Dokumentation Machine Learning und Intelligente Datenanalyse

Projekt 7 - Raphael Kluge

Vorbereitung:

Daten

- Datei laser.mat wird eingelesen
- Konvertieren von y zu einem Vektor
- Plotten von Beispielen zum näheren Verständnis der Daten
- Aufteilen der Daten in Trainings- und Testdaten

Methoden:

Kernel-Funktionen

- SVM mit regressivem ERM
- Regularisierer: L2-Regularisierer = C $||\theta||^2$
- Verlustfunktion: Hinge loss (L= max(0, 1 – y*f(x)):
  - gut für SVM, binäre Daten
- Mit Kernel-Funktion anpassen und trainieren
- Plotten der Training- und Test-Scores

Linearer Kernel: $k(x_i, x_j) = x_i^T x_j$
Polynomischer Kernel: $k(x_i, x_j) = (x_i^T x_j + 1)^p$
RBF Kernel: $k(x_i, x_j) = \exp(-\gamma ||x1 – x2||^2)$

→ Anpassen der Hyperparameter wie C und p

Random-Forest

- RandomForestClassifier aus sklearn
- Funktion, die unterschiedliche Anzahl an Estimatorn nimmt
- gibt Classifier sowie Training Accuracy und Test Accuracy zurück
- Plotten der Ergebnisse mit unterschiedlicher Anzahl an Estimatorn

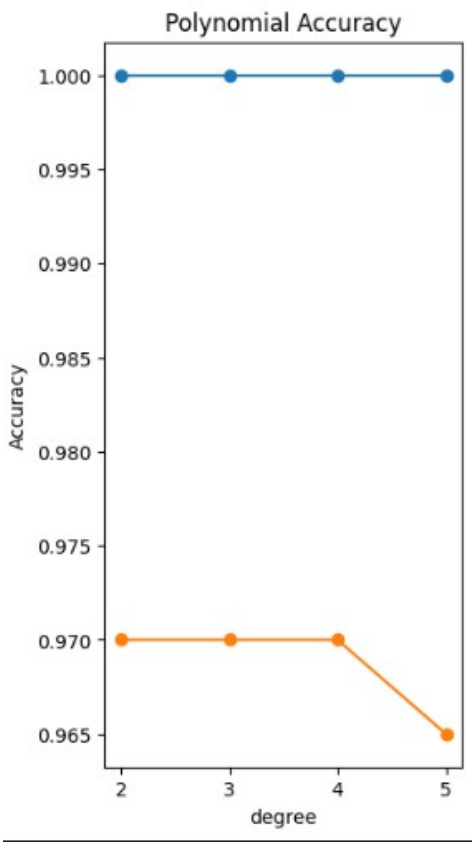
Bewertung

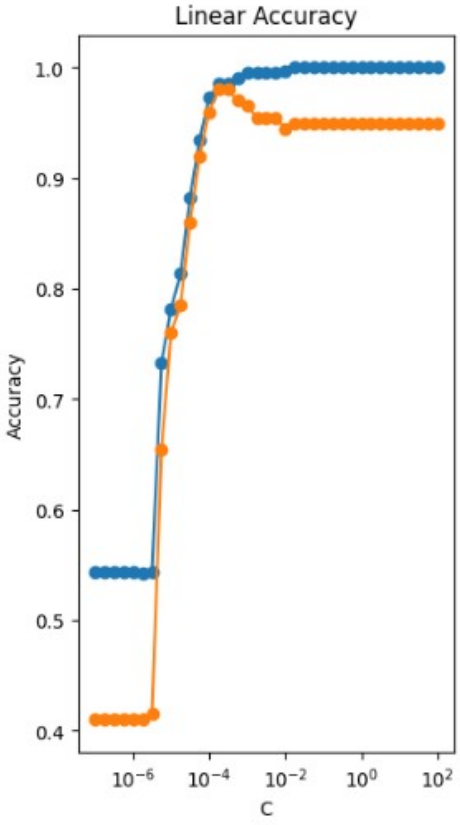
Alle Modelle wurden für ihre Hyperparameter mit einem K-Fold-Cross-Validation mehrfach getestet und der Mittelwert für Empirisches Risiko, Test-/ Trainingsgenauigkeit und allgemein die CV-Score zurückgegeben. Zusätzlich wurde die Nested CV-Score berechnet.
Für alle Modelle wurden außerdem Precision, Recall und die ROC in einem k-fold Cross-Validation

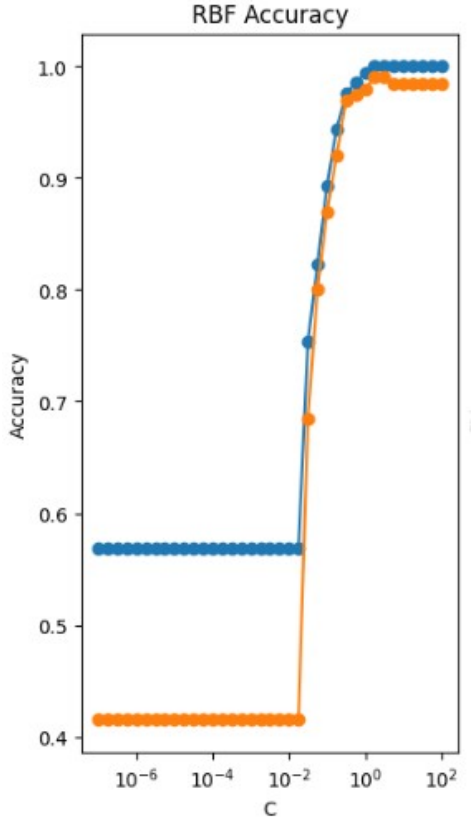
ermittelt und damit die Precision-Recall-Curve und die ROC geplottet sowie für das letztere auch die AUC ausgerechnet.

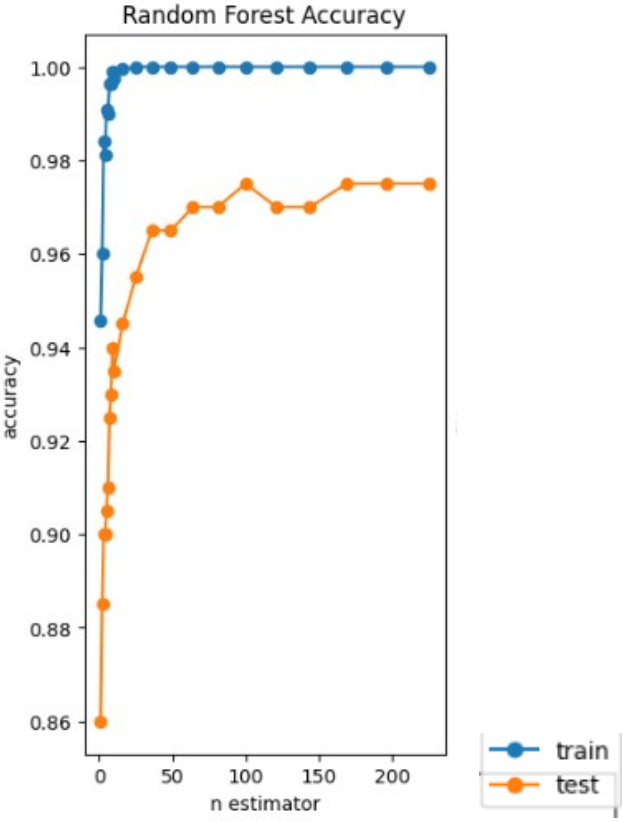
Ergebnisse

Accuracy





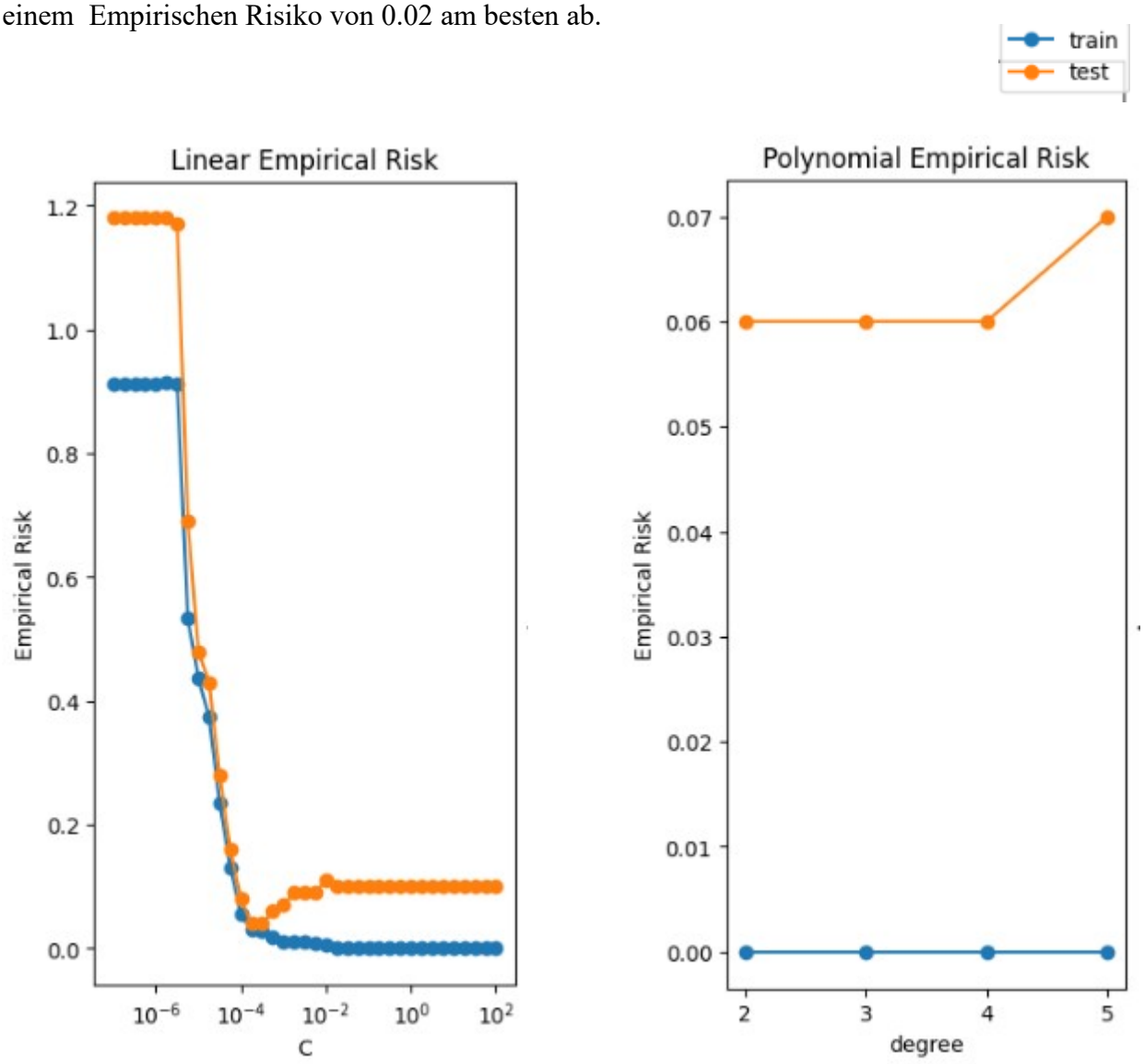


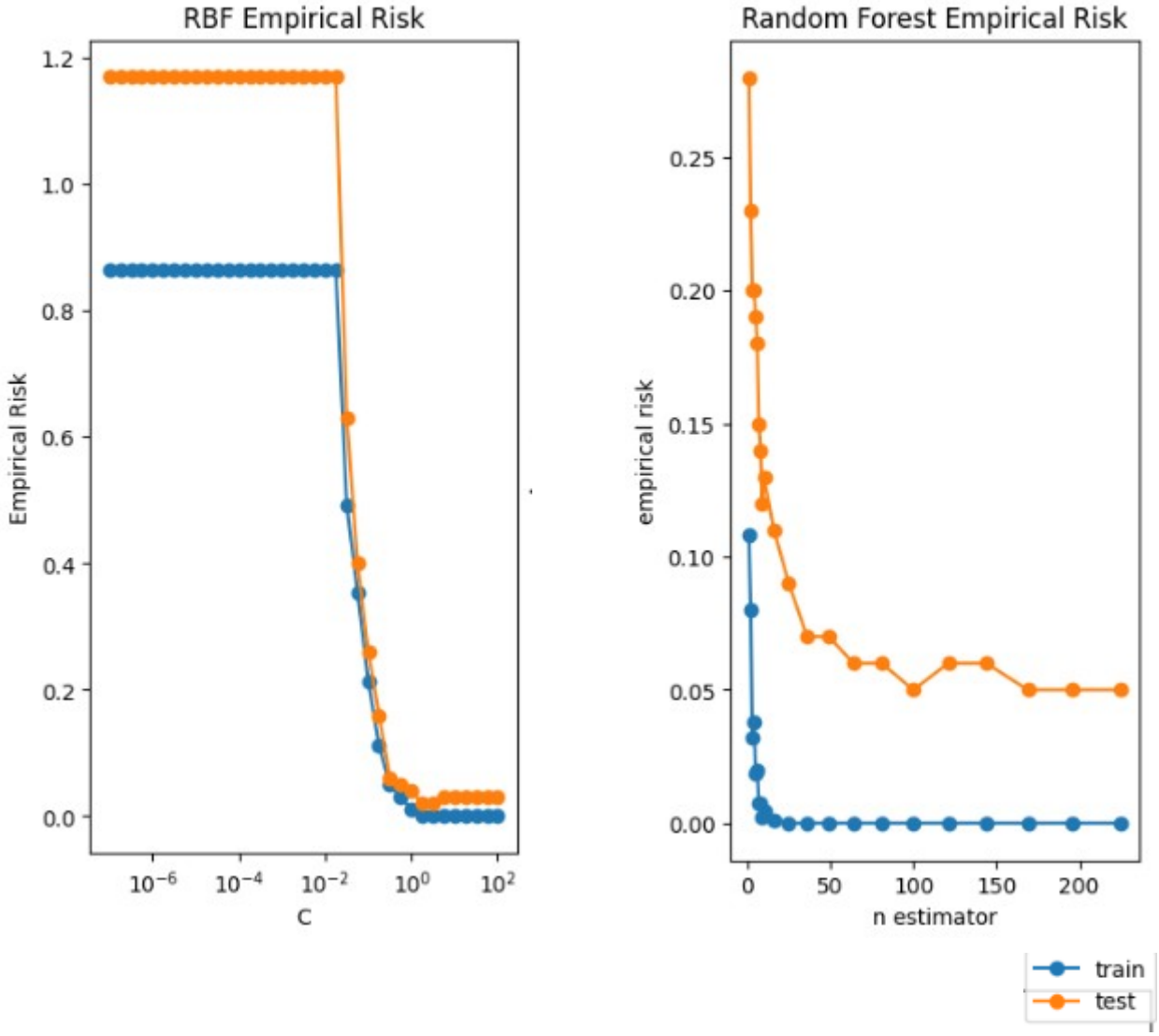


Alle Modelle haben mit den richtigen Hyperparametern eine sehr hohe Genauigkeit. Für den RBF-Kernel ist diese am höchsten. Mit C = 1.7782794100389228 erreicht es eine Genauigkeit von 99,0%. Die Trainingsgenauigkeit ist für alle Modelle nahezu 100%.

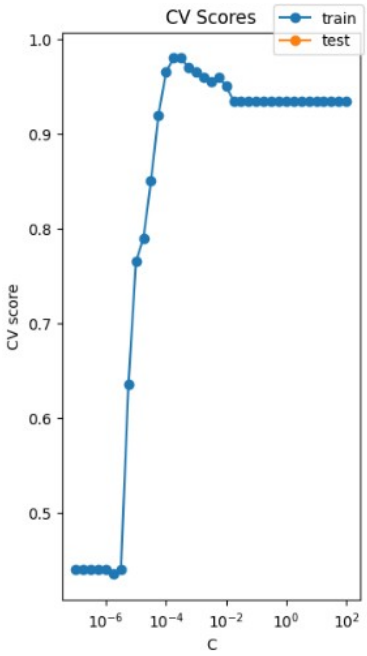
Empirical Risk

Das Empirische Risiko wurde mit dem Hinge-loss berechnet. Auch hier schneidet der RBF mit einem Empirischen Risiko von 0.02 am besten ab.

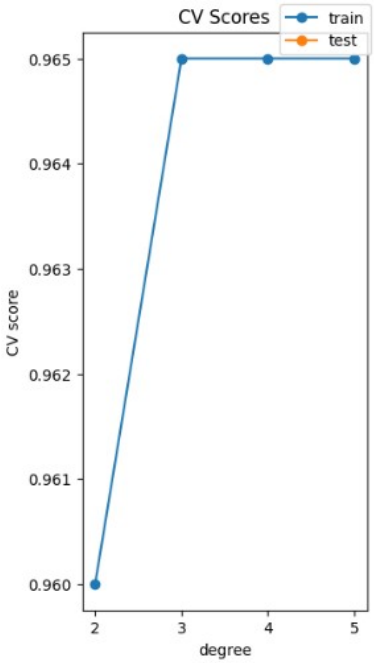




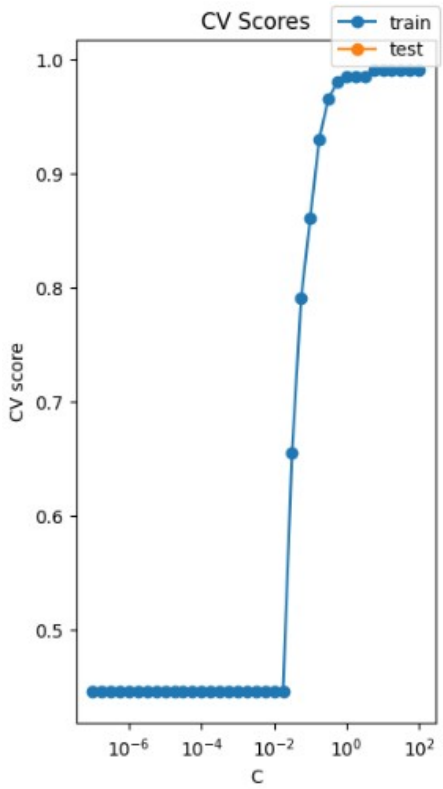
Cross Validation Score



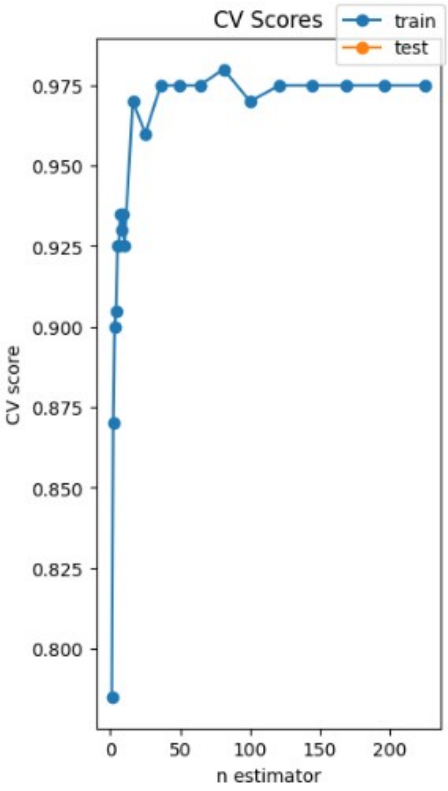
linear



polynomial

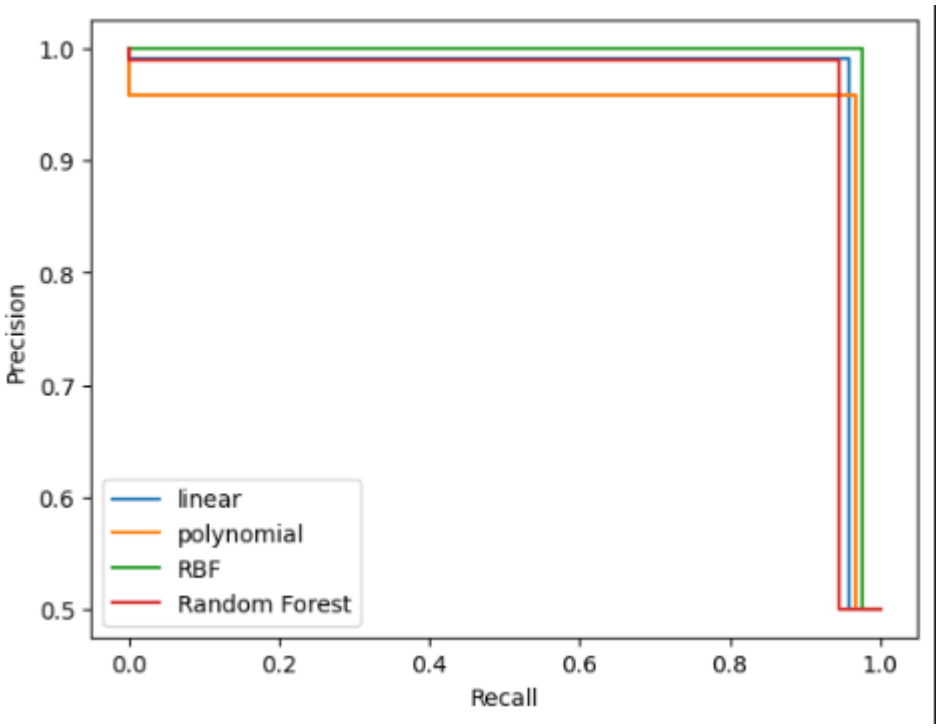


RBF

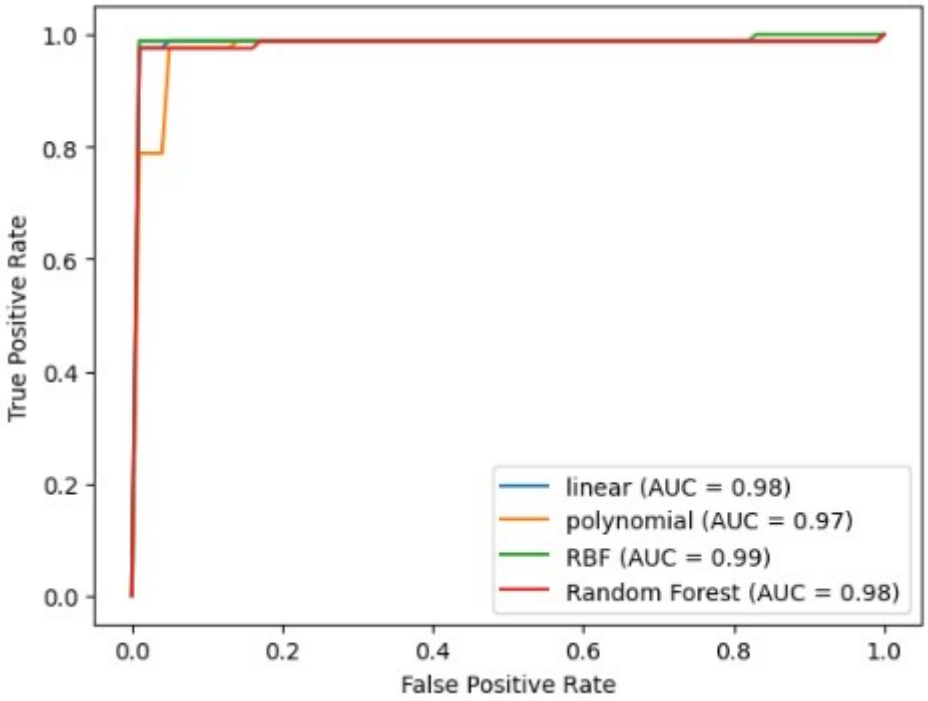


Random Forest

Precision-Recall-Curve



ROC


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
linear (AUC = 0.98)
polynomial (AUC = 0.97)
RBF (AUC = 0.99)
Random Forest (AUC = 0.98)