**"کاوش تهدیدات صنعتی: تحلیل و طبقه‌بندی بدافزارها با روش‌های آماری" با دیتاست Malware Dataset**

**نام دیتاست در سایت کگل :Microsoft Malware Classification Challenge (BIG2015)**

**الزامات:**

**-** پروژه شامل استخراج ویژگی و انتخاب ویژگی باشد.

- ارزیابی روشها به صورت مقایسهای انجام شود.

-گزارش پروژه در قالب یک مقاله علمی ژورنال یا گزارش علمی و فنی و همچنین اسلایدها برای ارائه باشد.

-گزارش پروژه بایستی شامل عنوان، نویسنده، چکیده، مقدمه، تئوری و توضیح الگوریتم، نتایج و مراجع باشد. حتماً چندمقاله علمی را مطالعه و استفاده کرده باشید تا بتوانید ارجاع دهید.

فایل تحویلی پروژه باید شامل اسلایدها ، گزارش، کدها و همچنین توضیحات کارهای انجام شده به صورت صوتی یا نوشتاری باشد، ضمنا شرح ایجاد اسلایدها در بخش 7 با عنوان سمینار به صورت کامل توضیح داده شده است.

شرح پروژه آن چیزی که باید انجام شود را به صورت کامل در ادامه توضیح داده شده است.

(راهکار به صورت خلاصه – شرح کامل در ادامه)

1. **آماده‌سازی داده‌ها**:
	* پیش‌پردازش: پاکسازی، نرمال‌سازی.
	* تحلیل اکتشافی.
2. **کاوش الگوهای پنهان**:
	* استفاده از GMM با الگوریتم EM برای خوشه‌بندی.
	* کاهش ابعاد با PCA یا t-SNE.
	* تفسیر خوشه‌ها (مثل بدافزارهای مبهم‌شده).
3. **طبقه‌بندی**:
	* روش‌های آماری: Naive Bayes، رگرسیون لجستیک.
	* SVM: کرنل خطی یا RBF.
	* شبکه‌های عصبی: 1-3 لایه مخفی.
	* ترکیب ویژگی‌ها و نتایج EM.
4. **ارزیابی**:
	* خوشه‌بندی: Silhouette Score, BIC.
	* طبقه‌بندی: Accuracy, F1-Score, ROC-AUC.
5. **کاربرد صنعتی**:
	* ربط الگوها به تهدیدات.
	* راهکار برای حفاظت.
6. **ارائه**:
	* نمایش الگوها و نتایج.
	* تأکید بر نقش EM.

**1.تعریف اهداف پروژه**

* **کاوش الگوهای پنهان**: شناسایی زیرگروه‌های بدافزار (مثل باج‌افزار، تروجان) یا رفتارهای خاص (مثل مبهم‌سازی) با استفاده از روش‌های آماری و الگوریتم EM.
* **طبقه‌بندی**: استفاده از الگوهای کشف‌شده برای طبقه‌بندی دودویی (بدافزار/بی‌خطر) با روش‌های آماری، SVM و شبکه‌های عصبی.
* **کاربرد صنعتی**: تحلیل الگوها و نتایج طبقه‌بندی برای حفاظت از زیرساخت‌های صنعتی.
* **مقایسه مدل‌ها**: بررسی عملکرد روش‌های آماری، SVM و شبکه‌های عصبی.

**2. آماده‌سازی و تحلیل اولیه داده‌ها**

* **درک دیتاست**:
	+ دیتاست Malware Dataset شامل ویژگی‌هایی مثل تعداد بخش‌ها، اندازه کد، آنتروپی، توابع واردشده و برچسب (بدافزار/بی‌خطر) است.
	+ مستندات را مطالعه کنید.
* **پیش‌پردازش داده‌ها**:
	+ **مدیریت داده‌های گمشده**: مقادیر گمشده را با میانگین یا میانه پر کنید.
	+ **نرمال‌سازی/استانداردسازی**: از Standard Scaling برای SVM و شبکه‌های عصبی استفاده کنید.
	+ **بررسی تعادل داده‌ها**: اگر داده‌ها نامتوازن هستند، از SMOTE یا وزن‌دهی کلاس‌ها استفاده کنید.
	+ **تقسیم داده‌ها**: 70% آموزشی، 15% اعتبارسنجی، 15% آزمایشی.
* **تحلیل اکتشافی داده‌ها (EDA)**:
	+ توزیع ویژگی‌ها (هیستوگرام، نمودار جعبه‌ای).
	+ ماتریس همبستگی برای شناسایی ویژگی‌های کلیدی.
	+ بررسی اولیه الگوها (مثل آنتروپی بالا برای مبهم‌سازی).

**3.کاوش الگوهای پنهان با الگوریتم EM**

* **مدل مخلوط گاوسی (GMM) با الگوریتم EM**:
	+ **چرا مناسب است؟**: GMM داده‌ها را به صورت مخلوطی از توزیع‌های گاوسی مدل‌سازی می‌کند و از الگوریتم EM برای کشف الگوهای پنهان (مثل زیرگروه‌های بدافزار) استفاده می‌کند.
	+ **نقش EM**:
		- **مرحله E**: احتمال عضویت هر فایل PE در هر خوشه (مثلاً باج‌افزار، تروجان) را تخمین می‌زند.
		- **مرحله M**: پارامترهای خوشه‌ها (میانگین، واریانس، وزن) را به‌روز می‌کند تا بهترین تطابق با داده‌ها حاصل شود.
	+ **کاربرد**: فایل‌های PE را به خوشه‌هایی تقسیم کنید (مثلاً خوشه 1: بدافزارهای مبهم‌شده، خوشه 2: بدافزارهای شبکه‌ای، خوشه 3: بی‌خطر).
	+ **تنظیمات**: تعداد خوشه‌ها را با معیارهای BIC یا Silhouette Score تعیین کنید.
* **کاهش ابعاد**:
	+ **PCA**: ابعاد را کاهش دهید تا ویژگی‌های اصلی را شناسایی کنید.
	+ **t-SNE یا UMAP**: خوشه‌ها را در فضای دو یا سه‌بعدی مصورسازی کنید.
* **تحلیل خوشه‌ها**:
	+ هر خوشه را بررسی کنید (مثلاً خوشه با آنتروپی بالا ممکن است بدافزارهای مبهم‌شده باشد).
	+ از تست‌های آماری (t-test، ANOVA) برای مقایسه ویژگی‌ها بین خوشه‌ها استفاده کنید.

**4.طبقه‌بندی با مدل‌های مختلف**

* **الف. روش‌های آماری**:
	+ **Naive Bayes**:
		- ساده و سریع.
		- Gaussian Naive Bayes برای ویژگی‌های پیوسته (مثل آنتروپی).
		- Bernoulli Naive Bayes برای ویژگی‌های باینری (مثل وجود تابع).
	+ **رگرسیون لجستیک**:
		- مدل خطی برای طبقه‌بندی دودویی.
	+ **کاربرد**: ویژگی‌های اصلی و احتمال عضویت در خوشه‌های GMM (از EM) را ترکیب کنید.
* **ب. ماشین بردار پشتیبان (SVM)**:
	+ **چرا مناسب است؟**: هیپرصفحه بهینه را برای جداسازی کلاس‌ها پیدا می‌کند.
	+ **کرنل‌ها**: خطی (برای داده‌های ساده)، RBF (برای روابط غیرخطی).
	+ **تنظیمات**: پارامترهای C (جریمه خطا) و Gamma (انعطاف‌پذیری) را با اعتبارسنجی متقاطع تنظیم کنید.
	+ **کاربرد**: ویژگی‌ها و نتایج خوشه‌بندی EM را برای طبقه‌بندی استفاده کنید.
* **ج. شبکه‌های عصبی**:
	+ **چرا مناسب است؟**: روابط پیچیده و غیرخطی را مدل‌سازی می‌کند.
	+ **ساختار**:
		- لایه ورودی: برابر با تعداد ویژگی‌ها و احتمال عضویت در خوشه‌ها.
		- لایه‌های مخفی: 1-3 لایه با نورون‌های 64، 32، 16 و تابع ReLU.
		- لایه خروجی: 1 نورون با تابع Sigmoid.
	+ **تنظیمات**:
		- تابع هزینه: Binary Cross-Entropy.
		- بهینه‌ساز: Adam.
		- جلوگیری از بیش‌برازش: Dropout و Early Stopping.
	+ **کاربرد**: ویژگی‌ها و نتایج EM (احتمال عضویت در خوشه‌ها) را به عنوان ورودی استفاده کنید.
* **مدیریت داده‌های نامتوازن**:
	+ از SMOTE یا وزن‌دهی کلاس‌ها در مدل‌ها استفاده کنید.

**5. ارزیابی و تحلیل نتایج**

* **ارزیابی کاوش الگوهای پنهان**:
	+ **معیارها**:
		- Silhouette Score: کیفیت خوشه‌بندی.
		- BIC: تعداد بهینه خوشه‌ها در GMM با EM.
	+ **مصورسازی**: خوشه‌ها را با t-SNE یا UMAP نشان دهید.
	+ **تحلیل**: الگوهای پنهان (مثل بدافزارهای مبهم‌شده) را تفسیر کنید.
* **ارزیابی طبقه‌بندی**:
	+ **معیارها**:
		- دقت (Accuracy)
		- دقت جزئی (Precision)
		- یادآوری (Recall)
		- امتیاز F1
		- ROC-AUC
	+ **اعتبارسنجی متقاطع**: از k-fold برای پایداری مدل‌ها.
* **مقایسه مدل‌ها**:
	+ عملکرد Naive Bayes، رگرسیون لجستیک، SVM و شبکه‌های عصبی را مقایسه کنید.
	+ نقش EM در بهبود نتایج (از طریق خوشه‌بندی) را بررسی کنید.

**6.ارتباط با تهدیدات صنعتی**

* **تحلیل الگوها**:
	+ خوشه‌های کشف‌شده با EM را به تهدیدات ربط دهید:
		- خوشه با آنتروپی بالا: بدافزارهای مبهم‌شده برای نفوذ.
		- خوشه با توابع شبکه‌ای: تهدیدات علیه سیستم‌های صنعتی.
* **کاربرد**:
	+ توضیح دهید چگونه الگوهای پنهان و طبقه‌بندی دقیق:
		- تهدیدات را سریع شناسایی می‌کند.
		- خسارات به زیرساخت‌های صنعتی (مثل SCADA) را کاهش می‌دهد.

**7. آماده‌سازی برای سمینار**

* **ساختار ارائه**:
	1. **مقدمه**: اهمیت بدافزارها در صنعت، هدف پروژه.
	2. **روش‌شناسی**:
		+ پیش‌پردازش داده‌ها.
		+ کاوش الگوهای پنهان با GMM و الگوریتم EM.
		+ طبقه‌بندی با روش‌های آماری، SVM و شبکه‌های عصبی.
	3. **نتایج**:
		+ نمایش الگوهای پنهان (خوشه‌ها).
		+ مقایسه معیارهای مدل‌ها (Accuracy, F1-Score, ROC-AUC).
	4. **بحث و نتیجه‌گیری**:
		+ تفسیر الگوها و نقش EM.
		+ نقاط قوت: EM (کشف الگو)، SVM (دقت)، شبکه عصبی (پیچیدگی).
		+ پیشنهادات: داده بیشتر، روش‌های ترکیبی.
* **مصورسازی**:
	1. توزیع ویژگی‌ها.
	2. خوشه‌های EM با t-SNE.
	3. جدول مقایسه مدل‌ها.