Case study - recruitment for Junior Data Scientist position

In the document below you will find three tasks - please complete all of them.

Task 1

Your customer is importing flowers (irises) from China. He always buys the same three species: Virginica, Versicolor oraz Setosa. Unfortunately, in the last order all the flowers got mixed up and nobody knows how to distinguish them (what is surprising, even the florist experts were not able to identify them). Luckily, there are still 150 flowers left from the previous order which are tagged properly. The client instructed his workers to measure the length and width of their sepals and petals so they could help to classify the newest batch correctly.

**Goal**: Based on attached data build a model which will classify flower species from the last order. Prepare a report where you describe your way of approaching the problem and the steps you took to solve it. Don’t forget to assess the quality of the model you have prepared.

Don’t focus on getting the highest score possible - the more important thing is to show your line of thought.

*If you have any doubts or something seems inconclusive based on the task description, write it down in the report in the way you would ask the customer to clarify.*

**What you need to provide for this task:**

1. The report (pdf/html/markdown) which includes the description of the model used.
2. The source code written in Python or R (published on one of the git platforms such as github or bitbucket).

Task 2

Enclosed dataset contains data about press headlines – their content and type. Aim of this task is to create binary classification model for headline type (sarcastic / not sarcastic) based on headline content.

**Goal**: Based on attached data build a model that will classify headline types. Prepare a report where you describe your way of approaching the problem and the steps you took to solve it. Don’t forget to assess the quality of the model you have prepared.

Don’t focus on getting the highest score possible - the more important thing is to show your line of thought and your approach to the problem - especially NLP techniques you used.

*If you have any doubts or something seems inconclusive based on the task description, write it down in the report in the way you would ask the customer to clarify.*

**What you need to provide for this task:**

1. The report (pdf/html/markdown) which includes the description of the model used.
2. The source code written in Python or R (published on one of the git platforms such as github or bitbucket).

Task 3

Write a SQL query, in which you will retrieve the information about the students who scored a 4 and above on their algebra exam. Use the following tables:

table: **students**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **student\_id** | **name** | **surname** | **birth\_date** | **faculty** |
| 34577 | John | Peters | 1993-02-15 | Mathematics |
| 34578 | Kate | Smith | 1992-10-29 | Economy |

table: **exam\_results**

|  |  |  |  |
| --- | --- | --- | --- |
| **student\_id** | **class\_id** | **exam\_date** | **grade** |
| 34577 | 6743 | 2019-06-19 | 2 |
| 34577 | 2876 | 2019-06-06 | 3 |
| 34578 | 6743 | 2019-06-19 | 4.5 |

table: **class\_catalogue**

|  |  |  |  |
| --- | --- | --- | --- |
| **class\_id** | **class\_name** | **professor\_id** | **semester** |
| 6743 | algebra | 4352 | 2019\_summer |
| 2876 | calculus | 5325 | 2019\_summer |
| 3466 | statistics | 6253 | 2019\_winter |