 **به نام خدا**

سری سوم تمارین درس داده کاوی

استاد درس:

دکتر مریم امیر مزلقانی

نیم سال اول 1402-1403

راه ارتباطی:

[Aut.DataMining.Fall@gmail.com](mailto:Aut.DataMining.Fall@gmail.com)

**بخش عملی:**

ﻫﺪﻑ ﺍﺯ ﺍﻳﻦ بخش، ﭘﻴﺎﺩﻩ ﺳﺎﺯﻱ ﻭ ﻣﻘﺎﻳﺴﻪ ﻋﻤﻠﻜﺮﺩ مدل های مختلف دسته‎بندی ﻣﻲ ﺑﺎﺷﺪ. ﺑﺮﺍﻱ ﭘﻴﺎﺩﻩ ﺳﺎﺯﻱ ﺍﻳﻦ ﻣﺪﻝ ﻫﺎ، ﻣﺠﻤﻮﻋﻪ ﺩﺍﺩﻩ ﺍﻱ ﺩﺭﻧﻈﺮ ﮔﺮﻓﺘﻪ ﺷﺪﻩ ﺍﺳﺖ ﻛﻪ ﻫﺪﻑ ﺁﻥ ﭘﻴﺶ ﺑﻴﻨﻲ ﺩﺭﺁﻣﺪ ﺍﻓﺮﺍﺩ ﺑﺎ ﺗﻮﺟﻪ ﺑﻪ ﻭﻳﮋﮔﻲ ﻫﺎﻱ ﺍﻓﺮﺍﺩ ﺩﺭ ﺍﻳﻦ ﻣﺠﻤﻮﻋﻪ ﺩﺍﺩﻩ ﺍﺳﺖ. ﺍﻳﻦ ﻣﺠﻤﻮﻋﻪ ﺩﺍﺩﻩ ﺩﺭ ﭘﻮﺷﻪ Practical ﻗﺮﺍﺭ ﺩﺍﺩﻩ ﺷﺪﻩ ﺍﺳﺖ. ﻣﺠﻤﻮﻋﻪ ﺩﺍﺩﻩ Adult\_TrainDataset ﺑﺮﺍﻱ ﺁﻣﻮﺯﺵ ﻣﺪﻝ ﻫﺎ ﺩﺭ ﻧﻈﺮ ﮔﺮﻓﺘﻪ ﺷﺪﻩ ﺍﺳﺖ ﻛﻪ ﺣﺎﻭﻱ32561 ﻧﻤﻮﻧﻪ ﻣﻲ ﺑﺎﺷﺪ ﻛﻪ ﺩﺭ ﺩﻭ ﻛﻼﺱ ﺩﺳﺘﻪ ﺑﻨﺪﻱ ﺷﺪﻩ ﺍﻧﺪ. ﺩﺭ ﺍﻳﻦ ﻣﺠﻤﻮﻋﻪ ﺩﺍﺩﻩ، ﺳﺘﻮن Income ﺣﺎﻭﻱ ﺑﺮﭼﺴﺐ ﻫﺎ ﻭ ﻣﺎﺑﻘﻲ ﺳﺘﻮﻥ ﻫﺎ، ﻭﻳﮋﮔﻲ ﻫﺎﻳﻲ ﻫﺴﺘﻨﺪ ﻛﻪ ﺑﺎﻳﺴﺘﻲ ﺑﺮﺍﻱ ﭘﻴﺎﺩﻩ ﺳﺎﺯﻱ ﻣﺪﻝ، ﺍﺯ ﺁﻥ ﻫﺎ ﺍﺳﺘﻔﺎﺩﻩ ﻧﻤﺎﻳﻴﺪ.

* ﺩﺭ ﺍﻳﻦ ﭘﺮﻭﮊﻩ ﺑﺮﺍﻱ ﭘﻴﺎﺩﻩ ﺳﺎﺯﻱ ﻣﺪﻝ ﻫﺎ ﻣﻲ ﺗﻮﺍﻧﻴﺪ ﺍﺯ ﻣﺪﻝ ﻫﺎﻱ ﺁﻣﺎﺩﻩ ﻛﺘﺎﺑﺨﺎﻧﻪ sklearnﺍﺳﺘﻔﺎﺩﻩ ﻛﻨﻴﺪ.
* ﻣﺠﻤﻮﻋﻪ ﺩﺍﺩﻩ ﻫﺎﻱ ﺁﻣﻮﺯﺵ ﺭﺍ ﺑﻪ ﻛﻤﻚ ﺣﺪﺍﻗﻞ ﺩﻭ ﻣﺪﻝ ﻧﻤﻮﺩﺍﺭ، ﺑﺼﺮﻱ ﺳﺎﺯﻱ ﻛﻨﻴﺪ ﻭ ﺑﻪ ﺗﺤﻠﻴﻞ ﺁﻥ ﻫﺎ ﺑﭙﺮﺩﺍﺯﻳﺪ. ﺑﺎ ﺗﻮﺟﻪ ﺑﻪ ﻭﺟﻮﺩ ﻣﻘﺎﺩﻳﺮ ﺍﺯ ﺩﺳﺖ ﺭﻓﺘﻪ (Null) ﺩﺭ ﺍﻳﻦ ﻣﺠﻤﻮﻋﻪ ﺩﺍﺩﻩ، ﺑﺎﻳﺪ ﻣﻘﺎﺩﻳﺮ ﺍﺯ ﺩﺳﺖ ﺭﻓﺘﻪ ﺭﺍ ﺑﺎ ﻣﻘﺎﺩﻳﺮ ﻣﻨﺎﺳﺐ ﺟﺎﻳﮕﺰﻳﻦ ﻛﻨﻴﺪ ﻭ ﻳﺎ ﺳﺘﻮﻥ ﻣﺮﺑﻮﻃﻪ ﺭﺍ ﺣﺬﻑ ﻧﻤﺎﻳﻴﺪ.
* ﺍﺯ ﺁﻧﺠﺎﻳﻲ ﻛﻪ ﺑﺮﺧﻲ ﺍﺯ ﺳﺘﻮﻥ ﻫﺎ ﺣﺎﻭﻱ ﻣﻘﺎﺩﻳﺮ Categorical ﻫﺴﺘﻨﺪ، ﺩﺭ ﺻﻮﺭﺕ ﺍﺳﺘﻔﺎﺩﻩ ﺍﺯ ﺁﻧﻬﺎ ﺑﺮﺍﻱ ﺁﻣﻮﺯﺵ ﻣﺪﻝ، ﺑﺎﻳﺪ ﻣﻘﺎﺩﻳﺮ ﺍﻳﻦ ﺳﺘﻮﻥ ﻫﺎ ﺭﺍ ﺑﺎ ﺍﺳﺘﻔﺎﺩﻩ ﺍﺯ ﺭﻭﺵ ﻫﺎﻱ ﻣﻮﺟﻮﺩ، Encode ﻛﺮﺩﻩ ﻭ ﺑﻪ ﻣﻘﺎﺩﻳﺮ ﻋﺪﺩﻱ ﺗﺒﺪﻳﻞ ﻛﻨﻴﺪ. ﺗﻮﺟﻪ ﺩﺍﺷﺘﻪ ﺑﺎﺷﻴﺪ ﻛﻪ ﺭﻭﺵ ﻫﺎﻱ ﻣﺨﺘﻠﻔﻲ ﺟﻬﺖ Encode ﻛﺮﺩﻥ ﺩﺍﺩﻩ ﻫﺎﻱCategorical وجود ﺩﺍﺭﺩ ﻛﻪ ﺍﺯ ﺟﻤﻠﻪ ﺁﻧﻬﺎ ﻣﻲ ﺗﻮﺍﻥ ﺑﻪOne hot Encoding ﻭLabel Encoding ﺍﺷﺎﺭﻩ ﻛﺮﺩ. ﺍﻳﻦ ﺩﻭ ﺭﻭﺵ ﺭﺍ ﺑﺎ ﻫﻢ ﻣﻘﺎﻳﺴﻪ ﻛﺮﺩﻩ ﻭ ﺗﻮﺿﻴﺢ ﺩﻫﻴﺪ ﺑﺎ ﺍﻧﺘﺨﺎﺏ ﻛﺪﺍﻡ ﺭﻭﺵ ﻣﻲ ﺗﻮﺍﻥ ﺑﻪ ﻧﺘﺎﻳﺞ ﺑﻬﺘﺮﻱ ﺩﺳﺖ ﻳﺎﻓﺖ ﻭ ﺑﺎ ﺭﻭﺵ ﺍﻧﺘﺨﺎﺏ ﺷﺪﻩ، ﺑﻪEncode ﻛﺮﺩﻥ ﺩﺍﺩﻩ ﻫﺎ ﺑﭙﺮﺩﺍﺯﻳﺪ.
* ﺩﺭ ﺻﻮﺭﺕ ﻧﻴﺎﺯ ﻣﻲ ﺗﻮﺍﻧﻴﺪ ﻗﺒﻞ ﺍﺯ ﺁﻣﻮﺯﺵ ﻣﺪﻝ ﻫﺎ، ﭘﻴﺶ ﭘﺮﺩﺍﺯﺵ ﻫﺎﻱ ﻣﺨﺘﻠﻔﻲ ﻫﻤﭽﻮﻥ ﻧﺮﻣﺎﻝ ﺳﺎﺯﻱ، ﺍﺳﺘﺎﻧﺪﺍﺭﺩﺳﺎﺯﻱ ﻭ... ﺭﺍ ﺭﻭﻱ ﺍﻳﻦ ﻣﺠﻤﻮﻋﻪ ﺩﺍﺩﻩ ﺍﻋﻤﺎﻝ ﻛﻨﻴﺪ.
* ﺩﺭ ﺁﻣﻮﺯﺵ ﻫﺮ ﻣﺪﻝ ، ﺩﺭ ﺻﻮﺭﺕ ﺗﻨﻈﻴﻢ ﻫﺎﻳﭙﺮﭘﺎﺭﺍﻣﺘﺮﻫﺎﻱ ﻣﺪﻝ ﺩﺭ ﻛﺘﺎﺑﺨﺎﻧﻪsklearn ، ﺩﻻﻳﻞ ﺍﻳﻦ ﺍﻧﺘﺨﺎﺏ ﻫﺎ ﻭ ﺗﺤﻠﻴﻞ ﺧﻮﺩﺗﺎﻥ ﺭﺍ ﺍﺯ ﻧﺘﺎﻳﺢ ﺣﺎﺻﻞ ﺷﺪﻩ، ﮔﺰﺍﺭﺵ ﺩﻫﻴﺪ.
* ﺩﺭ ﻧﻬﺎﻳﺖ ﺑﻌﺪ ﺍﺯ ﺁﻣﻮﺯﺵ ﻣﺪﻝ ﻫﺎ، ﺑﺮﺍﻱ ﺍﺭﺯﻳﺎﺑﻲ ﻋﻤﻠﻜﺮﺩ ﻣﺪﻝ ﻫﺎ، ﺍﺯ ﻣﺠﻤﻮﻋﻪ ﺩﺍﺩﻩ Adult\_TestDataset ﺍﺳﺘﻔﺎﺩﻩ ﻧﻤﺎﻳﻴﺪ ﻭ ﺑﺎ ﺍﺳﺘﻔﺎﺩﻩ ﺍﺯ ﻣﺎﺗﺮﻳﺲ ﺩﺭﻫﻢ ﺭﻳﺨﺘﮕﻲ (confusion matrix) ﻭ ﻣﻌﻴﺎﺭﻫﺎﻱ ﺍﺭﺯﻳﺎﺑﻲrecall ،precision ،accuracy ﻭf1-score، ﺑﻪ ﺗﺤﻠﻴﻞ ﻧﺘﺎﻳﺞ ﺑﺪﺳﺖ ﺁﻣﺪﻩ ﺑﭙﺮﺩﺍﺯﻳﺪ. ﺗﻮﺟﻪ ﺩﺍﺷﺘﻪ ﺑﺎﺷﻴﺪ ﻛﻪ ﺑﺮﺍﻱ ﺳﺎﺩﮔﻲ ﻛﺎﺭ، ﻣﻲ ﺗﻮﺍﻧﻴﺪ ﺑﺎ ﺍﺳﺘﻔﺎﺩﻩ ﺍﺯ ﺗﻮﺍﺑﻊ ﺁﻣﺎﺩﻩ sklearn ، ﻣﺎﺗﺮﻳﺲ ﺩﺭﻫﻢ ﺭﻳﺨﺘﮕﻲ ﻭ ﻣﻌﻴﺎﺭﻫﺎﻱ ﺍﺭﺯﻳﺎﺑﻲ ﺭﺍ ﺑﺪﺳﺖ ﺁﻭﺭﻳﺪ.
* ﺗﻮﺟﻪ ﺩﺍﺷﺘﻪ ﺑﺎﺷﻴﺪ ﻛﻪ ﺍﺭﺍﻳﻪ ﮔﺰﺍﺭﺵ ﻫﺮ ﺑﺨﺶ ﺍﻟﺰﺍﻣﻲ ﺍﺳﺖ ﻭ ﺑﺎﻳﺴﺘﻲ ﺩﺭ ﻓﺎﻳﻞ ﻧﻮﺕ ﺑﻮﻙ ﺧﻮﺩ، ﺑﻌﺪ ﺍﺯ ﻫﺮ ﺳﻠﻮل مرﺑﻮﻁ ﺑﻪ ﻛﺪ، ﮔﺰﺍﺭﺵ ﻫﺮ ﺑﺨﺶ ﺭﺍ ﺩﺭ ﻳﻚ ﺳﻠﻮﻝ ﻣﺘﻨﻲ ﺑﻨﻮﻳﺴﻴﺪ.