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

Санкт-Петербург

2023

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.

2. Задание

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

3. Ход работы

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

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

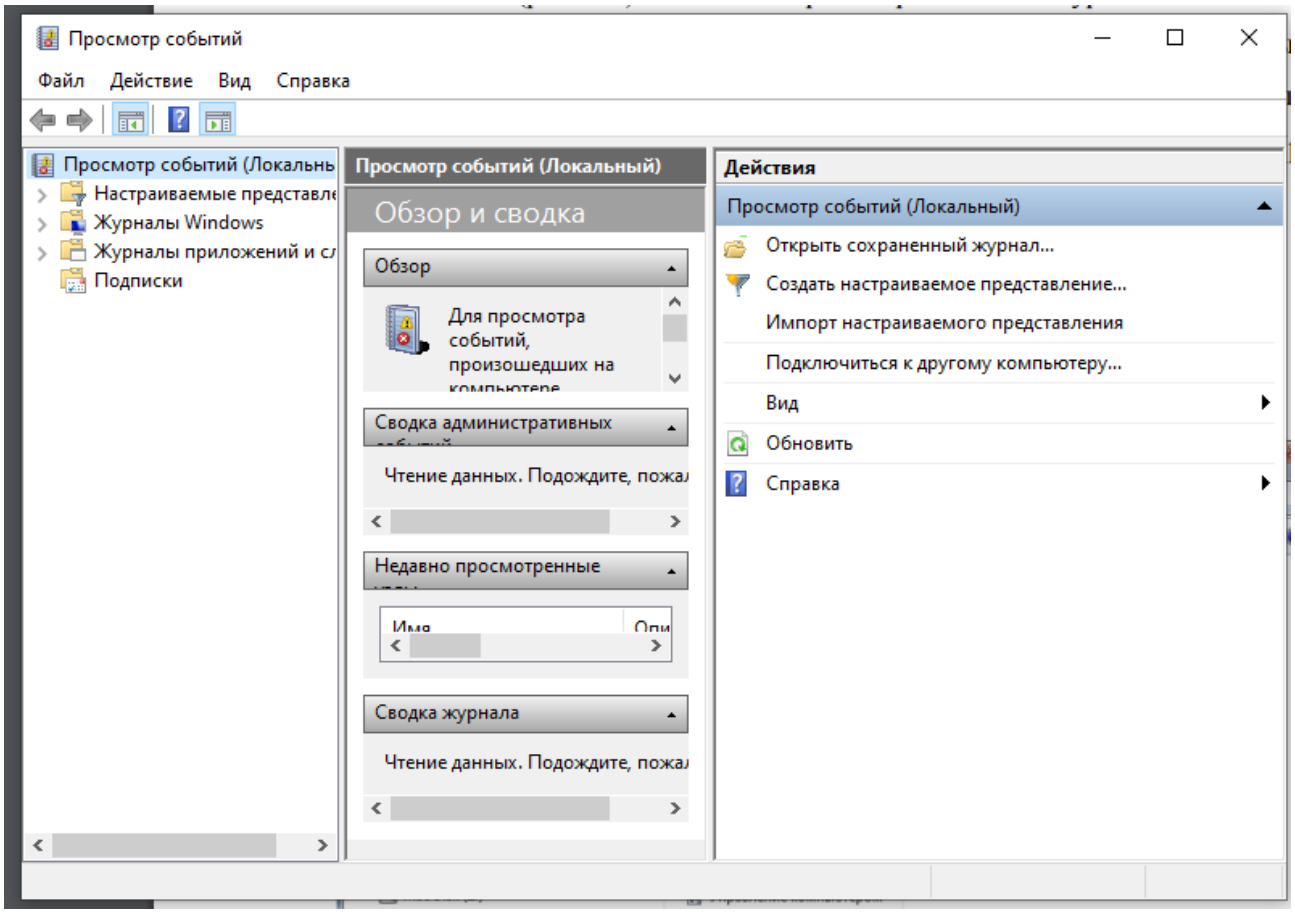

Рисунок 1. Доступ к журналу событий безопасности в Windows 10

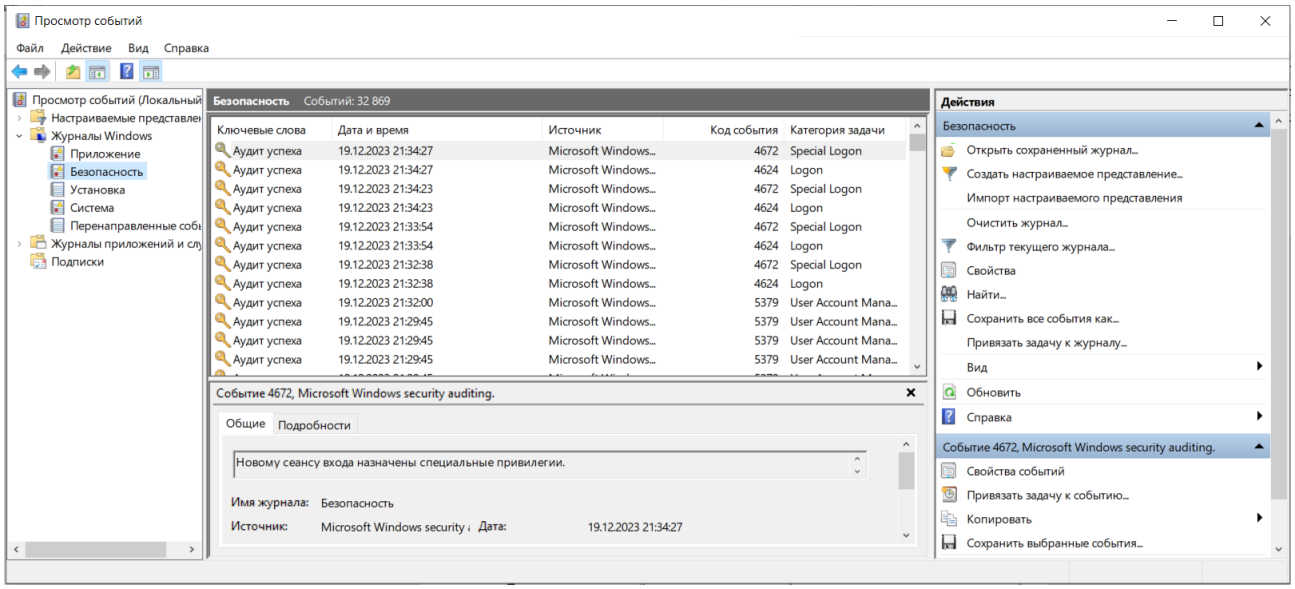

Рисунок 2. Журнал событий безопасности

	A	B	C	D	E
1	19.12.2023 21:50	5379	User Account Management		
2	19.12.2023 21:50	5379	User Account Management		
3	19.12.2023 21:50	5379	User Account Management		
4	19.12.2023 21:50	5379	User Account Management		
5	19.12.2023 21:50	5379	User Account Management		
6	19.12.2023 21:50	5379	User Account Management		
7	19.12.2023 21:50	5379	User Account Management		
8	19.12.2023 21:50	5379	User Account Management		
9	19.12.2023 21:50	5379	User Account Management		
10	19.12.2023 21:50	5379	User Account Management		
11	19.12.2023 21:50	5379	User Account Management		
12	19.12.2023 21:50	5379	User Account Management		
13	19.12.2023 21:50	5379	User Account Management		
14	19.12.2023 21:50	5379	User Account Management		
15	19.12.2023 21:50	5379	User Account Management		
16	19.12.2023 21:50	5379	User Account Management		
17	19.12.2023 21:50	4672	Special Logon		
18	19.12.2023 21:50	4624	Logon		
19	19.12.2023 21:50	4672	Special Logon		
20	19.12.2023 21:50	4624	Logon		
21	19.12.2023 21:47	4672	Special Logon		
22	19.12.2023 21:47	4624	Logon		
23	19.12.2023 21:41	4672	Special Logon		
24	19.12.2023 21:41	4624	Logon		
25	19.12.2023 21:40	5379	User Account Management		
26	19.12.2023 21:34	4672	Special Logon		
27	19.12.2023 21:34	4624	Logon		
28	19.12.2023 21:34	4672	Special Logon		
29	19.12.2023 21:34	4624	Logon		
30	19.12.2023 21:33	4672	Special Logon		
31	19.12.2023 21:33	4624	Logon		
32	19.12.2023 21:32	4672	Special Logon		
33	19.12.2023 21:32	4624	Logon		
34	19.12.2023 21:32	5379	User Account Management		
35	19.12.2023 21:29	5379	User Account Management		
36	19.12.2023 21:29	5379	User Account Management		
37	19.12.2023 21:29	5379	User Account Management		
38	19.12.2023 21:29	5379	User Account Management		

Рисунок 3. Таблица с событиями безопасности за 3 дня

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

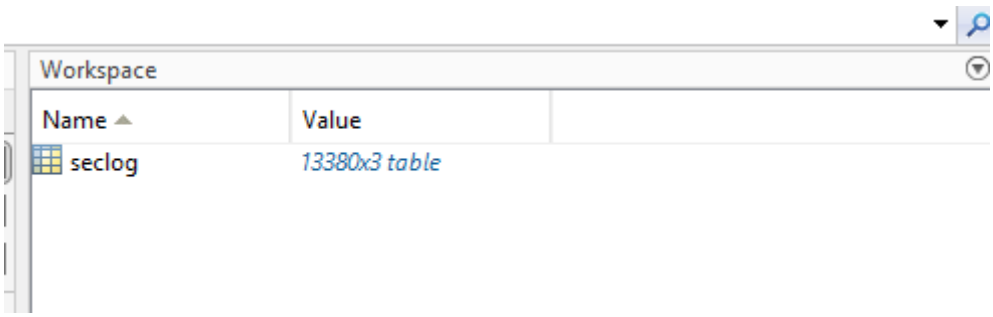

Рисунок 4. Перемещение таблицы с данными в рабочую зону MatLab

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.

```
RowTimes:

    datetime: 13381×1 datetime
        Values:
            Min         16-Dec-2023 23:29:32
            Median      18-Dec-2023 15:43:48
            Max         19-Dec-2023 21:50:35

Variables:

    EventCode: 13381×1 double

        Values:

            Min         1100
            Median      5379
            Max         5382

    EventCategory: 13381×1 cell array of character vectors
```

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.

	datetime	EventCode	EventCategory	hr_of_day
1	19-Dec-2023 21:50:35	5379	'User Account Management'	21
2	19-Dec-2023 21:50:35	5379	'User Account Management'	21
3	19-Dec-2023 21:50:35	5379	'User Account Management'	21
4	19-Dec-2023 21:50:35	5379	'User Account Management'	21
5	19-Dec-2023 21:50:35	5379	'User Account Management'	21
6	19-Dec-2023 21:50:35	5379	'User Account Management'	21
7	19-Dec-2023 21:50:35	5379	'User Account Management'	21
8	19-Dec-2023 21:50:35	5379	'User Account Management'	21
9	19-Dec-2023 21:50:35	5379	'User Account Management'	21

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.

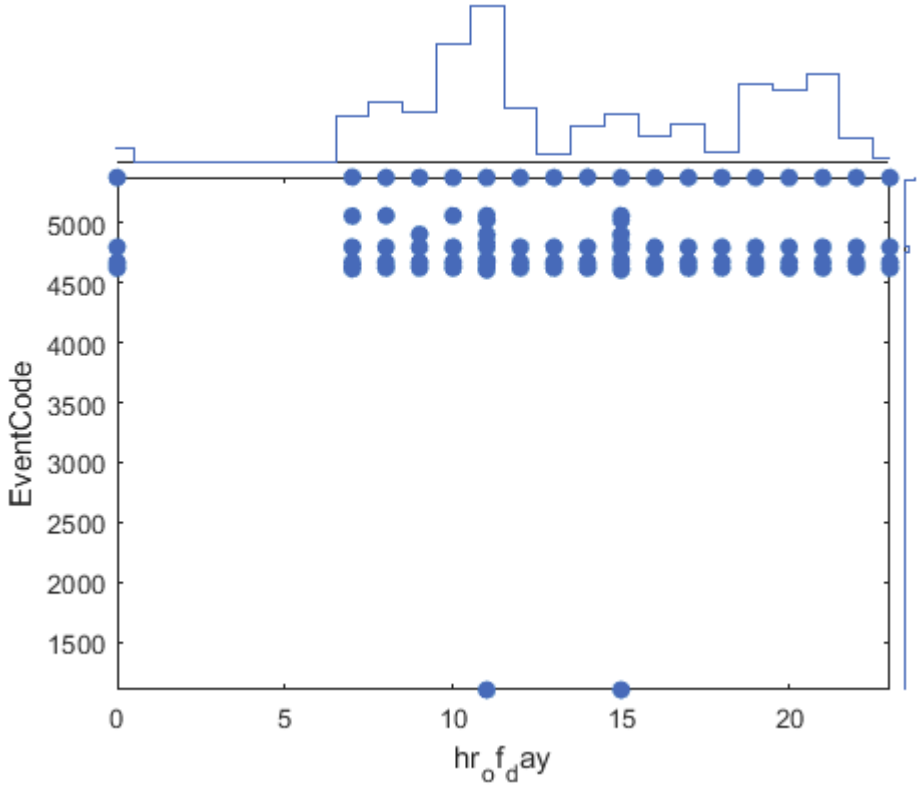

Рисунок 7. Scatter-гистрограмма

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 H – 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").

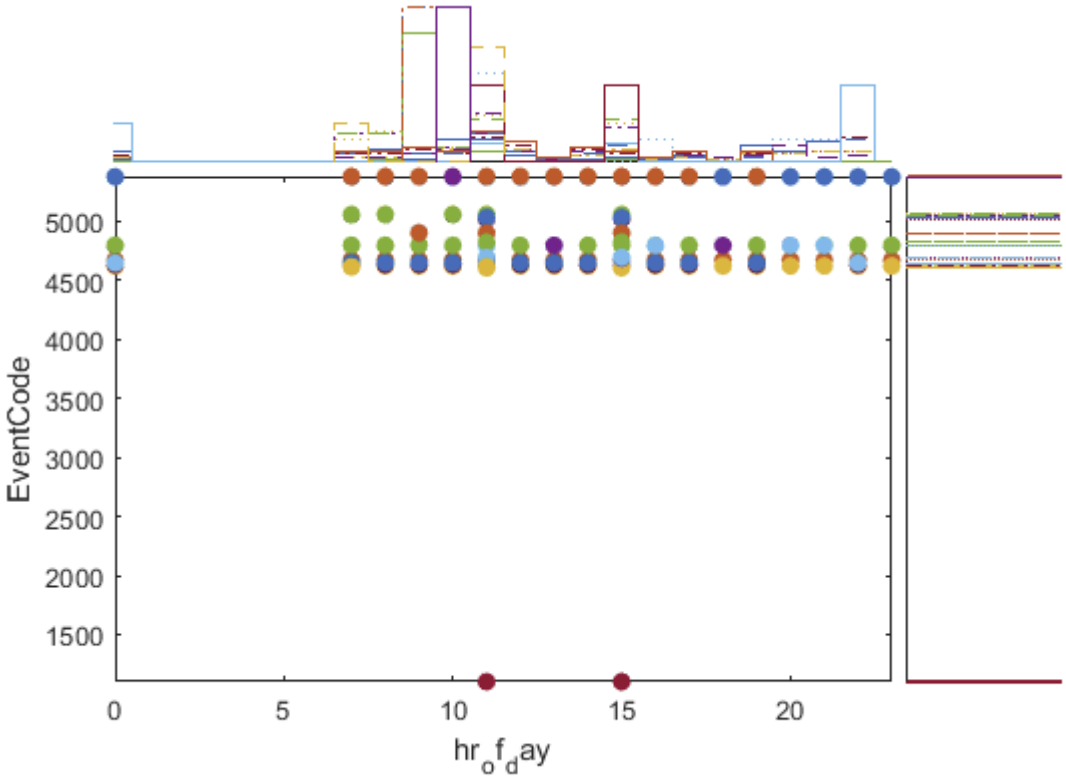

Рисунок 8. Доработанная Scatter-гистограмма

H = 450×3 table

	hr_of_day	EventCode	GroupCount
1	0	1100	0
2	0	4608	0
3	0	4616	0
4	0	4624	24
5	0	4634	3
6	0	4647	2
7	0	4648	4
8	0	4672	22
9	0	4688	0

Рисунок 9. Фрагмент матрицы H

H1 = 18×2 table

	hr_of_day	GroupCount
1	0	206
2	7	656
3	8	864
4	9	732
5	10	1674
6	11	2215
7	12	767
8	13	135
9	14	516

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 X, в который будут группироваться события безопасности ("EventCode") и уже подсчитанное их количество "GroupCount" из таблицы H. 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"])

H2 = 18×4 table

	hr_of_day	GroupCount	fun1_EventCode	fun1_GroupCount
1	0	25	1×25 double	1×25 double
2	7	25	1×25 double	1×25 double
3	8	25	1×25 double	1×25 double
4	9	25	1×25 double	1×25 double
5	10	25	1×25 double	1×25 double
6	11	25	1×25 double	1×25 double
7	12	25	1×25 double	1×25 double
8	13	25	1×25 double	1×25 double
9	14	25	1×25 double	1×25 double

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):

counts = 18×25

0	0	0	24	3	2	4	22	0	0	0 ...
0	0	1	31	4	0	4	27	0	0	0
0	0	0	27	4	0	4	23	0	0	0
0	0	0	49	6	0	3	46	0	0	0
0	0	0	72	4	0	3	69	0	0	0
1	1	3	156	12	1	10	146	11	1	20
0	0	0	65	8	0	4	61	0	0	0
0	0	0	16	2	0	1	15	0	0	0
0	0	0	37	6	0	3	34	0	0	0
1	1	0	65	4	1	6	59	11	1	0
⋮										

codes = 1×25

1100	4608	4616	4624	4634	4647	4648	4672	4688	4696	4797 ...

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.

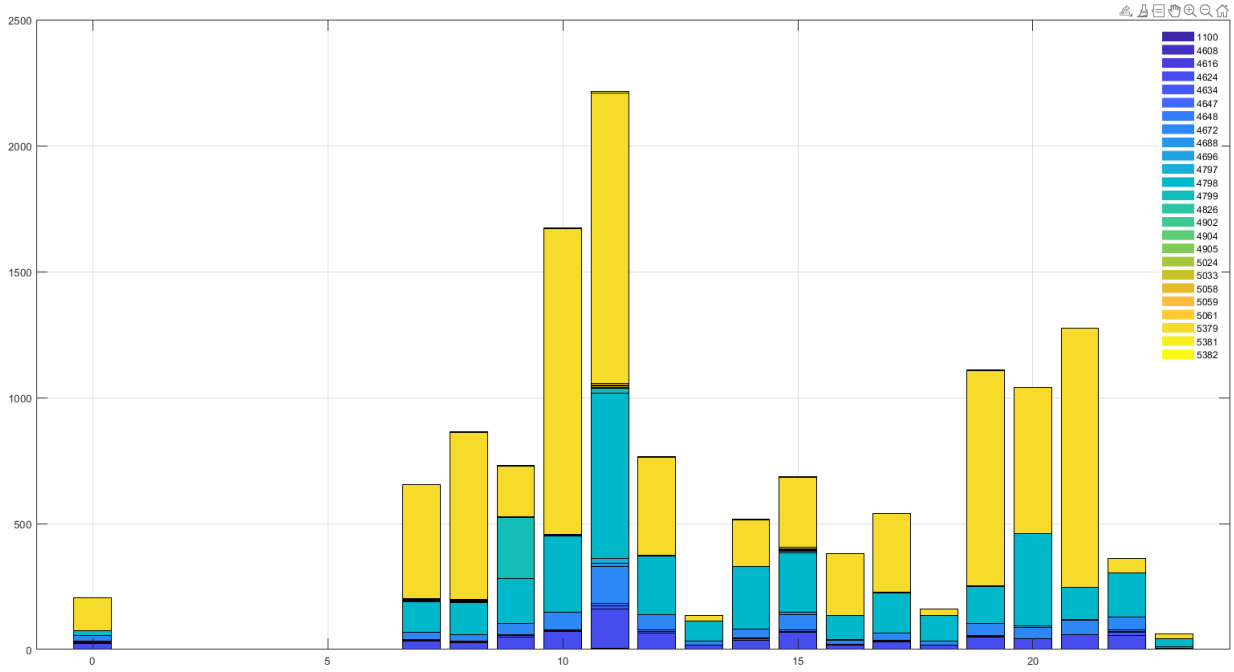

Рисунок 13. Цифровой проект пользователя

4. 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)))
```