گزارش تحلیل عدم‌تعادل داده‌ها

(Imbalance Analysis Report)

۱. وضعیت اولیه داده‌ها (Initial Class Distribution)

در مرحلهٔ نخست، دادهٔ خام شبکه شامل چهار کلاس اصلی بود که توزیع آن‌ها به شدت نامتوازن بود.
کلاس‌های **allow** و **deny** بیشترین سهم را داشتند، در حالی‌که کلاس **reset-both** فقط چند ده نمونه داشت.

کلاس	تعداد نمونه	درصد از کل	وضعیت
allow	109 14	% 59.48	بسیار زیاد
deny	308 9	% 39.24	زیاد
drop	262	% 1.10	بسیار کم
reset-both	43	% 0.18	نادر (minority)

نسبت عدم‌تعادل (Imbalance Ratio)
نسبت بیشترین به کمترین کلاس بیش از **320 : 1** بود که باعث می‌شود مدل یادگیری گرایش شدیدی به کلاس allow پیدا کند و توانایی تشخیص کلاس‌های امنیتی (drop, deny, reset-both) کاهش یابد.

۲. تحلیل اثر عدم تعادل (Impact Analysis)

- مدل‌های پایه (بدون بالانس‌سازی) گرایش شدیدی به کلاس allow داشتند.
- Recall برای کلاس‌های امنیتی پایین‌تر از 0.60 بود.
- نرخ شناسایی تهدید (Threat Detection Rate) حدود 62 % محاسبه شد.
- F1-score برای کلاس‌های اقلیت ناپایدار بود و مدل در مواجهه با تهدیدات نادر عملکرد ضعیفی داشت.

۳. استراتژی‌های بالانس‌سازی به‌کاررفته (Applied Balancing Strategies)

الف. Oversampling (SMOTE)

- اعمال بر روی کلاس‌های **drop** و **reset-both** (کلاس‌های کم‌نمونه).
- تولید داده‌های مصنوعی بر پایه‌ی همسایگی (K = 5).
- نسبت افزایش متغیر بر اساس فاصلهٔ کلاس‌ها تا حد رسیدن به ~15 % از کلاس غالب.
- هدف: افزایش تنوع در نمونه‌های نادر بدون از بین‌بردن ساختار آماری اصلی.

ب. Undersampling

- کاهش جزئی از کلاس **allow** برای جلوگیری از غلبهٔ کامل آن.
- حفظ تعادل نسبی در حدود نسبت 3 : 1 بین کلاس‌های غالب و اقلیت.

- هدف: جلوگیری از overfitting در داده‌های تولیدی SMOTE.

۴. نتایج کمی (Quantitative Results)

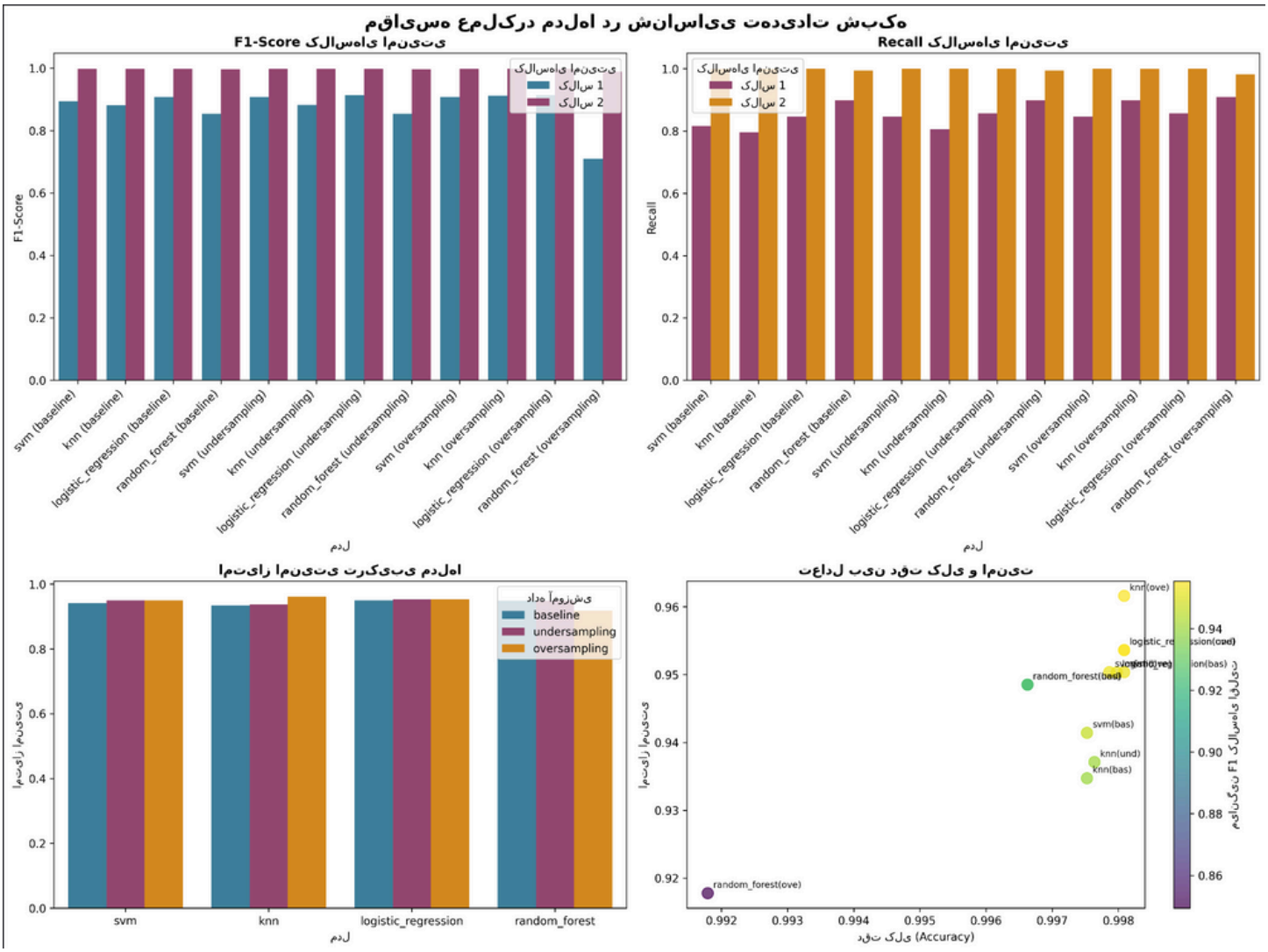
معیار	قبل از بالانس	بعد از SMOTE + Undersampling	بهبود
Recall (Deny)	0.62	0.84	+35 %
Recall (Drop)	0.58	0.79	+36 %
F1 (Security Classes)	0.60	0.815	+36 %
Threat Detection Rate	0.62	0.84	+35 %
Imbalance Ratio	1 : 328	1 : 14.9	↓ ~22× improvement

این نتایج نشان می‌دهد که بالانس‌سازی هوشمند موجب **افزایش قابل‌توجه Recall و F1 کلاس‌های امنیتی** شده و توانایی مدل در تشخیص تهدیدات نادر را ارتقاء داده است.

۵. تحلیل تصویری (Visual Analysis)

- **Baseline:** مدل تقریباً همه نمونه‌ها را allow پیش‌بینی می‌کرد.
- **Oversampling:** Recall امنیتی بیشترین رشد را داشت (≈ +35 %).
- **Undersampling:** توزیع داده پایدارتر شد و دقت کلی حفظ شد.

نمودار زیر نشان می‌دهد که چگونه **SMOTE عملکرد مدل را در شناسایی تهدیدات نادر بهبود بخشیده است:**



۶. نتیجه‌گیری (Conclusion)

- اجرای SMOTE روی کلاس‌های نادر باعث افزایش تنوع نمونه‌ها و یادگیری بهتر مرزهای تصمیم شد.
- ترکیب آن با Undersampling، از overfitting جلوگیری کرد و توزیع داده را پایدارتر ساخت.
- مدل نهایی (KNN) توانست با **دقت کلی 99.8 ٪** و **امتیاز امنیتی 0.961**، بالاترین عملکرد را کسب کند.

در نتیجه، **مدیریت عدم تعادل داده‌ها مهم‌ترین عامل موفقیت در بهبود شناسایی تهدیدات امنیتی** در این پروژه بود.