

## بررسی تاثیر پیش پردازش بر روی کیفیت داده و مدل های پیش بینی عود مجدد سرطان پستان

زینب سجادینیا<sup>۱</sup>، سید رئوف خیامی<sup>۲</sup>، سید محمد رضا موسوی<sup>۳</sup>، مهدیه دینانی<sup>۴</sup>

۱- کارشناسی ارشد مهندسی کامپیوتر- نرم افزار، دانشگاه صنعتی شیراز، شیراز

۲- استادیار، دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی شیراز، شیراز

۳- استادیار، دانشکده مهندسی برق و کامپیوتر- بخش مهندسی و علوم کامپیوتر و فناوری اطلاعات دانشگاه شیراز، شیراز

۴- متخصص رادیوتراپی آنکولوژی، مدیر کمیته آموزش و پژوهش مرکز تخصصی رادیوتراپی آنکولوژی (رضاع)، مشهد

### چکیده

امروزه به دلیل شیوع انواع سرطان‌ها، محققان زیادی علاقه‌مند به کشف دانش مفید موجود در داده‌های سرطانی شده‌اند. داده کاوی یکی از ابزارهای قدرتمندی است که بدین منظور استفاده می‌شود اما نتایج آن زمانی قابل اعتماد و مفید است که داده‌های با کیفیتی در اختیار باشد. با توجه به اینکه اکثر داده‌های دنیای واقعی دارای نواقصی همچون ناسازگاری، خطا، نویز، مقادیر گمشده و ... می‌باشند، پیش‌پردازش داده‌ها امری جداناپذیر از داده کاوی است. در این مقاله بر روی یک مجموعه داده واقعی سرطان پستان کار شده که هدف بررسی تاثیر پیش‌پردازش بر روی کیفیت داده و نتایج مدل‌های پیش‌بینی عود سرطان پستان می‌باشد. در ابتدا با استفاده از سه الگوریتم طبقه بندی نزدیکترین همسایه، بیز ساده و بهینه سازی حداقل تربیتی، مدل‌های پیش‌بینی ساخته شده‌اند. سپس پیش‌پردازش‌هایی از قبیل رفع خطا و ناسازگاری، پرکردن مقادیر بدون‌محتوی و کاهش داده (انتخاب ویژگی) بر روی مجموعه داده اعمال و مجدداً مدل‌های پیش‌بینی ساخته شده‌اند. در پایان نتایج مدل‌های پیش‌بینی با استفاده از دو معیار صحت و حساسیت با یکدیگر مقایسه گردیده‌اند. نتایج ارزیابی نشان می‌دهد که مدل‌های پیش‌بینی بهینه سازی حداقل تربیتی از نظر صحت و حساسیت به ترتیب ۳,۵۶ و ۷۳,۵۹، بیز ساده ۴,۸۶ و ۶۶,۱۶ و نزدیکترین همسایه ۵,۴ و ۶۰,۹ درصد نسبت به قبل از پیش‌پردازش داده‌ها رشد داشته است. بنابراین پیش‌پردازش موجب بهبود نتایج طبقه بندی و کیفیت داده می‌شود.

**واژگان کلیدی:** پیش‌پردازش داده‌ها، سرطان سینه، داده کاوی، کیفیت داده، طبقه بندی