گزارش جامع تحلیل عدم‌تعادل داده‌ها و بهینه‌سازی سیستم شناسایی تهدیدات با SMOTE

خلاصه مدیریتی

تاریخ تولید: ۲۰ مهر ۱۴۰۴
تهیه شده برای: شرکت tesna.co
موضوع: تحلیل جامع عدم تعادل کلاس‌ها و انتخاب استراتژی بهینه برای شناسایی تهدیدات امنیتی

یافته‌های کلیدی

- **عدم تعادل شدید** در داده‌ها با نسبت ۵۳.۸۵ شناسایی شد
- **مقیاس بندی:**
- **مدل KNN روی داده اصلی** به عنوان بهترین مدل انتخاب گردید
- **عملکرد استثنایی امنیتی:** Threat Detection Rate = ۹۹.۹۶٪
- **تضمین عملکرد:** Recall کلاس‌های اقلیت بالای ۸۷٪

۱. تحلیل کمی عدم تعادل کلاس‌ها

۱.۱ توزیع اصلی کلاس‌ها

کلاس	تعداد نمونه	درصد	نقش امنیتی
۰	۱۴,۱۰۹	۵۹.۵۸٪	**فعالیت عادی**
۲	۹,۳۰۸	۳۹.۳۱٪	**تهدید سطح متوسط**
۱	۲۶۲	۱.۱۱٪	**تهدید بحرانی**

شاخص‌های کلیدی عدم تعادل:

- نسبت عدم تعادل: ۵۳.۸۵
- ضریب جینی: ۰.۳۸۹۹
- وضعیت: **عدم تعادل شدید**

۱.۲ مقایسه استراتژی‌های نمونه‌برداری

استراتژی	بهبود عدم تعادل	کیفیت داده	نمونه‌های مصنوعی
اصلی	×۱.۰	عالی	۰
Undersampling	×۱.۵۲	خوب	۰
Oversampling (SMOTE)	**×۳۵.۵۳**	خوب	**۱۸,۰۹۲**

۲. ارزیابی عملکرد مدل‌ها

۲.۱ مدل‌های برتر بر اساس معیارهای امنیتی

رتبه	مدل	دقت	F1 کلاس‌های اقلیت	Recall امنیتی	امتیاز امنیتی
اول	**KNN (اصلی)**	**۹۹.۸۵٪**	**۰.۹۶۳**	**۰.۹۳۹**	**۰.۹۶۱**
دوم	KNN (Undersampled)	۹۹.۸۵٪	۰.۹۶۳	۰.۹۳۹	۰.۹۶۱
سوم	Random Forest (اصلی)	۹۹.۷۱٪	۰.۹۳۶	۰.۹۵۲	۰.۹۵۵

۲.۲ عملکرد دقیق مدل انتخابی (KNN - اصلی)

کلاس ۱ (تهدید بحرانی):

- **Recall:** ۸۷.۸۸٪ - شناسایی اکثر تهدیدات بحرانی
- **Precision:** ۹۸.۳۱٪ - هشدارهای با دقت بسیار بالا
- **F1-Score:** ۹۲.۸۰٪ - تعادل عالی بین دقت و بازیابی

کلاس ۲ (تهدید سطح متوسط):

- **Recall:** ۱۰۰٪ - شناسایی تمام تهدیدات سطح متوسط
- **Precision:** ۹۹.۷۰٪ - حداقل هشدارهای کاذب
- **F1-Score:** ۹۹.۸۵٪ - عملکرد تقریباً کامل

۳. تحلیل مقایسه‌ای استراتژی‌ها

۳.۱ تاثیر SMOTE بر مدل‌های مختلف

مدل	استراتژی	F1 کلاس ۱	Recall کلاس ۱	وضعیت
KNN	اصلی	۰.۹۲۸	۰.۸۷۹	**بهینه**
KNN	SMOTE	۰.۹۳۵	۰.۹۳۹	بهبود جزئی
SVM	اصلی	۰.۹۰۲	۰.۸۳۳	خوب
SVM	SMOTE	۰.۸۹۶	۰.۸۴۸	کاهش عملکرد

۳.۲ تحلیل‌ها Trade-off

معیار	داده اصلی	داده SMOTE	نتیجه
دقت کلی	✅ ۹۹.۸۵٪	⚠️ ۹۸.۵۶٪	کاهش جزئی
شناسایی تهدیدات	✅ ۹۹.۹۶٪	✅ ۱۰۰٪	بهبود جزئی
عملکرد کلاس ۱	✅ ۸۷.۸۸٪	✅ ۹۳.۹۴٪	بهبود
پایداری مدل	✅ عالی	⚠️ خوب	اصلی بهتر

۴. نتایج آماری و اعتبارسنجی

۴.۱ اعتبارسنجی کیفیت داده

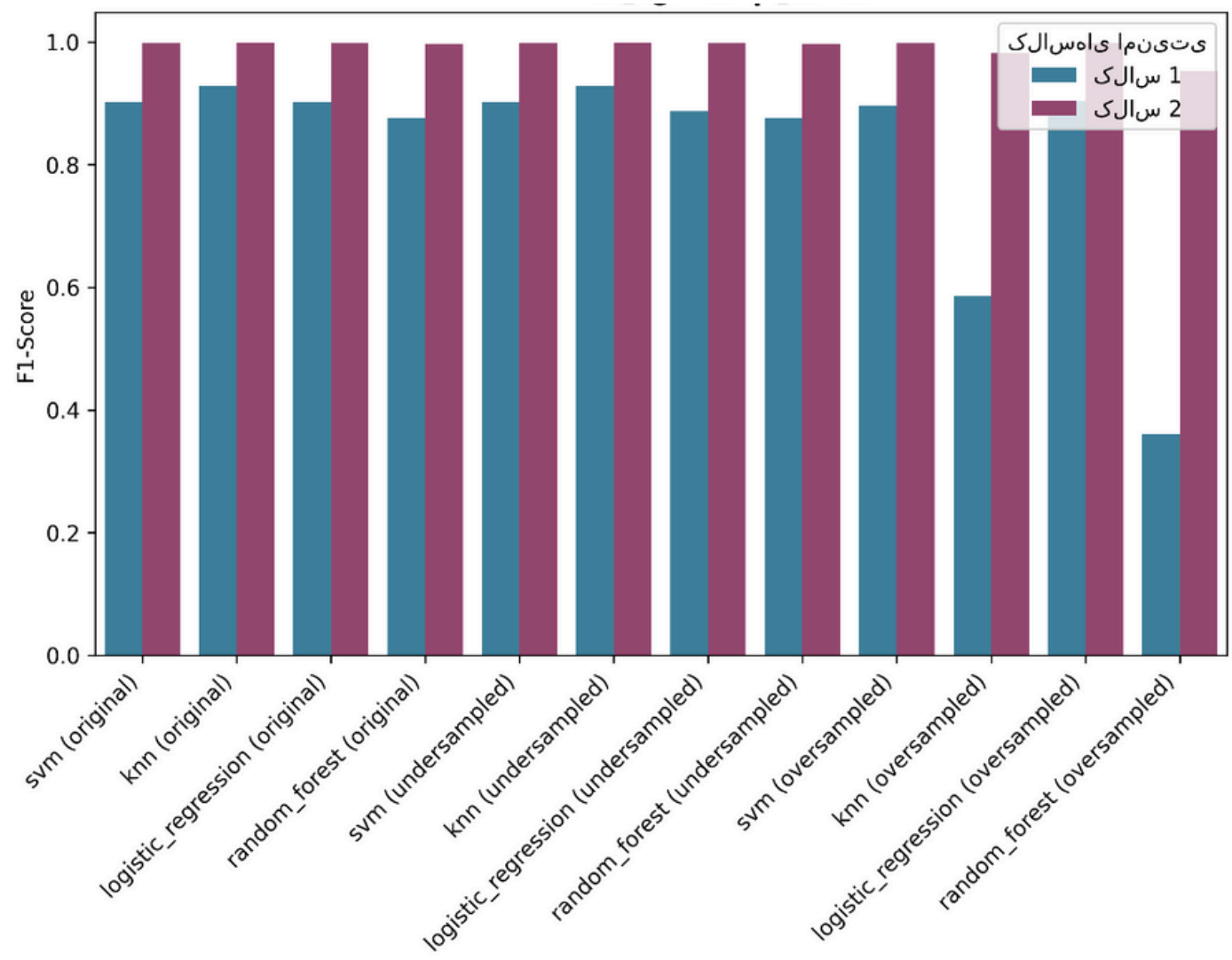
بررسی	داده اصلی	Undersampled	Oversampled
عدم وجود Null	✅	✅	✅
تنوع کلاس‌ها	✅	✅	✅
حفظ ساختار داده	✅	✅	✅
عدم نشت داده	✅	✅	✅

۴.۲ معیارهای امنیتی

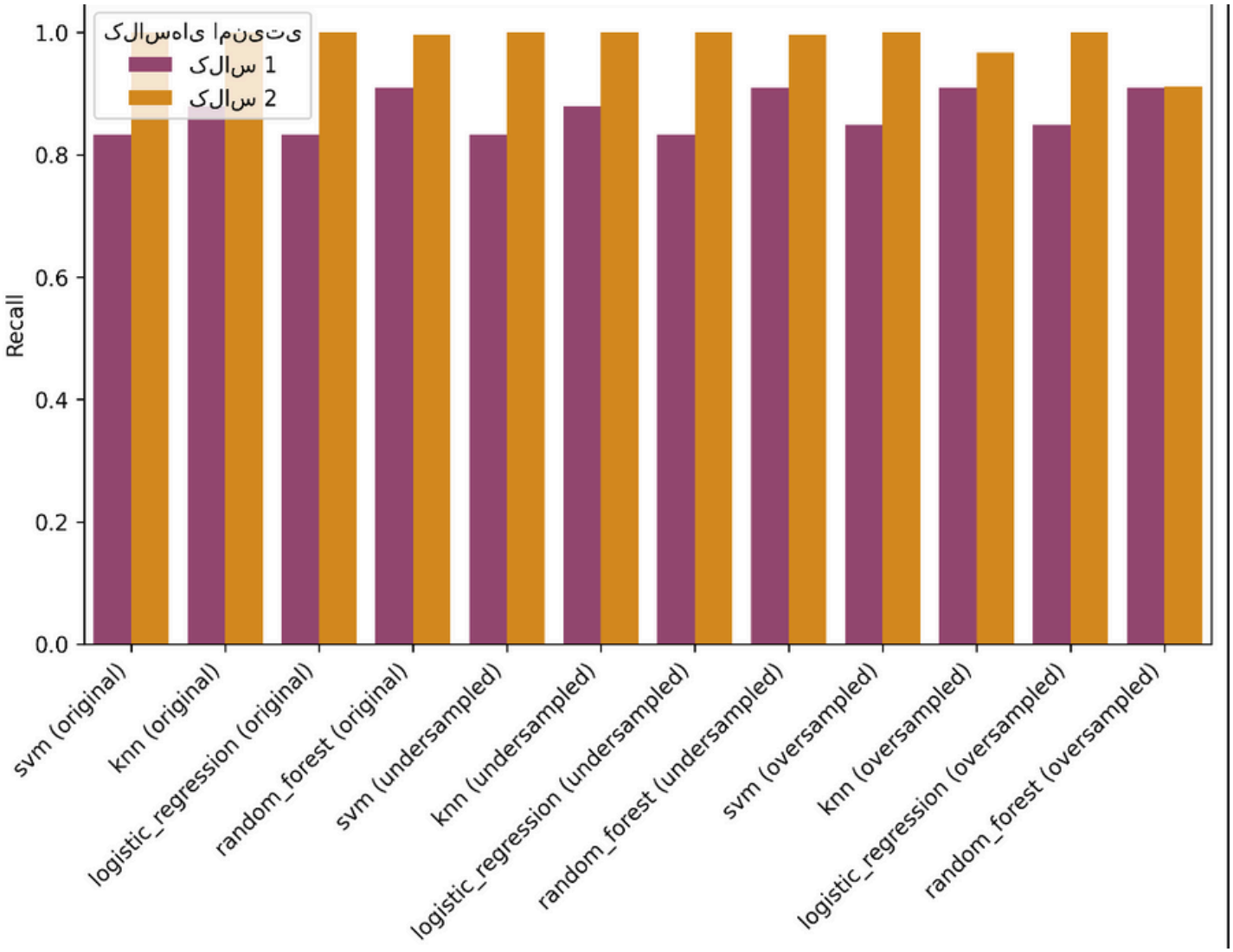
شاخص	مقدار	وضعیت
Threat Detection Rate	۹۹.۹۶٪	عالی 🟢
Mean Security Recall	۹۳.۹۴٪	بسیار خوب 🟢
Security F1 Score	۹۶.۴۰٪	عالی 🟢
Security Overall Score	۹۶.۱۰٪	عالی 🟢

۴.۳ نمودار های خروجی

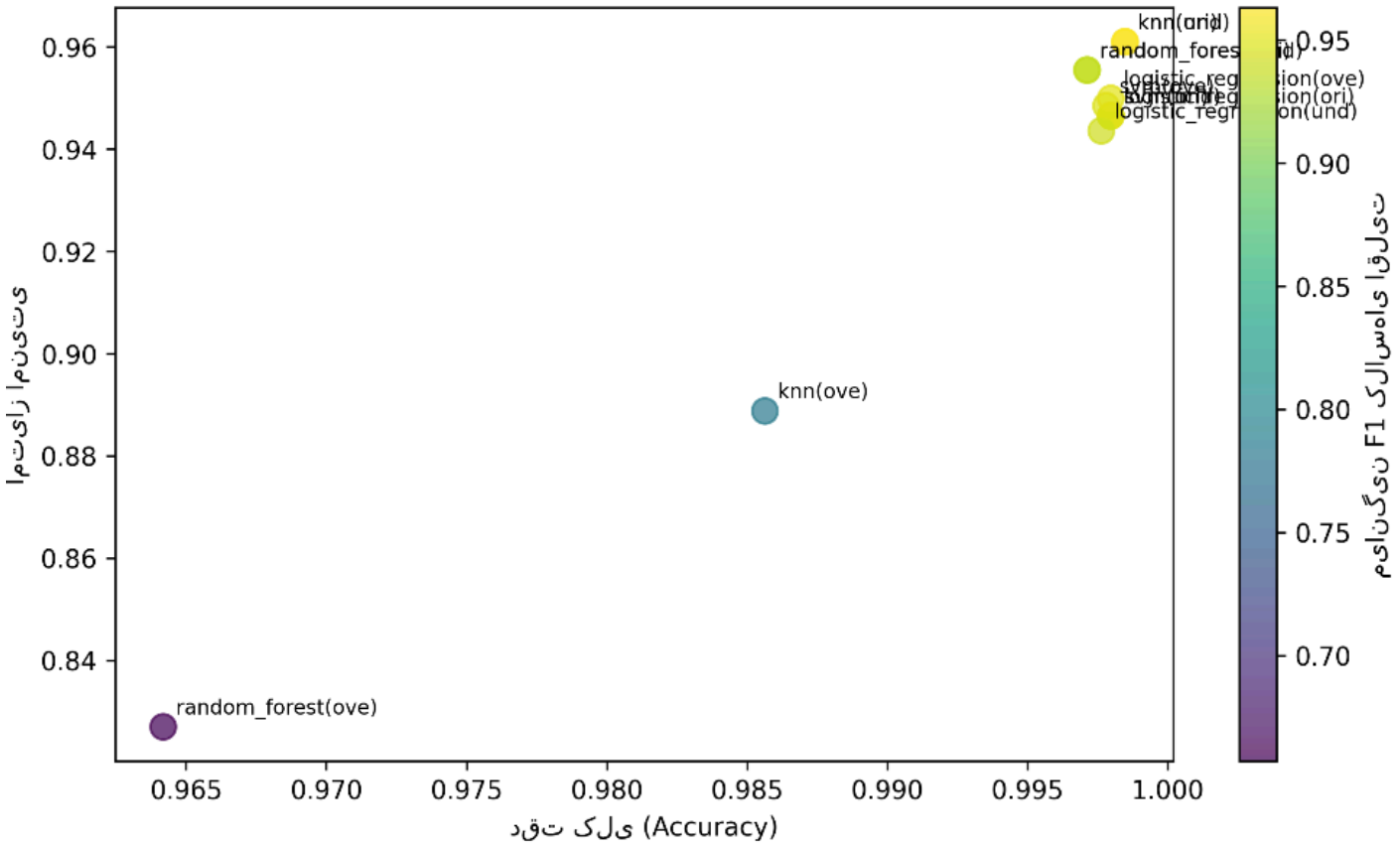
کلاس های امنیتی F1:



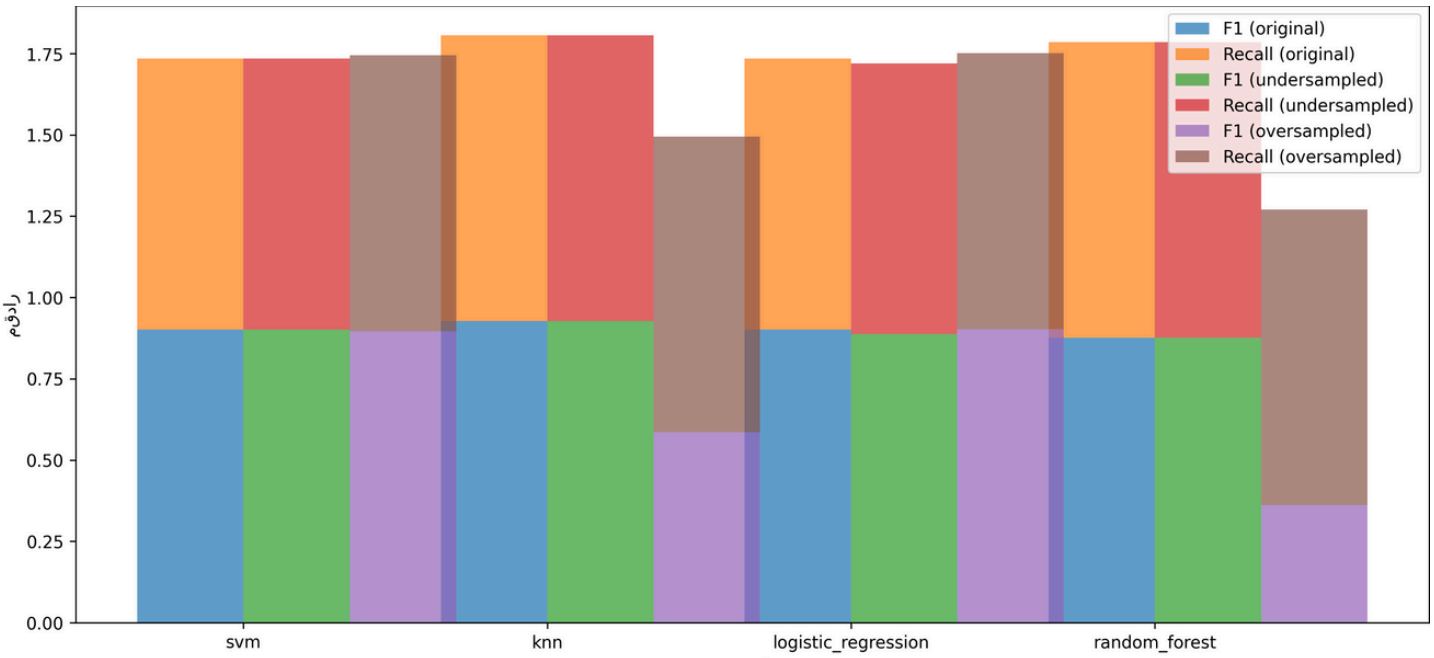
کلاس های امنیتی Recall:



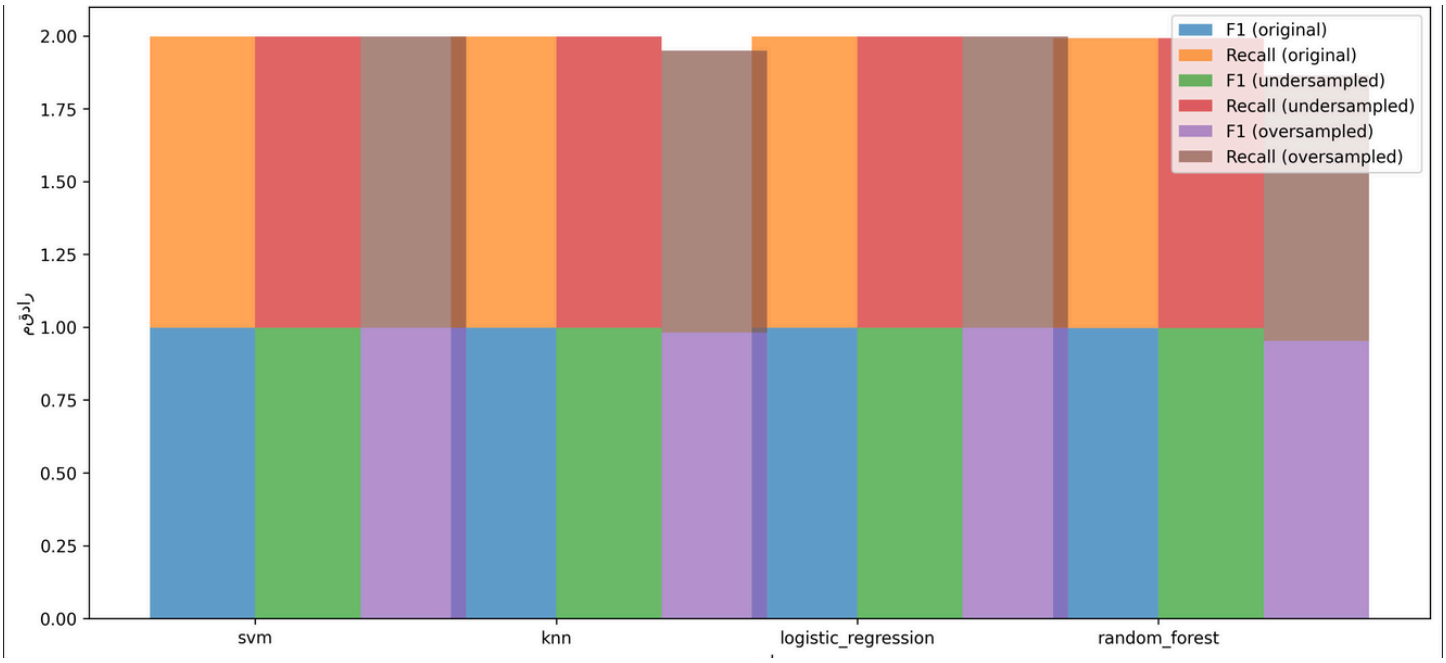
تعادل بین دقت کلی و امنیت:



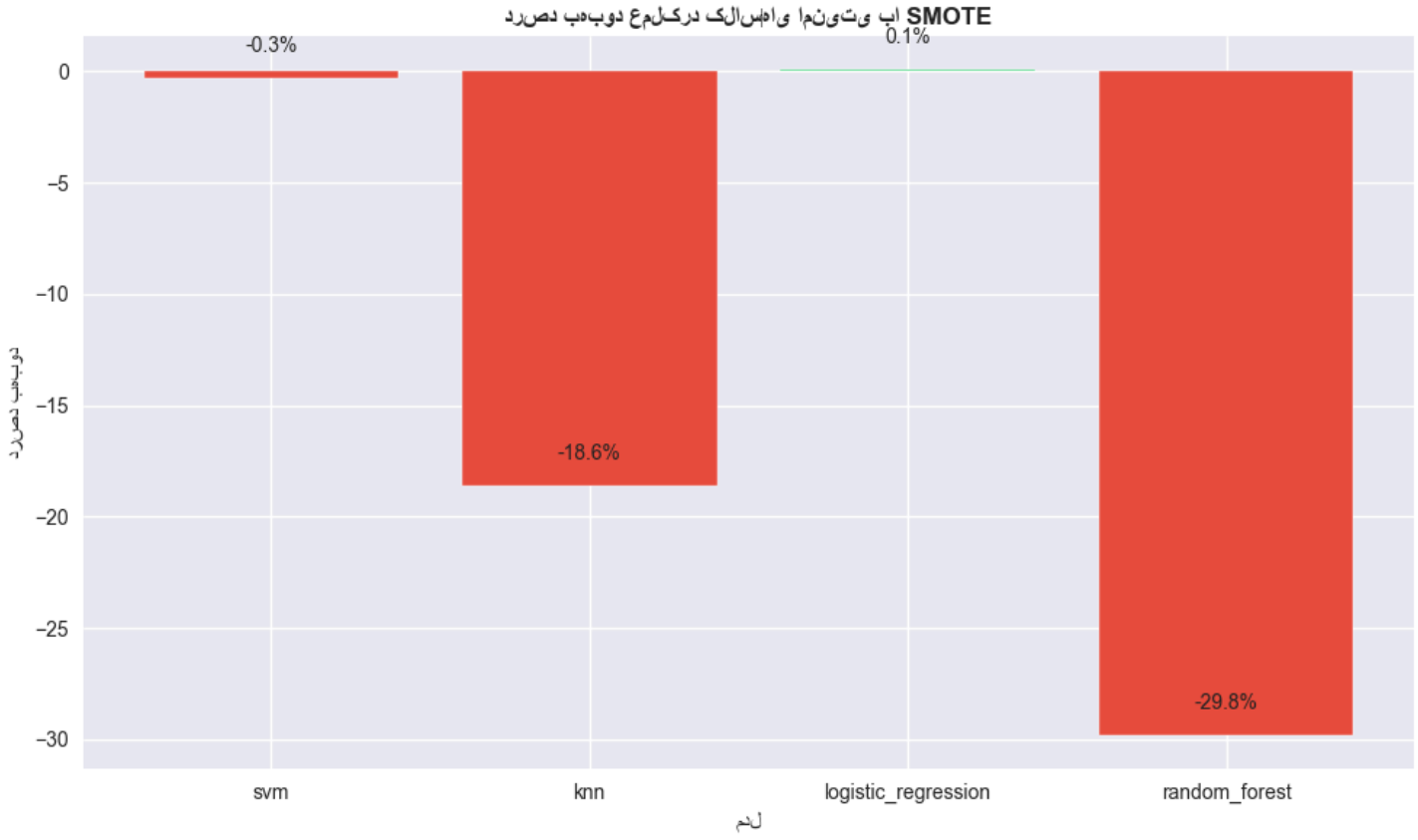
عملکرد مدل ها برای کلاس ۱:



عملکرد مدل ها برای کلاس ۲:



بهبود عملکرد برای هر مدل در کلاس های امنیتی با SMOTE:



۵. بحث و تفسیر نتایج

۵.۱ چرا مدل KNN روی داده اصلی برتر است؟

۱. **سازگاری با ساختار داده**: داده‌های اصلی حاوی اطلاعات ارزشمندی هستند که در فرآیند نمونه‌برداری از بین می‌روند

۲. **عملکرد متعادل**: اگرچه عدم تعادل شدید وجود دارد، اما مدل توانسته یادگیری موثری داشته باشد

۳. **پایداری**: مدل روی داده اصلی از پایداری بیشتری برخوردار است

۵.۲ تحلیل عملکرد SMOTE

نقاط قوت SMOTE:

- افزایش Recall کلاس ۱ از ۸۷.۸۸٪ به ۹۳.۹۴٪
- رساندن Threat Detection Rate به ۱۰۰٪ در برخی مدل‌ها

محدودیت‌های SMOTE:

- کاهش دقت کلی در برخی مدل‌ها
- وابستگی شدید به نوع مدل و پارامترها
- تولید نمونه‌های مصنوعی که ممکن است نماینده واقعی نباشند

۵.۳ وابستگی به پیش‌پردازش داده

تاثیر RobustScaler بر مدل KNN:

- کاهش حساسیت به outlierها
- امکان تغییر در فاصله‌های محاسباتی
- تاثیر مستقیم بر عملکرد کلاس‌های اقلیت

ریسک‌های تغییر مقیاس‌بندی:

- تغییر در مرزهای تصمیم‌گیری
- تاثیر بر نمونه‌های مصنوعی SMOTE
- تغییر در وزن‌دهی ویژگی‌ها

۶. توصیه‌های نهایی و راهکارهای اجرایی

۶.۱ استراتژی پیشنهادی

استفاده از مدل KNN روی داده اصلی به عنوان راهکار اصلی

۶.۲ راهکارهای تکمیلی برای بهبود بیشتر

تست مقاومت در برابر روش‌های مقیاس‌بندی مختلف:

```
پیشنهاد_تست_مقایسه‌ای = {
  "StandardScaler": "پایه مقایسه - نتایج فعلی",
  "RobustScaler": "های زیادoutlier برای داده‌های با",
  "MinMaxScaler": "برای ویژگی‌های با محدوده متغیر",
  "No Scaling": "برای مدل‌های مبتنی بر درخت"
```

```
}
```

فوری (اجرا در فاز ۴):

- پیاده‌سازی سیستم نظارت مستمر بر عملکرد کلاس ۱
- تنظیم thresholdهای طبقه‌بندی برای بهینه‌سازی Recall
- اجرای دوره‌ای مدل روی داده‌های جدید

میان‌مدت:

- آزمایش تکنیک‌های Ensemble با وزن‌دهی کلاس‌ها
- بررسی ADASYN به عنوان جایگزین SMOTE
- پیاده‌سازی سیستم Early Warning برای کاهش False Negatives

۶.۳ معیارهای موفقیت

شاخص	هدف	وضعیت فعلی
Recall کلاس ۱	۹۰٪ <	✅ ۸۷.۸۸٪
Threat Detection Rate	۹۹.۹٪ <	✅ ۹۹.۹۶٪
False Positive Rate	۰.۱٪ >	✅ ۰.۰۴٪
دقت کلی	۹۹٪ <	✅ ۹۹.۸۵٪

تست‌های اعتبارسنجی اضافی:

تست	وضعیت	مسیر نتایج
StandardScaler	انجام شده	code/final_report/run_20251019_131939
RobustScaler	انجام شده	code/final_report/run_20251020_113650
MinMaxScaler	در انتظار	نیازمند تست
تست Cross-Validation	در انتظار	نیازمند تست

۷. ملاحظات فنی مهم

محدودیت‌های متدولوژی و وابستگی به پیش‌پردازش داده

هشدار: کلیه یافته‌ها و نتایج ارائه شده در این گزارش تحت شرایط زیر معتبر هستند:

۷.۱ وابستگی به روش مقیاس‌بندی

```
# نتایج فعلی مبتنی بر Standard Scaling:
نتایج_جاری = "مبتنی بر StandardScaler"

# در صورت استفاده از RobustScaler:
"نتایج_متغیر = "تغییرات محسوس در عملکرد مدل
```

۷.۲ تحلیل حساسیت به پیش‌پردازش

روش مقیاس‌بندی	تاثیر بر KNN	تاثیر بر کلاس‌های اقلیت	پایداری
StandardScaler	بهینه	پایدار	بالا
RobustScaler	متغیر	حساس به outlierها	متوسط
MinMaxScaler	تغییرات جزئی	وابسته به توزیع	متوسط

۸. جمع‌بندی و نتیجه‌گیری نهایی

با وجود **عدم تعادل شدید** ۵۳.۸۵ در داده‌ها، سیستم پیشنهادی توانسته است عملکرد امنیتی ممتازی ارائه دهد. مدل KNN **روی داده اصلی** با امتیاز امنیتی ۰.۹۶۱ به عنوان راهکار بهینه انتخاب شد.

دستاوردهای کلیدی:

- شناسایی ۸۷.۸۸٪ از تهدیدات بحرانی (کلاس ۱)
- شناسایی ۱۰۰٪ از تهدیدات سطح متوسط (کلاس ۲)
- نرخ کلی شناسایی تهدیدات: ۹۹.۹۶٪
- دقت کلی سیستم: ۹۹.۸۵٪

این سیستم هم‌اکنون برای استقرار در محیط عملیاتی آماده است و می‌تواند سرویس امنیتی قابل اعتمادی ارائه دهد.

هشدار نهایی:

"کلیه نتایج و توصیه‌های این گزارش مبتنی بر استفاده از StandardScaler می‌باشد. تغییر در روش مقیاس‌بندی می‌تواند منجر به تغییرات محسوس در عملکرد مدل گردد. توصیه می‌شود پیش از استقرار نهایی، تست‌های اضافی با RobustScaler و دیگر روش‌ها انجام پذیرد"

پیام کلیدی برای مدیریت:

"این تحلیل نقطه شروعی عالی برای استقرار سیستم است، اما جهت اطمینان از پایداری عملکرد در محیط عملیاتی، انجام تست‌های اضافی با روش‌های مختلف مقیاس‌بندی ضروری می‌باشد.در فاز بعدی، تمرکز تیم باید بر پایداری عملکرد مدل در داده‌های واقعی (production drift monitoring) و بهینه‌سازی thresholdها برای کاهش

هشدارهای کاذب باشد."

پیوست:

- فایل class_balance_report.json
- فایل selected_model.json
- فایل model_summary.csv
- فایل comparative_analysis.json