Week5

資工三 4110056030

1. **下表為某公司的員工資料表，請採用漢明距離進行相似度的衡量**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 員工代號 | 性別 | 部門 | 薪資 | 配偶 |
| 01 | 男 | 市場部 | 24000 | 無 |
| 02 | 女 | 研發部 | 45000 | 有 |
| 03 | 女 | 會計部 | 25000 | 有 |
| 04 | 男 | 研發部 | 40000 | 無 |
| 05 | 女 | 會計部 | 24000 | 有 |
| 06 | 男 | 研發部 | 40000 | 無 |
| 07 | 男 | 市場部 | 24000 | 有 |

ANS:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 01 | 02 | 03 | 04 | 05 | 06 | 07 |
| 01 |  | 0/4 | 0/4 | 2/4 | 1/4 | 2/4 | 3/4 |
| 02 |  |  | 2/4 | 1/4 | 2/4 | 1/4 | 1/4 |
| 03 |  |  |  | 0/4 | 3/4 | 0/4 | 1/4 |
| 04 |  |  |  |  | 0/4 | 4/4 | 1/4 |
| 05 |  |  |  |  |  | 0/4 | 2/4 |
| 06 |  |  |  |  |  |  | 1/4 |

1. **根據上表，利用漢明距離和熵衡量來進行特性排序**

移除性別:

|  |  |  |  |
| --- | --- | --- | --- |
| 員工代號 | 部門 | 薪資 | 配偶 |
| 01 | 市場部 | 24000 | 無 |
| 02 | 研發部 | 45000 | 有 |
| 03 | 會計部 | 25000 | 有 |
| 04 | 研發部 | 40000 | 無 |
| 05 | 會計部 | 24000 | 有 |
| 06 | 研發部 | 40000 | 無 |
| 07 | 市場部 | 24000 | 有 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 01 | 02 | 03 | 04 | 05 | 06 | 07 |
| 01 |  | 0/3 | 0/3 | 1/3 | 1/3 | 1/3 | 2/3 |
| 02 |  |  | 1/3 | 1/3 | 1/3 | 1/3 | 1/3 |
| 03 |  |  |  | 0/3 | 2/3 | 0/3 | 1/3 |
| 04 |  |  |  |  | 0/3 | 3/3 | 0/3 |
| 05 |  |  |  |  |  | 0/3 | 2/3 |
| 06 |  |  |  |  |  |  | 0/3 |

移除部門:

|  |  |  |  |
| --- | --- | --- | --- |
| 員工代號 | 性別 | 薪資 | 配偶 |
| 01 | 男 | 24000 | 無 |
| 02 | 女 | 45000 | 有 |
| 03 | 女 | 25000 | 有 |
| 04 | 男 | 40000 | 無 |
| 05 | 女 | 24000 | 有 |
| 06 | 男 | 40000 | 無 |
| 07 | 男 | 24000 | 有 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 01 | 02 | 03 | 04 | 05 | 06 | 07 |
| 01 |  | 0/3 | 0/3 | 2/3 | 1/3 | 2/3 | 2/3 |
| 02 |  |  | 2/3 | 0/3 | 2/3 | 0/3 | 1/3 |
| 03 |  |  |  | 0/3 | 2/3 | 0/3 | 1/3 |
| 04 |  |  |  |  | 0/3 | 3/3 | 1/3 |
| 05 |  |  |  |  |  | 0/3 | 2/3 |
| 06 |  |  |  |  |  |  | 1/3 |

移除薪資:

|  |  |  |  |
| --- | --- | --- | --- |
| 員工代號 | 性別 | 部門 | 配偶 |
| 01 | 男 | 市場部 | 無 |
| 02 | 女 | 研發部 | 有 |
| 03 | 女 | 會計部 | 有 |
| 04 | 男 | 研發部 | 無 |
| 05 | 女 | 會計部 | 有 |
| 06 | 男 | 研發部 | 無 |
| 07 | 男 | 市場部 | 有 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 01 | 02 | 03 | 04 | 05 | 06 | 07 |
| 01 |  | 0/3 | 0/3 | 2/3 | 0/3 | 2/3 | 2/3 |
| 02 |  |  | 2/3 | 1/3 | 2/3 | 1/3 | 1/3 |
| 03 |  |  |  | 0/3 | 3/3 | 0/3 | 1/3 |
| 04 |  |  |  |  | 0/3 | 3/3 | 1/3 |
| 05 |  |  |  |  |  | 0/3 | 1/3 |
| 06 |  |  |  |  |  |  | 1/3 |

移除配偶:

|  |  |  |  |
| --- | --- | --- | --- |
| 員工代號 | 性別 | 部門 | 薪資 |
| 01 | 男 | 市場部 | 24000 |
| 02 | 女 | 研發部 | 45000 |
| 03 | 女 | 會計部 | 25000 |
| 04 | 男 | 研發部 | 40000 |
| 05 | 女 | 會計部 | 24000 |
| 06 | 男 | 研發部 | 40000 |
| 07 | 男 | 市場部 | 24000 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 01 | 02 | 03 | 04 | 05 | 06 | 07 |
| 01 |  | 0/3 | 0/3 | 1/3 | 1/3 | 1/3 | 3/3 |
| 02 |  |  | 1/3 | 1/3 | 1/3 | 1/3 | 0/3 |
| 03 |  |  |  | 0/3 | 2/3 | 0/3 | 0/3 |
| 04 |  |  |  |  | 0/3 | 3/3 | 1/3 |
| 05 |  |  |  |  |  | 0/3 | 1/3 |
| 06 |  |  |  |  |  |  | 1/3 |

原先亂度:

9{-[1/4\*log21/4 + (1-1/4)\*log2(1-1/4)]} + 5{-[2/4\*log22/4 + (1-2/4)\*log2(1-2/4)]} = 12.301503120132194

移除性別後:

12{-[1/3\*log21/3 + (1-1/3)\*log2(1-1/3)]} = 11.019550008653878

移除部門後:

12{-[1/3\*log21/3 + (1-1/3)\*log2(1-1/3)]} = 11.019550008653878

移除薪資後:

12{-[1/3\*log21/3 + (1-1/3)\*log2(1-1/3)]} = 11.019550008653878

移除配偶後:

11{-[1/3\*log21/3 + (1-1/3)\*log2(1-1/3)]} = 10.101254174599388

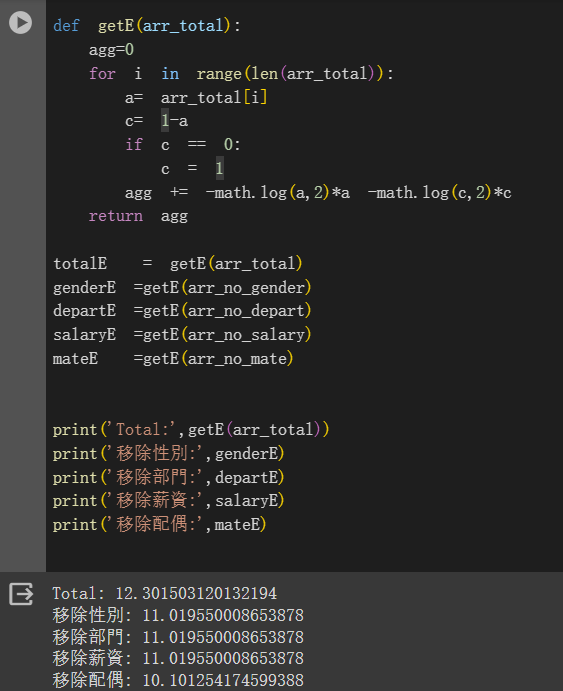
12.301503120132194 - 11.019550008653878

= 1.281953111478316

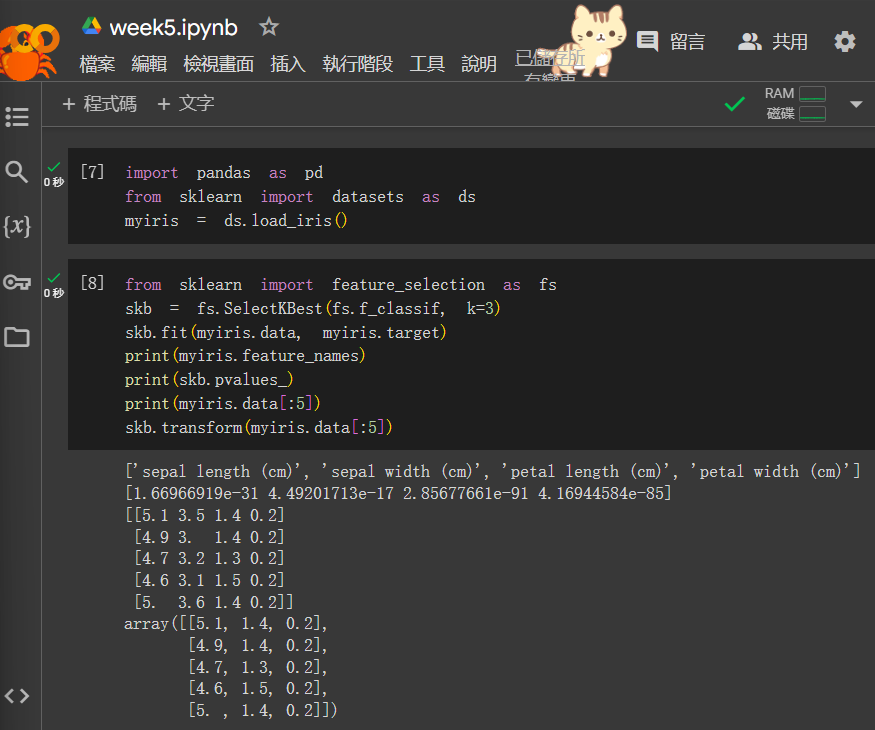
12.301503120132194 - 10.101254174599388

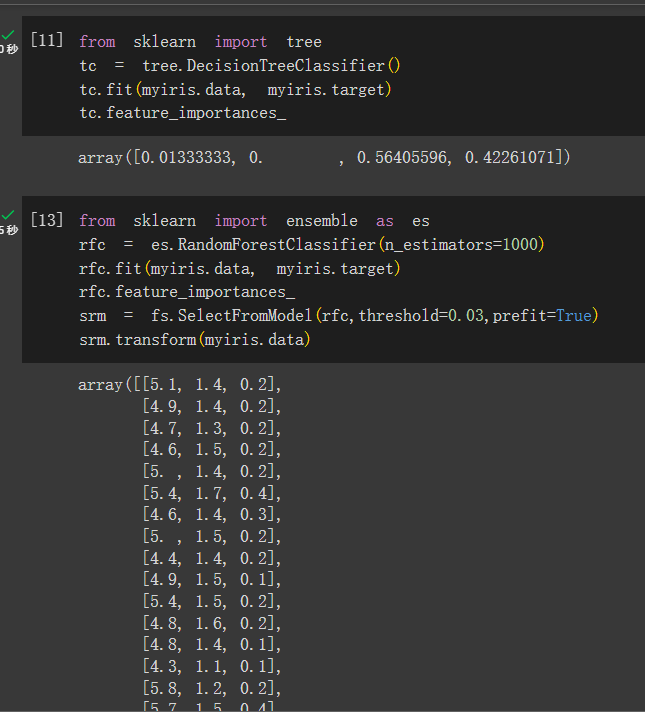
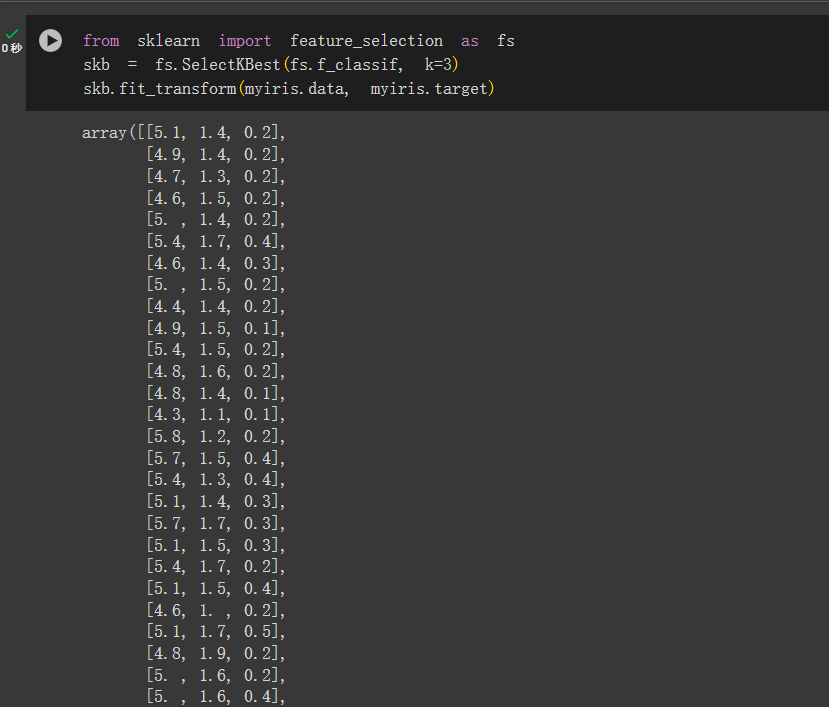
= 2.200248945532806

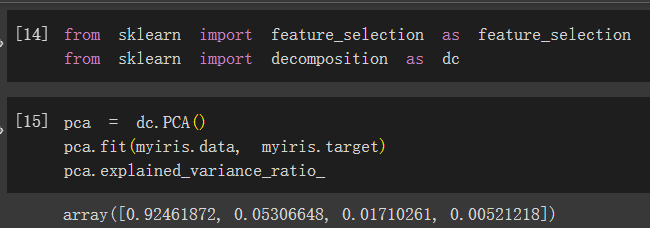
* So, 性別,部門,薪資與原先亂度差得最少,要先被移除的feature是: 性別or部門 or 薪資

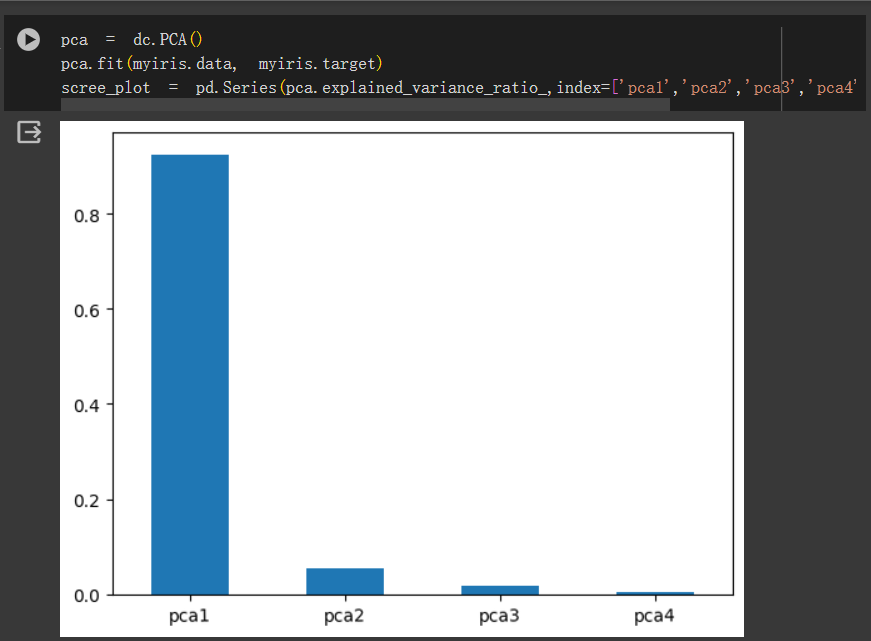


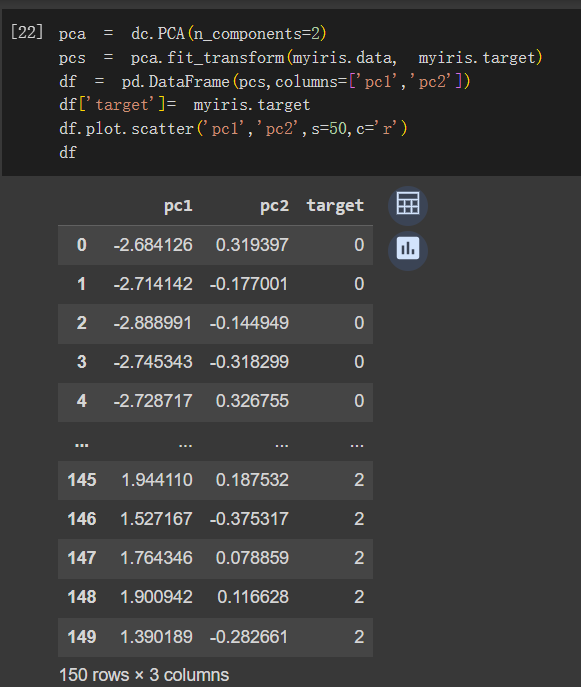
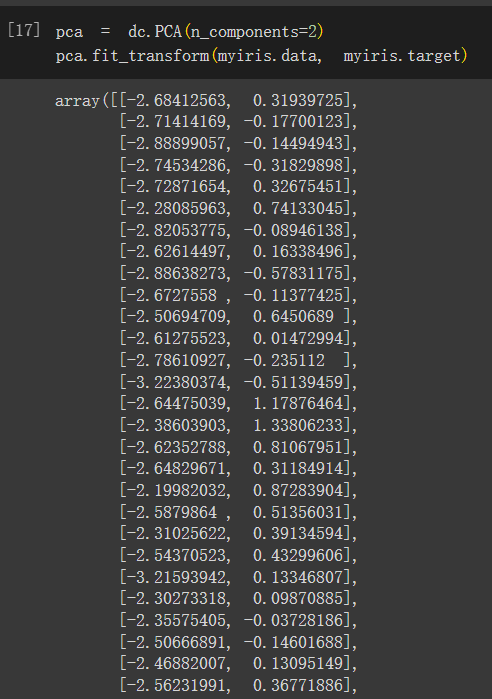
**Python exercise**

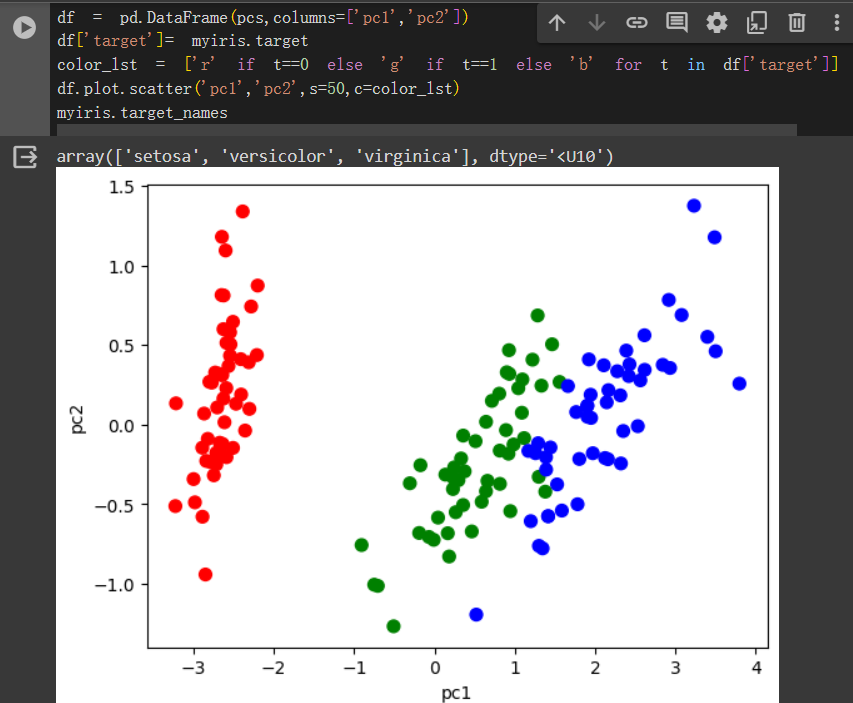
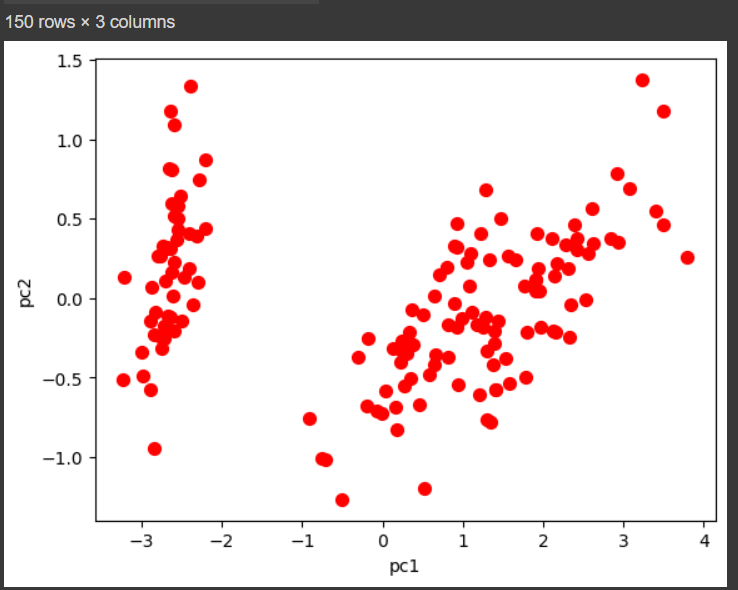












**Python exercise(t-sne)**

