به نام خدا

فاطمه صاذقی

پریا یوسفی نوالدین وند

پروژه 1

دکتر حسین حق بین

**مقدمه**

در عصر حاضر که دیتا به عنوان جواهری بی‌قیاس در دسترس ماست، داده کاوی به عنوان یکی از ابزارهای اصلی برای استفاده از این دارایی غنی به شمار می‌آید. این هنر، امکان تحلیل داده‌ها به گونه‌ای است که الگوها، روابط، و اطلاعات ارزشمند مختلف به سطح آشکاری برسند. در این زمینه، درس داده کاوی به عنوان یک رشته کلیدی در علوم داده به شمار می‌رود.

این گزارش به عنوان یک نمونه کاربرد عملی از درس داده کاوی، به تجزیه و تحلیل دقیق داده‌های آتش‌سوزی جنگل پرداخته است. به وسیله تکنیک‌های پیشرفته داده کاوی، سعی شده است تا پیش‌بینی‌ها و الگوهای نهفته در این حوادث طبیعی را شناسایی کنیم. این تحلیل به ما کمک خواهد کرد تا به درک عمیقی از علل و پیامدهای آتش‌سوزی جنگل برسیم.

هدف اصلی این گزارش، خلاقیت و تفکر تحلیلی را ترکیب کرده و نه تنها به شناخت دقیق الگوها بپردازد، بلکه با ارائه داده‌های آماری قابل اعتماد، ارتباطات بین عوامل مختلف در آتش‌سوزی جنگل را بیان کند. این اطلاعات به عنوان پایه‌ای برای ایجاد استراتژی‌های بهبود یافته در زمینه‌های پیشگیری و مدیریت آتش‌سوزی خدمت می‌کند.

در ادامه این گزارش، با استفاده از داده‌های به دست آمده و تحلیل‌های انجام شده، به گسترده‌ترین نگرش به موضوع اتش‌سوزی جنگل پرداخته می شود.

**توضیح دیتاست**

در ابتدا به بررسی ترجمه از سایت معروف UCI که در این زمینه معروف هست پرداخته می شود.

**ترجمه متن قرار گرفته در سایت UCI:**

در [Cortez و Morais، 2007]، خروجی 'area' ابتدا با یک تابع ln(x+1) تبدیل شد. سپس، چندین روش داده کاوی به کار گرفته شد. پس از انطباق مدل‌ها، خروجی‌ها با معکوس تبدیل ln(x+1) پردازش شدند. چهار تنظیم ورودی مختلف مورد استفاده قرار گرفت. آزمایشات با استفاده از یک 10-fold (تنها 10 بخش) x 30 اجرا (30 تکرار) انجام شد. دو معیار رگرسیون اندازه‌گیری شدند: MAD و RMSE. یک ماشین بردار پشتیبانی گاوسی (SVM) با استفاده از تنها 4 شرایط آب و هوایی مستقیم (دما، RH، باد و باران) بهترین مقدار MAD را کسب کرد: 12.71 ± 0.01 (میانگین و بازه اطمینان در 95% با استفاده از توزیع دانشجو). بهترین RMSE توسط پیش‌بینی‌کننده میانگین نادان به دست آمد. تحلیلی روی منحنی خطای رگرسیون (REC) نشان می‌دهد که مدل SVM نمونه‌های بیشتری را با یک خطا کمتر پیش‌بینی می‌کند. به عبارت دیگر، مدل SVM بهتر پیش‌بینی می‌کند که آتش‌های کوچک، که اکثریت را تشکیل می‌دهند، را پیش‌بینی کند.

**توضیحات کلی**

دیتا ست که در سال 2007 منتشر شده است با گرفتن اطلاعات هواشناسی و آنالیز آن ها قصد بر پیش بینی آتش سوزی جنگل داشته است که با استفاده از روش های داده کاوی به جواب خوبی هم رسیده است(MAD: 12.71 ± 0.01).

**این دیتا ست شامل 13 ویژگی (12 ورودی + 1 خروجی) می باشد:**

1. X: مختصات فضایی محور x در نقضه پارک مونتزینیو: 1 – 9
2. Y: مختصات فضایی محور y در نقشه پارک مونتزینیو: 2 – 9
3. Month: ماه سال میلادی ژانویه – دسامبر
4. Day: روز هفته 1-7
5. FFMC: یکی از شاخص های هواشناسی 18.7 – 96.20
6. DMC: یکی از شاخص های هواشناسی 1.1 – 291.3
7. DC: یکی از شاخص های هواشناسی 7.9 – 860.6
8. ISI: یکی از شاخص های هواشناسی 0.0 – 56.10
9. Temp: دما به درجه سلسیوس 2.2 – 33.30
10. RH: رطوبت نسبی در درصد 15.0 - 100
11. Wind: سرعت باد به کیلومتر بر ساعت 0.40 – 9.40
12. Rain: باران خارج از مساحت در میلی متر بر مربع 0.0 – 6.4
13. Area: مساحت جنگل سوخته (به هکتار) 0.00 – 1090.84

(این متغیر خروجی به شدت انحراف شده به‌طوری که ممکن است با تبدیل لگاریتمیک مدل کردن منطقی باشد)