
**دانشگاه تهران**

**پردیس دانشکده‌های فنی**

**دانشکده مهندسی برق و کامپیوتر**


---

# ترکیب داده/ اطلاعات

## تمرین سری پنجم

---


**محمدحسین بدیعی**

**شماره دانشجویی 810199106**

**گرایش : کنترل – هوش مصنوعی و رباتیک**

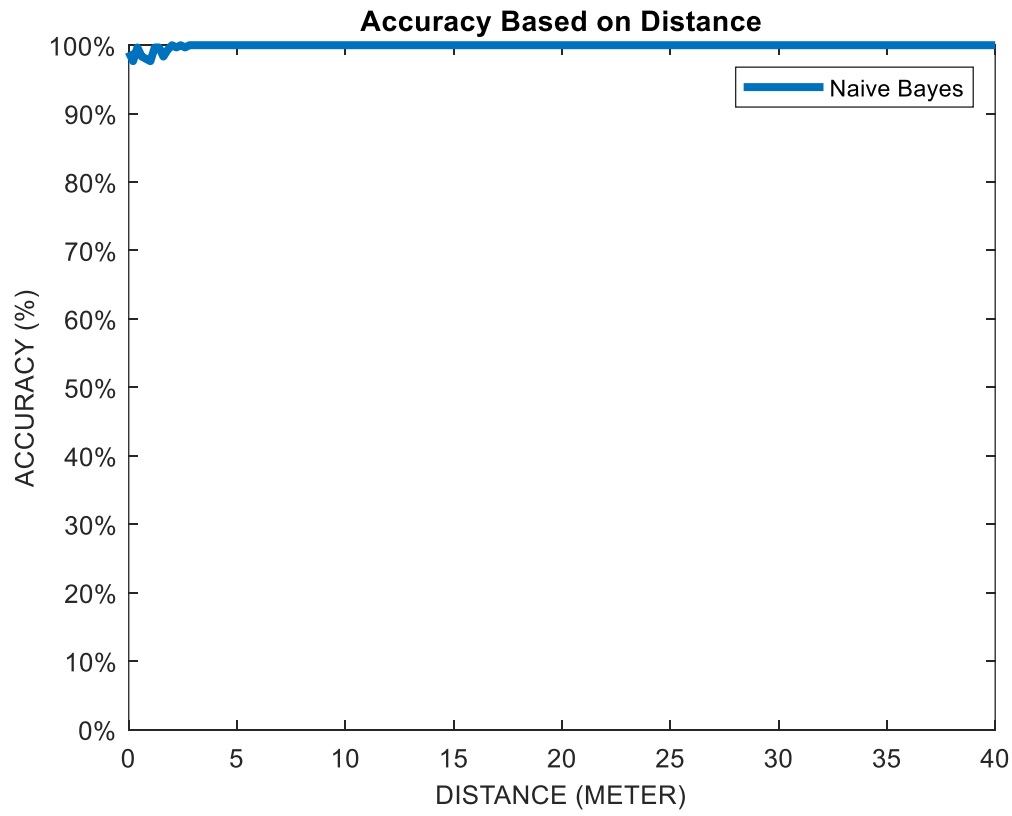
**استاد : دکتر بهزاد مشیری**


**بهار 1399-1400**

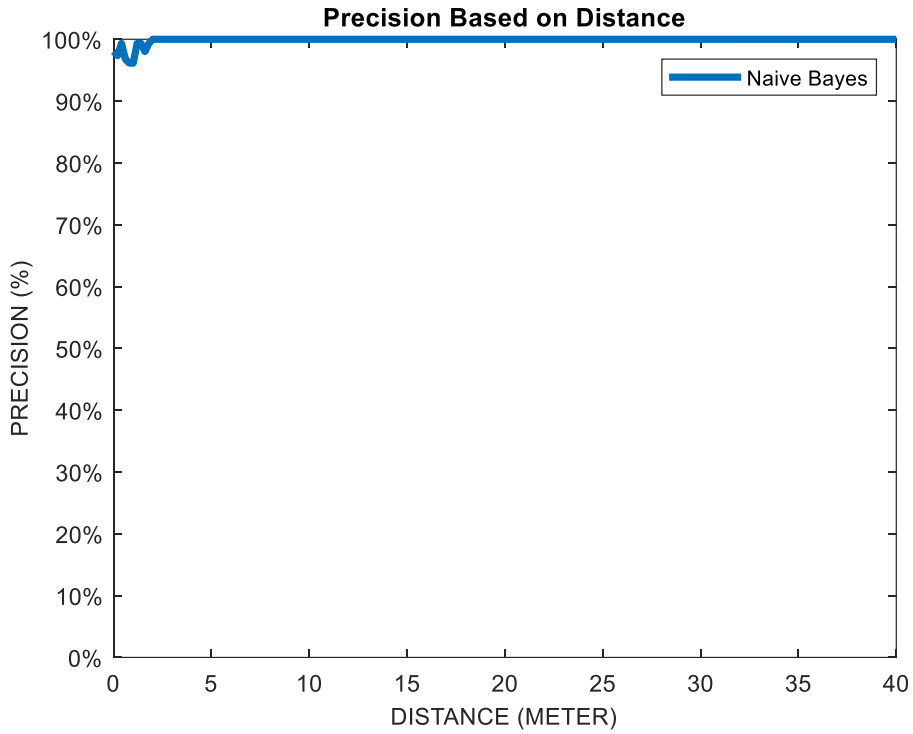
**بخش a و b )** مطابق با خواسته‌ی مسأله دو کلاسِ داده که هر کدام دارای تعدادِ 500 ریکورد با ابعادِ 100 می‌باشند را تولید نمودیم و سپس سیگماها را مطابق با صورت سوال تنظیم کرده و توسطِ همه‌ی پنج کلاسیفایرِ موجود در صورتِ سوال، مدل‌های مربوط به هر یک را توسطِ داده‌های آموزش ساخته و توسطِ داده‌های تست موردِ ارزیابی قرار دادیم. طبق قسمت دومِ سوال تعداد 70% داده‌ها را به عنوان train و بقیه را بع عنوان تست گرفتیم.

**بخش c)** در قسمتِ C که از ما خواسته شده است فواصل اقلیدسی بین کلاس‌ها را تغییر داده و رابطه‌ی آن با صحت و دقت را بیابیم لذا نتایج برای هر کلاسیفایر به قرار زیر است.

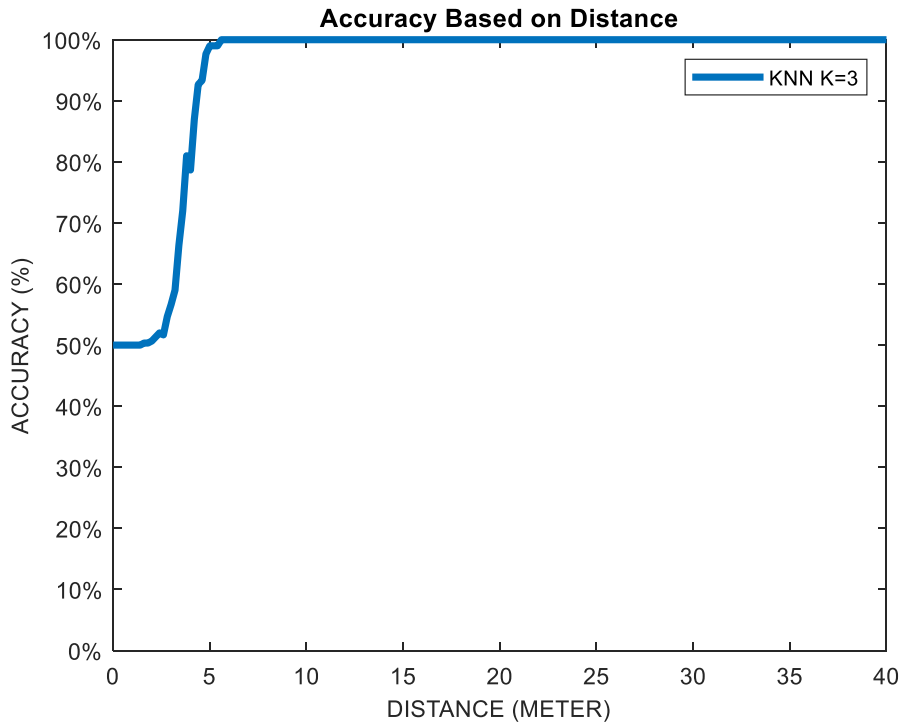
**نمودارِ صحت بر اساسِ فاصله برای کلاسفایرِ Naïve Bayes**



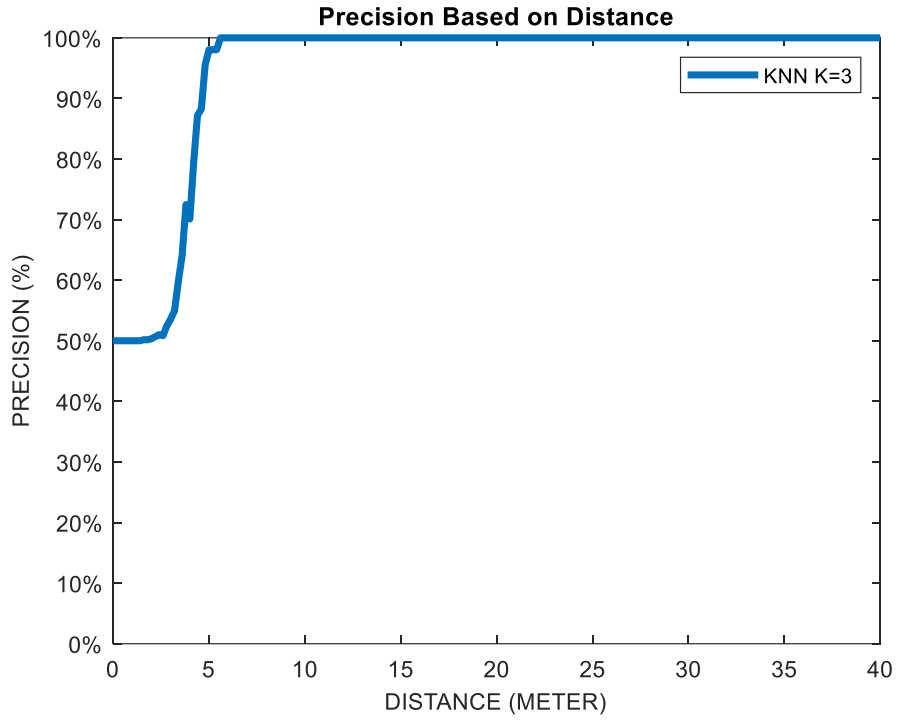
نمودارِ دقت بر اساسِ فاصله برای کلاسفایرِ **Naïve Bayes**



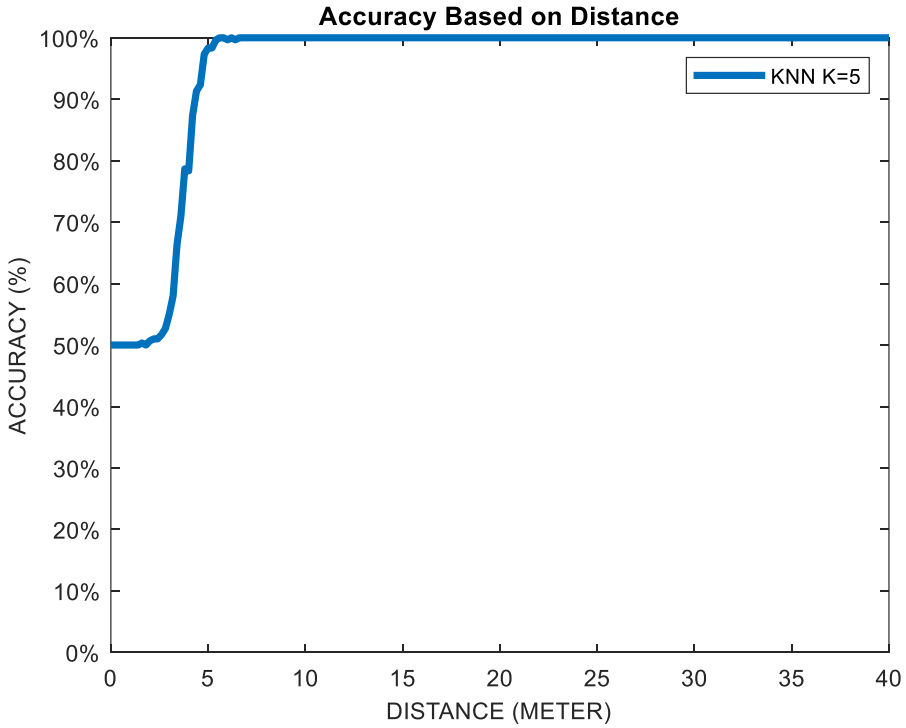

نمودارِ صحت بر اساسِ فاصله برای کلاسفایرِ **KNN** برای **K=3**



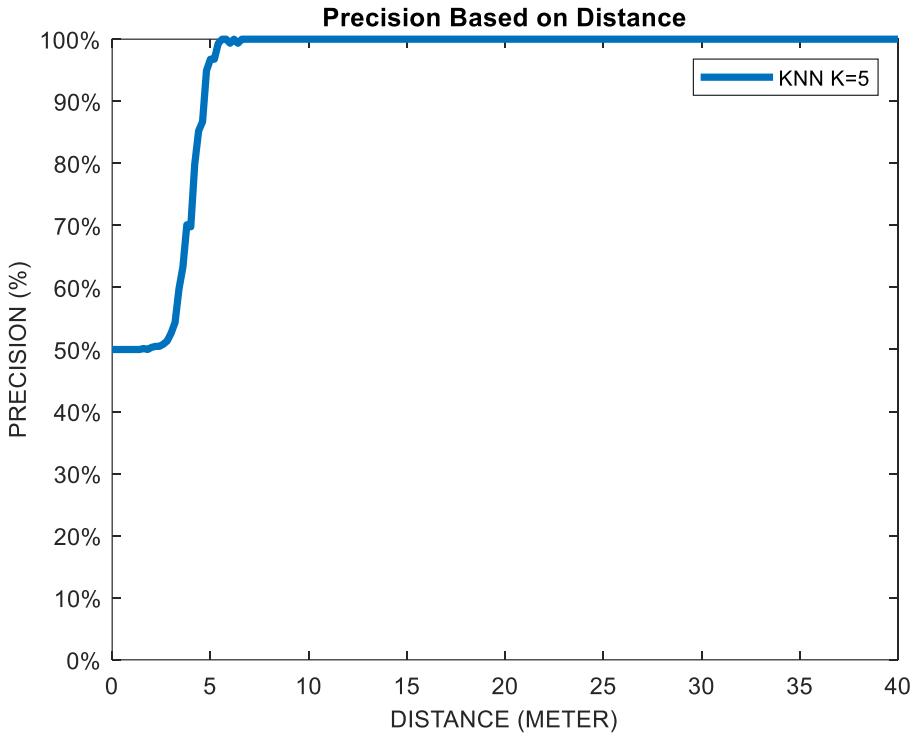
نمودارِ دقت بر اساسِ فاصله برای کلاسفایرِ KNN برای K=3



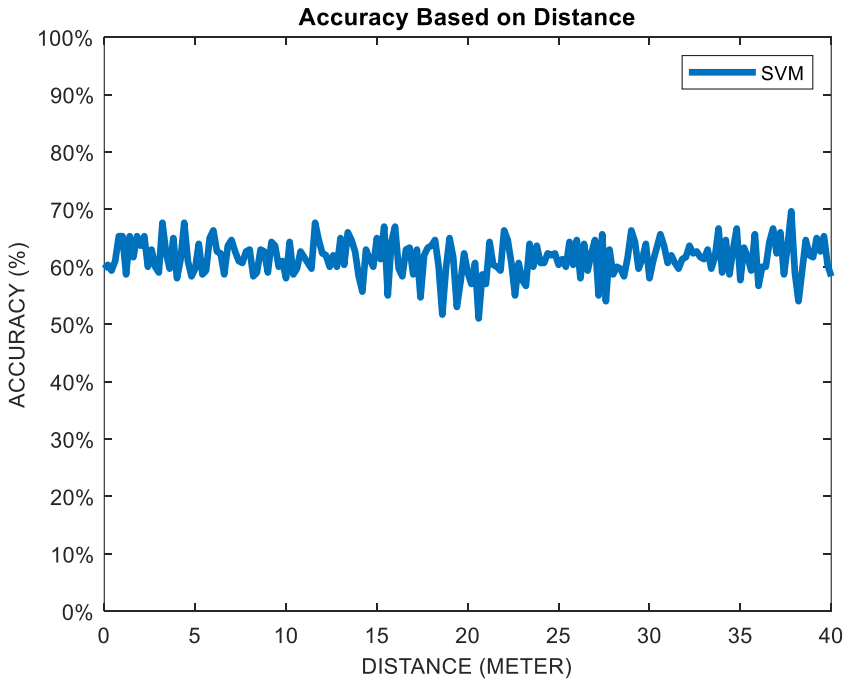

نمودارِ صحت بر اساسِ فاصله برای کلاسفایرِ KNN برای K=5



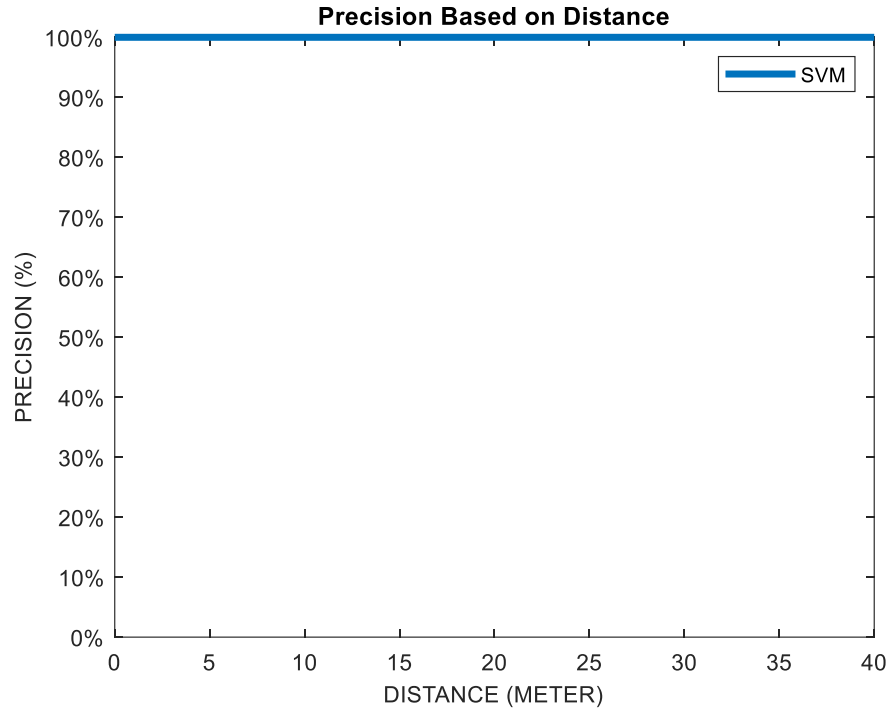
نمودارِ دقت بر اساسِ فاصله برای کلاسفایرِ KNN برای K=5



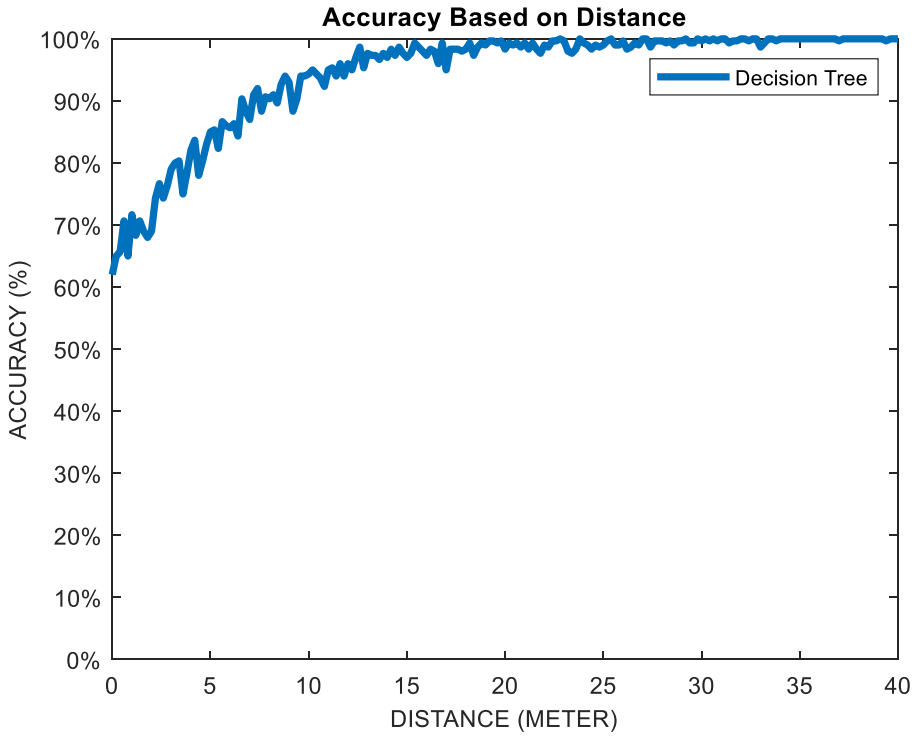

نمودارِ صحت بر اساسِ فاصله برای کلاسفایرِ SVM



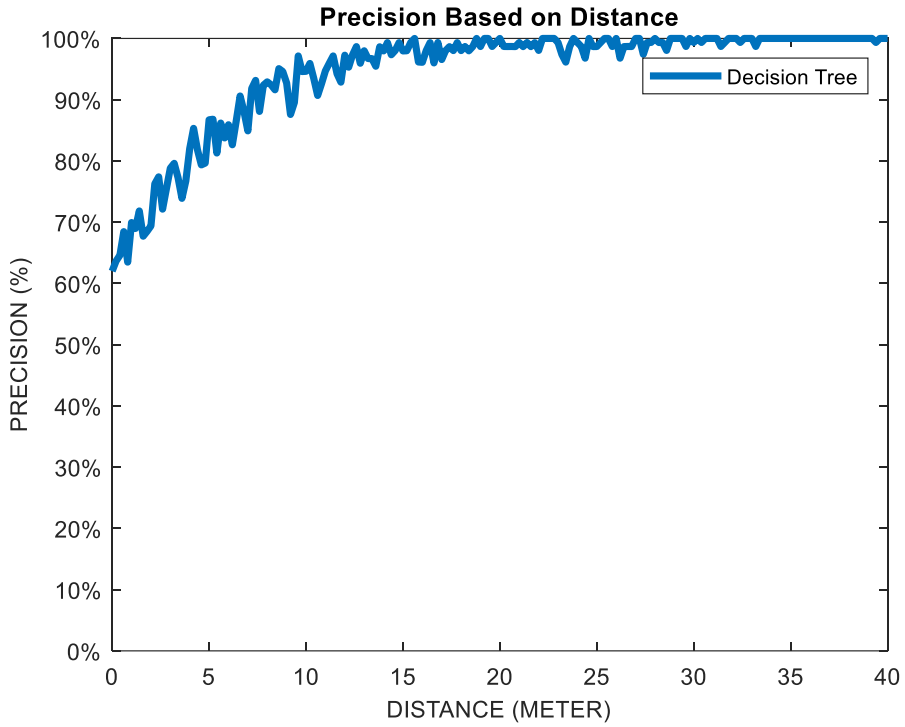
نمودارِ دقت بر اساسِ فاصله برای کلاسفایرِ SVM




نمودارِ صحت بر اساسِ فاصله برای کلاسفایرِ Decision Tree



نمودارِ دقت بر اساسِ فاصله برای کلاسفایرِ **Decision Tree**

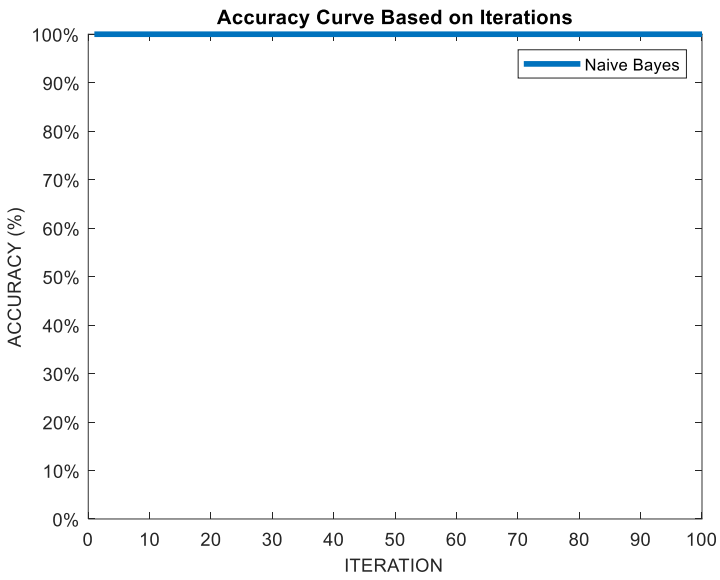



همانطور که از نتایج برمی‌آید، دقت و صحت در دسته‌بندیِ داده‌ها با فاصله رابطه‌ی عکس دارد؛ بدین صورت که هر اندازه که فاصله بیشتر باشد، دقت و صحتِ کلاسیفایرها در دسته بندی بیشتر خواهد شد و هر چه فاصله کمتر باشد نیز صحت و دقتِ کلاسیفایرها در کلاس بندی کمتر است که از شکل‌های فوق نیز کاملا این موضوع مشهود است.

**بخش d)** مطابق خواسته‌ی این بخش داده‌ها را به صورتِ رندوم به داده‌های آموزش و تست تقسیم کردیم به گونه‌ای که 0.7 داده‌ها برای آموزش و مابقی برای تست بکار گرفته شوند و در نهایت طبق خواسته‌ی سوال انحراف معیار و میانگین را برای صحت و دقتِ این داده‌ها محاسبه نمودیم. همچنین توجه داشته باشید که محورِ افقی در این بخش iteration های هر بار اجرا می‌باشد که در هر یک از این iteration ها داده‌ها را متناسب با نسبتی که عرض کردم به دو گروهِ آموزش و تست تقسیم کردیم.

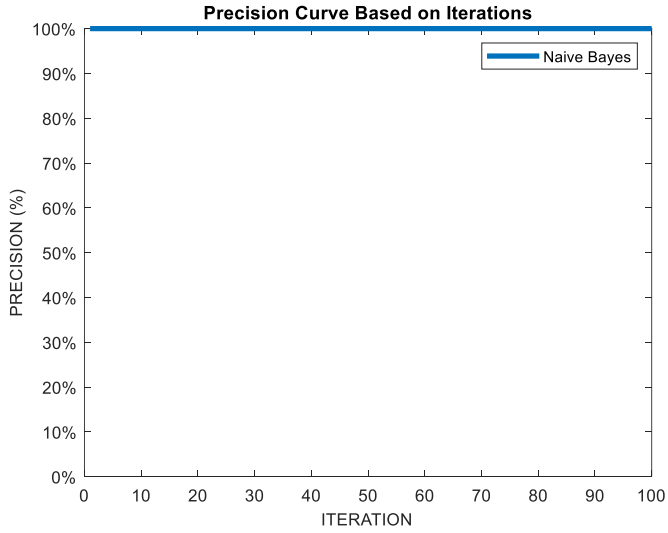
توجه داشته باشید که فاصله‌ی بین کلاس‌ها را 4 (متر) در نظر گرفته‌ایم که اغلبِ کلاسیفایرها صحت و دقت خوبی را برای مقایسه داشتند. در نهایت نتایج به صورت زیر بدست آمد.

## منحنیِ صحت برای داده‌های متغیرِ آموزش و تست از یک دیتاستِ ثابت در کلاسیفایرِ Naïve Bayes



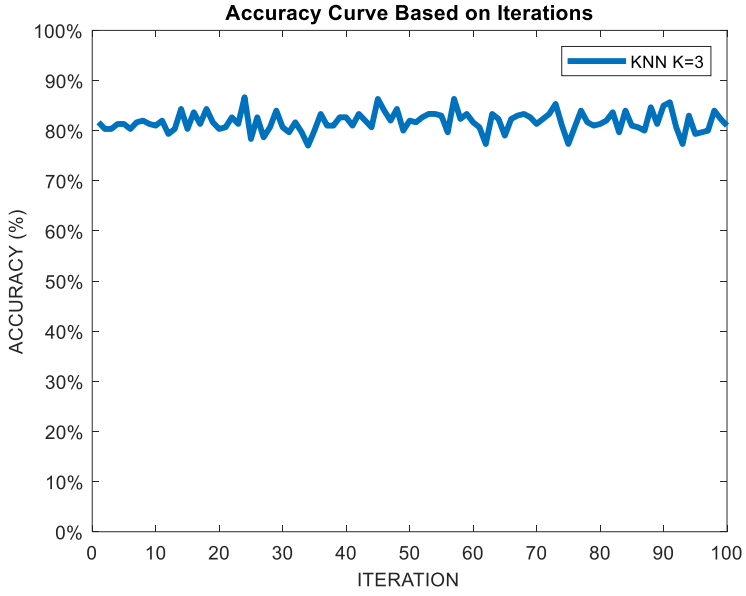

| | |
|---|---|
| Average accuracy | 1 |
| Standard deviation | 0 |

## منحنیِ دقت برای داده‌های متغیرِ آموزش و تست از یک دیتاستِ ثابت در کلاسیفایرِ Naïve Bayes



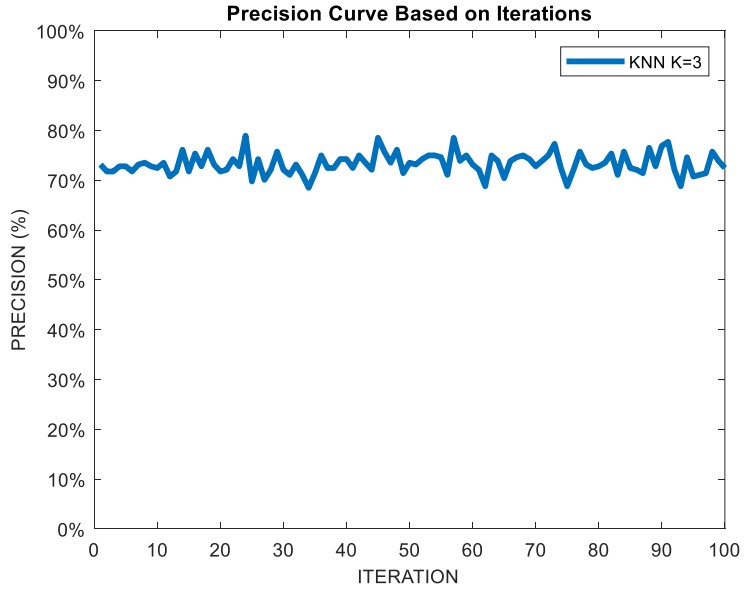

| | |
|---|---|
| Average precision | 1 |
| Standard deviation | 0 |

منحنیِ صحت برای داده‌های متغیرِ آموزش و تست از یک دیتاستِ ثابت در کلاسیفایرِ KNN برای K=3



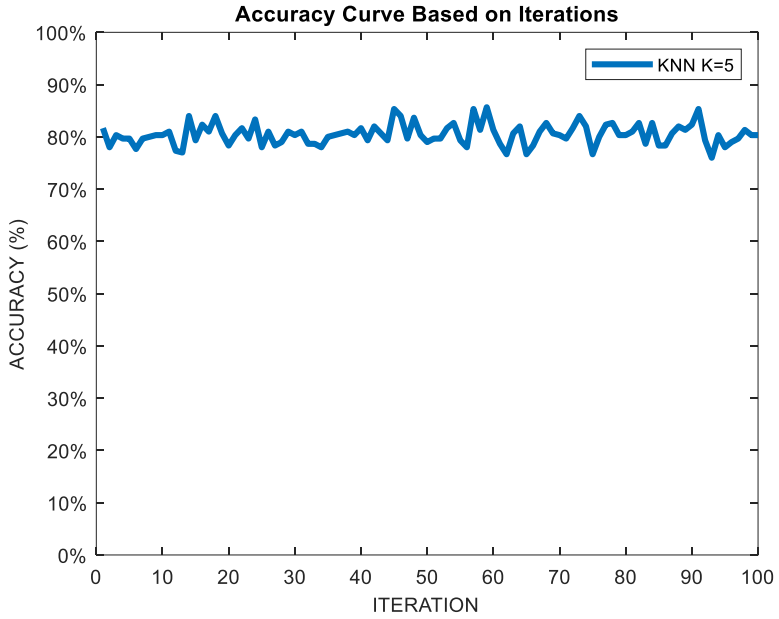

| | |
|---|---|
| Average accuracy | 0.8175 |
| Standard deviation | 0.0196 |

منحنیِ دقت برای داده‌های متغیرِ آموزش و تست از یک دیتاستِ ثابت در کلاسیفایرِ KNN برای K=3



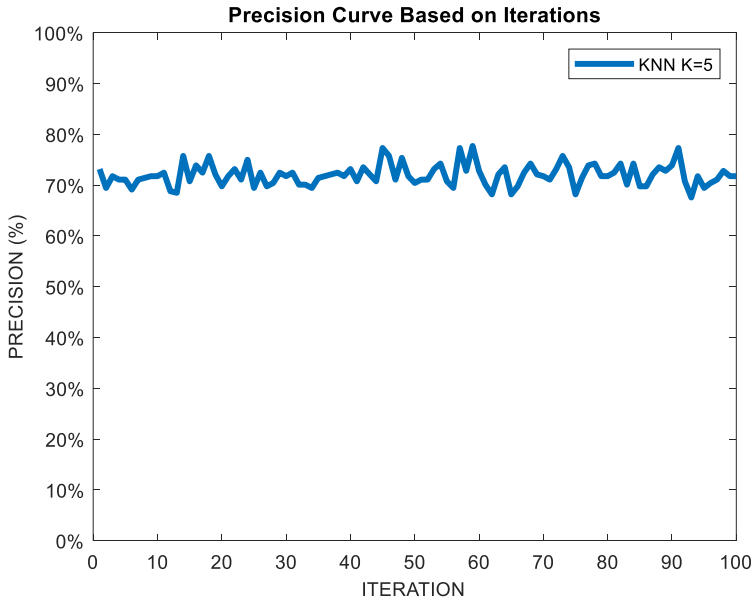

| | |
|---|---|
| Average precision | 0.7332 |
| Standard deviation | 0.0212 |

منحنیِ صحت برای داده‌های متغیرِ آموزش و تست از یک دیتاستِ ثابت در کلاسیفایرِ KNN برای K=5



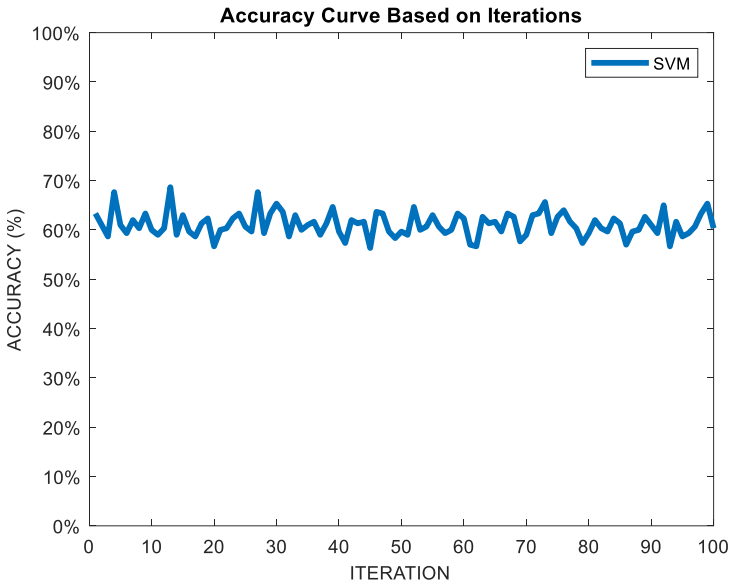

| | |
|---|---|
| Average accuracy | 0.8048 |
| Standard deviation | 0.0201 |

منحنیِ دقت برای داده‌های متغیرِ آموزش و تست از یک دیتاستِ ثابت در کلاسیفایرِ KNN برای K=5



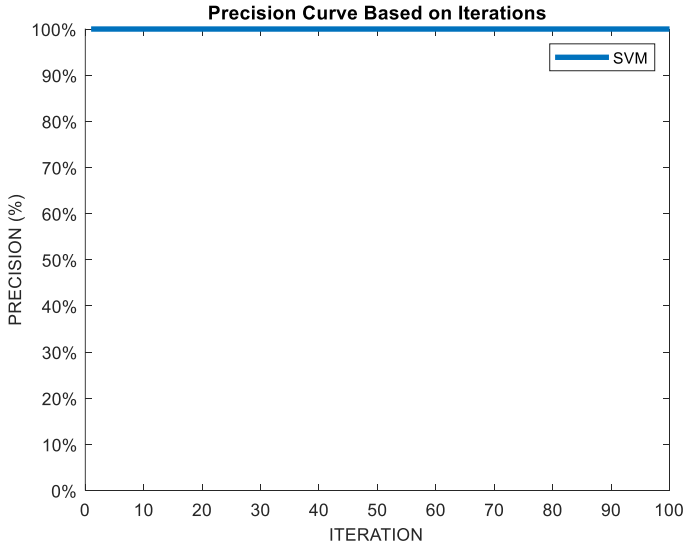

| | |
|---|---|
| Average precision | 0.7199 |
| Standard deviation | 0.0211 |

منحنیِ صحت برای داده‌های متغیرِ آموزش و تست از یک دیتاستِ ثابت در کلاسیفایرِ SVM



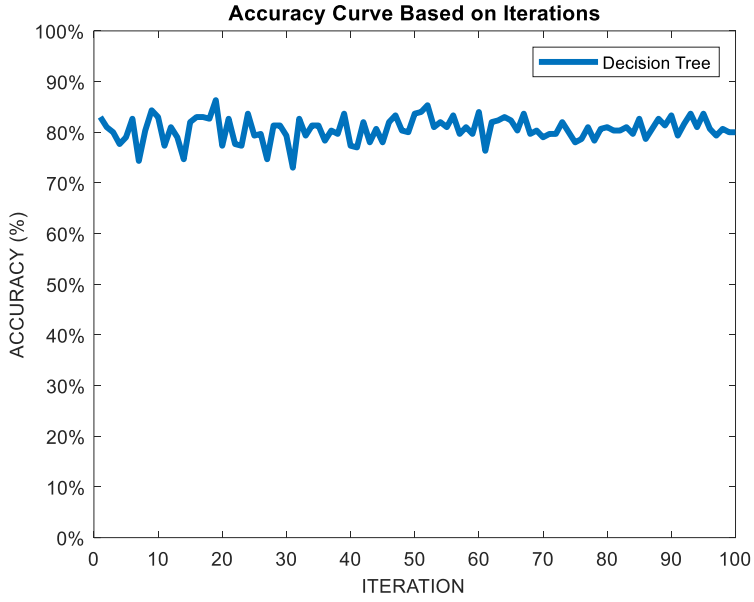

| | |
|---|---|
| Average accuracy | 0.6111 |
| Standard deviation | 0.0246 |

منحنیِ دقت برای داده‌های متغیرِ آموزش و تست از یک دیتاستِ ثابت در کلاسیفایرِ SVM



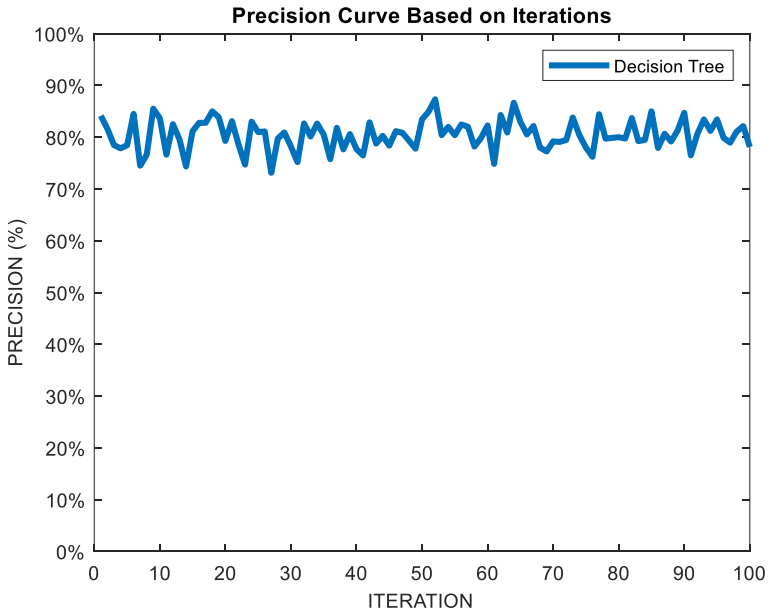

| | |
|---|---|
| Average precision | 1 |
| Standard deviation | 0 |

منحنیِ صحت برای داده‌های متغیرِ آموزش و تست از یک دیتاستِ ثابت در کلاسیفایرِ **Decision Tree**




| | |
|---|---|
| Average accuracy | 0.8060 |
| Standard deviation | 0.0240 |

منحنیِ دقت برای داده‌های متغیرِ آموزش و تست از یک دیتاستِ ثابت در کلاسیفایرِ **Decision Tree**

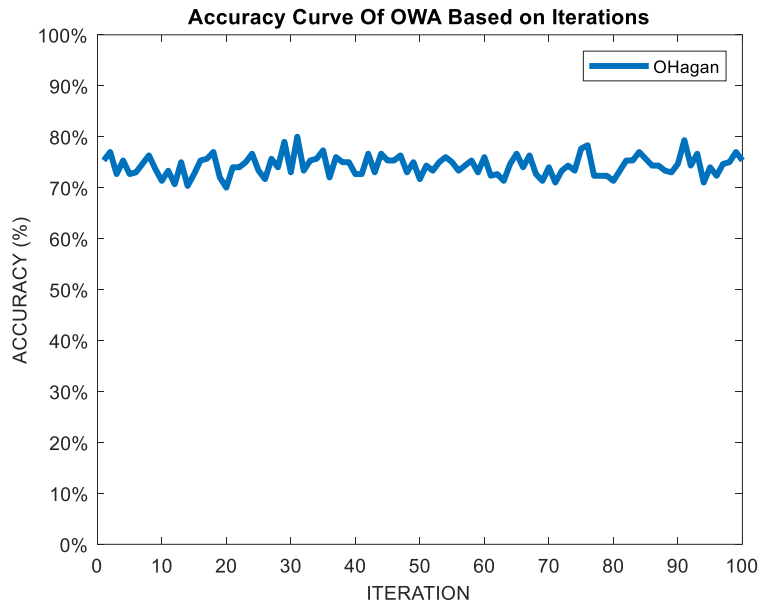


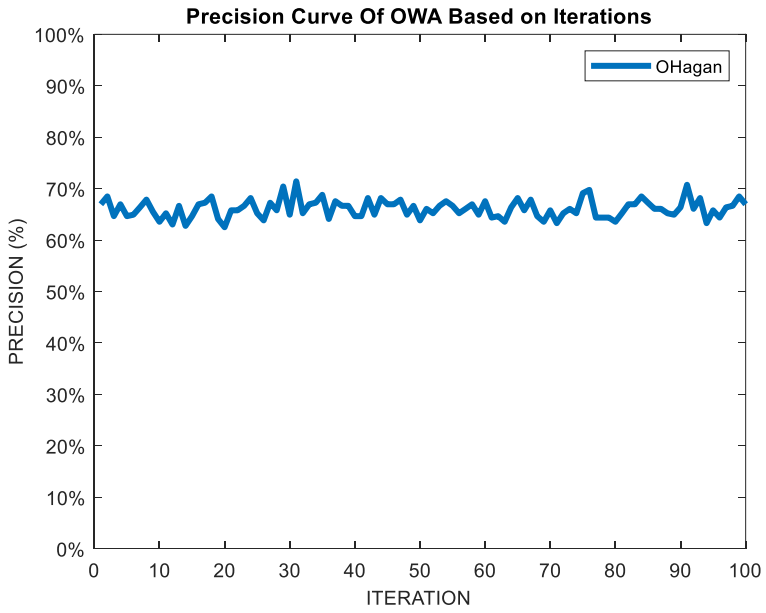

| | |
|---|---|
| Average precision | 0.8045 |
| Standard deviation | 0.0289 |

**بخش e)** در این بخش کلاسیفایرهای بخش قبل را با یکدیگر ادغام کردیم.

نتایج به صورت زیر در آمد:

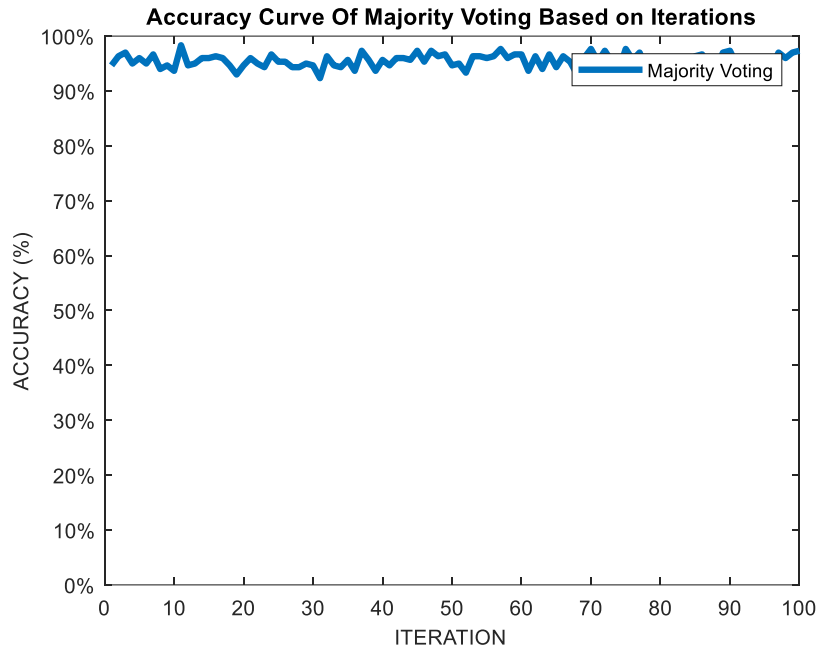
ترکیب کلاسیفایرها OWA به روش O'Hagan :

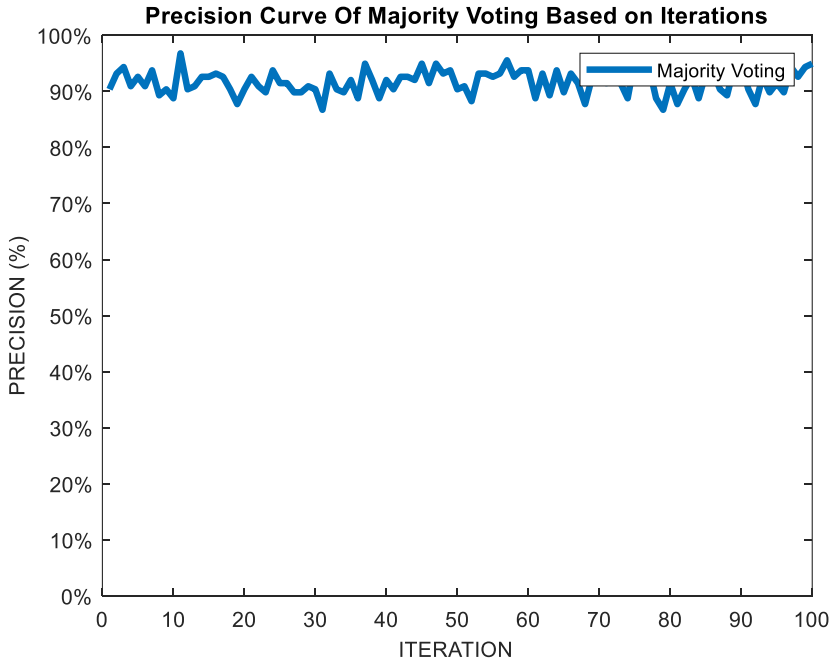



| | |
|---|---|
| Average accuracy | 0.7432 |
| Standard deviation of accuracy | 0.0202 |
| Average precision | 0.6612 |
| Standard deviation of precision | 0.0178 |

ترکیبِ کلاسیفایرها با روش Majority voting:

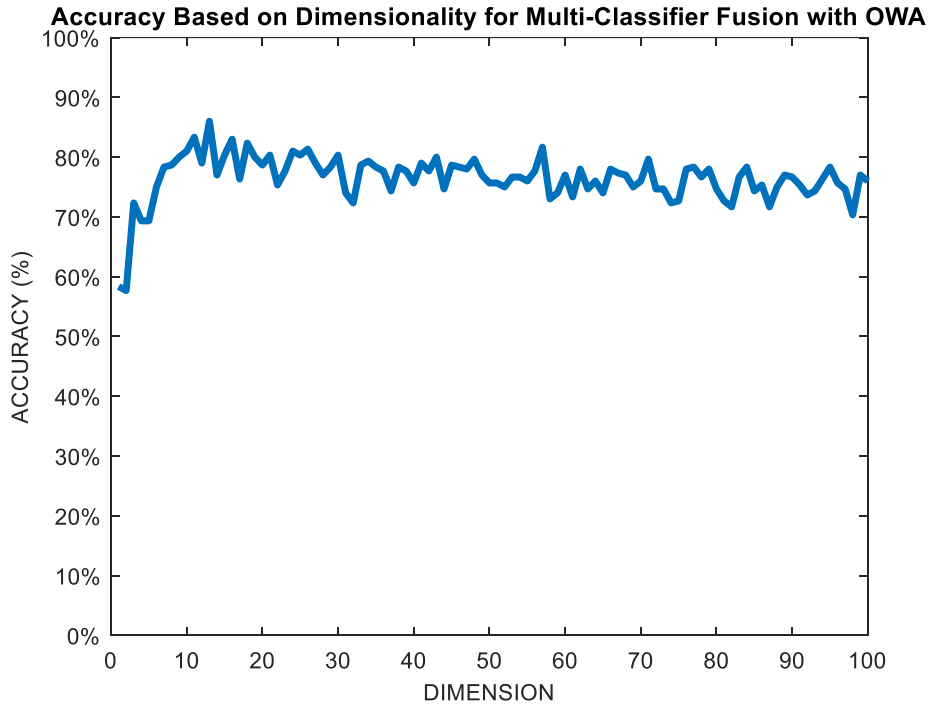


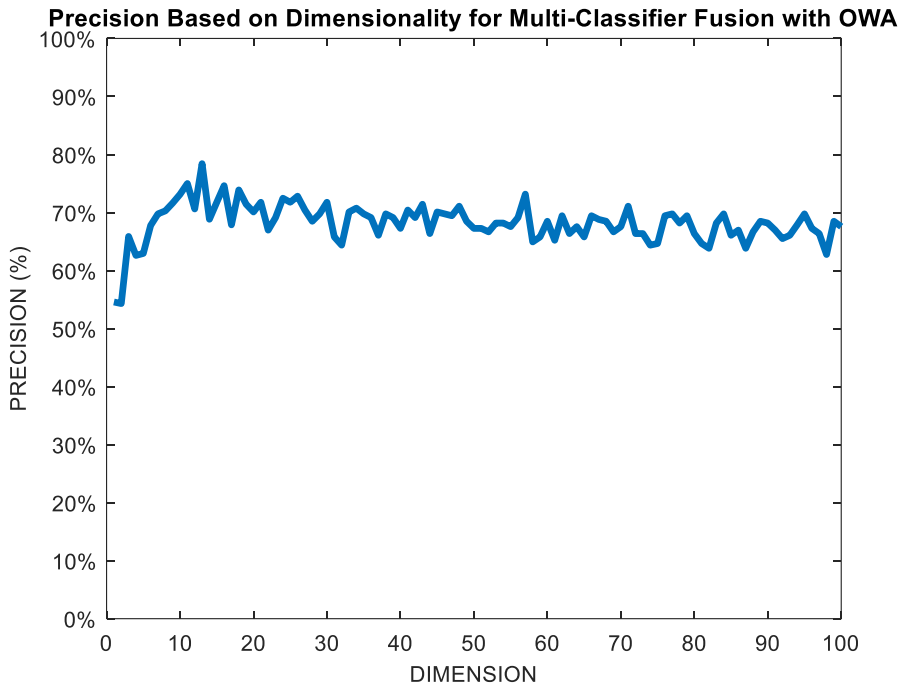



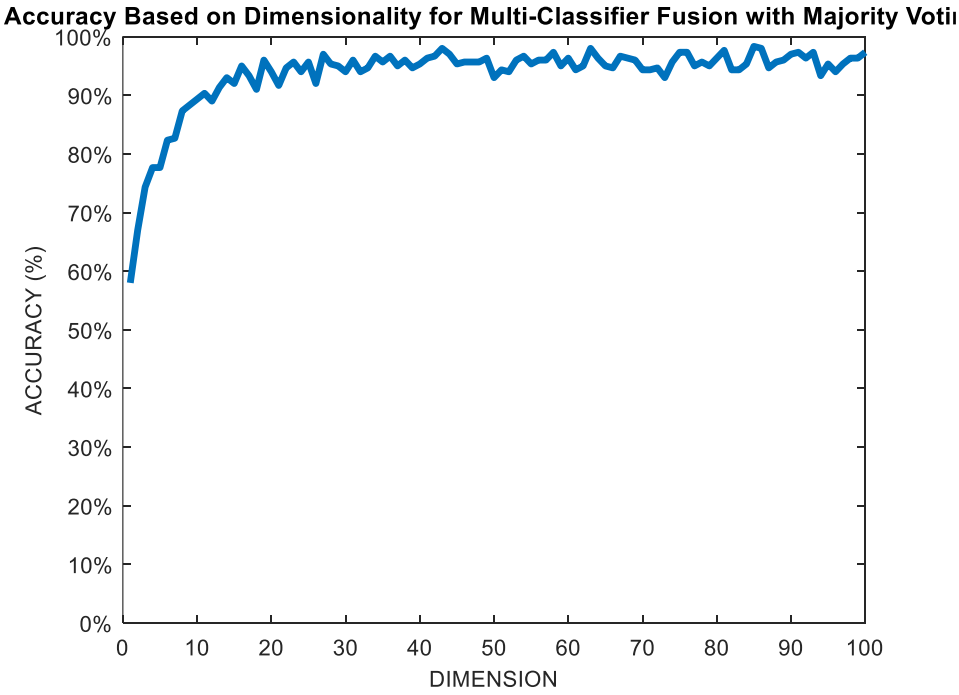

| | |
|---|---|
| Average accuracy | 0.9568 |
| Standard deviation of accuracy | 0.0114 |
| Average precision | 0.9208 |
| Standard deviation of precision | 0.0192 |

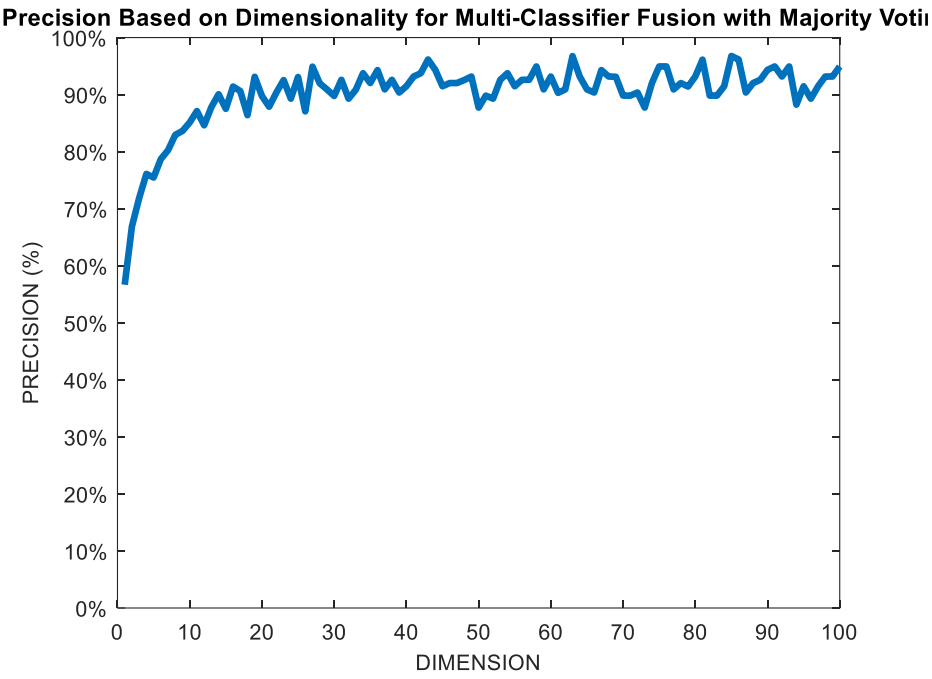
همانطور که از نمودارهای فوق مشاهده می‌کنید با استفاده از ترکیب کردنِ کلاسیفایرها با یکدیگر توانستیم به یک حد مطلوب تری از دسته بندی داده‌ها برسیم. در واقع روش Majority voting بسیار خوب عمل کرده و میانگین صحت و دقتِ بالایی را در تشخیص به وجود آورده ولی در روشِ OWA این بهبودی کمتر مشهود است. به نظر من دلیلِ این امر آن است که احتمالِ تشخیص در هر یک کلاسیفایرها مشخص نیست و لذا بیشترین وزن به ان کلاسیفایری تخصیص می‌یابد که برچسب را اشتباه تشخیص داده است.

**بخش f)** در این بخش خواسته شده است که تغییر ایعاد (فیچرها) داده‌ها را بر روی نمودار بررسی کنیم که نتایج این بخش نیز به صورت زیر می‌باشد. (فاصله را 4 متر گرفتیم)




Precision Based on Dimensionality for Multi-Classifier Fusion with OWA
PRECISION (%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
10
20
30
40
50
60
70
80
90
100
DIMENSION


Accuracy Based on Dimensionality for Multi-Classifier Fusion with Majority Voti
ACCURACY (%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
10
20
30
40
50
60
70
80
90
100
DIMENSION




همانطور که از نمودارهای خروجی گرفته شده در این بخش مشاهده میکنید، هر اندازه که ابعاد بیشتر می‌شود و به عبارتی دیگر فیچرهای یک داده افزایش می‌یابند، به مرور دقت و صحت افزایش پیدا می‌کند تا به یک حدِ مشخصی برسد که اضافه کردنِ ابعاد خیلی کارایی نداشته باشد. به نظر من این بحثِ underfitting را به گونه‌ای بیان می‌کند به گونه‌ای که هر چه ابعاد کم باشد مشخصه‌ها برای شناساییِ دسته‌ی داده مافی نخواهند بود و هر چه زیاد باشند نیز از یکجا به بعد تاثیرِ زیادی را در افزایشِ صحت و دقت نخواهد گذاشت و اگر برچسب ها زیاد باشند نیز ممکن است به گونه‌ای overfit هم کند.

در کل می‌توان روند نمودار را از منظر کارایی اینطور بیان کرد که با افزایشِ ابعاد ابتدا کارایی افزایش می‌یابد و در صورتِ بیشتر شدن از یک حدی کاهش خواهد یافت.

با تشکر - بدیعی