حال رشد و نیز تنوع آنها به شکل داده‌های متنی، اعداد، گرافیک‌ها، نقشه‌ها، عکس‌ها، تصاویر ماهواره‌ای و عکس‌های گرفته شده با اشعه ایکس نمایانگر پیچیدگی کار تبدیل داده‌ها به اطلاعات است[۴].

در پاسخ به این نیاز، بحث *داده‌کاوی* با تأکید بر الگوریتم‌هایی جهت استخراج اطلاعات جدید از داده‌ها مطرح گشت. این فرآیند به‌صورت منظم و متوالی و با بهره‌گیری از ابزارهای تجزیه و تحلیل داده‌ها مانند مدل‌های آماری، الگوریتم‌های ریاضی و روش‌های یادگیری ماشین[1] [۲۸]، به دنبال کشف الگوها و روابط معتبری است که تاکنون ناشناخته بوده‌اند. داده‌کاوی بر خلاف روش‌های سنتی تجزیه و تحلیل داده‌ها، کشف محور است و به‌صورت خودکار تجربه‌ای را از طریق ابزارهایی مانند شبکه‌های عصبی[2] یا درخت‌های تصمیم‌گیری[3] کسب کرده و و بدون نیاز به دخالت انسان آن را بهبود می‌بخشد. مقایسه تجزیه و تحلیل سنتی و داده‌کاوی نشان می‌دهد امروزه داده‌کاوی سهم بزرگی در بررسی و گرفتن نتایج جدید از داده‌ها دارد.

داده‌کاوی شامل روش‌هایی است که می‌تواند از داده‌های کمی، متنی، یا چندرسانه‌ای برای استحصال موارد زیر مورد استفاده قرار گیرد:

- قوانین وابستگی: الگوهایی که در آن وجود یک آیتم دلالت بر وجود آیتم دیگر دارد.
- کلاس‌بندی: انتساب الگوها به یک مجموعه کوچک از کلاس‌های از قبل تعریف شده به وسیله کشف بعضی روابط بین ویژگی‌ها.
- خوشه‌ندی: گروه‌بندی مشتریان یا مجموعه الگوهایی که ویژگی‌های مشابهی دارند.

1. Machine Learning Methods
2. Neural Networks
3. Decision Trees

– پیش‌گویی: کشف الگوها برای پیش‌گویی منطقی درباره آینده.
– تحلیل مسیر یا الگوهای ترتیبی: الگوهایی که در آن یک رخداد منجر به وقوع رخداد دیگر می‌شود.

داده‌کاوی یک تکنولوژی جدید نیست ولی کاربرد آن به‌طور معناداری در بخش‌های مختلف خصوصی و عمومی روبه‌رشد بوده و عموما صنایعی چون بانک، بیمه، پزشکی و خرده‌فروشی از داده‌کاوی به هدف کاهش هزینه‌ها، افزایش تحقیقات و افزایش فروش استفاده می‌کنند [۴].

یکی از مسائلی که همواره مورد تأکید برنامه‌ریزان و صاحب‌نظران بوده است میزان مصرف انرژی و توصیه‌هایی جهت کاهش مصرف آن است. چراکه امروزه مصرف انرژی در تمام جنبه‌های زندگی رو به افزایش است در حالی که منابع انرژی محدود هستند. مصرف بی‌رویه انرژی سبب صدمه زدن به محیط زیست، هدر دادن سرمایه ملی و به خطر افتادن زندگی آینده بشر می‌شود. از سوی دیگر، استفاده نادرست از انرژی خسارات جبران‌ناپذیری بر بودجه سالانه کشور تحمیل می‌کند. این واقعیات، لزوم برنامه‌ریزی منطقی و دیدگاه همه جانبه‌نگری به انرژی را به روشنی نمایان می‌سازد.

یکی از موارد مصرف انرژی، که باید مورد توجه قرار گیرد و با توصیه‌های مناسب برای آن فرهنگ‌سازی شود، میزان مصرف انرژی خانگی است. باتوجه به جمعیت حدود ۸۰ میلیونی کشور ایران و گسترش روز افزون خانواده‌ها و استفاده از وسایل برقی خانگی به طور گسترده، لازم است نسبت به بهینه‌سازی مصرف انرژی تجدید نظر جدی صورت گیرد.

هدف این پژوهش پیش‌بینی عوامل مؤثر بر انرژی مصرفی خانگی می‌باشد که برای نیل به آن، از روش‌ها و الگوریتم‌های داده کاوی برای پیش‌بینی عوامل محیطی مؤثر بر انرژی مصرفی خانگی استفاده شده است. روش‌های مورد استفاده در این تحقیق، شامل K ,M5Rules– نزدیک‌ترین همسایه[1] و جنگل تصادفی[2] می‌باشند.

در ادامه این مقاله، در بخش (۲) برخی از پژوهش‌های مرتبط با انرژی بررسی خواهد شد. سپس در بخش ۳ کار انجام شده تشریح خواهد شد و در بخش آخر نیز نتایج گزارش شده و نتیجه‌گیری عنوان خواهد شد.

1. K- Nearest Neighbors
2. Random Forest (RF)

۲. مرور ادبیات

از مهم‌ترین مسائلی که جامعه امروز با آن مواجه است، کاهش منابع انرژی می‌باشد. انرژی مصرفی در ساختمان‌ها یکی از مباحث مهم در این زمینه است که موضوع مطالعات متعددی می‌باشد [۵، ۲۳، ۱۹، ۱۸، ۱۱، ۸]. با توجه به هزینه‌های بالای انرژی مصرفی و طولانی بودن زمان بازگشت این انرژی‌ها به چرخه زندگی، امروزه بهینه‌سازی مصرف انرژی ساختمان‌ها بسیار مورد توجه قرار گرفته است [۲]. به عنوان نمونه، در یک مطالعه در بریتانیا مشخص شد لوازم خانگی مانند تلویزیون و لوازم الکترونیکی مصرفی که در حالت آماده‌به‌کار[1] هستند، منجر به افزایش مصرف انرژی برق به میزان ۱۰/۲٪ می‌شوند [۱۴].

مدل‌های رگرسیون برای استفاده از انرژی می‌توانند به درک روابط بین متغیرهای مختلف و اندازه‌گیری تأثیر آنها کمک کنند. بنابراین، مدل‌های پیش‌بینی مصرف انرژی الکتریکی در ساختمان‌ها می‌توانند برای تعدادی از برنامه‌های کاربردی مفید باشند؛ از جمله برای تعیین اندازه‌گیری مناسب فتوولتائیک[2] و ذخیره انرژی برای کاهش جریان برق به گره [۲۶]، برای تشخیص الگوهای غیرطبیعی مصرف انرژی [۲۴]، به عنوان بخشی از سیستم مدیریت انرژی برای کنترل بار [۲۷، ۱۰، ۷]، برای مدل‌سازی برنامه‌های پیش‌بینی کنترل بارها [۹]، برای مدیریت سمت تقاضا[3] و پاسخ سمت تقاضا[4] [۲۱، ۱۳، ۱۱] و به عنوان یک ورودی برای تجزیه و تحلیل عملکرد ساخت و ساز [۲۳، ۲۲، ۱۷]. در ادامه چند مقاله در حوزه پیش‌بینی انرژی مصرفی ساختمان مورد بررسی قرارگرفته است.

لوئیس و همکاران [۲۰] در سال ۲۰۱۷ مجموعه داده‌های پیش‌بینی انرژی لوازم خانگی را مورد مطالعه قرار دادند که در این مطالعه از چهار مدل آماری زیر استفاده شده است:

۱. رگرسیون خطی چندگانه[5]

۲. ماشین بردار پشتیبان با هسته رادیال[6]

۳. جنگل تصادفی

1. Standby mode
2. Photovoltaics
3. Demand side management
4. Demand side response
5. Multiple linear regression (LM)
6. Support vector machine with radial kernel (SVM radial)

۴. ماشین‌های تقویت گرادیان[1]

نتایج حاصل از این پژوهش به کمک چهار مدل فوق نشان می‌دهد که ماشین‌های تقویت گرادیان بهترین کارایی را داشته است. این مدل منجر به ۹۷٪ واریانس در مجموعه آموزش و ۵۷٪ واریانس در مجموعه آزمایش شده است. این مقادیر بیانگر دقت بالای این مدل نسبت به سایر روش‌ها است.

یکی از تکنیک‌های رایج برای پیش‌بینی میزان مصرف انرژی ساختمان‌ها، مدل‌سازی معکوس است که اغلب به منظور شناسایی تأثیر رفتار ساکنان یک ساختمان بر دقت مدل مورد استفاده قرار می‌گیرد. البته از آن‌جا که مجموعه داده‌های محدودی در این زمینه در اختیار است امکان ارائه راهنمایی و پیشنهادات در این زمینه زیاد نیست. هدف اصلی چنین مطالعاتی، یافتن علل ناهنجاری در داده‌های مصرف انرژی و تعیین اینکه آیا حذف علل ناهنجاری‌ها بر عملکرد مدل تأثیری خواهد داشت یا خیر، می‌باشد.

در این راستا مطالعه‌ای در سال ۲۰۱۸ توسط ستین و همکاران [۱۲] بر روی یک مجموعه داده که از ۱۲۸ ساختمان مجتمع مسکونی در ایالات متحده جمع‌آوری شده بود صورت گرفت. در این پژوهش برای شناسایی عوامل ناهنجاری و کشف تأثیر آنها از سه روش استفاده شد. نتایج این مطالعه نشان داد که داده‌های حدود ۱۹٪ از خانه‌ها خارج از محدوده بودند. با استفاده از داده‌های جمع‌آوری شده، دلایل این ناهنجاری در داده‌ها عمدتاً نورپردازی و الکترونیک بوده است. همچنین در ۲۰٪ از خانه‌هایی که در معرض حذف از مجموعه داده‌ها قرار داشتند با صرف‌نظر کردن از این عوامل ناهنجاری، عملکرد مدل بهبود یافت.

با توجه به این‌که طبق مقررات سخت‌گیرانه ساخت‌وساز در کشور دانمارک، ساختمان‌های تازه ساخته شده کم‌مصرف هستند، پژوهشی توسط فوتیناکی و همکاران [۱۵] در سال ۲۰۱۸ با هدف بررسی میزان انعطاف‌پذیری انرژی در ساختمان‌های کم‌مصرف در دانمارک انجام شد. انعطاف پذیری انرژی باعث می‌شود یک ساختمان بتواند با توجه شرایط آب و هوا و توانایی‌های شبکه انرژی، تقاضای ساکنان خود را بدون کاستن از آسایش حرارتی آنها پاسخ دهد. جهت شناسایی نقش ساختمان‌های کم‌مصرف در سیستم انرژی، ظرفیت ذخیره‌سازی ذاتی جرم‌های ساختمانی مورد بررسی قرار گرفت. در این مطالعه دو نوع ساختمان «خانه تک خانواده» و «بلوک آپارتمان» مورد بررسی قرار گرفت. هدف، کمی سازی میزان

1. Gradient boosting machines (GBM)

انرژی‌ای است که در هر بازه زمانی می‌تواند به ساختمان اضافه یا کم شود بدون آن‌که آسایش حرارتی ساکنان آن به خطر بیفتد. برای این مسئله ویژگی‌های طراحی ساختمان، شرایط مرزی و سناریوهای مختلف از جمله شروع زمان و مدت زمان این پژوهش مورد مطالعه و بحث قرار گرفت. نتایج نشان داد که ساختمان‌های کم‌مصرف بسیار مقاوم هستند و می‌توانند برای چندین ساعت استقلال خود را حفظ کنند. همچنین انعطاف‌پذیری انرژی برای ساختمان‌های متراکم نسبت به ساختمان‌های تکی بیشتر می‌باشد. این تجزیه و تحلیل وابستگی بالقوه انعطاف‌پذیری انرژی به شرایط مرزی (درجه حرارت محیط، تابش خورشیدی، دستاوردهای داخلی) را نشان می‌دهد و بر اهمیت عایق‌کاری تأکید می‌کند.

کنترل پیش‌بین مدل[1] (MPC) یک روش مبتنی بر مدل است که به طور گسترده و با موفقیت در سال‌های گذشته به منظور بهبود عملکرد سیستم‌های کنترل استفاده شده است. شناسایی یک مدل پیش‌بینی‌کننده برای یک ساختمان، یک عامل کلیدی است که برای سیستم‌های پیچیده مانند ساختمان‌هایی که با مشکلات هزینه و محدودیت زمانی مواجه‌اند، مانع پذیرش گسترده این مدل می‌شود. برای غلبه بر این مشکل اسمار و همکارانش [۲۵]، یک ایده جدید برای کنترل پیش‌بینی براساس الگوریتم‌های یادگیری ماشین مانند درخت‌های رگرسیون و جنگل‌های تصادفی ارائه کرده‌اند. آنها این رویکرد را کنترل پیش‌بین داده‌محور[2] (DPC) نامیدند و از آن برای سه مطالعه موردی مختلف استفاده کردند. این سه مطالعه برای نشان دادن عملکرد، مقیاس پذیری و کارایی کنترل پیش‌بین داده‌محور انجام شد.

آنها در اولین مطالعه موردی، معیار کنترل‌کننده MPC را با استفاده از یک مدل ساختمان دوقطبی[3] در نظر گرفتند. سپس DPC را به یک مجموعه داده‌ای که از این مدل دوقطبی شبیه‌سازی شده بود، اعمال کردند و یک کنترل‌کننده براساس داده‌ها ایجاد کردند. نتایج آنها نشان داد که DPC می‌تواند عملکرد قابل مقایسه‌ای را نسبت به MPC که به‌صورت یک مدل ریاضی کاملا شناخته شده است، نشان دهد.

در دومین مطالعه موردی، DPC را به یک مدل که دارای ۲۲ ساختمان ۶ طبقه بود، اعمال کردند که مشکلات اقتصادی و عملیاتی که از پیچیدگی شدید مدل نشات می‌گیرد را نداشت و توانست مسئله

1. Model Predictive Control (MPC)
2. Data-driven model Predictive Control (DPC)
3. Bilinear building model

پاسخگویی تقاضا (پاسخگویی بار) را حل کند. نتایج به‌دست آمده از این مدل نشان داد که مقیاس‌پذیری و کارایی DPC توان مطلوب را با خطای متوسط ۳٪ فراهم می‌کند.

در سومین مطالعه موردی، DPC بر روی داده‌های واقعی که از یک خانه خارج از شبکه در شهر لاکویلای[1] ایتالیا جمع‌آوری شده بود پیاده‌سازی و آزمایش شد. آنها کل مقدار انرژی ذخیره شده را باتوجه به کنترل‌کننده بنگ-بنگ کلاسیک[2] مقایسه کردند. این مقایسه نشان داد که صرفه‌جویی در انرژی را می‌توان تا ۴۹/۲٪ افزایش داد.

در نهایت آنها نتیجه گرفتند در شرایطی که داده‌های واقعی به اندازه کافی یا قابل اطمینان در اختیار نیست و همچنین پیش‌بینی مطمئنی از آب و هوا وجود ندارد مدل DPC روش مؤثر و کارایی می‌باشد.
در ادامه به بیان جزئیات مطالعه انجام شده و مدل‌های استفاده شده در پژوهش حاضر پرداخته شده است.

۳. بررسی روش‌های داده‌کاوی برای پیش‌بینی عوامل مؤثر در مصرف انرژی خانگی به کمک مجموعه داده‌های انرژی

با توجه به مطالب فوق و نظر به اهمیت میزان انرژی که در ساختمان‌ها مصرف می‌شود، این پژوهش با هدف پیش‌بینی انرژی مصرفی خانگی انجام شده است و برای نیل به این هدف به مطالعه عوامل مؤثر بر میزان انرژی مصرفی و همچنین میزان اهمیت این عوامل پرداخته شده است. در این مطالعه از تکنیک‌ها و الگوریتم‌های داده‌کاوی که در نرم‌افزار weka موجود می‌باشد استفاده شده است و سپس نتایج حاصل از الگوریتم‌ها با معیارهای ضریب همبستگی[3]، میانگین خطای مطلق[4]، ریشه میانگین مربع خطا[5]، خطای مطلق نسبی[6] و خطای مربع ریشه[7] مورد ارزیابی قرار گرفته‌اند. داده‌های مورد مطالعه در این پژوهش، «مجموعه داده‌های پیش‌بینی انرژی لوازم خانگی»[8] می‌باشد که در وبسایت UCI[9] قابل مشاهده است [۲۸].

1. L'Aquila
2. Classical bang-bang controller
3. Correlation coefficient (Cc)
4. Mean absolute error (MAE)
5. Root mean squared error (RMSE)
6. Relative absolute error (RAE)
7. Root relative squared error (RRSE)
8. Appliances energy prediction Data set
9. University of California Irvine

الگوریتم‌های مورد استفاده در این پژوهش، شامل K ,M5Rules– نزدیک‌ترین همسایه و جنگل تصادفی می‌باشند. همچنین با توجه به این‌که نمونه‌هایی که در این مجموعه داده وجود دارند، اعدادی بسیار بزرگ با رقم اعشار بالایی می‌باشند، برای سهولت و افزایش دقت از تکنیک‌های نرمال‌سازی استفاده شده است. مجموعه داده مورد استفاده در اینجا دارای ۲۹ ویژگی است که برخی از آنها تأثیر چندانی بر نتیجه ندارند بنابراین از الگوریتم ارزیابی همبستگی ویژگی‌ها برای تعیین میزان اهمیت عوامل استفاده شده است. در ادامه جزئیات این مراحل بیان خواهد شد.

جدول ۱. ویژگی‌های مجموعه داده‌ها

توضیحات	ویژگی	توضیحات	ویژگی
درجه حرارت و دما در اتاق اتو	T7	تاریخ	Date
رطوبت در اتاق اتو	RH_7	انرژی مصرفی لوازم خانگی	Appliances
درجه حرارت و دما در اتاق نوجوان	T8	انرژی مصرفی چراغ‌ها	Lights
رطوبت در اتاق نوجوان	RH_8	درجه حرارت و دما در آشپزخانه	T1
درجه حرارت و دما در اتاق پدر و مادر	T9	رطوبت در آشپزخانه	RH_1
رطوبت در اتاق پدر و مادر	RH_9	درجه حرارت و دما در اتاق نشیمن	T2
درجه حرارت و دما در خارج از ایستگاه	T_out	رطوبت در اتاق نشیمن	RH_2
رطوبت خارج از ایستگاه	RH_out	درجه حرارت و دما در محل لباسشوئی	T3
فشار	Press_mm_hg	رطوبت در محل لباسشوئی	RH_3
سرعت وزیدن باد	Windspeed	درجه حرارت و دما در اتاق اداری	T4
قابلیت مشاهده	Visibility	رطوبت در اتاق اداری	TH_4
نقطه شبنم	Tdewpoint	درجه حرارت و دما در حمام	T5
اولین متغیر تصادفی	rv1	رطوبت در حمام	RH_5
دومین متغیر تصادفی	rv2	درجه حرارت و دما در خارج از ساختمان	T6
***	***	رطوبت در خارج از ساختمان	RH_6

مأخذ: یافته‌های پژوهش

«مجموعه داده‌های پیش‌بینی انرژی لوازم خانگی» در مدت ۴/۵ ماه از یک منزل شخصی در بلژیک جمع‌آوری شده که این جمع‌آوری در هر شبانه روز و هر ۱۰ دقیقه یکبار صورت گرفته است. شرایط دما

و رطوبت خانه با یک شبکه حسگر بی‌سیم به نام زیگ‌بی[1] اندازه‌گیری شده است. هر گره بی‌سیم درجه حرارت و رطوبت را حدود ۳/۳ دقیقه منتقل کرده است. سپس اطلاعات بی‌سیم به مدت ۱۰ دقیقه به طور میانگین محاسبه شده است. داده‌های انرژی هر ۱۰ دقیقه با مترهای انرژی اندازه‌گیری شده‌اند.

آب و هوا از نزدیک‌ترین ایستگاه هواشناسی واقع در فرودگاه (فرودگاه چیورس، بلژیک) دانلود گردیده و به همراه ستون تاریخ و زمان، به مجموعه داده‌ها اضافه شده است. آب و هوا به فاصله هر یک ساعت از ایستگاه هواشناسی دانلود شده است ولی به دلیل این‌که بازه زمانی نمونه‌گیری هر ۱۰ دقیقه است به کمک درونیابی، وضعیت آب و هوا در زمان‌های مورد نیاز محاسبه و به مجموعه داده‌ها افزوده می‌شود. در نهایت، مجموعه داده‌های مورد استفاده دارای ۲۹ ویژگی و ۱۹۷۳۵ نمونه می‌باشد که توصیف این ویژگی‌ها در جدول(۱) آمده است [۲۹].

هدف این مطالعه پیش‌بینی عوامل مؤثر بر انرژی مصرفی خانگی می‌باشد. در حین انجام پژوهش مشخص شد برخی از ویژگی‌هایی که در جدول (۱) معرفی گردید تأثیر چندانی بر نتیجه ندارند و می‌توان از آنها صرف‌نظر نمود. بنابراین به منظور شناسایی مؤثرترین ویژگی‌ها، از الگوریتم انتخاب ویژگی به نام ارزیابی همبستگی ویژگی‌ها استفاده شده است. این الگوریتم با توجه به اهمیت و تأثیر ویژگی‌ها، به هر یک از آنها وزن‌هایی اختصاص می‌دهد. بدین ترتیب امکان حذف ویژگی‌هایی که وزن آنها بسیار کم است و اهمیت چندانی ندارند را فراهم می‌کند.

همان‌طور که در ابتدا گفته شد نرمال‌سازی داده‌ها به کوچکتر شدن داده‌ها و افزایش دقت الگوریتم‌ها کمک می‌کند. مجموعه داده‌های مورد استفاده در این پژوهش اعداد بسیار بزرگ با ارقام اعشار بالایی هستند. بنابراین برای سهولت در کار و افزایش دقت از تکنیک‌های نرمال‌سازی استفاده شده است. در ادامه سه روش نرمال‌سازی مورد استفاده به طور خلاصه بیان شده است:

۳-۱. نرمال‌سازی دسیمال[2]

در این روش ارقام اصلی عدد بدون تغییر می‌مانند و فقط نقطه اعشار آن به گونه‌ای جابه‌جا می‌شود که عدد حاصل درون بازه [۱و۰] قرار گیرد. نحوه انجام این نرمال‌سازی در رابطه (۱) نشان داده شده است

1. ZigBee
2. Decimal

که در آن k به گونه‌ای تعیین می‌شود که تمام مقادیر پس از نرمال‌سازی در بازه [۰،۱] قرار گیرند (به عبارت دیگر، k تعداد ارقام صحیح بزرگترین عدد در بین تمام مقادیر است).

$$V(i) = V(i)/10^{k}$$

۳-۲. نرمال‌سازی مین-مکس[1]

در این روش هدف، قرار گیری داده‌ها در بازه [۱و۱−] می‌باشد، که با استفاده از فرمول (۲) به‌دست می‌آید که در آن min(V) و max(V) به ترتیب کوچکترین و بزرگترین مقدار از بین تمامی مقادیر می‌باشند.

$$V(i)=[V(i)-min(V)]/[(max(V]-min(V)]$$

۳-۳. نرمال سازی انحراف معیار

برای این روش از فرمول (۳) استفاده می‌شود و در این روش نیز هدف آن است که داده‌ها در بازه [۱و۱−] قرارگیرند. منظور از sd(V) انحراف معیار و mean(V) میانگین است.

$$V(i)=[V(i)-mean(V)]/sd(V)$$

در ادامه به توضیح سه الگوریتم M5Rules, K−نزدیک‌ترین همسایه و جنگل تصادفی که در این پژوهش مورد استفاده قرار گرفته است، پرداخته می‌شود.

M5Rules

روش M5Rules یکی از تکنیک‌های یادگیری ماشین است و بر اساس یادگیری درختی بنا نهاده شده است که قوانین خود را از یک سری درخت یادگرفته شده استخراج می‌کند [۱]. این مدل برای حل مسائل مختلف طبقه‌بندی، پیش‌بینی و رگرسیون قابل استفاده می‌باشد.

در M5Rules یک درخت به کمک داده‌های آموزشی ساخته می‌شود. سپس یک قانون از بهترین برگ درخت (گرهی با بیشترین پوشش نمونه‌ها) استخراج شده و درخت حذف می‌شود. در ادامه تمام نمونه‌هایی که توسط این قانون پوشش داده می‌شوند از مجموعه داده‌های آموشی حذف می‌گردند. این روند تا زمانی که تمام نمونه‌ها لااقل توسط یک قانون پوشش داده شوند یا به عبارت دیگر مجموعه داده‌های آموزشی تهی گردد، تکرار می‌شود.

1. Min_max

مشابه یادگیری درخت تصمیم، در روش M5Rules یک درخت در دو مرحله ایجاد می‌شود؛ در مرحله اول که فاز رشد نام دارد، الگوریتم با یک گره برگ شروع می‌شود و به‌صورت تکراری تلاش می‌کند تا هر گره برگ را بر اساس یکی از ویژگی‌ها تقسیم کند و نمونه‌ها با توجه به مقدار ویژگی انتخاب شده، درون یکی از فرزندان این گره قرار گیرند. در پایان مرحله اول، یک درخت کامل ساخته می‌شود. پس از این مرحله، برای هرس (اختیاری)، گره‌های درخت تا زمانی که خطای آنها از حدی بیشتر نباشد با هم ادغام می‌شوند [۱۶].

برخلاف روش درختان تصمیم جزئی[1]، که یک درخت را به‌صورت غیرکامل تولید می‌کنند و سرعت و دقت بالاتری نسبت به درختان کامل دارند، M5Rules یک سری درختان کامل تولید می‌کند که از هر یک از این درختان بهترین گره به‌صورت یک قانون باقی می‌ماند. لازم به ذکر است که استخراج قوانین از بهترین برگ درخت در هر مرحله، خطر هرس بیش از حد را از بین می‌برد.

K- نزدیک‌ترین همسایه

الگوریتم K- نزدیک‌ترین همسایه روشی غیر پارامتری است که برای طبقه‌بندی و رگرسیون استفاده می‌شود[5]. در هر دو حالت، ورودی شامل K- نزدیک‌ترین مثال‌های آموزشی در فضای ویژگی است. خروجی این الگوریتم بسته به این که برای طبقه‌بندی استفاده می‌شود یا رگرسیون، به ترتیب به‌صورت گسسته یا پیوسته محاسبه می‌شود.

الگوریتم K – نزدیک‌ترین همسایه در واقع فاز آموزش و ساخت مدل را ندارد زیرا در زمان آموزش تنها به ذخیره نمونه‌ها می‌پردازد و تمام محاسبات را تا زمان طبقه‌بندی یا رگرسیون به تعویق می‌اندازد. از این جهت، گاهی به آن *یادگیری مبتنی بر نمونه*[2] یا *یادگیری تنبل*[3] گفته می‌شود. بدین ترتیب، فاز آموزش بسیار سریع است و در واقع زمانی لازم ندارد و تمام بار محاسباتی به مرحله آزمایش موکول می‌گردد. علی‌رغم این‌که الگوریتم K – نزدیک‌ترین همسایه عملیات طبقه‌بندی و رگرسیون را به سادگی انجام می‌دهد، اما در بسیاری از کاربردها نتایجِ قابل اطمینانی به‌عنوان پیش‌بینی ارائه می‌دهد [۳۰].

1. Partial decision trees (PART)
2. Instance-based learning
3. lazy

اگر از K– نزدیک‌ترین همسایه به منظور طبقه‌بندی استفاده شود، با توجه به این‌که خروجی باید برچسب یک کلاس باشد، شیء جدید به کلاسی از خروجی تعلق می‌گیرد که در بین K نزدیک‌ترین همسایه این شی، بیشترین تکرار (فراوانی) را داشته باشد که K یک عدد صحیح مثبت و معمولاً کوچک است. بدیهی است که اگر K = 1 باشد، شیء مورد سئوال، به سادگی به کلاس نزدیک‌ترین همسایه خود اختصاص داده می‌شود. در کاربرد رگرسیون از این الگوریتم، میانگین مقادیر K– نزدیک‌ترین همسایگان شی مورد سئوال که مقداری پیوسته است به عنوان خروجی در نظر گرفته می‌شود.

در مرحله آزمایش، این الگوریتم باید فاصله نمونه مورد سئوال با کلیه نمونه‌های اولیه را محاسبه نموده و سپس K تا از نزدیک‌ترین نمونه‌ها را ملاک محاسبات خود قرار دهد. واضح است که برای مجموعه نمونه‌های بزرگ این محاسبات بسیار سنگین خواهد بود که البته با وجود الگوریتم‌های تقریبی برای جستجوی نزدیک‌ترین همسایه‌ها می‌توان سربار این محاسبات را برای مجموعه‌های داده‌های بزرگ، کاهش داد. یکی از رایج‌ترین روش‌های محاسبه فاصله، استفاده از روش اقلیدسی است که طبق فرمول (۴) محاسبه می‌شود:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (۴)$$

طبق آنچه در بالا توصیف شد، تمام K نزدیک‌ترین همسایه نقش یکسانی در تولید خروجی خواهند داشت. اما می‌توان برای برای افزایش دقت، به هر یک از K نزدیک‌ترین همسایه وزنی را تخصیص داد به گونه‌ای که همسایگان نزدیک‌تر نسبت به داده‌های دورتر تأثیر بیشتری داشته باشند و سپس متداول‌ترین کلاس یا میانگین وزنی خروجی را با توجه به این وزن‌ها محاسبه نمود. به عنوان مثال یک روش رایج، تخصیص وزن 1/d به هر همسایه است که d فاصله آن همسایه با نمونه مورد سئوال است [۳۰].

الگوریتم K– نزدیک‌ترین همسایه به‌صورت شبه کد در شکل (۱) بیان شده است.

Algorithm: K-Nearest Neighbors

Input: D , a chunk of the original distance matrix

, dimension of the chunk$n_{chunksize}$

Split, index of split

Chunk, index of chunk

Maxk, an array to hold the farthest neighbors for each row index in the chunk

〖row〗 ^'←blockIdx.x×blockDim.x+threadIdx.x

If $row' < n_{chunksize}$ then

- $row' \leftarrow split \times n_{chunksize} + row'$
- For $column' \leftarrow 1$ to $n_{chunksize}$ do
  - $column \leftarrow chunk \times n_{chunk\ size} + column'$
  - If $row = column$ or $row > n_{row}$ or $column > n_{col}$ then

Continue

- If $[row' \times n_{chunksize} + column'] < Gk'[Maxk[row']].weight$ then
  - $Gk'[Maxk[row']].source \leftarrow row$
  - $Gk'[Maxk[row']].target \leftarrow column$
  - $Gk'[Maxk[row']].weight \leftarrow D[row' \times n_{chunksize} + column']$

Search the new maximum element in $row'(D)$ and store the index in $Maxk[row']$

End

شکل ۱. شبه کد الگوریتم K-نزدیک‌ترین همسایه

جنگل تصادفی

این الگوریتم یکی از روش‌های یادگیری مرکب است که تعداد زیادی درخت تصمیم [۳] را به‌صورت تصادفی و اغلب به روش کیسه‌گذاری[1] تولید می‌کند. نخستین الگوریتم برای جنگل‌های تصمیم تصادفی را «تین کم هو[2]» با بهره‌گیری از روش زیر فضاهای تصادفی پدید آورد [۳۱]. ایده این روش بدین‌صورت است که چندین درخت تصمیم مستقل روی زیرمجموعه‌های مختلفی از داده‌های آموزشی و با استفاده از زیرنمونه‌هایی از کل ویژگی‌های موجود ساخته می‌شوند. واضح است که این کار درختان تصمیم متنوعی تولید می‌کند.

برای ساخت خروجی نهایی جنگل تصادفی، خروجی هر کدام از درختان تصمیم محاسبه شده و با هم تجمیع می‌شوند. در یک مسئله دسته‌بندی هر درخت تصمیم موجود در جنگل تصادفی به نمونه مورد سئوال یک رأی می‌دهد و در نهایت کلاسی که بیشترین رأی را از بین تمام درختان تصمیم داشته باشد به نمونه مورد سئوال نسبت داده می‌شود. در مسئله رگرسیون نیز خروجی نهایی برابر با میانگین خروجی تمام درخت‌های تصمیم است. یعنی [۳۱]:

$$F(x) = \frac{\sum_{i=1}^{K} T_i(x)}{K} \quad (۵)$$

که $T_i(x)$ خروجی درخت تصمیم i ام، x نمونه مورد سئوال و K تعداد درختان جنگل تصادفی هستند.

ترکیب پاسخ صدها درخت تصمیم تصادفی موجب می‌شود که خروجی جنگل تصادفی واریانس کمتری نسبت به درخت تصمیم تکی داشته باشد. در عمل نیز این الگوریتم نیاز به تنظیم تعداد کمی ابرپارامتر دارد برخی از ابرپارامترهایی که در این روش باید تنظیم شوند عبارتند از: تعداد زیرنمونه‌های ویژگی‌ها، تعداد درختان تصمیم و عمق هر درخت تصمیم است تا از بیش‌برازش جلوگیری کند.

درنتیجه جنگل تصادفی قادر است برای مسائل دچار بیش‌برازش یا برای داده‌های نویزی، دقت قابل قبولی را ارائه دهد. مشابه درخت تصمیم تکی، نیازی به مقیاس[3] کردن داده‌ها ندارد اما مقاومت خوبی نسبت به نحوه انتخاب مجموعه آموزشی و وجود نویز در داده‌های آموزشی نشان می‌دهد. متاسفانه جنگل

1. Bagging
2. Tin Kam Ho
3. Scale

تصادفی نسبت به درخت تصمیم، شفافیت و قابلیت توضیح کمتری دارد. همچنین ممکن است برای مجموعه داده‌های بسیار بزرگ، آموزش تعداد زیادی درخت تصمیم باعث کندی الگوریتم جنگل تصادفی گردد. که برای بهبود کارایی پیشنهاد می‌شود درختان تصمیم به‌صورت موازی تولید و مدیریت شوند. شکل (۲) یک شبه کد از مراحل الگوریتم جنگل تصادفی را نشان می‌دهد.

Algorithm : Random Forest

Input: Sequence of N examples $< (x1, y1), \dots, (xN, yN) >$ with lables $y_i \in Y = \{1, \dots, L\}$

Distribution D over the N examples

Integer K specifying number of iterations

number of variables $=P$

Do for $t = 1,2, \dots, K$ (number of trees)

- Choose bootstrap sample D_i from D to construct tree T_i
- Select m input variables at random from P
  - $(m \ll P)$
  - To determine the decision tree at a node of the tree
- Calculate the best split based on these m variables in the training set

End

شکل ۲. شبه کد الگوریتم جنگل تصادفی

جهت ارزیابی وتعیین میزان دقت الگوریتم‌ها و روش‌های داده‌کاوی از معیارهای مختلفی استفاده می‌شود. در ادامه این بخش، برخی از این معیارها به‌صورت خلاصه بیان می‌شود.

برای سنجش شدت رابطه بین متغیر وابسته و مستقل می‌توان از *ضریب همبستگی* استفاده کرد. هر چه ضریب همبستگی به ۱ یا ۱- نزدیک‌تر باشد، شدت رابطه خطی بین متغیرهای مستقل و وابسته شدیدتر است. البته اگر ضریب همبستگی نزدیک به ۱ باشد جهت تغییرات هر دو متغیر یکسان است که به آن رابطه مستقیم گویند. در صورتی که ضریب همبستگی به ۱- نزدیک باشد، جهت تغییرات متغیرها معکوس یکدیگر خواهد بود و به آن رابطه عکس گفته می‌شود. ولی در هر دو حالت امکان پیش‌بینی

مقدار متغیر وابسته برحسب متغیر مستقل وجود دارد. هرچند ضریب همبستگی راهی برای نشان دادن رابطه بین دو متغیر مستقل و وابسته است ولی مدل، رابطه بین این دو متغیر را نشان نمی‌دهد. با رگرسیون می‌توان قانونی که بین داده‌ها وجود دارد را کشف کرده و به کار بست. فرمول سایر معیارهای ارزیابی نیز به‌صورت زیر می‌باشد:

$$\text{MAE} = \frac{\sum_{i=1}^{n}|y(i)-y'(i)|}{n} \quad (۶)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(|y(i)-y'(i)|)^2}{n}} \quad (۷)$$

$$\text{RAE} = \frac{\sum_{i=1}^{n}|y(i)-y'(i)|}{y(i)} \quad (۸)$$

$$\text{RRSE} = \sqrt{\frac{\sum_{i=1}^{n}(|y'(i)-y(i)|)^2}{\sum_{i=1}^{n}(y(i)-(mean(y)))}} \quad (۹)$$

که y(i) مقدار اندازه‌گیری واقعی (مصرف انرژی) و y'(i) پیش‌بینی ارزش و n تعداد اندازه‌گیری است. همچنین mean (y) مقدار میانگین می‌باشد. محاسبه این مقادیر ارزیابی، توسط نرم‌افزار weka صورت می‌گیرد.

٤. نتایج عملی

در این بخش خطای حاصل از اجرای سه الگوریتم M5Rules, K– نزدیک‌ترین همسایه و جنگل تصادفی برروی مجموعه داده‌های پیش‌بینی انرژی لوازم خانگی، مورد مقایسه و بررسی قرار گرفته است که نتایج آنها در جدول (۲) نمایش داده شده است.

جدول ۲. خطای حاصل از اجرای الگوریتم‌ها بر روی مجموعه داده‌ها

الگوریتم/ معیارهای ارزیابی	M5Rules	K– نزدیک‌ترین همسایه	جنگل تصادفی
RMSE	۹۱/۸۹۰۶	۱۰۲/۴۰۱۴	۷۰/۸۵۵
MAE	۵۱/۰۱۹۳	۴۵/۳۵۴۷	۳۳/۴۴۹۴
RRSE	۸۹/۱۱۷۵	۹۹/۳۱۱۲	۶۸/۶۶۸۷
RAE	۸۴/۲۳۴۱	۷۴/۸۸۱۷	۵۵/۲۲۵۸
Cc	۰/۴۵۷۲	۰/۴۹۷۹	۰/۷۳۰۸

مأخذ: یافته‌های پژوهش

با توجه به داده‌های جدول(۲) مشاهده می‌شود که درصد خطاها بسیار بالا است و نمی‌توان آنها را با معیارهای ارزیابی معرفی شده مقایسه کرد. از این رو ابتدا داده‌ها را نرمال و سپس سه الگوریتم ذکر شده بر روی داده‌های نرمال شده نیز اجرا می‌شود. باتوجه به اینکه RMSE معیاری حیاتی و تعیین‌کننده می‌باشد، تنها از این معیار برای مقایسه سه روش نرمال‌سازی استفاده می‌گردد. نتایج در جدول (۳) به نمایش درآمده است.

جدول۳. خطای RMSE حاصل از اجرای الگوریتم‌ها بر روی داده‌های نرمال شده

نرمالسازی/ الگوریتم	Decimal	Min_max	انحراف معیار
M5Rules	۰/۰۰۸۹	۰/۰۸۷	۰/۸۸۵۲
K– نزدیک‌ترین همسایه	۰/۰۱۰۶	۰/۰۹۶۴	۰/۹۹۸۸
RandomForest	۰/۰۰۷۱	۰/۰۶۶۱	۰/۶۸۹۷

مأخذ: یافته‌های پژوهش

همان‌طور که در جدول(۳) مشاهده می‌شود کمترین خطا مربوط به داده‌های نرمال‌سازی دسیمال با استفاده از الگوریتم جنگل تصادفی می‌باشد.

مجموعه داده مورد استفاده دارای ۲۹ ویژگی می‌باشد که ممکن است برخی از آنها تأثیر ناچیزی بر روی میزان دقت خروجی و خطای مدل داشته باشد. همان‌طور که قبلا گفته شد می‌توان از الگوریتم ارزیابی همبستگی ویژگی‌ها برای انتخاب بهترین و مؤثرترین عوامل استفاده نمود. این الگوریتم با توجه به اهمیت و تأثیر ویژگی‌ها به آنها وزن‌هایی اختصاص می‌دهد. به این ترتیب ویژگی‌هایی که وزن آنها بسیار کم است و اهمیت چندانی ندارند حذف می‌شوند. در این پژوهش حد آستانه برای حذف ویژگی‌های کم اهمیت، ۰.۰۸ در نظر گرفته می‌شود یعنی ویژگی‌هایی که وزن‌شان کمتر از ۰.۰۸ باشد حذف می‌شوند. در جدول (۴) مهمترین این ویژگی‌ها به همراه وزن اختصاص یافته به آنها نشان داده شده است.

جدول ۴. مهمترین ویژگی‌ها

Ranked Attributes	Weight
1. lights	۰/۱۹۷۲۷۸
4. T2	۰/۱۲۰۰۷۳
12. T6	۰/۱۱۷۶۳۸
20. T_out	۰/۰۹۹۱۵۵
23. Windspeed	۰/۰۸۷۱۲۲
3.RH_1	۰/۰۸۶۰۳۱
6. T3	۰/۰۸۵۰۶

مأخذ: یافته‌های پژوهش

جدول (۴) بیانگر این است که ویژگی چراغ‌ها و وسایل روشنایی با وزن ۰/۱۹۷۲۷۸ بیشترین تأثیر و درجه حرارت و دما در محل لباسشوئی با وزن ۰/۰۸۵۰۶ کمترین تأثیر را در پیش بینی انرژی مصرفی خانگی داشته است.

در ادامه با در نظر گرفتن ویژگی‌های ذکر شده در جدول (۴) و اجرای سه الگوریتم M5Rules, K نزدیک‌ترین همسایه و جنگل تصادفی برروی مجموعه داده‌های نرمال شده پیش‌بینی انرژی لوازم خانگی، خطای الگوریتم‌ها با استفاده از معیار RMSE در جدول (۵) بیان شده است.

جدول ۵. خطای RMSE حاصل از انتخاب ۷ ویژگی و روش‌های نرمال‌سازی مختلف

نرمالسازی/ الگوریتم	main_data	Decimal	min_max	انحراف معیار
M5Rules	۹۷/۲۷۸۸	۰/۰۰۹۷	۰/۰۹۱۴	۰/۹۴۶۳
K– نزدیک‌ترین همسایه	۹۱/۰۸۲۱	۰/۰۰۹۱	۰/۰۸۶	۰/۸۸۸۴
جنگل تصادفی	۷۷/۲۷۱۵	۰/۰۰۷۶	۰/۰۷۲۴	۰/۷۵۶۱

مأخذ: یافته‌های پژوهش

همان‌طور که در جدول(۵) مشاهده می‌شود، ابتدا ۷ ویژگی مهم انتخاب و داده‌ها نرمال‌سازی می‌شوند. سپس سه الگوریتم M5Rules, K – نزدیک‌ترین همسایه و جنگل تصادفی بر روی آنها اعمال و نتایج به‌دست آمده مورد بررسی قرار می‌گیرد. مشاهده می‌شود الگوریتم M5Rules به دلیل خطاهای

زیاد برای این بررسی مناسب نمی‌باشد. الگوریتم K– نزدیک‌ترین همسایه عملکرد بهتری نسبت به الگوریتم M5Rules دارد و درصد خطاهای حاصل از آن کاهش یافته است. الگوریتم جنگل تصادفی بهترین الگوریتم برای این بررسی می‌باشد و درصد خطاهای آن در مقایسه با دو الگوریتم K ,M5Rules – نزدیک‌ترین همسایه کمتر بوده و عملکرد دقیق‌تری در این پیش‌بینی دارد.

۵. نتیجه‌گیری

با توجه به محدودیت منابع انرژی و اهمیت حفظ آن‌ها، در این پژوهش به مطالعه عوامل مؤثر بر میزان انرژی مصرفی خانگی پرداخته شد. مجموعه داده مورد استفاده دارای ۲۹ ویژگی و ۱۹۷۳۵ نمونه می‌باشد که در مدت ۴/۵ ماه به فاصله هر ۱۰ دقیقه یکبار از یک منزل شخصی جمع‌آوری شده است. این پژوهش به دنبال یافتن یک الگوریتم داده‌کاوی دقیق، یک روش نرمال‌سازی مناسب و همچنین تشخیص مهمترین ویژگی‌ها برای این مجموعه داده است.

باتوجه به نتایج ارزیابی شده و توضیحاتی که ذکر شد، می‌توان نتیجه گرفت که اگر داده‌های اصلی به روش دسیمال نرمال شوند، با اجرای الگوریتم جنگل تصادفی و در نظر گرفتن ۷ ویژگی مهمتر، میزان خطا به ۰/۰۰۷۶ کاهش می‌یابد و مشخص می‌شود ویژگی چراغ‌ها و وسایل روشنایی تأثیر بسیار زیادی بر روی انرژی مصرفی خانگی دارد.

به عنوان توصیه برای صرفه‌جویی در مصرف انرژی خانگی، مهمترین ویژگی‌هایی که می‌توان به ناظران پیشنهاد داد عبارتند از: چراغ‌ها و وسایل روشنایی، درجه حرارت و دما در اتاق نشیمن، درجه حرارت و دما در خارج از ساختمان، درجه حرارت و دما در خارج از ایستگاه هواشناسی (چیورس)، سرعت وزیدن باد، رطوبت در منطقه آشپزخانه و درجه حرارت و دما در محل لباسشوئی. چنان که انتظار می‌رفت، الگوریتم جنگل تصادفی نسبت به سایر الگوریتم‌های مورد بررسی دقیق‌تر است.

منابع

[۱] جماعت، داده‌کاوی با نرم افزار weka.

[۲] زمردیان، زهراسادات و محمد تحصیلدوست (۱۳۹۴). "اعتبار سنجی نرم افزارهای شبیه‌سازی انرژی در ساختمان: با رویکرد تجربی و مقایسه‌ای". نشریه انرژی ایران، دوره هجدهم، شماره ۴ (پیاپی ۵۶).

[۳] کانتاردزیک، مهمد (۱۳۹۲). *داده‌کاوی (Data Mining)*. امیر علی‌خان‌زاده (مترجم). نشر علوم رایانه، بابل.

[۴] هژبر، ابراهیم (۱۳۹۳). "داده‌کاوی، مفاهیم و کاربرد". پروژه مهندسی نرم افزار کامپیوتر.

[5] Altman N.S. (1992). "An Introduction to Kernel and Nearest-neighbor Nonparametric Regression". The American Statistician. No. 46(3), pp.175-185.
[6] Arghira N., Hawarah L., Ploix S. and M. Jacomino (2012). "Prediction of Appliances Energy use in Smart Homes", Energy, No. 48 (1), pp. 128–134.
[7] Barbato A., Capone A., Rodolfi M. and D. Tagliaferri (2011). "Forecasting the Usage of Household Appliances Through Power Meter Sensors for Demand Management in the Smart Grid". IEEE International Conference on Smart Grid Communications (SmartGridComm), IEEE, pp. 404–409.
[8] Basu K., Hawarah L., Arghira N., Joumaa H. and S. Ploix (2013). "A Prediction System for Home Appliance Usage", Energy Build, No. 67, pp. 668–679.
[9] Candanedo A., Dehkordi V.R. and M. Stylianou (2013). "Model-based Predictive Control of an Ice Storage Device in a Building Cooling System". Appl. Energy, No. 111, pp. 1032–1045.
[10] Castillo-Cagigal M., Caama˜no-Martín E., Matallanas E., Masa-Bote D., Gutiérrez A., Monasterio-Huelin F. and J. Jiménez-Leube (2011). "PV self-consumption optimization with storage and active DSM for the residential Sector", Solar Energy No. 85 (9), pp. 2338–2348.
[11] Cetin K.S. (2016). "Characterizing Large Residential Appliance Peak Load Reduction Potential Utilizing a Probabilistic Approach", Sci. Technol. Built Environ, 22(6), pp. 720–732.
[12] Ceti K.S., Do H. (2018). "Evaluation of the Causes and Impact of Outliers on Residential Building Energy Use Prediction Using Inverse Modeling", No. 138, pp. 194-206.
[13] D'hulst R., Labeeuw W., Beusen B., Claessens S., Deconinck G., and K. Vanthournout K. (2015). "Demand Response Flexibility and Flexibility Potential Ofresidential Smart Appliances: Experiences from Large Pilot Test in Belgium", Appl. Energy, No. 155. pp. 79–90.
[14] Firth S., Lomas K., Wright A. and R. Wall (2008). "Identifying Trends in the Use of Domestic Appliances from Household Electricity Consumption Measurements", Energy Build. 40(5), pp. 926–936.
[15] Foteinaki K., Li R., Heller A. and C. Rode (2018). "Heating System Energy ßexibility of low-energy Residential Buildings", Energy and Buildings, No. 180, pp. 95-108.
[16] Jekabsons G. (2016). "M5'regression tree, Model Tree and Tree Ensemble Toolbox for Matlab". Octave ver, 1(0).
[17] Johnson G. and I. Beausoleil-Morrison (2017). "Electrical-end-use data from 23 houses Sampled Each Minute for Simulating Micro-generation Systems. Applied Thermal Engineering", No.114, pp.1449-1456.
[18] Jones R.V. and K.J. Lomas (2016). "Determinants of High Electrical Energy Demand in UK Homes: Appliance Ownership and Use", Energy Build, No. 117, pp. 71–82.
[19] Kavousian A., Rajagopal R. and M. Fischer (2015). "Ranking Appliance Energy Efficiency in Households: Utilizing Smart Meter Data and Energy

Efficiency Frontiers to Estimate and Identify the Determinants of Appliance Energy Efficiency in Residential Buildings", Energy Build, No. 99, pp. 220–230.
[20] Luis M., Candanedo, Véronique Feldheim, Dominique Deramaix. (2017). "Data Driven Prediction Models of Energy Use of Appliances in alow-energy House", No. 140, pp. 81-97.
[21] Mitchell S., Sarhadian R., Guow S., Coburn B., Lutton J., Chisti I., Rauss D. and C. Haiad (2014). "Residential Appliance Demand Response Testing. 2014 ACEEE Summer Study on Energy Efficient Buildings, Pacific Grove", CA, August 17, 22, p. 2014.
[22] Muratori M., Roberts M.C., Sioshansi R., Marano V. and G. Rizzoni (2013). "A Highlyresolved Modeling Technique to Simulate Residential Power Demand", Appl.Energy, No. 107, pp. 465–473.
[23] Ruellan M., Park H. and R. Bennacer (2016). "Residential Building Energy Demand and Thermal Comfort: Thermal Dynamics of Electrical Appliances and Their Impact", Energy Build, pp. 46–54.
[24] Seem J.E. (2007). "Using Intelligent Data Analysis to Detect Abnormal Energy Consumption in Buildings", Energy Build, No. 39 (1), pp. 52–58.
[25] Smarra F., Di Girolamo G.D., De Iuliis, V., Jain A., Mangharam R. and A. D'Innocenzo (2020). "Data-driven switching modeling for MPC using Regression Trees and Random Forests". Nonlinear Analysis: Hybrid Systems, 36, 100882.
[26] Spertino F., Leo P.D. and V. Cocina (2014). "Which are the Constraints to the Photovoltaic Grid-parity in the Main European Markets?", Solar Energy, No. 105, pp. 390–400.
[27] Zhao P., Suryanarayanan S. and M.G. Simoes (2013). "An Energy Management System for Building Structures Using a Multi-agent Decision-Making Control Methodology", IEEE Trans. Ind. Appl. No. 49 (1), pp. 322–330.
[28] https://archive.ics.uci.edu/ml/machine-learning-databases/00374 /.
[29] https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction.
[30] http://sciencewise.info/resource/Ibk_algorithm/Ibk_algorithm_by_Wikipedia.