### **پاکسازی دیتاست ناسا**

آقای پتریک و همکاران (Petrić, Bowes, Hall, Christianson, & Baddoo, 2016) برای پاک‌سازی دیتاست ناسا، نمونه‌هایی دارای ویژگی‌های غیرمحتمل (برای مثال نمونه‌ای که number of operator آن -1 است) را حذف می-کنند. می‌توان مقادیر این‌چنین ویژگی‌هایی را با استفاده از ویژگی‌های دیگر به دست آورد. برای مثال با استفاده از N=N\_1+N\_2 و داشتن مقدار N و N2 مقدار N1 که یک مقدار غیرممکن است را به دست آورد.

در زمینه پاکسازی دیتاست ناسا آقای شپرد و همکاران این دیتاست را بررسی کرده و درنهایت دو نسخه D’ و D” از آن را برای استفاده در مطالعات در این زمینه منتشر کرده¬اند(Shepperd, Song, Sun, & Mair, 2013). در دو نسخه ذکرشده، ویژگی‌های دارای مقدار ثابت (برای مثال تمامی نمونه‌های پروژه p دارای ویژگی x برابر 10 هستند.) و ویژگی‌هایی که مقدار دقیقاً مشابه دارند (برای مثال ویژگی x1 و x2 از پروژه p دقیقاً مشابه می‌باشند.) حذف می‌شوند. حذف چنین ویژگی‌هایی در محیط cpdp باعث پایین آمدن دقت می‌شود زیرا پروژه‌های مختلف که ویژگی با مقدار ثابت یا مشابه دارند، آن ویژگی تنها در پروژه موردنظر ثابت یا مشابه با ویژگیِ دیگری است و در پروژه¬های دیگر این¬چنین نخواهد بود. برای مثال تمامی نمونه‌های پروژه p دارای ویژگی x برابر 10 هستند اما این ویژگی در پروژه‌های دیگر مقادیر دیگری دارد و ازآنجایی‌که در cpdp دیتاست داده آموزشی بقیه پروژه‌ها هستند حذف چنین ویژگی‌ای باعث از دست رفتن بخشی از اطلاعات مفید خواهد شد.

در بررسی‌ای که آقای جو و همکاران انجام داده‌اند(Zhou et al., 2018) 35 تحقیق از 71 تحقیق انجام‌شده درزمینه‌ی cpdp، از دیتاست ناسا استفاده کرده‌اند(حدود 48%). این دیتاست نسخه‌های مختلفی دارد که تعداد رکوردها و ویژگی‌های آن‌ها متفاوت‌اند. ازآنجایی‌که کیفیت داده‌ها یک معیار بسیار مهم درزمینه‌ی یادگیری ماشین می‌باشد مقایسه دیتاست های موجود برای یافتن مناسب‌ترین دیتاست برای پیش‌بینی خطای بین پروژه‌ای امری ضروری است.

یکی از مشکلات دیتاست ناسا، تعداد زیاد مقادیر ازدست‌رفته می‌باشد. حذف رکورد و جانهی معمولاً برای این­گونه مقادیر استفاده می‌شود.(آقای شپرد و همکاران برای پاکسازی دیتاست ناسا از روش حذف رکورد استفاده کرده­اند(Shepperd, Song, Sun, & Mair, 2013)) حذف رکوردها باعث از دست رفتن بخشی از اطلاعات می‌شود که ممکن است اطلاعت مفیدی باشند. استفاده از جانهی نیز با توجه به اینکه داده‌ها با استفاده از یک مدل پیش‌بینی می‌شوند واقعی نبوده و احتمال اینکه اشتباه باشند وجود دارد. برای چنین مقادیری می توان از یک متغیر نماینده استفاده کرد. بدین صورت که برای صفات خاصه ای که مقادیر از دست رفته دارند یک صفت خاصه جدید ایجاد می شود.برای نمونه هایی که مقادیرشان موجود است 0 و برای نمونه هایی که مقادیرشان از دست رفته 1 در صفت خاصه جدید قرار می گیرد. به جای مقادیر از دست رفته صفت خاصه اصلی نیز از جانهی یا مقدار 0 استفاده می شود. این روش بدین صورت عمل می کند: برای مثال فرض کنید ویژگی k برای نیمی از نمونه¬ها ازدست‌رفته باشد، دو مدل یکی برای نمونه‌هایی که ویژگی k در آن‌ها مقدار دارد و دیگری برای نمونه‌هایی که مقدار ازدست‌رفته است ساخته می‌شود. هنگام پیش‌بینی، کافی است مدل مناسب انتخاب شود.