# МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

# федеральное государственное автономное образовательное учреждение высшего образования

# «САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

# АЭРОКОСМИЧЕСКОГО ПРИБОРОСТРОЕНИЯ»

КАФЕДРА 33

ОТЧЕТ ЗАЩИЩЕН С ОЦЕНКОЙ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ПРЕПОДАВАТЕЛЬ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ассистент |  |  |  | Н.С. Красников |
| должность, уч. степень, звание |  | подпись, дата |  | инициалы, фамилия |

ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ № 4

ПРЕДОБРАБОТКА ДАННЫХ СИСТЕМЫ ЖУРНАЛОВ БЕЗОПАСНОСТИ

по курсу: ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ

|  |  |  |  |
| --- | --- | --- | --- |
| СТУДЕНТ ГР. № | 3031 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | М.В. Вдовин |
|  | номер группы | подпись, дата | инициалы, фамилия |

1. Цель работы:

Проведение анализа данных системных журналов безопасности операционной системы Windows 7/10/11. Предварительная подготовка данных для их интеллектуального анализа «цифрового портрета» пользователя.

1. Задание
2. Изучить теоретические материалы формирования log-журналов различных операционных систем, приведенные в данном лабораторном практикуме.
3. Провести предварительную подготовку и анализ собственных системных журналов безопасности.
4. Сделать выводы о "цифровом портрете" пользователя в точки зрения событий безопасности.
5. Создать датасет из растровых изображений-гистограмм цифровых портретов студентов группы (самостоятельный групповой проект)
6. Сформулировать (самостоятельно) задачу выявления аномалий цифрового профиля
7. Решить задачу бинарной классификации (выявления аномалий) в поведении пользователей, используя в качестве исходных данных изображения scatterгистограмм и/или их векторные представления.
8. Решить задачу многоклассовой классификация – Определить принадлежность цифрового портрета конкретному пользователю (студенту группы)
9. Рассчитать метрики качества классификации.
10. Отчет по лабораторной работе представить в виде "живого" скрипта с результатами анализа собственных журналов безопасности, а также в виде pdf документа данного скрипта.
11. Ход работы

В качестве источников данных используем системный журнал безопасности Windows 10. Для получения «сырой» выборки воспользуемся консолью управления (ММС) - Рисунок 1.

Объектом исследования является журнал безопасности - Рисунок 2. Выберем события за последние три дня и скопируем их в таблицу Excel - Рисунок 3.

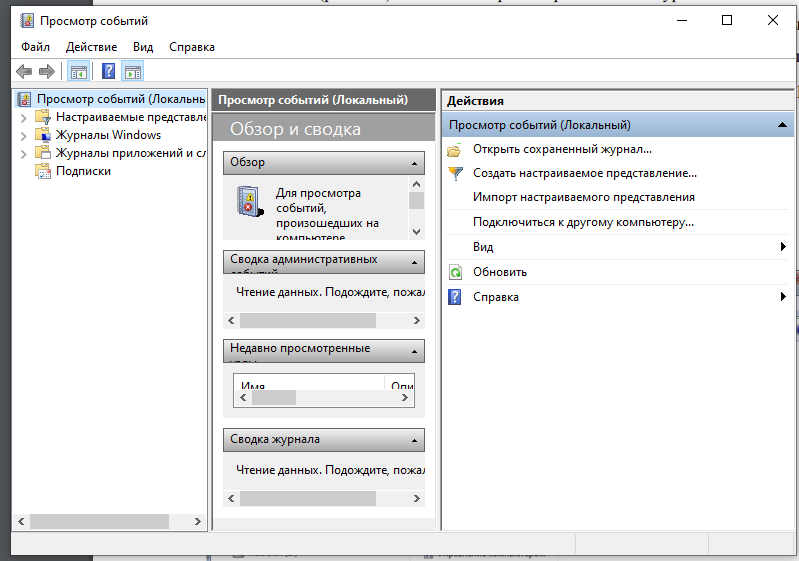


Рисунок 1. Доступ к журналу событий безопасности в Windows 10

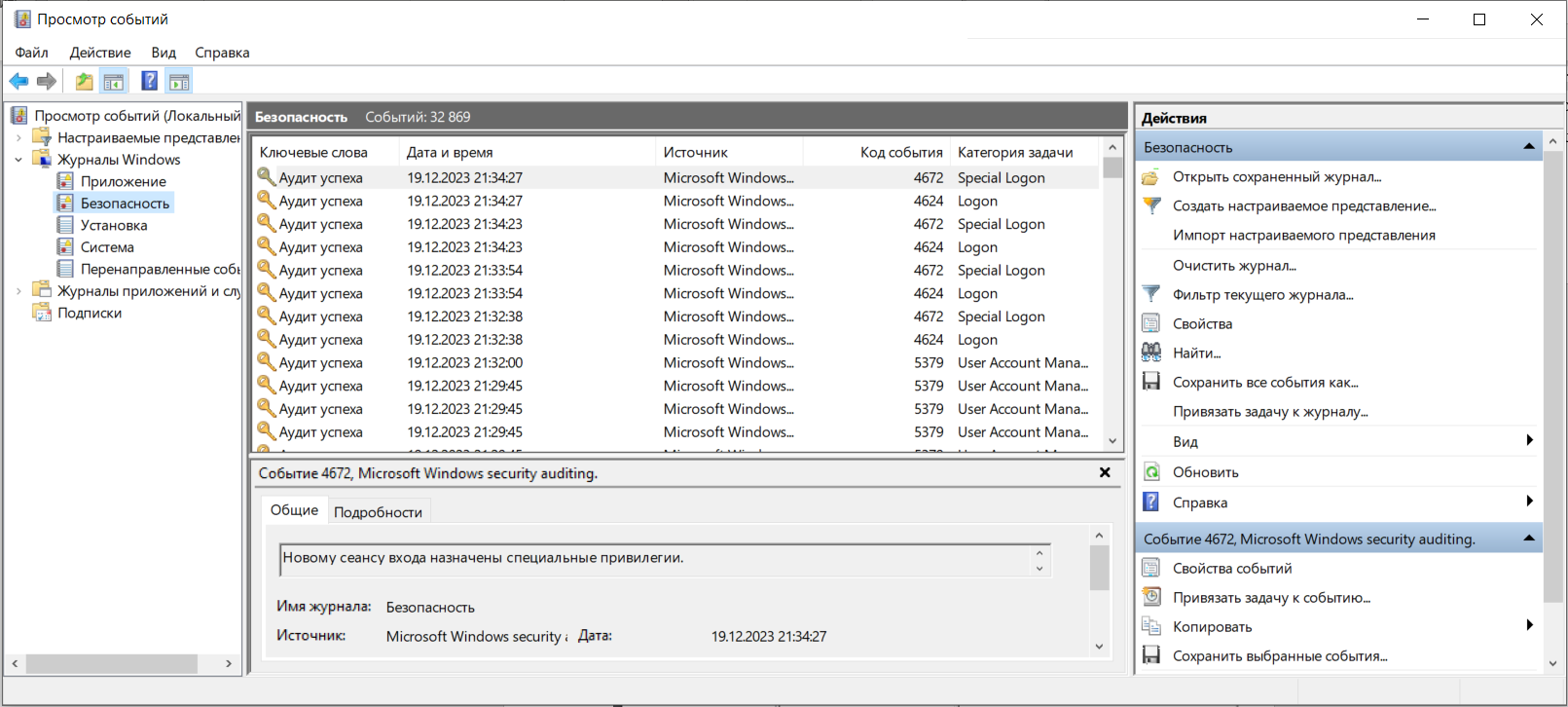


Рисунок 2. Журнал событий безопасности

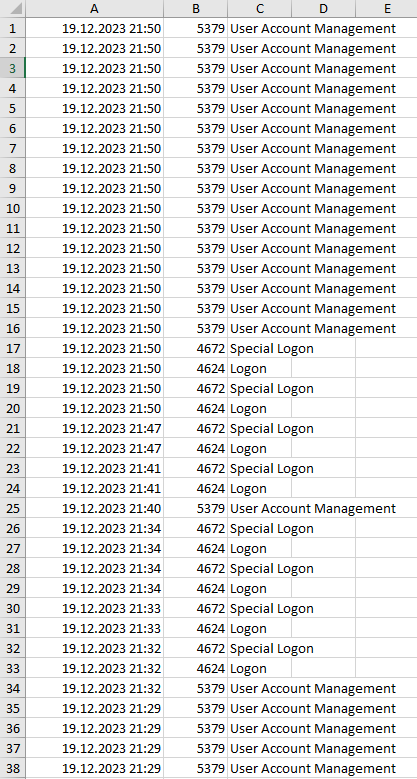


Рисунок 3. Таблица с событиями безопасности за 3 дня

Откроем среду Matlab и перенесем в рабочую зону получившуюся таблицу – Рисунок 4.

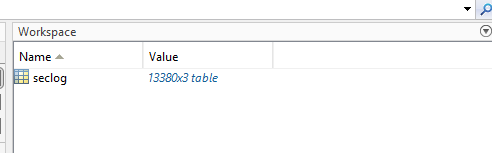


Рисунок 4. Перемещение таблицы с данными в рабочую зону MatLab

Поскольку требуется, чтобы значения столбца "Дата и время" мигрировали в состав первичного ключа таблицы, требуется выполнить код: T=readtable(“seclog.xlsx”)

При преобразовании таблицы русские наименования столбцов были подвергнуты принудительному преобразованию и потеряли свой логический смысл. Поэтому следует выполнить их переименование:

T.Properties.DemensionName(1, 1) = “datetime”;

T.Properties.DemensionName(1) = “EventCode”;

Аналогично переименуем второй столбец и проведем анализ табличной переменной T с помощью функции summary: T.Properties.VariableNames(2) = “EventCategory”;

Summary(T) – Рисунок 5.

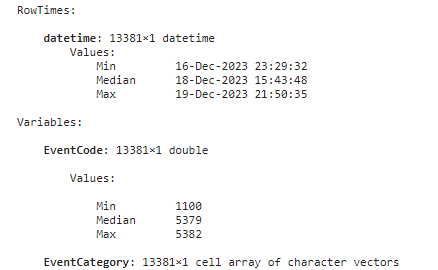


Рисунок 5. Сырая выборка

Можно сделать следующие выводы: в данную сырую выборку перенесено 13381 события безопасности с кодами событий от 1100 до 5382, а данные из временного интервала с 16.12.2023 по 19.12.2023.

Теперь требуется узнать, как распределены события безопасности в течении суток.

По этой информации можно делать выводы о поведении пользователя/ей данной вычислительной системы. Поведение пользователей формирует их "цифровой портрет", который в дальнейшем потребуется для интеллектуального анализа и. возможно, предсказания их действий. В частности, анализ событий системных журналов безопасности принимаются в суде в качестве доказательной базы. Итак, добавим в таблицу T новый столбец hr\_of\_day, который будет хранить, сколько событий безопасности зарегистрировано в течении конкретного часа суток – Рисунок 6.

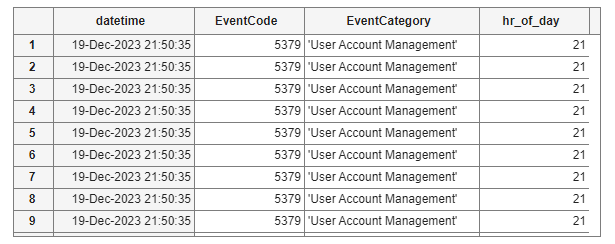


Рисунок 6. Количество событий, зарегистрированных в конкретный час суток

Чтобы произвести предварительный анализ данных построим scatter-гистограмму командной scatterhistogram(T, “hr\_of\_day”, “EventCode”) – Рисунок 7.

Из данный гистограммы видно:

1. Пиковыми часами являются часы с 6 утра по 12 ночи, когда выполняется наибольшее количество действий, связанный с безопасностью. А в период с 12 ночи до 6 утра наступает затишье.

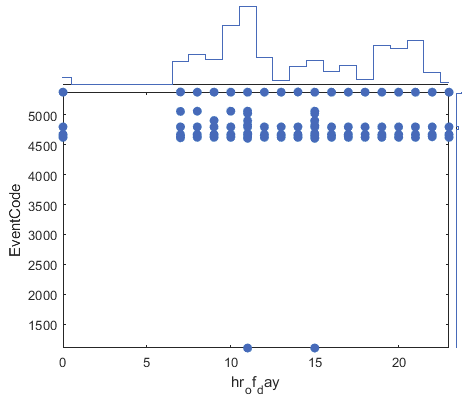


Рисунок 7. Scatter-гистрограмма

Чтобы уточнить характер (категорию) событий, происходящих в каждый час суток, модернизируем scatter-гистограмму указав ей необходимость группировать данные (параметр GroupVariable) по полю "EventCode": Scatterhistogram (T,"hr\_of\_day","EventCode",'GroupVariable',"EventCode") – Рисунок 8. Из данной гистограммы видно, что с 8 по 11 часов происходит самое широкое по спектру событий безопасности множество событий. Делать конкретные выводы, с чем это связано, пока что рано.

Зафиксируем полученные гистограммой срезы событий безопасности по конкретным часам суток в отдельной таблице Н – Рисунок 9. Важно, что для фиксирования "пустых" событий (когда в конкретные часы суток не происходило ничего) нужно использовать параметр "IncludeEmptyGroups" = true: H=groupsummary(T,["hr\_of\_day","EventCode"], 'IncludeEmptyGroups', true).

Чтобы зафиксировать ОБЩЕЕ количество событий безопасности в конкретные часы суток, создадим еще одну таблицу H1 – Рисунок 10, используя функцию groupsummary: H1=groupsummary(T,"hr\_of\_day").

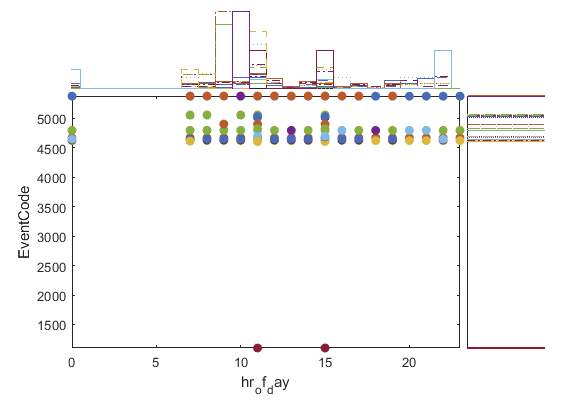


Рисунок 8. Доработанная Scatter-гистограмма

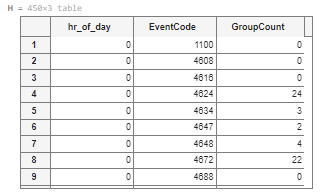


Рисунок 9. Фрагмент матрицы H

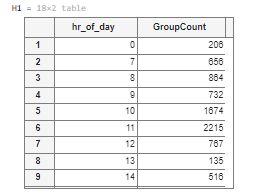


Рисунок 10. Фрагмент матрицы H1

Для отображения этой таблицы в виде гистограммы используем код (убрав %\*)

%\* bar(H1.hr\_of\_day, H1.GroupCount)

Поскольку в данном случае наиболее информативной будет "гистограмма с накоплением", для этого произведем ряд преобразований:

1. В качестве агрегирующей (накапливающей) переменной будем использовать вектор Х, в который будут группироваться события безопасности ("EventCode") и уже подсчитанное их количество "GroupCount" из таблицы Н. В новой таблице H2 – Рисунок 11 в результате выполнения функции groupsummary появится четыре столбца "hr\_of\_day", "GroupCount", fun1\_EventCode и fun1\_GroupCount, причем два последних появляются автоматически, как результат группировки: H2 = groupsummary(H, "hr\_of\_day", @(x) { x' }, ["EventCode", … "GroupCount"])

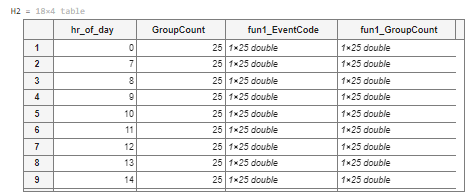


Рисунок 11. Фрагмент матрицы H2

Создадим промежуточные переменные hrs (вектор часов суток), counts (матрица количества событий безопасности для каждого кода события) и codes (вектор кодов событий безопасности):

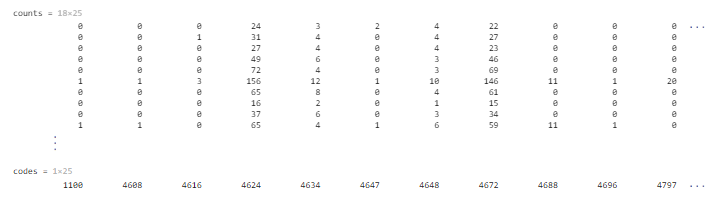


Рисунок 12. Вектор кодов событий безопасности

И теперь можем построить гистограмму с накоплением, а также сформировать легенду – Рисунок 13.

Построенная гистограмма с накоплением показывает, что в 11 часов дня происходит наибольшее количество событий безопасности, таких как:

5397 - когда пользователь выполняет операцию чтения учетных данных, сохраненных в диспетчере учетных данных Windows (WCM).

4799 – перечисление процессом локальных групп безопасности пользователя на компьютере или устройстве.

4688 – создание новых процессов.

4616 – изменение системного времени.

Данные события являются наиболее часто встречаемыми для данного пользователя. Наибольшая активность пользователя наблюдается в первой половине дня, с 6 утра до 12 дня, затем активность снижается и возрастает только в вечернее время с 19:00 по 22:00. В часы с 12:00 по 6:00 активность не наблюдается. Данная особенность пользователя может помочь выявить аномальное поведение работы под данным пользователем.

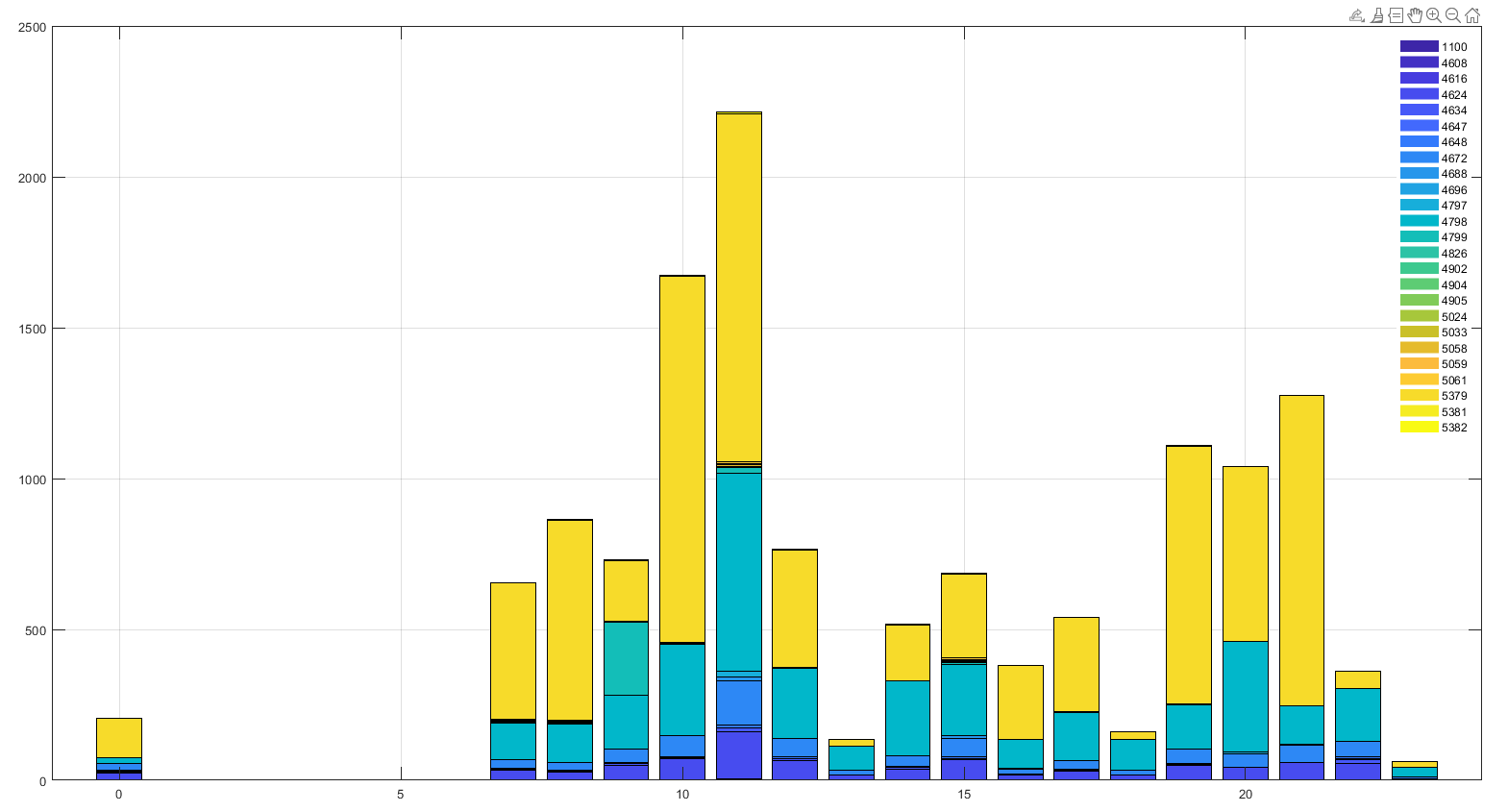


Рисунок 13. Цифровой проект пользователя

1. Вывод

В ходе лабораторной работы был осуществлен анализ данных системных журналов безопасности ОС Windows 10, который включал в себя изучение различных событий и записей, предоставляемых системой для отслеживания действий пользователя.

Листинг

T=readtimetable("seclog.xlsx")

T.Properties.DimensionNames(1,1) = "datetime";

T.Properties.VariableNames(1) = "EventCode";

T.Properties.VariableNames(2) = "EventCategory";

summary(T)

T.hr\_of\_day = hour(T.datetime)

scatterhistogram(T,"hr\_of\_day","EventCode" )

scatterhistogram(T,"hr\_of\_day","EventCode",'GroupVariable',"EventCode" )

H=groupsummary(T,["hr\_of\_day","EventCode"], 'IncludeEmptyGroups', true)

H1=groupsummary(T,"hr\_of\_day")

%bar(H1.hr\_of\_day, H1.GroupCount)

H2 = groupsummary(H, "hr\_of\_day", @(x) { x' }, ["EventCode", "GroupCount"])

hrs = H2.hr\_of\_day;

counts = cell2mat(H2.fun1\_GroupCount(:))

codes = H2.fun1\_EventCode{1}

b = bar(hrs, counts, 'stacked','FaceColor','flat');

for k = 1:size(counts,2)

b(k).CData = k;

end

grid on

legend(num2cell(string(codes)))