**Introduction**

Soccer is a team sport that include multiple roles for players such as goalkeeper, defender, midfielder, and striker. Each position also requires different physical abilities and skills to perform their duties effectively. For example, a striker is responsible for scoring goals and requires better attributes on pace, shooting, and attacking positioning. On the other hand, a defender needs to stop opposing teams' attacks and requires attributes such as tackling, marking, and interceptions. Seeking ways to gain a competitive edge, the use of data analytic techniques to fit and predict a player’s possible performances is an emerging trend in soccer. For instance, the German national team utilized data analytics to inform their training and match strategies, which led to their success in the 2014 World Cup.

In this research, the classification of players into strikers or defenders based on their physical attributes and skills is performed, which is crucial in player recruitment, team formation planning, and match strategy development. For instance, a team may use this analysis to recruit a striker who fits their attacking style and has a high probability of scoring goals. Additionally, coaches can use this analysis to develop a formation that maximizes the strengths of their players and neutralizes the opposition's strengths. Therefore, we need a quick way to judge whether a player should be placed in an attacking position or at a defending position on the field given some physical data such as max speed, passing ability, dribbling skills, and pace that could be quantified and measured for each player. Also, in modern soccer, the role of whether a defending player should act as a traditional back (LB / RB) or as a modern wing back (LWB / RWB), which would require more attacking abilities, is getting more important, as teams are seeking a way to break through the traditional tactics. This research also analyzes on how should we determine whether a player is better as back or wing back in this game, given the physical quantified data.

**Statistical Method**

1. Data description

The original FIFA20 dataset is obtained from Kaggle (FIFA 20 Complete Player Dataset, n.d.). This dataset contains FIFA20 player data. We pre-process the dataset with python to only obtain the data columns that we want to use in this final project and removed all data of Goalkeepers, which has a different record form than other players. The updated dataset (Players\_20\_edited.csv, 2023) contains 16242 non-goalkeeper soccer players and their physical data, scores on skills, and positions.

2. Classification Method

We used Support Vector Machines (SVM) as the main statistical method to classify whether a player is better to play as a striker or as a defender on the field. SVM is a well-established method for classification, which has been used in many fields, including image recognition, bioinformatics, and finance (Cortes & Vapnik, 1995; Cristianini & Shawe-Taylor, 2000). We selected the SVM method because it can effectively deal with high-dimensional data, and it is capable of finding nonlinear decision boundaries (Hsu & Lin, 2002).

In the two classifications, we applied simple Linear SVM, tuned Linear SVM, Radial Kernel SVM, and Polynomial Kernel SVM and compares in their effects of classification for the training data and k-fold cross validation is applied during training to prevent overfitting of the model and provide a more accurate estimate of the model's performance on unseen data. (Hastie et al., 2001)

**Results**

(1) Classification – Striker vs Defender

In this classification task, we try to train a classifier that classifies a player to be Striker (ST) or Center Back (CB). There are 7835 players for CB and 8407 players for ST in the whole dataset. After splitting the dataset into a Training set and a Testing set with ratio 8:2, we also tested the performance of the models with some famous players deliberately picked out to show the accuracy of the models. The famous players include three STs, three CBs, LW, RW, LB, RB, and several midfield players.

(a) Simple Linear SVM

In this part, we fit a Simple Linear SVM (Cost = 1) on the training data and train with 10-fold cross-validation. The training results in an accuracy of 0.9938. Table 1 below shows the statistics of testing on Test Set and Famous Players.

**Table 1**

*Test Results on Simple Linear SVM (Cost = 1)*

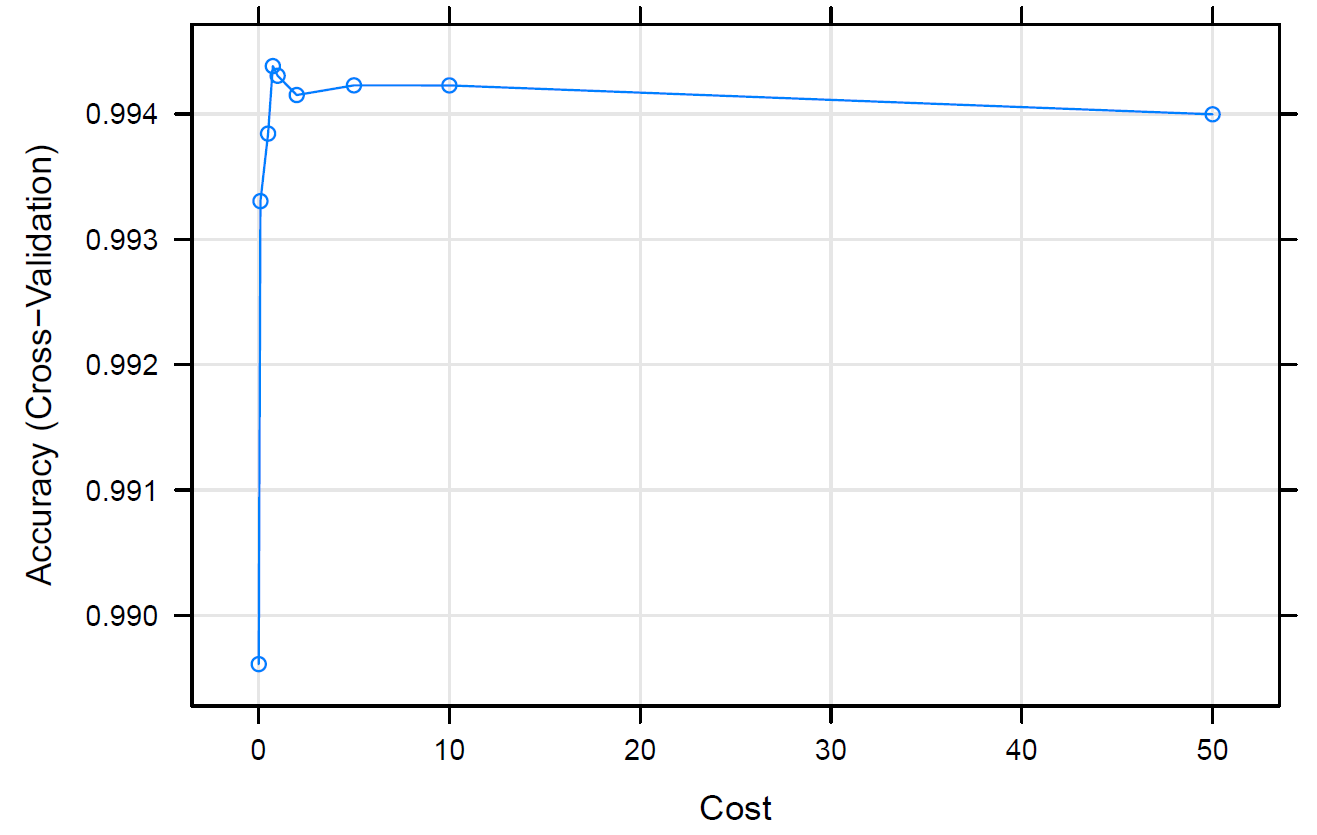
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Sensitivity | Specificity | Precision | F-1 Score |
| Test Set | 0.9926 | 0.9929 | 0.9923 | 0.9929 | 0.9929 |
| Famous Players | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

(b) Tuned Linear SVM

In this part, we tune a Linear SVM on the training data with the Cost and train with 10-fold cross-validation. The training picks a best Cost of 0.75 with accuracy of 0.9944. Figure 1 below shows the tuning curve with each cost values. below shows the statistics of testing on Test Set and Famous Players.

**Figure 1**

*Tune Results of Linear SVM on Training Set*



**Table 2**

*Test Results on Tuned Linear SVM (Cost = 0.75)*

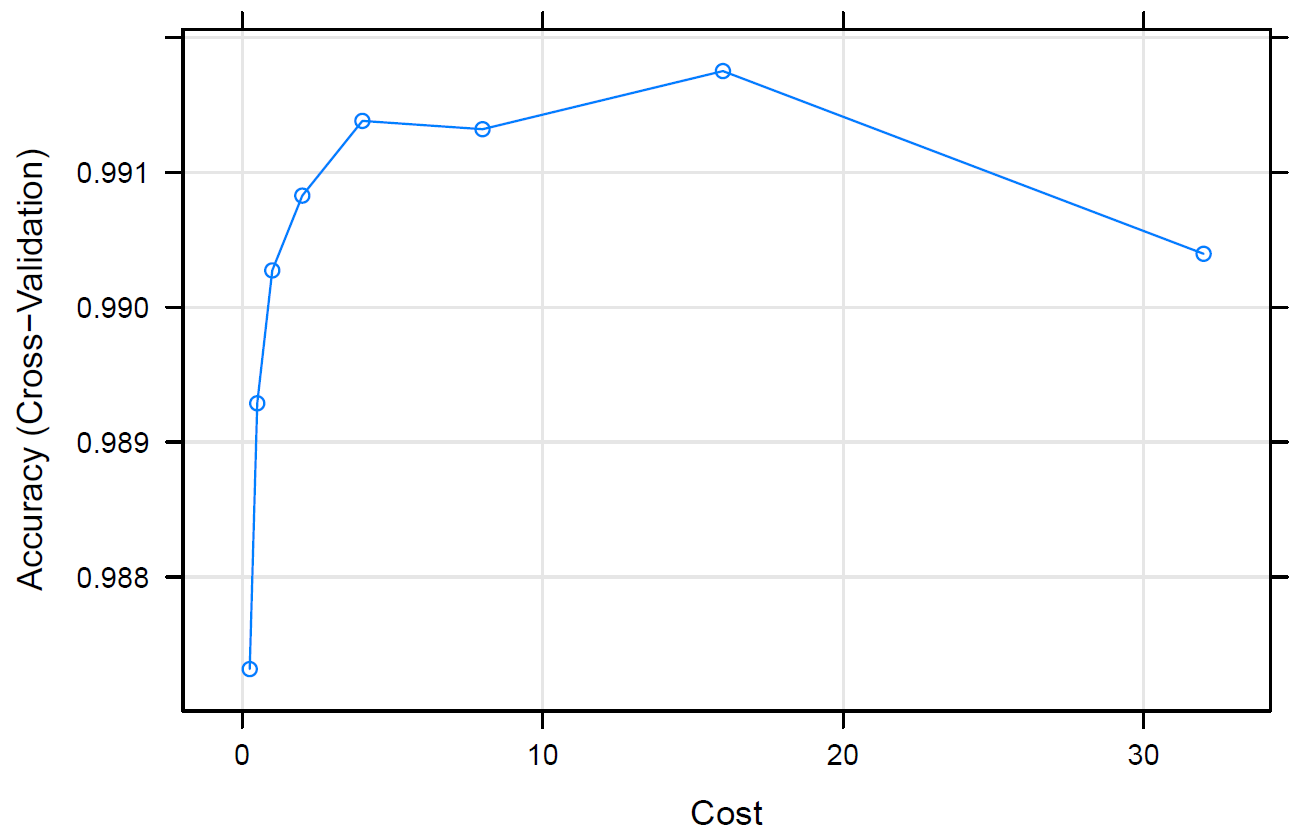
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Sensitivity | Specificity | Precision | F-1 Score |
| Test Set | 0.9926 | 0.9923 | 0.9917 | 0.9935 | 0.9929 |
| Famous Players | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

(c) Radial SVM

In this part, we train a Radial Kernel SVM on the training data and train with 10-fold cross-validation. The training picks a best Cost of 32 with accuracy of 0.9900. Figure 2 below shows the tuning curve with each cost values. Table 3 below shows the statistics of testing on Test Set and Famous Players.

**Figure 2**

*Tune Results of Radial SVM on Training Set*



**Table 3**

*Test Results on Tuned Radial Kernel SVM (Cost = 32)*

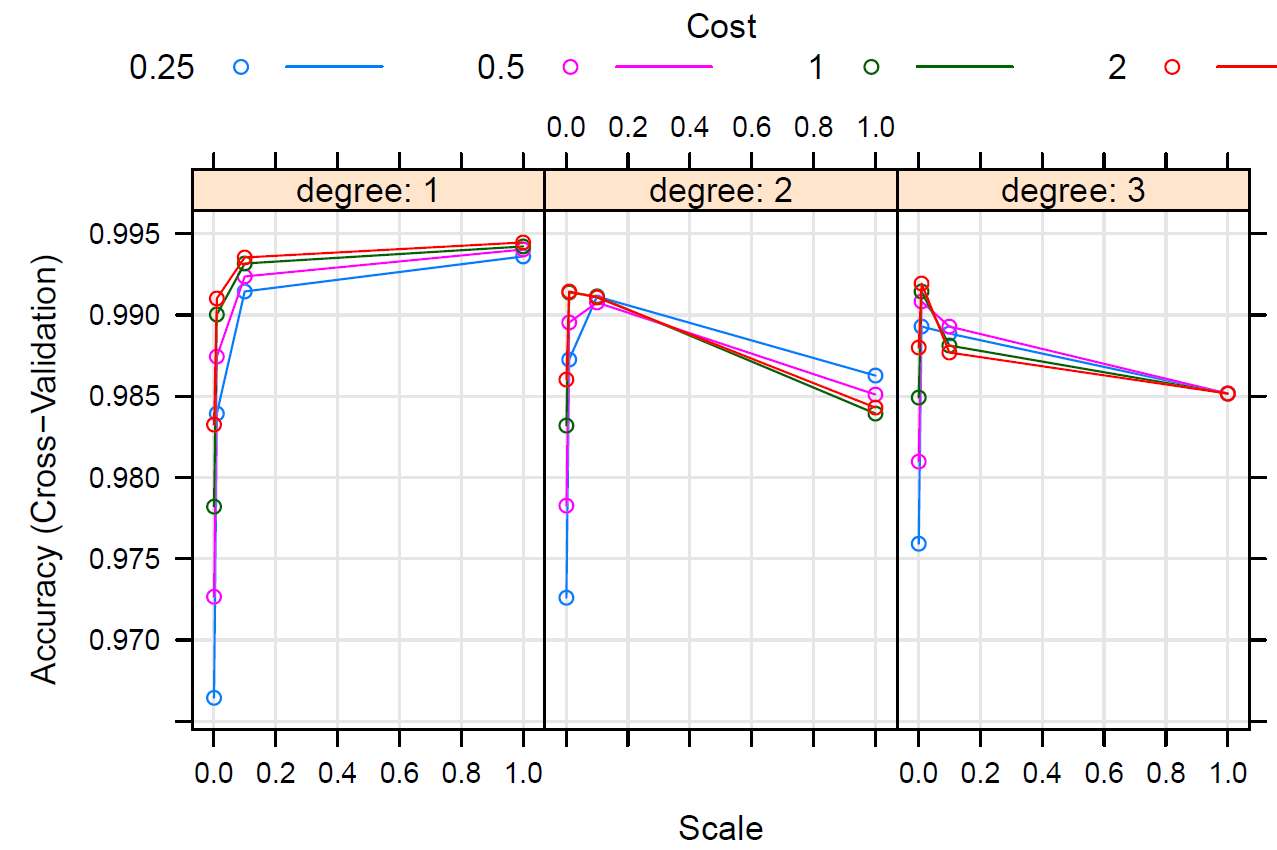
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Sensitivity | Specificity | Precision | F-1 Score |
| Test Set | 0.9994 | 0.9994 | 0.9994 | 0.9994 | 0.9994 |
| Famous Players | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

(d) Polynomial SVM

In this part, we train a Polynomial Kernel SVM on the training data and train with 10-fold cross-validation. Figure 3 below shows the tuning curve with each cost and scale. Table 4 below shows the statistics of testing on Test Set and Famous Players.

**Figure 3**

*Tune Results of Polynomial SVM on Training Set*



**Table 4**

*Test Results on Tuned Radial Kernel SVM (Cost = 32)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Sensitivity | Specificity | Precision | F-1 Score |
| Test Set | 0.9951 | 0.9964 | 0.9936 | 0.9941 | 0.9953 |
| Famous Players | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

(e) Report

Based on the training performances and test results, we choose to go with the Radial Kernel SVM to classify on the position of a player.

(2) Classification – Back vs Wing Back

In this classification task, we try to train a classifier that classifies a player to be traditional Back (LB/RB) or modern Wing Back (LWB/RWB). There are 12492 players better as LWB/RWB and 3750 players for LB/RB in the whole dataset.

**Table 5**

*Test Results on Test Set*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Sensitivity | Specificity | Precision | F-1 Score |
| Simple Linear SVM (Cost = 1) | 0.9594 | 0.9708 | 0.9185 | 0.9770 | 0.9739 |
| Tuned Linear SVM (Cost = 0.5) | 0.9606 | 0.9709 | 0.9237 | 0.9786 | 0.9747 |
| Radial Kernel SVM | 0.9606 | 0.9705 | 0.9249 | 0.9790 | 0.9747 |
| Polynomial Kernel SVM | 0.9603 | 0.9701 | 0.9248 | 0.9790 | 0.9745 |

(a) Simple Linear SVM

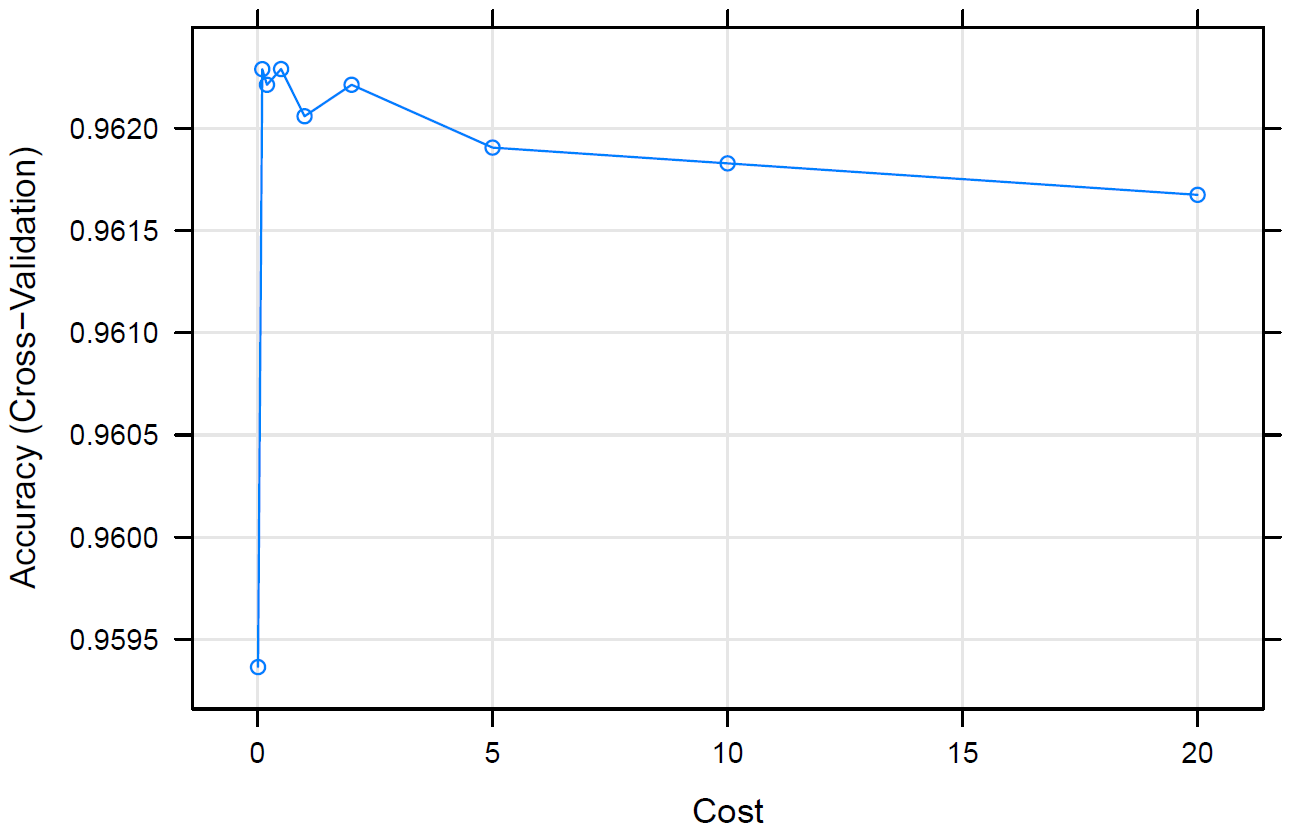
In this part, we fit a Simple Linear SVM (Cost = 1) on the training data and train with 10-fold cross-validation. The training results in an accuracy of 0.9609. Table 5 shows the statistics of testing on Test Set.

(b) Tuned Linear SVM

In this part, we tune a Linear SVM on the training data with the Cost and train with 10-fold cross-validation. The training picks a best Cost of 0.5 with accuracy of 0.9623. Figure 4 shows the tuning curve with each cost values. Table 5 shows the statistics of testing on Test Set.

**Figure 4**

*Tune Results of Linear SVM on Training Set*

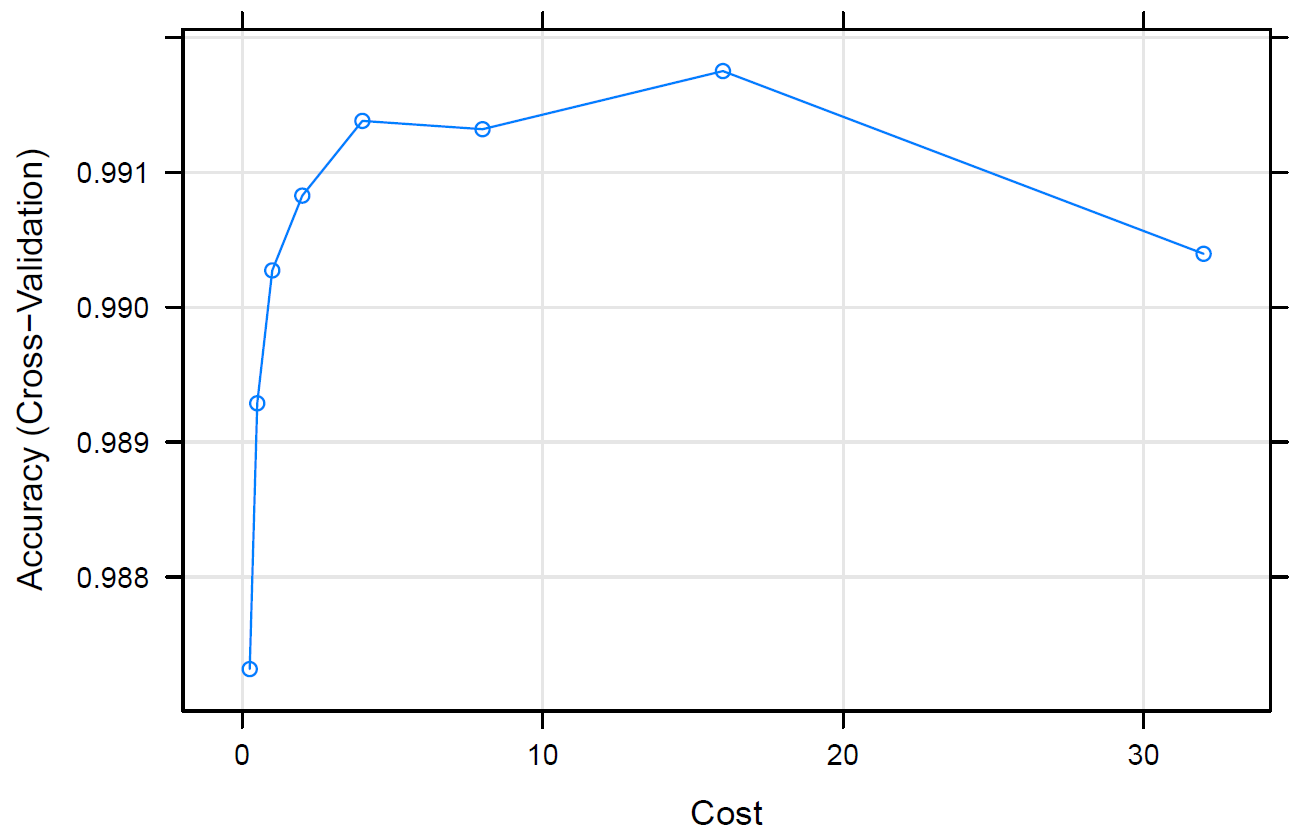


(c) Radial SVM

In this part, we train a Radial Kernel SVM on the training data and train with 10-fold cross-validation. The training picks a best Cost of 8 with accuracy of 0.9580. Figure 5 shows the tuning curve with each cost values. Table 5 shows the statistics of testing on Test Set.

**Figure 5**

*Tune Results of Radial SVM on Training Set*

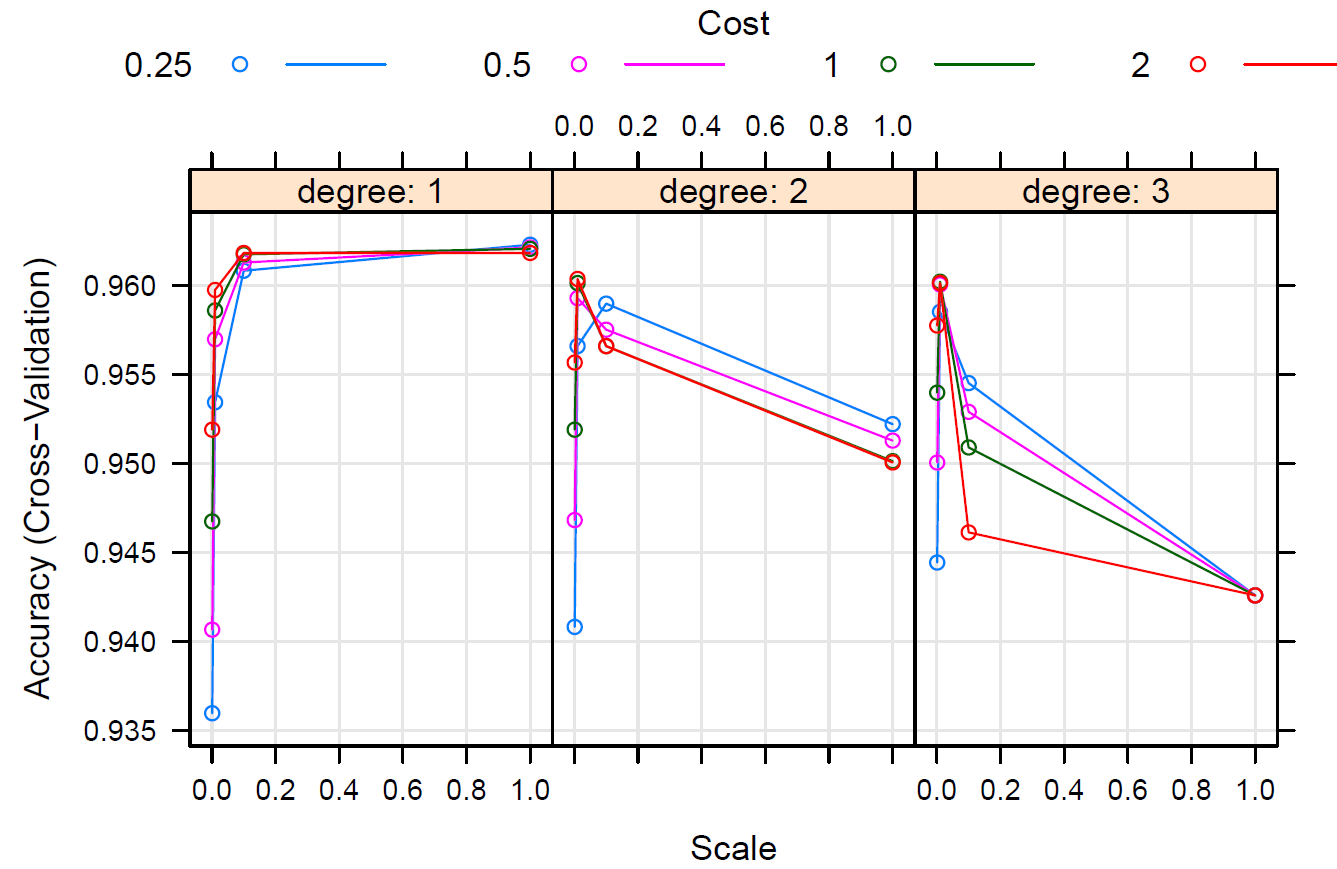


(d) Polynomial SVM

In this part, we train a Polynomial Kernel SVM on the training data and train with 10-fold cross-validation. Figure 6 shows the tuning curve with each cost and scale. Table 5 shows the statistics of testing on Test Set.

**Figure 6**

*Tune Results of Polynomial SVM on Training Set*



(e) Report

Based on the training performances and test results, we choose to go with the Radial Kernel SVM to classify on the position of a player. Though the accuracy of Tuned Linear SVM and Radial Kernel SVM are basically the same, we should still go with Radial Kernel SVM as it provide us with better interpretation on high-dimensional data, such as the FIFA dataset.

**Discussion**

(1) Classification

From the results of the two classifications, we both choose Radial Kernel SVM as the final classifier, and they both achieve a good performance on the test set. Radial Kernel SVM can capture non-linear relationships between the features and the target variable. It’s also effective with high-dimensional data like the FIFA dataset, while it is robust to noise. Therefore, it’s theoretically and practically right to use a Radial Kernel SVM.

One problem we need to fix in future researches is that we directly train the classifiers on all feature variables, which could lead to overfitting. In the later researches, we might apply variable selection methods such as Lasso or BIC selection before the actual training of models. Other details that might be improved includes a theoretical derivation of to what extent should we tune the parameters so that it’s both mathematically make sense and doesn’t take a lot of computations to run. This would make the training faster and more accurate.

**Reference**

Bauer, E., & Kohavi, R. (1999). Machine Learning, 36(1/2), 105–139. <https://doi.org/10.1023/a:1007515423169>

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. <https://doi.org/10.1007/bf00994018>

Cristianini, N., & Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. <https://doi.org/10.1017/cbo9780511801389>

Chih-Wei Hsu, & Chih-Jen Lin. (2002). A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks, 13(2), 415–425. <https://doi.org/10.1109/72.991427>

FIFA 20 complete player dataset. (n.d.). Www.kaggle.com. Retrieved May 6, 2023, from <https://www.kaggle.com/datasets/stefanoleone992/fifa-20-complete-player-dataset?resource=download&select=players_20.csv>

Hastie, T., Friedman, J., & Tibshirani, R. (2001). The Elements of Statistical Learning : Data Mining, Inference, and Prediction. Springer.

players\_20\_edited.cs Retrieved May 6, 2023, from <https://uofi.app.box.com/s/91m9qt18v7zhxh00d18edv7fx6x95fsw>