**Тетрадь исследователя**

**ФИО студента: Строев Павел Васильевич**

**Группа: ИВТ 22-2**

**ЛР 6.1 «****Машинное обучение no-code на платформе BigML»**

**Задание 1.** Решение задачи классификации ирисов

|  |  |  |
| --- | --- | --- |
| **№** | **Вопрос** | **Ответ** |
| А | Сколько данных в наборе?  Что прогнозируется?  Является ли данная задача многоклассовой задачей классификации? | 1) 5 данных  2)Прогнозируются параметры цветка (ширина и длина лепестка и бутона)  3)Да |
| B | Какие атрибуты ирисов взяты за основу при поиске закономерностей (составляют features)?  Какой атрибут в нашем датасете выбран целевым значением (является label)?  Что означают слова features и label применительно к понятию «датасет»? | Features – sepal and petal’s length and width  Label –species(Iris-virginica, Iris-versicolor, Iris-setosa)  Features – данные в которых находится закономерность  Label – то, к чему относятся данные( в этом случается вид цветка) |
| C | Какой вид ириса преобладает в корневой вершине дерева?  Какова его метрика Confidence? | Iris virginica  Confidence = 27.30% |
| D | Какой атрибут ириса является наиболее важным для определения его разновидности?  Каким условиям должны удовлетворять ирисы, чтобы модель отнесла их к виду Iris-setosa?  А к двум остальным видам? | Petal’s length и width  Iris-setosa.  Petal length <=2.6  Iris-virginica . 2.6< Petal length <=4.95  Petal width >1.75  Iris-versicolor Petal length >2.6 & Petal width <=1.75 |
| E | По какой формуле рассчитывается доля правильных ответов?  Чему равно ее значение для вашего тестового набора? | (TP+TN)/(TP+TN+FP+FN)  TP- True Positive, TN- True Negative, FP- False Positive, FN- False Negative  93.33% |
| F | Какие данные вы вводили и какие ответы были получены?  Какова была при этом вероятность принадлежности введенных данных к прогнозному классу ирисов?  Приведите два примера. | Пример 1:  1.71  Ответ: Iris-virginica  Вероятность = 67.41%  Пример 2:  3.84  0.9  Ответ: Iris-versicolor  Вероятность =98.48% |

**Задание 2.** Решение задачи регрессии

|  |  |  |
| --- | --- | --- |
| **№** | **Вопрос** | **Ответ** |
| А | Какой датасет вы выбрали?  Почему?  Приведите краткое описание датасета.  Каково признаковое пространство этого набора данных?  Что выбрано вами в качестве целевой функции и почему? | Я выбрал dataset о спросе на аренду велосипедов в Сеуле.  Dataset показался довольно интересным. В нём содержана информация о погоде, количество арендованных велосипедов в час и дата аренды. Целевой функцией является обеспечить минимально необходимым количеством велосипедов в любое время.(количество арендуемых велосипедов) |
| B | Какая работа по предварительной подготовке датасета была вами проведена?  Убирали ли вы при настройке какие-либо признаки?  Если да, то какие признаки не использовали и почему? | Я не учитывал такие признаки, как выходные дни, видимость и осадки, т.к они не сильно влияли на результат. Нет, не убирал |
| C | Какие метрики были использованы в этот раз?  Какую метрику можно считать наиболее подходящей для вашей задачи?  Опишите выбранную метрику – в каких случаях она используется и каким образом рассчитывается? |  |
| D | Какие признаки оказались наиболее важными? | Сезон, рабочий/нерабочий день, время аренды |
| E | Какие данные вы использовали при прогнозировании и почему выбрали именно такие данные?  Какие результаты получили?  Приведите два примера. |  |
| F | Считаете ли вы выбранный датасет качественным?  К каким выводам вы пришли, проведя исследование?  Что можете сказать о следующих ситуациях в машинном обучении:   * много качественных данных, * много некачественных данных, * мало качественных данных, * мало некачественных данных | Много качественных данных –  Много некачественных данных –  Мало качественных данных –  Мало некачественных данных – |

**ЛР 6.2 «Кластеризация студентов»**

**Задание 1.** Решение задачи кластеризации методом *k*-средних при *k* = 4

|  |  |  |
| --- | --- | --- |
| **№** | **Вопрос** | **Ответ** |
| L39 =ПОИСКПОЗ(L38;L34:L37;0) | | |
| А | Что в формуле для L39 означает параметр 0?  Как вы объясните полученный результат? | Функция ПОИСКПОЗ находит первое значение, равное искомому значению(L38=3), вывело 1, т.к. L34=3, а он первый элемент в массиве.  Просматриваемый массив может быть не упорядочен.(1-по возрастанию, (-1)-по убыванию. |
| Н2 =СУММЕСЛИ('4КС'!$L$39:$DG$39;'4КС-Рейтинг'!H$1;'4КС'!$L2:$DG2) | | |
| B | Что в формуле для H2 означает каждый из трех параметров?  Как можно интерпретировать полученный результат? | '4КС'!$L$39:$DG$39 – диапазон ячеек, оцениваемых на соответствие условиям( от L39 до DG39 как пример)  '4КС-Рейтинг'!H$1 – это ссылка на ячейку, определяющая, какие ячейки необходимо суммировать.( Суммируем значения из кластеров, например H$1 – первый кластер)  '4КС'!$L2:$DG2 – это диапазон суммирования – ячейки, значения из которых суммируются. |
| C | Какие курсы наиболее популярны в кластере 1?  Что в них общего?  Как можно охарактеризовать студентов в этом кластере?  *Ответьте на эти же вопросы для каждого кластера 2-4.* | Кластер 1:  три наиболее популярных курса: 31,22,7 Общее направление в экономику, право, предпринимательство и маркетинг. Т. е направления, связанные с ведением бизнеса. Язык преподавания – английский. Автомат получить довольно сложно(в 2 из 3 курсов нельзя). Средний балл > 4,4  Кластер 2:  три наиболее популярных курса: 26,14,6 Общее направление в экономику и математику. Язык преподавания – английский. Автомат получить нельзя. Средний балл > 4,3  Кластер 3:  три наиболее популярных курса: 30,29,8 Общее направление в экономику, предпринимательство и маркетинг. Т. е направления, связанные с ведением бизнеса. Язык преподавания – английский и русский Автомат получить возможно Средний балл > 4,2  Кластер 4 три наиболее популярных курса: 26,4 Общее направление в математику и экономику Т. е направления, связанные с вычислительными задачами Язык преподавания – английский. Автомат получить нельзя Средний балл > 4,4 |
| D | Явно ли прослеживаются отличия кластеров друг от друга?  Подходит ли 4 для значения *k* в кластеризации по *k‑*средним наилучшим образом? | Да, явно. Например, в кластере 3 гораздо легче выявить популярнейшие курсы среди студентов, а в кластере 2 выбрать популярнейшие довольно трудно. По моему мнению, подходит, т. к, имея 4 кластера, уже можно сформировать мнению и найти определённые закономерности. |
| E | Как можно оценить полученное значение силуэта?  Почему? |  |

**Задание 2.** Решение задачи кластеризации методом *k*-средних при *k* = 5

|  |  |  |
| --- | --- | --- |
| **№** | **Вопрос** | **Ответ** |
| А | Стало ли более логичным разделение?  Что теперь общего у участников кластеров? | Кластер 1:  Кластер 2:  Кластер 3:  Кластер 4:  Кластер 5: |
| B | Значение силуэта стало больше? Можно ли считать такой результат удовлетворительным? |  |
| C | Удаётся ли как-то улучшить силуэт? |  |

**Задание 3.** Решение задачи кластеризации методом *k*-медиан при *k* = 5

|  |  |  |
| --- | --- | --- |
| **№** | **Вопрос** | **Ответ** |
| А | Что общего теперь у участников каждого кластера?  Какие параметры сделок совпадают у участников кластеров?  Стало ли разделение на кластеры более логичным? | Кластер 1:  Кластер 2:  Кластер 3:  Кластер 4:  Кластер 5: |