گزارش پروژه سوم داده کاوی

محمدرضا صیدگر-97222055

این پروژه شامل 4 سرویس مختلف است که هر کدام از سرویس ها کار های مختلفی را می خواهد:

1) سرویس اول از ما میخواهد که یک درونیابی انجام بدیم روی داده های سری زمانی و ورودی این سرویس به صورت زیر خواهد بود :


```
{
  "data": {
    "time": {
      "0": "1390/11/14",
      "1": "1390/11/15",
      "2": "1390/11/17"
    },
    "vol": {
      "0": 20,
      "1": 40,
      "2": 100
    }
  },
  "config": {
    "type": "shamsi/miladi",
    "time": "daily/monthly",
    "interpolation": "linear/polynomial"
  }
}
```


ورودی به این شکل خواهد بود که تاریخ می تواند میلادی یا شمسی باشد و نوع زمان ها هم به صورت روزانه یا ماهانه باشد و از روش های درونیابی هم خطی و چندجمله ای میتواند باشد

برای ورودی بالا خروجی سرویس 1 من به صورت زیر بود:

```
{
   "data":
"{\"time\":{\"0\":1328227200000,\"1\":1328313600000,\"2\":1328400000000,\"3\":132848
6400000},\"vol\":{\"0\":20.0,\"1\":40.0,\"2\":70.0,\"3\":100.0}}"
}
```


همانطور که می بینیم خروجی ها به شکل میلی ثانیه بوده و برای روز 16 ام که موجود نبود عدد 70 که با درونیابی خطی بدست آمده درج شده است.

2) سرویس دوم ورودی را به صورت میلادی دریافت میکند و همان کار درونیابی را از ما میخواد با این تفاوت که این دفعه خروجی ما حتما باید به صورت شمسی باشد. ورودی زیر را به سرویس 2 دادیم:


```
{
  "data": {
    "time": {
      "0": "1999/11/14",
      "1": "1999/11/15",
      "2": "1999/11/17"
    },
    "vol": {
      "0": 20,
      "1": 40,
      "2": 100
    }
  },
  "config": {
    "type": "miladi",
    "time": "daily",
    "interpolation": "polynomial"
  }
}
```


خروجی سرویس ما به شکل زیر بود:


```
{
```

```
   "data":
"{\"time\":{\"0\":\"1378-08-23\",\"1\":\"1378-08-24\",\"2\":\"1378-08-25\",\"3\":\"1378-08-26\
"},\"vol\":{\"0\":20.0,\"1\":40.0,\"2\":66.6666666667,\"3\":100.0}}"
}
```


همانطور که می بینیم تاریخ ها به صورت شمسی در آمده و عدد 66.6 برای تاریخ 1378-08-25 به صورت چندجمله ای درجه 2 بدست آمده است.

3) سرویس سوم از ما میخواهد داده های پرت را برای یه سری داده عادی و همینطور برای داده های سری زمانی پیدا کنیم پس ما 2 تا ورودی را برای این سرویس انتخاب کردیم. ورودی اول به صورت زیر است:


```
{
  "data": {
    "id": {
      "0": 1,
      "1": 2,
      "2": 3,
      "3": 4,
      "4": 5,
      "5": 6
    },
    "feature": {
      "0": 100,
      "1": 20,
      "2": 35,
      "3": 67,
      "4": 89,
      "5": 90
    }
  },
  "config": {
    "time_series": false
  }
}
```


که خروجی سرویس ما برای این ورودی به شکل زیر است:

```
{
   "data":
"{\"id\":{\"0\":1,\"1\":2,\"2\":3,\"3\":4,\"4\":5,\"5\":6},\"feature\":{\"0\":100,\"1\":20,\"2\":35,\"3\
":67,\"4\":89,\"5\":90},\"method1\":{\"0\":true,\"1\":true,\"2\":false,\"3\":false,\"4\":false,\"5\":
false},\"method2\":{\"0\":\"false\",\"1\":\"false\",\"2\":\"false\",\"3\":\"false\",\"4\":\"false\",\"5\
":\"false\"}}"
}
```


همانطور که می بینیم در متود اول (متود sort است یعنی داده ها را از کوچک به بزرگ مرتب کرده و 10 درصد داده های بالا و 10 درصد داده های پایین را داده های پرت معرفی می کند) اعداد 100 و 20 به عنوان داده های پرت مشخص شدند و در متود دوم (داده هایی را که از میانگین بیش از 3 برابر انحراف معیار فاصله دارند داده پرت معرفی میکند) هیچ کدام از اعداد داده پرت نیستند.

ورودی دوم به صورت زیر است:


```
{
  "data": {
    "time": {
      "0": "1390-11-14",
      "1": "1390-11-15",
      "2": "1390-11-16",
      "3": "1390-11-17"
    },
    "feature": {
      "0": 20,
      "1": 40,
      "2": 70,
      "3": 100
    }
  },
  "config": {
    "time_series": true
  }
}
```


که خروجی سرویس ما برای این ورودی به شکل زیر است:


```
{
```

```
   "data":
"{\"time\":{\"0\":\"1390-11-14\",\"1\":\"1390-11-15\",\"2\":\"1390-11-16\",\"3\":\"1390-11-17\"
},\"feature\":{\"0\":20,\"1\":40,\"2\":70,\"3\":100},\"method1\":{\"0\":true,\"1\":false,\"2\":fals
e,\"3\":true},\"method3\":{\"0\":false,\"1\":false,\"2\":true,\"3\":false}}"
}
```


همانطور که می بینیم در متود اول (همان متود sort است) اعداد 100 و 20 در روز های اول و چهارم به عنوان داده های پرت مشخص شدند و در متود دوم (با مدل autoregression داده ها را پیشگویی کرده و داده هایی را که با مقدار پیشگویی شده فاصله زیادی داشتند داده پرت در نظر گرفتیم) عدد 70 در روز سوم داده پرت است.

4) سرویس چهارم میخواهد که داده ها را در کلاس های مختلف که نامتوازن است را متوازن کند. برای این سرویس از 4 متد SMOTE و Oversampling و UnderSampling و Cluster Centroids استفاده شد. یکی از ورودی ها به صورت زیر است:


```
{
  "data": {
    "id": {
      "0": 1,
      "1": 2,
      "2": 3,
      "3": 4,
      "4": 5,
      "5": 6
    },
    "feature1": {
      "0": 50,
      "1": 12,
      "2": 50,
      "3": 500,
      "4": 60,
      "5": 12
    },
```

```
    "class": {
      "0": 1,
      "1": 1,
      "2": 1,
      "3": 1,
      "4": 1,
      "5": 0
    }
  },
  "config": {
    "major_class": 1,
    "minor_class": 0,
    "method": "Oversampling"
  }
}
```


که خروجی سرویس ما برای این ورودی به شکل زیر است:


```
{
    "data":
"{\"index\":{\"0\":0,\"1\":1,\"2\":2,\"3\":3,\"4\":4,\"5\":5,\"6\":6,\"7\":7,\"8\":8,\"9\":9},\"id\":{\"0\
":1,\"1\":2,\"2\":3,\"3\":4,\"4\":5,\"5\":6,\"6\":7,\"7\":8,\"8\":9,\"9\":10},\"feature1\":{\"0\":50,\"1
\":12,\"2\":50,\"3\":500,\"4\":60,\"5\":12,\"6\":12,\"7\":12,\"8\":12,\"9\":12},\"class\":{\"0\":1,\"
1\":1,\"2\":1,\"3\":1,\"4\":1,\"5\":0,\"6\":0,\"7\":0,\"8\":0,\"9\":0}}"
}
```


همانطور که می بینیم در ورودی اولیه 5 داده از کلاس یک و 1 داده از کلاس صفر داشتیم ولی در خروجی می بینیم که 5 تا از هر کدام کلاس ها داریم پس داده ها متوازن شده اند.