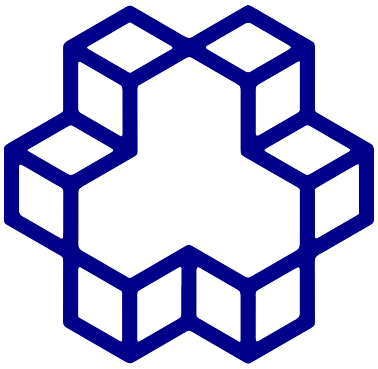


دانشگاه صنعتی خواجه نصیرالدین طوسی

دانشگاه صنعتی خواجه نصیرالدین طوسی

final course project

Course: Machine Learning

Student Name: Hadi Fathipour

Student Number: 40411334

خلاصه پروژه

هدف این پروژه، تشخیص آلارم آتش‌سوزی از روی داده‌های حسگرهای IoT است. برای این منظور، یک مدل شبکه عصبی چندلایه و یک مدل SVM آموزش داده شده‌اند و عملکرد آن‌ها با معیارهای مختلف مقایسه شده است. در این گزارش، مدل نهایی، نحوه آموزش، نمودارهای اصلی، جدول مقایسه و تحلیل خطا ارائه می‌شود.

منبع داده دیتاست مورد استفاده از مجموعه Smoke Detection Dataset در Kaggle تهیه شده است که بر پایه پروژه Real-time Smoke Detection (AI-based Sensor Fusion) گردآوری شده و شامل حدود ۶۰ هزار رکورد حسگری با نرخ نمونه‌برداری 1Hz است. برای هر رکورد، زمان UTC ثبت شده و سناریوهای متنوعی مانند محیط‌های داخلی/خارجی عادی، آتش چوب و گاز در محیط آموزشی آتش‌نشانی، کباب در فضای باز، و رطوبت بالا پوشش داده شده‌اند. همچنین ویژگی‌های حسگری شامل دما، رطوبت، TVOC، eCO2، Raw H2، Raw Ethanol، فشار هوا، ذرات معلق PM1.0/PM2.5 و غلظت عددی ذرات NC0.5/NC1.0/NC2.5 هستند.

معرفی داده‌ها

داده‌ها شامل ویژگی‌های حسگرهای محیطی (دما، رطوبت، TVOC، eCO2، اتانول خام و فشار) و برچسب Fire Alarm هستند. برچسب برابر ۱ نشان‌دهنده رخداد آتش و ۰ نشان‌دهنده حالت عادی است.

آمار کلی داده‌ها تعداد کل نمونه‌ها قبل از پاکسازی برابر 62630 است. پس از پاکسازی داده‌ها (حذف پرت‌ها و آماده‌سازی)، تعداد نمونه‌ها به 47457 کاهش یافته است.

توزیع برچسب‌ها تعداد نمونه‌های با برچسب 1 (آتش‌سوزی) برابر 33431 است که حدود 70.44% از کل داده‌ها را تشکیل می‌دهد. تعداد نمونه‌های با برچسب 0 (عدم حریق) برابر 14026 است که حدود 30% از داده‌ها است. این توزیع نشان می‌دهد داده‌ها نامتوازن هستند و بنابراین معیارهایی مثل Recall و F1 برای ارزیابی کلاس آتش‌سوزی اهمیت بیشتری دارند.

پیش‌پردازش

مراحل اصلی پیش‌پردازش به صورت زیر انجام شده است:

- بررسی همبستگی ویژگی‌ها با برچسب Fire Alarm.
- حذف ویژگی‌های کم‌اثر بر اساس تحلیل همبستگی و دانش مسئله.
- شناسایی و حذف داده‌های پرت با روش IQR.
- بررسی مقادیر گمشده و توزیع برچسب‌ها.

مدل نهایی

مدل نهایی یک شبکه عصبی چندلایه (MLP) است که به صورت زیر تعریف شده است:

- دو لایه مخفی با تعداد نرون‌های [40, 20]
- تابع فعال‌سازی سیگموید
- Batch Normalization و Dropout برای کاهش بیش‌برازش
- لایه خروجی با سیگموید برای طبقه‌بندی دودویی

مدل SVM در کنار شبکه عصبی، یک مدل SVM نیز برای مقایسه آموزش داده شد. این مدل با kernel = RBF و پارامتر C = 0.5 تنظیم شده است و به عنوان یک الگوریتم کلاسیک برای طبقه‌بندی دودویی مورد استفاده قرار گرفت.

نحوه آموزش

آموزش به دو روش انجام شده است:

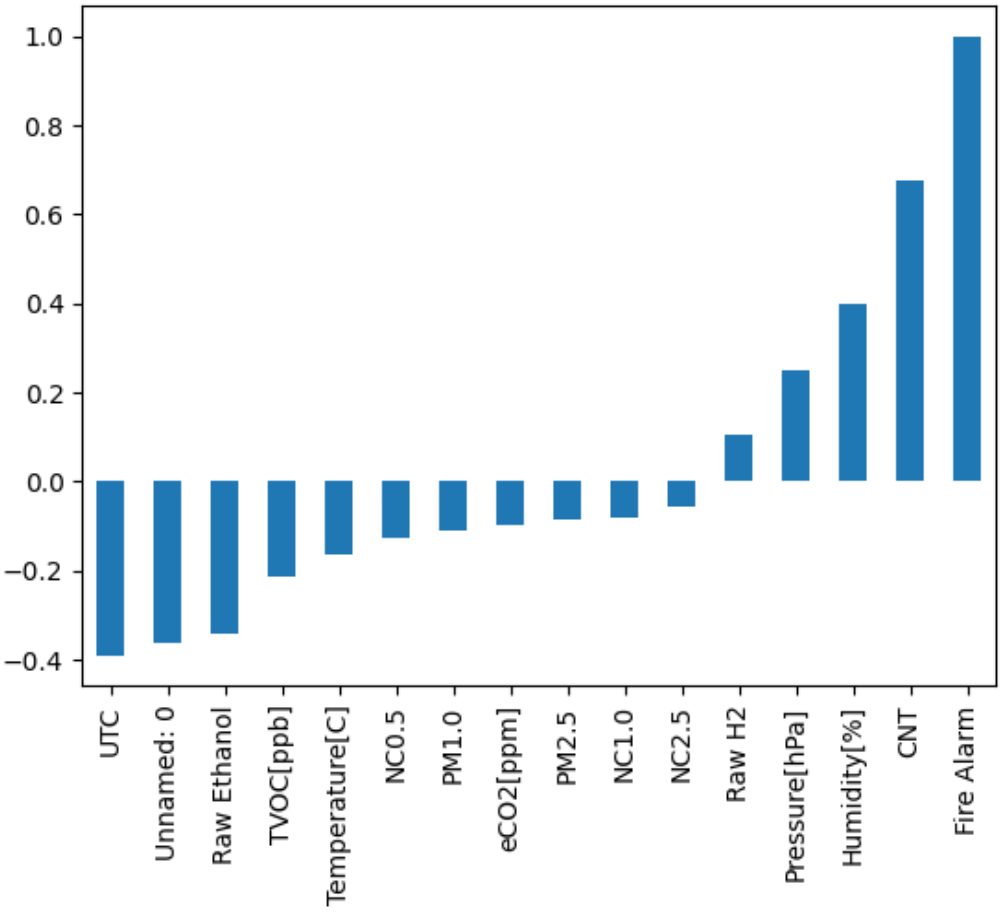
۱. **روش Holdout**: تقسیم داده به آموزش و آزمون و گزارش معیارها.

۲. **اعتبارسنجی K-Fold**: آموزش و ارزیابی در چند تکرار برای کاهش بایاس ارزیابی.

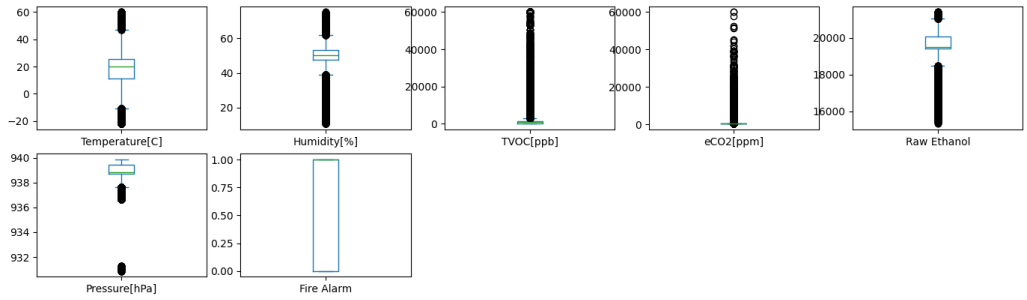
نمودارها

نمودارهای زیر برای تحلیل مدل و داده‌ها استفاده شده‌اند:

- نمودار همبستگی ویژگی‌ها با برچسب
- نمودار جعبه‌ای برای مشاهده داده‌های پرت
- نمودار تاریخچه آموزش (کاهش خطا و تغییر معیارها)
- Confusion Matrix
- منحنی ROC
- نمودار SHAP برای تفسیر ویژگی‌ها



شکل ۱: نمودار همبستگی ویژگی‌ها با برچسب Fire Alarm



شکل ۲: نمودار جعبه‌ای برای مشاهده داده‌های پرت

جدول مقایسه مدل‌ها

در جدول زیر معیارهای اصلی مقایسه شده‌اند.

Loss	AUC	F1	Recall	Precision	Accuracy	مدل
0.058116	0.998408	0.984837	0.998953	0.971114	0.978317	شبکه عصبی
		0.999791	0.999821	0.999761	0.999705	SVM

تحلیل خطا

تحلیل خطا بر اساس Confusion Matrix انجام می‌شود. موارد مهم:

- **False Negative**: آتش واقعی که مدل آن را تشخیص نداده است (ریسک بالا).
- **False Positive**: هشدار اشتباه (هزینه عملیاتی).

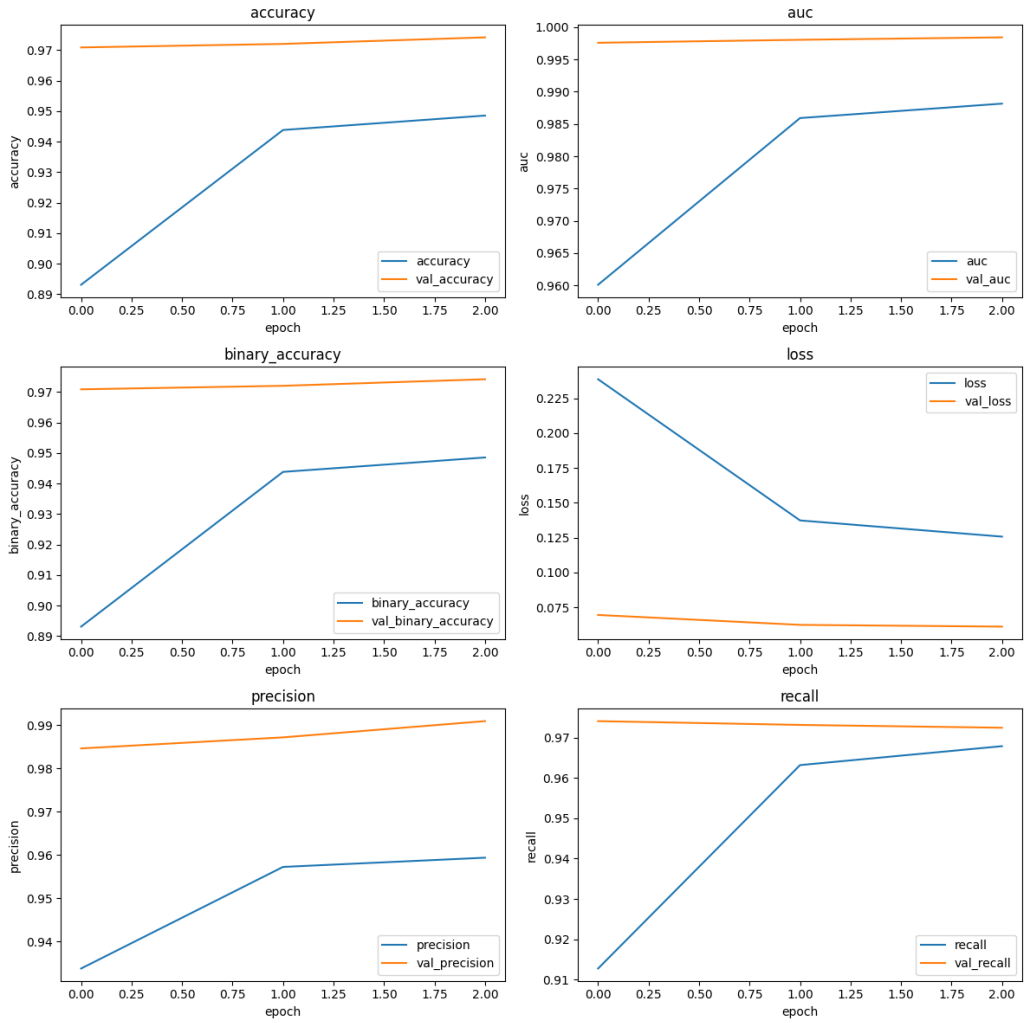
بررسی می‌شود که آیا مدل به سمت یکی از کلاس‌ها سوگیری دارد یا خیر، و آیا آستانه تصمیم‌گیری باید تنظیم شود.

نتیجه‌گیری

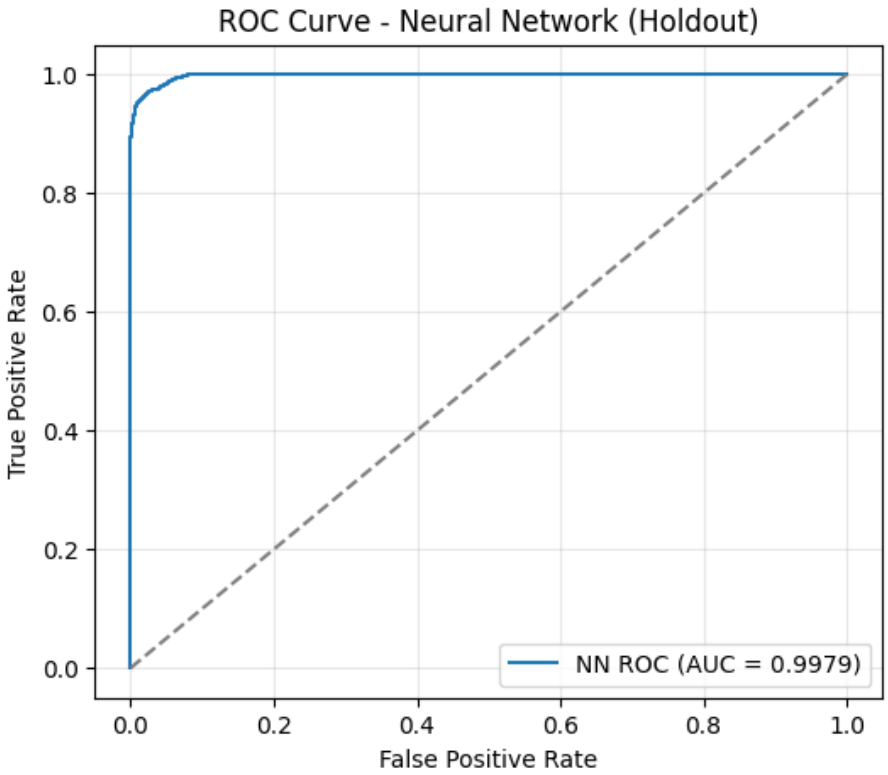
در این پروژه، مدل شبکه عصبی عملکرد بالایی در تشخیص آلارم نشان داد و با مدل SVM مقایسه شد. با توجه به نتایج، SVM عملکرد بهتری در معیارهای کلی داشت، اما تحلیل خطا نشان می‌دهد که باید روی کاهش False Negative تمرکز ویژه داشت. همچنین اجرای ارزیابی بدون نشت داده، تصویر دقیق‌تری از تعمیم‌پذیری مدل ارائه می‌دهد.

منبع دیتاست

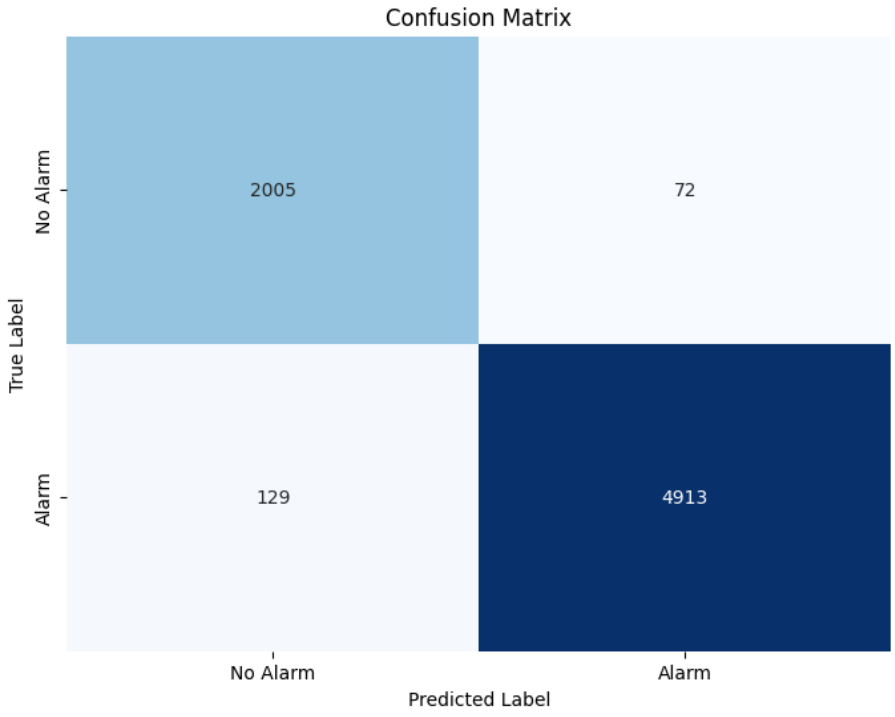
`https://www.kaggle.com/datasets/deepcontractor/smoke-detection-dataset`



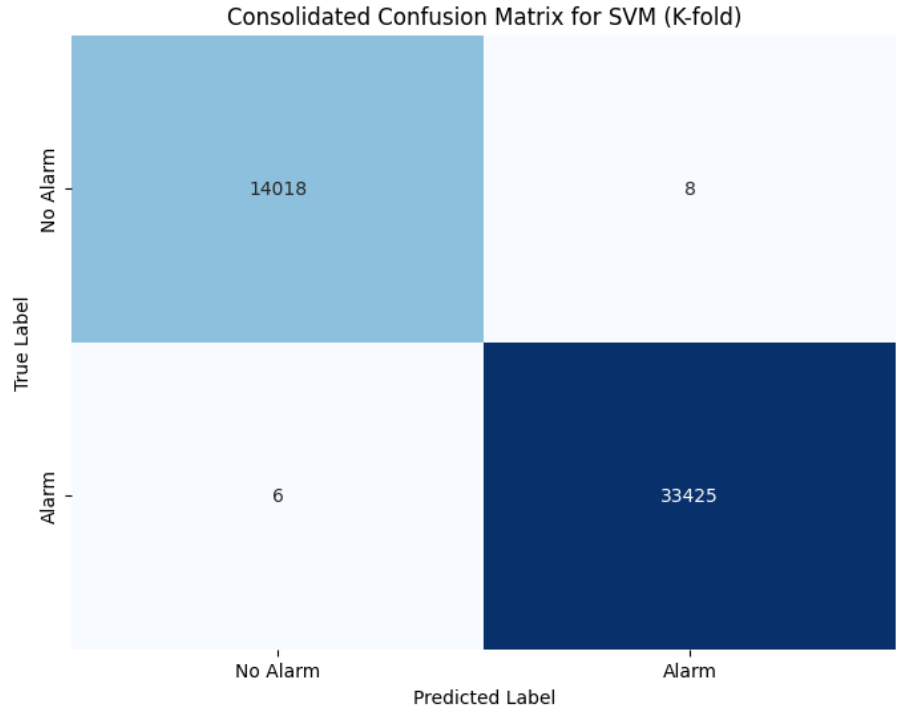
شکل ۳: نمودار تاریخچه آموزش شبکه عصبی (خطا و معیارها)



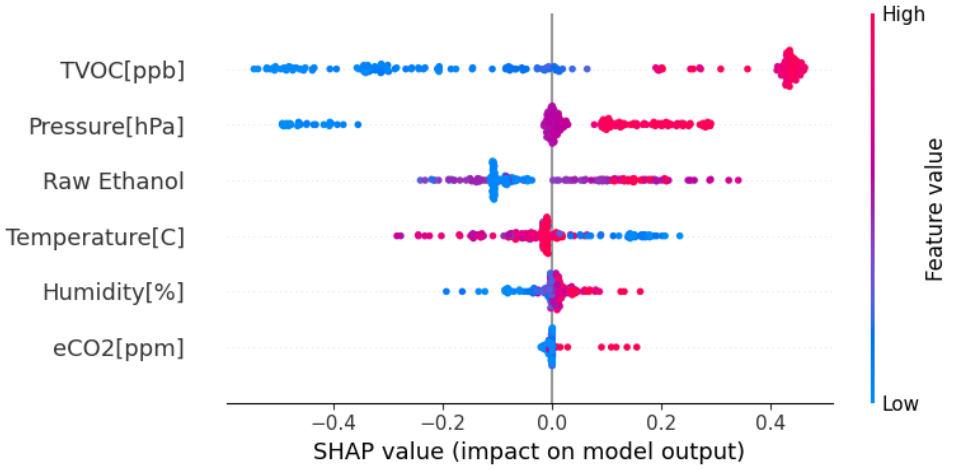
شکل ۴: منحنی ROC برای شبکه عصبی



شکل ۵: ماتریس درهم‌ریختگی (Confusion Matrix) برای شبکه عصبی



شکل ۶: ماتریس درهم‌ریختگی (Confusion Matrix) برای SVM



شکل ۷: نمودار SHAP برای تفسیر اهمیت ویژگی‌ها

	Model	Accuracy	Precision	Recall	AUC	F1	Loss
0	Neural Network	0.978317	0.971114	0.998953	0.998408	0.984837	0.058116
1	SVM	0.999705	0.999761	0.999821	NaN	0.999791	NaN

شکل ۸: خلاصه نتایج و مقایسه نهایی