

Приложение Д
(Справочное)
Руководство программиста

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ
ОПРЕДЕЛЕНИЯ АВТОРСТВА ИСХОДНОГО КОДА ПРОГРАММ НА ЯЗЫКЕ C/C++
«WHOSECPPCODE»

Руководство программиста

Листов 16

Аннотация

Требования настоящего документа применяются при:

- предварительных испытаниях системы;
- опытной эксплуатации;
- приемочных испытаниях;
- эксплуатации на предприятиях.

Настоящий документ содержит назначение, условия применения, характерные особенности и описание возможностей программного обеспечения «WhoseCppCode». Предъявлены требования к квалификации программиста. Указана последовательность действий программиста, обеспечивающих установку, настройку, загрузку, запуск, выполнение и завершение программы, приведено описание входных и выходных данных, а также сообщения, выводимые программой, и соответствующие им действия.

Содержание

1 Назначение программы.....	4
2 Уровень подготовки программиста.....	5
3 Условия применения программы.....	6
4 Характеристика программы.....	7
5 Инструкция по установке и настройке программной среды.....	8
6 Обращение к программе.....	10
6.1 Сбор и обработка данных с ресурса GitHub.....	10
6.2 Вычисление матрицы объектов-признаков.....	11
6.3 Классификация, входные данные и вывод программы.....	12
7 Сообщения.....	19

1 Назначение программы

Программное обеспечение (ПО) «WhoseCppClassCode» предназначено для определения авторства программ на языке C/C++ по исходному коду и может быть использовано организациями, занимающимися решением вопросов информационной безопасности, лицензирования ПО, интеллектуальной собственности и расследования инцидентов, связанных с применением вредоносного ПО. Программа состоит из программного модуля, реализующего заявленный функционал, и интерфейса, предназначенного для визуализации ввода и вывода данных, а также удобной работы с возможностями основного модуля. При этом интерфейс не является обязательным, программа может быть использована в качестве модуля при разработке иных автоматизированных систем.

ПО «WhoseCppClassCode» предоставляет следующие возможности:

- обработка файлов исходного кода на языке C/C++;
- сбор данных с ресурса GitHub;
- построение модели классификации авторов программного обеспечения;
- формирование отчетности в форматах *.json и *.csv;
- визуализация результатов классификации.

Работа с ПО «WhoseCppClassCode» доступна всем пользователям с доступом к предварительно установленной и настроенной рабочей программной среде, реализованной на ПЭВМ, специально предназначенном сервере или с помощью средств виртуализации.

2 Уровень подготовки программиста

Программист, использующий ПО «WhoseCppClassCode», должен иметь опыт работы с ОС Linux, базовые навыки программирования, а также обладать следующими знаниями:

- знать соответствующую предметную область;
- понимать основы машинного обучения, построения и оценки моделей классификации.

Квалификация программиста должна позволять осуществлять установку и настройку программной среды для работы системы, а также сбор и анализ данных.

3 Условия применения программы

Для работы с ПО «WhoseCppClassCode» необходимо следующее программное обеспечение:

- ОС Linux (тестирование программы производилось на ОС Ubuntu 17.04) с доступом к глобальной сети Интернет;
- программная среда с установленными зависимостями (библиотеками);
- веб-обозреватель.

Инструкция по установке и настройке программной среды приведена в разделе 5 настоящего документа.

Требования к характеристикам автоматизированного рабочего места:

- 2,00 Гб ОЗУ или выше;
- CD-ROM;
- 1 Гб на жестком диске.

4 Характеристика программы

ПО «WhoseCppCode» состоит из двух основных частей:

- программного модуля, реализующего все необходимые функции для сбора, анализа и обработки данных, а также построения модели классификации авторов исходного кода
- программного интерфейса на основе технологии Jupyter Notebook, предназначенного для визуализации полученных в ходе классификации результатов, сбора необходимых данных с ресурса GitHub, построения матрицы объектов-признаков на основе входных данных, проведения вычислительных экспериментов.

Интерфейс основан на веб-технологиях, может использоваться для демонстрации возможностей программ на языке Python. ПО «WhoseCppCode» использует ряд сторонних библиотек, реализующих алгоритмы классификации, используемых в процессе классификации авторов исходного кода, обеспечивающих работу интерфейса, а также визуализацию полученных результатов. Подробное описание особенностей программы приведено в пояснительной записке к ПО «WhoseCppCode».

Основной модуль программы «WhoseCppCode» может быть использован отдельно от Jupyter Notebook при разработке различного рода программ, систем и интерфейсов лицами, заинтересованными в задаче классификации программистов.

Время выполнения программы зависит от вычислительных мощностей автоматизированного рабочего места, количества данных, подвергаемых обработке, а также количества циклов работы классификатора (число итераций вычислительных экспериментов), задаваемых пользователем.

5 Инструкция по установке и настройке программной среды

Ниже приведена последовательность терминальных команд по установке и настройке ПО «WhoseCppCode» в ОС Ubuntu:

1) Перейти в директорию проекта:

```
$ cd whose_cpp_code
```

2) Установить Python версии 3.6:

```
$ sudo apt-get install python3.6
```

3) Создать виртуальную оболочку, в которую будут установлены необходимые зависимости проекта:

```
$ virtualenv --python=python3.6
```

4) Активировать виртуальную оболочку:

```
$ source whose_cpp_code/bin/activate
```

5) Установить необходимые зависимости:

```
$ pip install -r requirements.txt
```

6) Запустить локальный сервер Jupyter Notebook:

```
$ jupyter notebook
```

После запуска локального сервера в веб-обозревателе автоматически открывается страница проекта. Необходимо запустить файл `user_notebook.ipynb` (рис. 5.1).

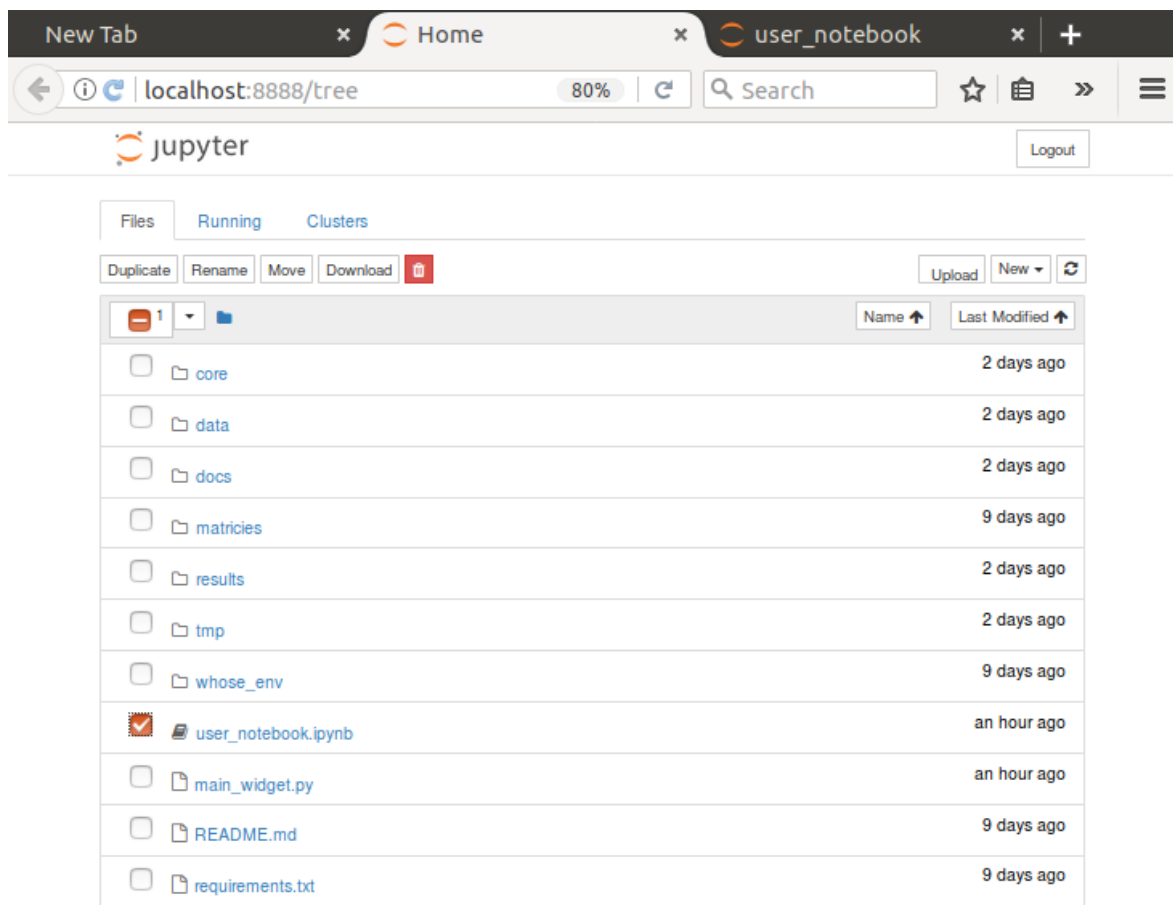


Рисунок 5.1 — Главное окно Jupyter Notebook

6 Обращение к программе

Для начала работы с программой необходимо открыть в веб-обозревателе адрес сервера JupyterNotebook, перейти в меню «Cell» и запустить выполнение ячеек командой «Run All», после чего дождаться загрузки всех ячеек (рис. 6.1).

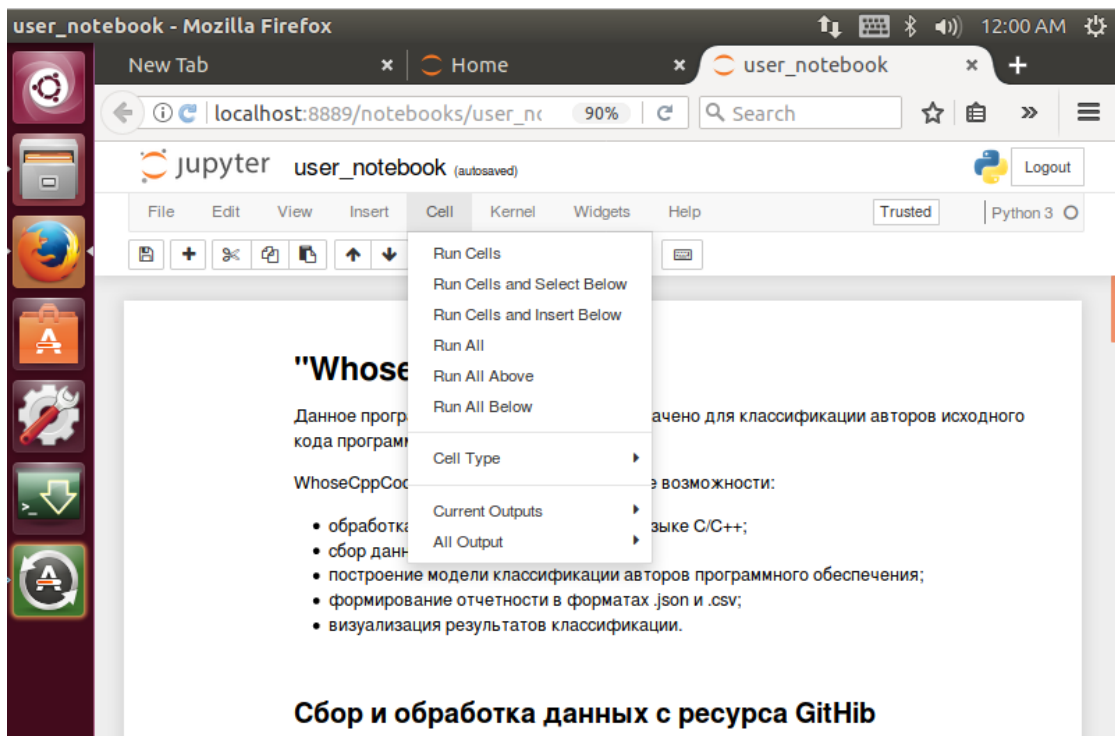


Рисунок 6.1 — Запуск программы

6.1 Сбор и обработка данных с ресурса GitHub

Форма, приведенная на рисунке 6.2, предназначена для сбора файлов исходного кода на языке C/C++ с веб-хостинга GitHub. При этом необходимо иметь аутентификационные данные, соответственно, пользователь должен быть зарегистрирован на сайте www.github.com. Помимо логина и пароля в форму через запятую вводится список пользователей, чьи файлы будут загружены и обработаны.

В процессе работы данной формы файлы указанных пользователей загружаются в корень проекта в директорию «data», все файлы конвертируются в кодировку UTF-8, удаляются пустые. Подобная обработка необходима для корректной дальнейшей работы с загруженными файлами.

▼ **Сбор и обработка данных с ресурса GitHub**

Введите список пользователей через запятую.

Удаление пустых файлов и изменение кодировки.

```
In [1]: from scrap_widget import display_scrapping_form
display_scrapping_form()
```

✕

Логин:	<input type="text" value="IvanovIvan"/>
Пароль:	<input type="text" value="duh73@!"/>
Найти:	<input type="text" value="paroj, Soyen, sipa, gavinandresen, theuni, luke-jr, ddunbar"/>

✓ Получить данные

Идет сбор файлов, пожалуйста, подождите...

Готово. Данные расположены в корне проекта в папке data.

Если данные не были загружены, проверьте правильность ввода логина, пароля, а также имен пользователей.

Рисунок 6.2 — Пример работы с формой «Сбор и обработка данных»

6.2 Вычисление матрицы объектов-признаков

На рисунке 6.3 приведен пример работы с формой, предназначенной для формирования на основе загруженных данных матрицы-объектов признаков, подаваемой в последствии на вход алгоритму классификации. В данном случае под объектами подразумеваются авторы программ на языке C/C++, каждому из которых соответствует вектор стилистических признаков, вычисленных на основе представленных файлов исходного кода.

Матрица объектов-признаков вычисляется отдельно во избежание повторения вычислений и для экономии времени, затрачиваемого на работу программы, поскольку процесс классификации предполагает несколько циклов вычислительных экспериментов, включающих обучение и тестирование классификатора.

▼ **Матрица объектов-признаков**

Формирование матрицы объектов-признаков для дальнейшей классификации. Объектами в данном случае являются авторы исходного кода, признаками - вычисленный для каждого автора на основе набора файлов исходного кода вектор значений признаков, характеризующих индивидуальный стиль разработчика.

Матрица объектов-признаков вычисляется отдельно во избежание повторения вычислений, а также сокращения времени, затрачиваемого на обучение и тестирование классификатора.

```
In [2]: from sample_matrix_widget import display_matrix_widget

# Путь к данным
path = './data/'
outpath = './data/matrices/'

display_matrix_widget(path, outpath)
```

✕

✓ Получить матрицу

Готово.

Рисунок 6.3 — Пример работы с формой «Матрица объектов-признаков»

6.3 Классификация, входные данные и вывод программы

Основной функцией пользовательского интерфейса ПО «WhoseCppCode» является демонстрация процесса классификации авторов исходного кода программ на языке C/C++ (рис. 6.4) на заранее сформированных наборах данных:

- «students» — лабораторные работы студентов ТУСУР кафедры КИБЭВС 1-го курса обучения по дисциплине «Основы программирования»;
- «Google Code Jam 2016» — работы участников ежегодной олимпиады по программированию от компании Google;
- «GitHub» — данные с веб-хостинга www.github.com.

Существует возможность классификации пользовательского набора данных. Для этого необходимо сформировать и обработать данные при помощи форм, описанных в разделах 6.1 и 6.2 настоящего документа, после чего в выпадающем меню «Данные» выбрать опцию «user_data» (рис. 6.5). При этом на вход классификатору будет подаваться сформированная пользователем матрица объектов-признаков, расположенная в корне проекта в директории data/matrices.

Все наборы данных, а также список GitHub-пользователей, чьи программы подвергались классификации, прилагаются к ПО «WhoseCppCode».

▼ Классификация

Построение модели классификации, ее обучение на выбранном наборе данных, визуализация результатов.

Отчеты по результатам работы программы в форматах .json и .csv располагаются в директории results в корне проекта.

Ввод:

- **Циклов** - количество итераций эксперимента
- **Данные** - данные для классификации:
 - students - лабораторные работы студентов каф. КИБЭВС по дисциплине "Основы программирования"
 - Google Code Jam 2016 - работы участников ежегодной олимпиады по программированию от компании Google
 - GitHub - данные с веб-хостинга GitHub
- **Алгоритм** - алгоритм классификации

```
In [3]: from main_widget import display_main_form  
display_main_form()
```

× Циклов:

Данные:

Алгоритм:

✓ Классифицировать

Рисунок 6.4 — Пример работы с формой «Классификация»

В форме «Классификация» существует возможность выбора числа циклов (количество итераций вычислительных экспериментов), набора данных, на которых будет производиться процесс классификации (рис. 6.5), алгоритма классификации (рис. 6.6).

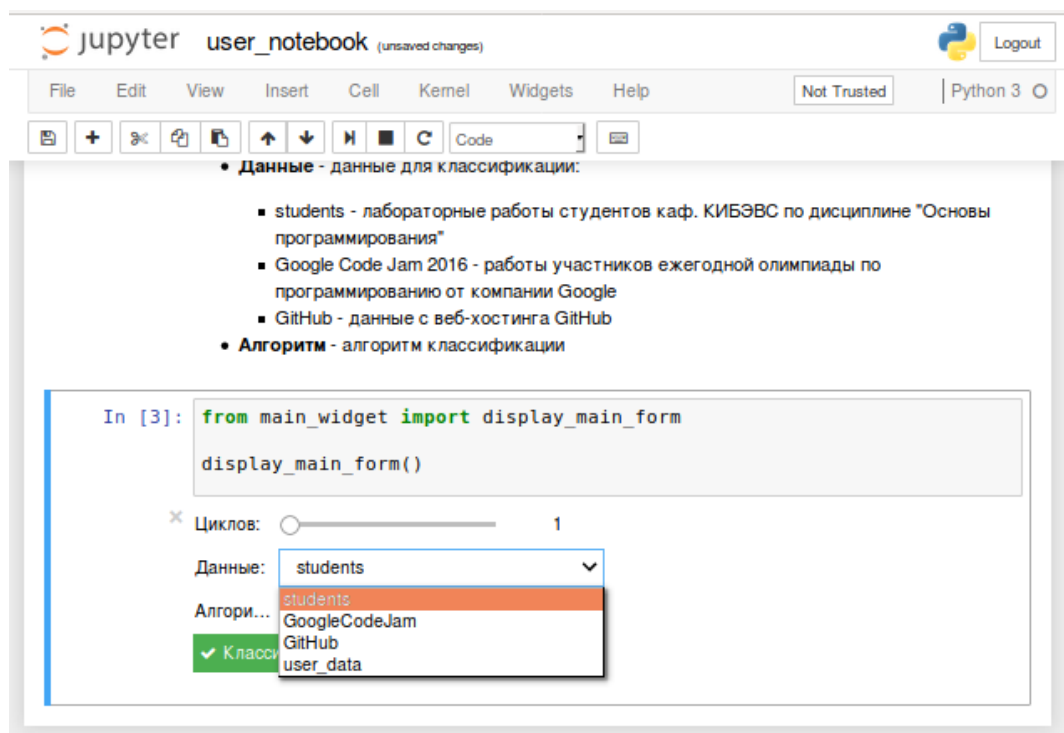


Рисунок 6.5 — Выбор набора данных для дальнейшей классификации

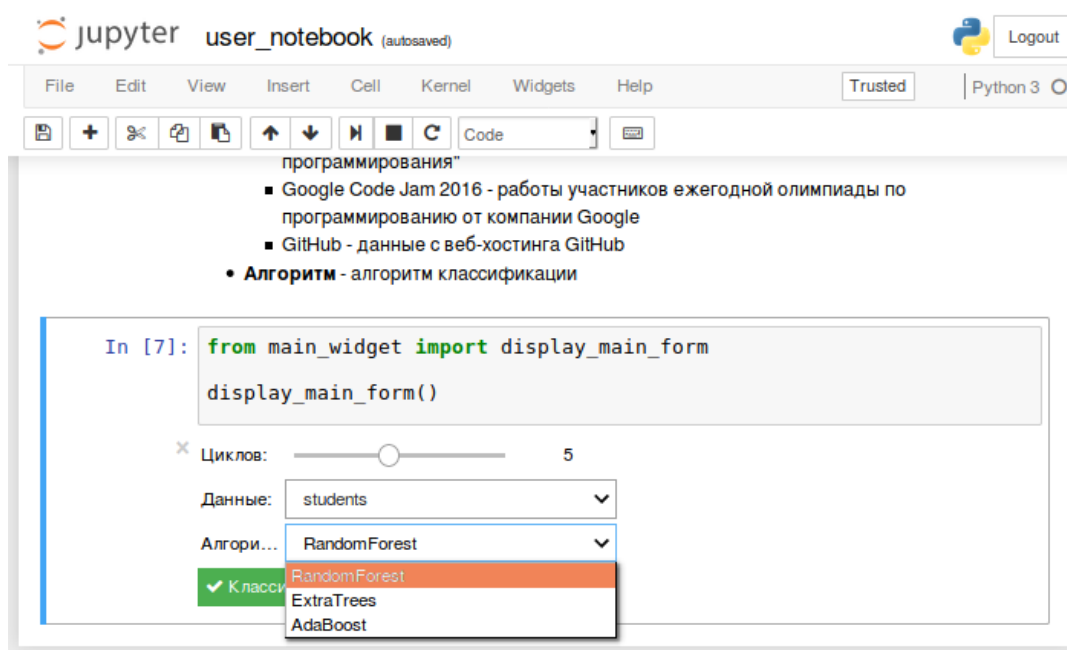
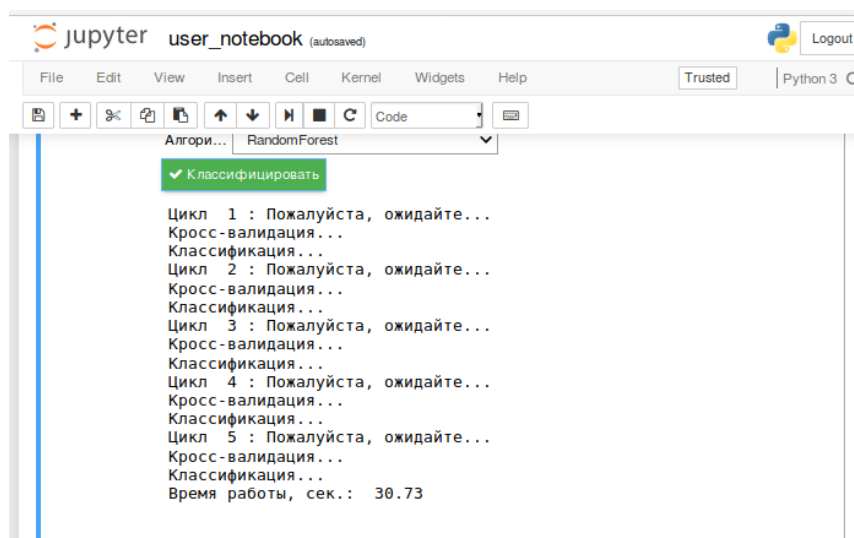


Рисунок 6.6 — Выбор алгоритма классификации

На рисунках 6.7, 6.8 и 6.9 представлен вывод результатов работы программы. Полученные диаграммы можно экспортировать в формате *.png с помощью команды «Download plot as png», как показано на рисунке 6.10.



```
Алгоритм: RandomForest
✓ Классифицировать
Цикл 1 : Пожалуйста, ожидайте...
Кросс-валидация...
Классификация...
Цикл 2 : Пожалуйста, ожидайте...
Кросс-валидация...
Классификация...
Цикл 3 : Пожалуйста, ожидайте...
Кросс-валидация...
Классификация...
Цикл 4 : Пожалуйста, ожидайте...
Кросс-валидация...
Классификация...
Цикл 5 : Пожалуйста, ожидайте...
Кросс-валидация...
Классификация...
Время работы, сек.: 30.73
```

Рисунок 6.7 — Вывод программы в процессе классификации

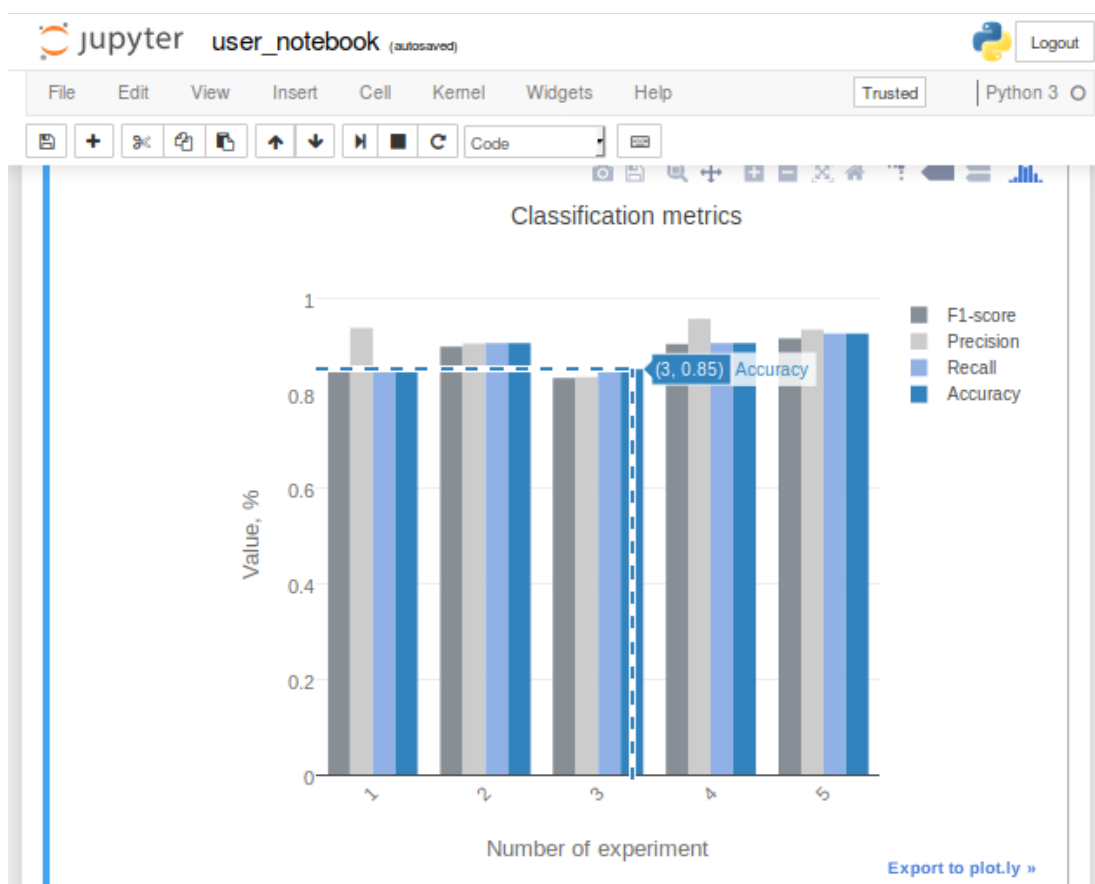


Рисунок 6.8 — Результаты работы программы: диаграмма значений метрик классификации

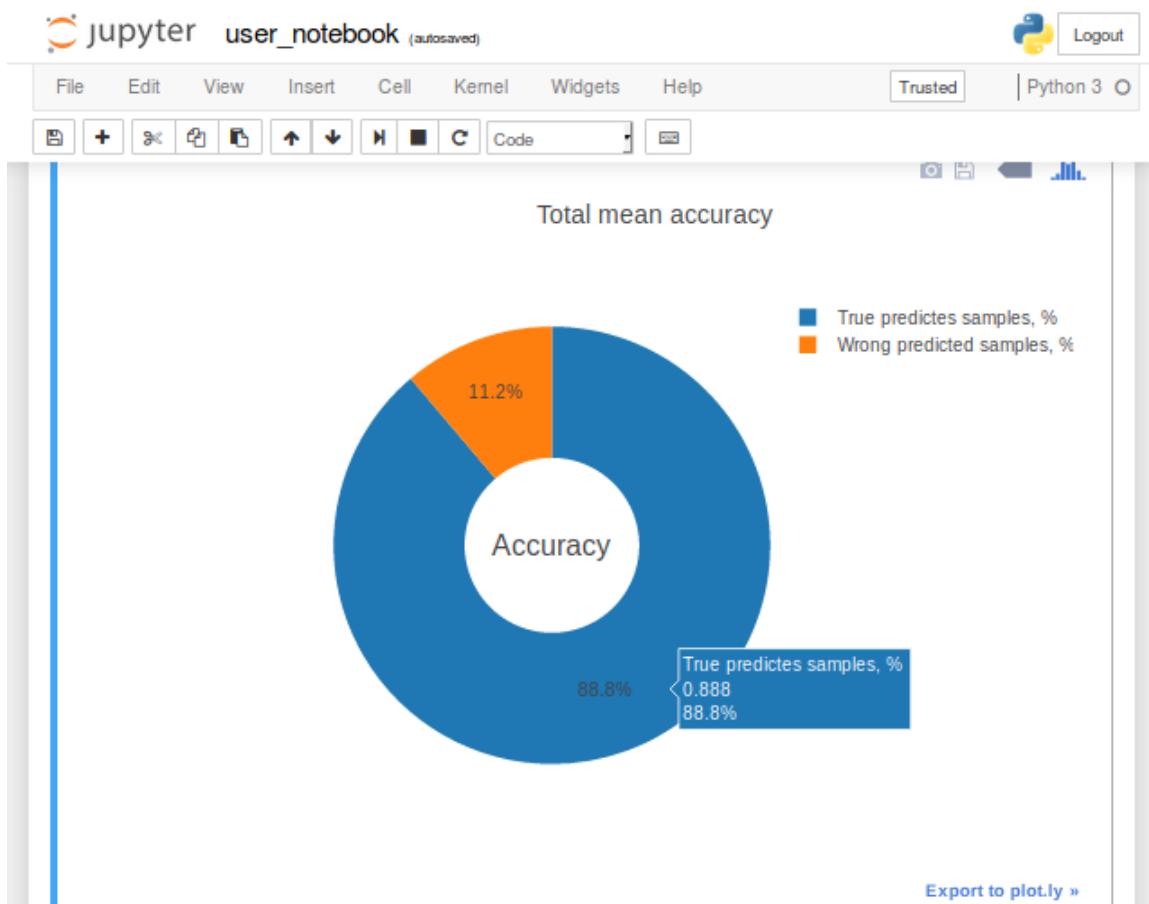


Рисунок 6.9 — Результаты работы программы: диаграмма значения средней точности классификации

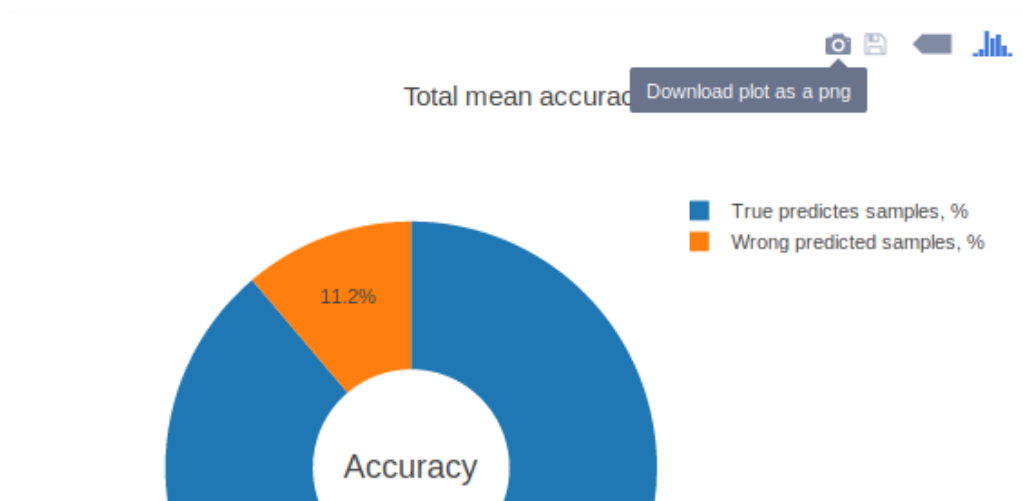


Рисунок 6.10 — Сохранение диаграммы в формате *.png

Выводимые сообщения и соответствующие им действия:

- «Пожалуйста, введите данные для аутентификации на сайте www.github.com». Заполнены не все поля формы для сбора данных, необходимо ввести логин и пароль GitHub-пользователя.
- «Идет сбор файлов, пожалуйста, подождите...». Сообщение о запуске процесса сбора и обработки данных с веб-хостинга www.github.com.
- «Готово. Данные расположены в корне проекта в папке data. Если данные не были загружены, проверьте правильность ввода логина, пароля, а также имен пользователей.». Завершение сбора и обработки данных с веб-хостинга www.github.com. При отсутствии результатов работы программы в указанных директориях, следует проверить корректность ввода логина и пароля для аутентификации на сайте www.github.com, а также имен GitHub-пользователей, чьи файлы будут подвержены сбору и обработке.
- «Готово». Завершение процесса преобразования данных в матрицу объектов-признаков.
- «Цикл 1: Пожалуйста, ожидайте... Кросс-валидация... Классификация...». Вывод во время процесса классификации для контроля за выполнением программы.
- «Время работы, сек.: 9.42». Классификация успешно завершена, вывод суммарного времени работы в секундах.

Для повторного запуска выполнения любой из форм программ, необходимо нажать на нужную ячейку щелчком мыши, после чего — комбинацию клавиш «Ctrl + Enter» и продолжить работу с формой. При сбое в работе программы необходимо перезапустить все ячейки, перейдя в меню «Cell Run — All», дождаться загрузки ячеек.