Early Stage Diabetes Risk Prediction

Wentao Jiang, Hai Xi

Our Goal

- Build a reliable and accurate model that can assist healthcare professionals in early diagnosis and management of diabetes.

Our Approach

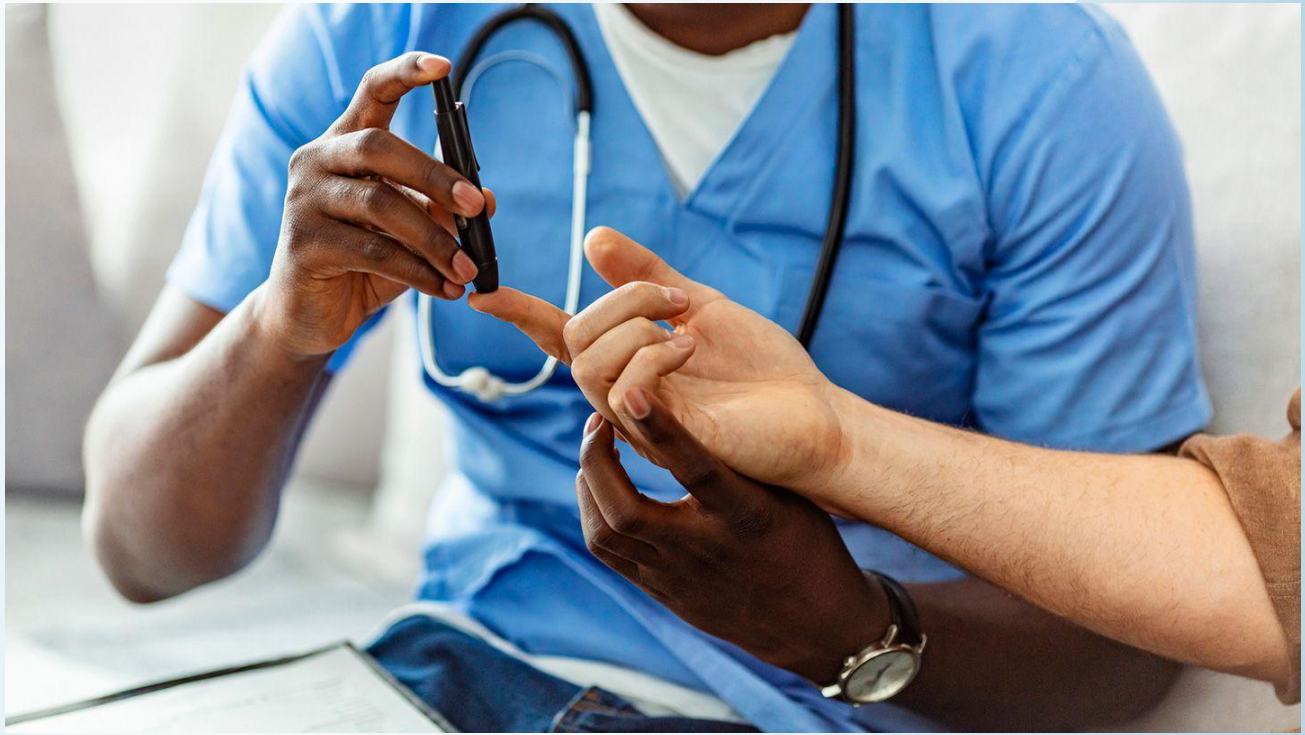
- The model will be trained on preprocessed data, and its performance will be evaluated using various metrics such as accuracy, confusion matrix, classification report, ROC curve, and cross-validation.

Sections

1. Data Overview
2. Data Analyzing and Preprocessing
3. Model Selection
4. Model Evaluation and Comparison
5. Conclusion and Real World Application

Dataset Overview

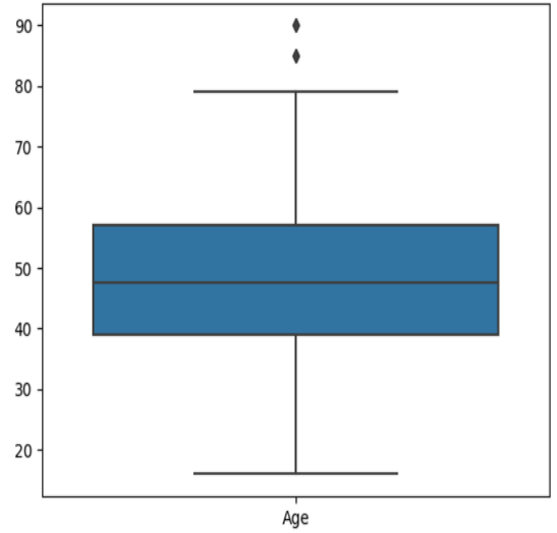
1	Age	16 - 90
2	Sex	Male / Female
3	Polyuria	Yes / No
4	Polydipsia	Yes / No
5	Sudden weight loss	Yes / No
6	Weakness	Yes / No
7	Polyphagia	Yes / No
8	Genital thrush	Yes / No
9	Visual blurring	Yes / No
10	Itching	Yes / No
11	Irritability	Yes / No
12	Delayed healing	Yes / No
13	Partial paresis	Yes / No
14	Muscle stiness	Yes / No
15	Alopecia	Yes / No
16	Obesity	Yes / No
18	Class	Positive / Negative



Dataset Overview

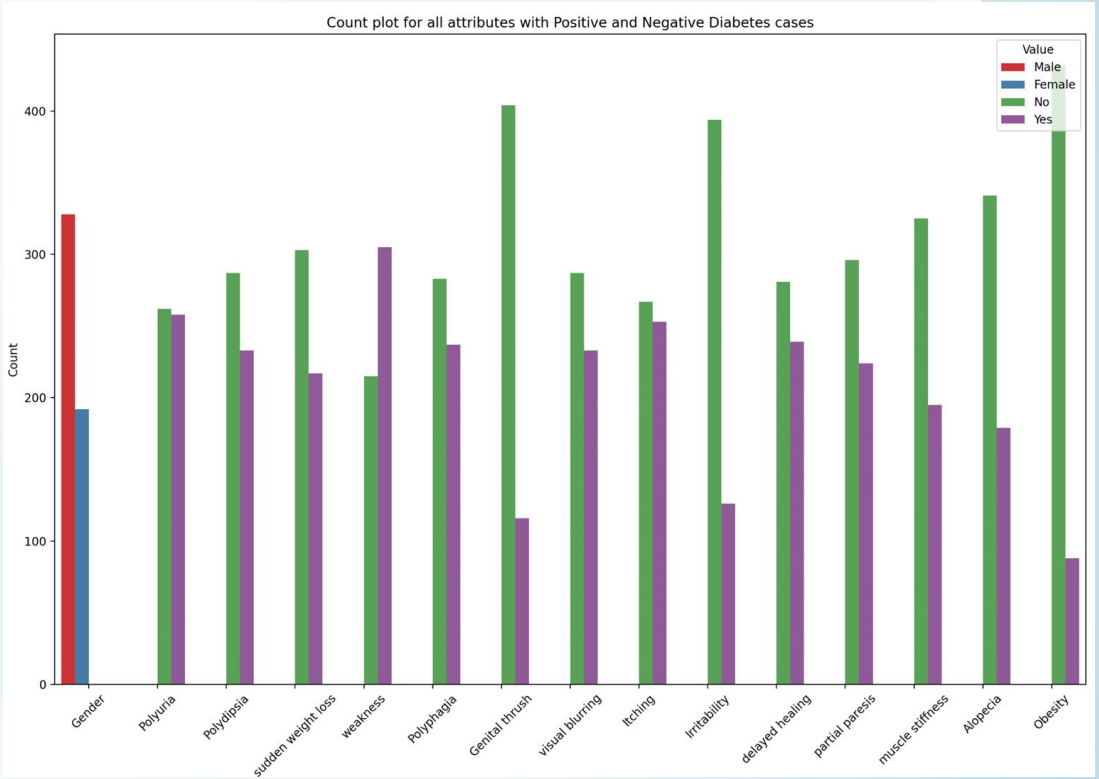
1	Age	16 - 90
2	Sex	Male / Female
3	Polyuria	Yes / No
4	Polydipsia	Yes / No
5	Sudden weight loss	Yes / No
6	Weakness	Yes / No
7	Polyphagia	Yes / No
8	Genital thrush	Yes / No
9	Visual blurring	Yes / No
10	Itching	Yes / No
11	Irritability	Yes / No
12	Delayed healing	Yes / No
13	Partial paresis	Yes / No
14	Muscle stiness	Yes / No
15	Alopecia	Yes / No
16	Obesity	Yes / No
18	Class	Positive / Negative

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 520 entries, 0 to 519
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age                 520 non-null    int64
 1   Gender              520 non-null    object
 2   Polyuria            520 non-null    object
 3   Polydipsia          520 non-null    object
 4   sudden weight loss  520 non-null    object
 5   weakness            520 non-null    object
 6   Polyphagia          520 non-null    object
 7   Genital thrush      520 non-null    object
 8   visual blurring     520 non-null    object
 9   Itching             520 non-null    object
 10  Irritability        520 non-null    object
 11  delayed healing     520 non-null    object
 12  partial paresis     520 non-null    object
 13  muscle stiffness    520 non-null    object
 14  Alopecia            520 non-null    object
 15  Obesity             520 non-null    object
 16  class               520 non-null    object
dtypes: int64(1), object(16)
memory usage: 69.2+ KB
```



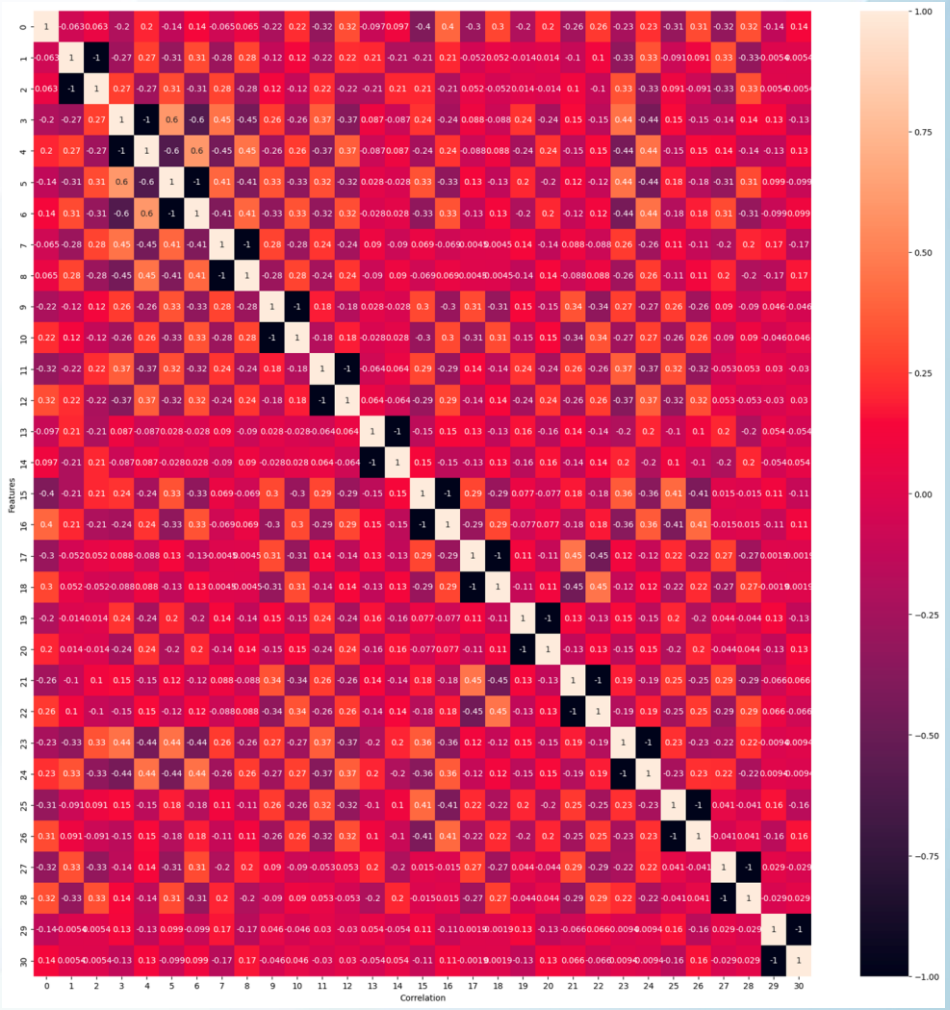
Dataset Overview

1	Age	16 - 90
2	Sex	Male / Female
3	Polyuria	Yes / No
4	Polydipsia	Yes / No
5	Sudden weight loss	Yes / No
6	Weakness	Yes / No
7	Polyphagia	Yes / No
8	Genital thrush	Yes / No
9	Visual blurring	Yes / No
10	Itching	Yes / No
11	Irritability	Yes / No
12	Delayed healing	Yes / No
13	Partial paresis	Yes / No
14	Muscle stiness	Yes / No
15	Alopecia	Yes / No
16	Obesity	Yes / No
18	Class	Positive / Negative



Correlation Matrix

```
Encoded X Value:
 [[0.32432432 0.         1.         1.         0.         0.
  1.         1.         0.         0.         1.         1.
  0.         1.         0.         1.         0.         0.
  1.         1.         0.         0.         1.         1.
  0.         0.         1.         0.         1.         0.
  1.        ]
 [0.56756757 0.         1.         1.         0.         1.
  0.         1.         0.         0.         1.         1.
  0.         1.         0.         0.         1.         1.
  0.         1.         0.         1.         0.         0.
  1.         1.         0.         0.         1.         1.
  0.        ]
 [0.33783784 0.         1.         0.         1.         1.
  0.         1.         0.         0.         1.         0.
  1.         1.         0.         1.         0.         0.
  1.         1.         0.         0.         1.         1.
  0.         0.         1.         0.         1.         1.
  0.        ]]
```



Model Selection

A. Train Test Split

- 80% Training Set
- 20% Testing Set

B. Models

- Logistic Regression (Baseline model)
- Support Vector Machine Classification **Linear** Kernel
- Support Vector Machine Classification **Polynomial** Kernel

Model Selection: Accuracy

1. Logistic Regression

```
Model Accuracy with 1000 max_iters for lr: 0.9230769230769231
```

1. SVM Classification
   a. “Linear” kernel, choose C=4
   b. “Polynomial” kernel, choose C=1

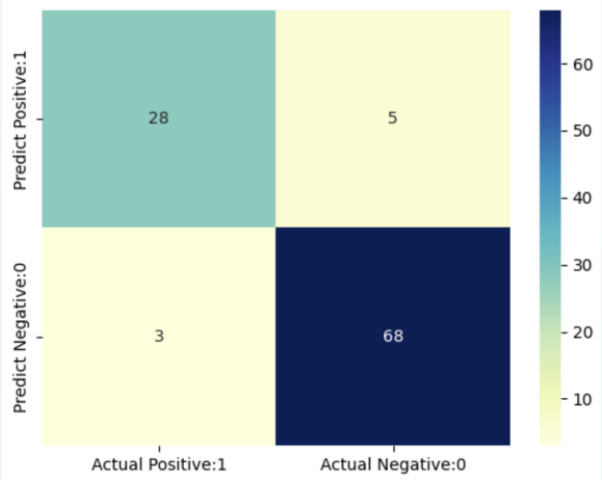
```
Model accuracy score with linear, C = 4, with the value of: 0.9423
Model accuracy score with polynomial, C = 1:, with the value of: 0.9904
```

Model Evaluation

1. Confusion Matrix
2. Classification Report
3. ROC_AUC Curve
4. Cross-Validation
5. KFold Validation ($k = 10$)

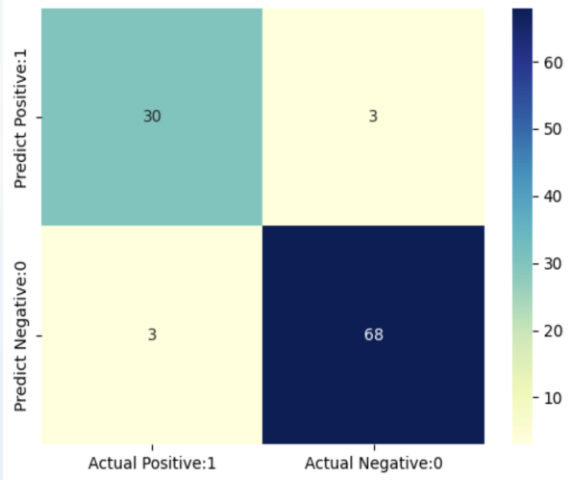
Confusion Matrix

Logistic Regression (max_iter=1000)



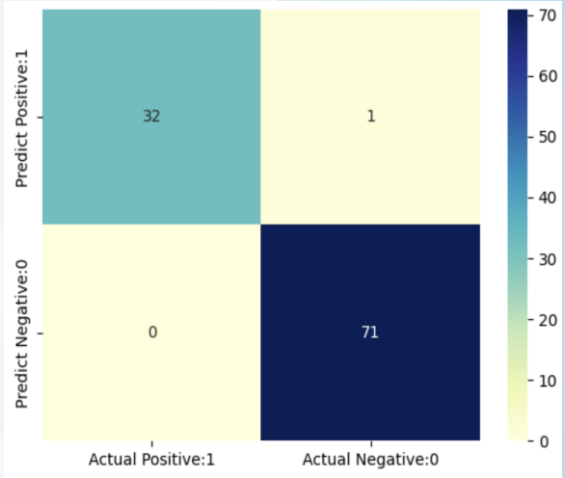

```
True Positives(TP) =  28
True Negatives(TN) =  68
False Positives(FP) =  5
False Negatives(FN) =  3
```

SVM Classifier (kernel: **linear**, C=4)




```
True Positives(TP) =  30
True Negatives(TN) =  68
False Positives(FP) =  3
False Negatives(FN) =  3
```

SVM Classifier (kernel: **polynomial**, C=1)




```
True Positives(TP) =  32
True Negatives(TN) =  71
False Positives(FP) =  1
False Negatives(FN) =  0
```

Classification Reports

Classification Report for Logistic Regression:

	precision	recall	f1-score	support
Negative	0.90	0.85	0.88	33
Positive	0.93	0.96	0.94	71
accuracy			0.92	104
macro avg	0.92	0.90	0.91	104
weighted avg	0.92	0.92	0.92	104

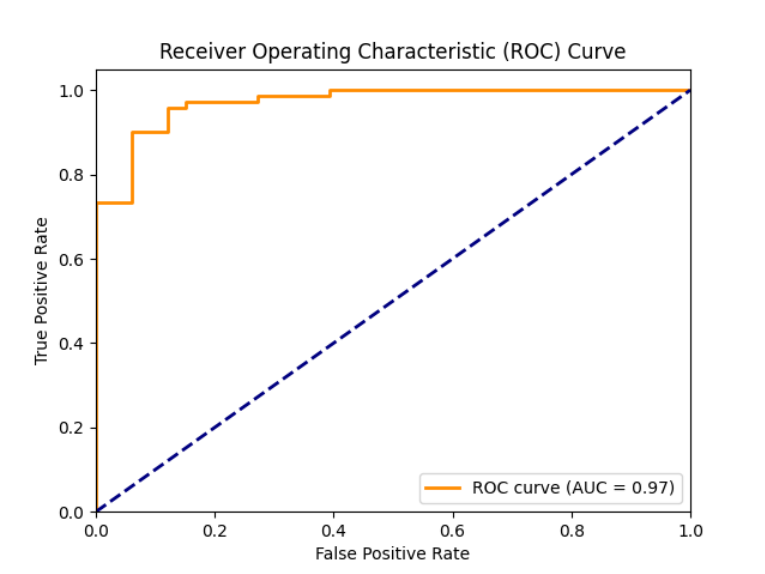
Classification Report for Linear SVM:

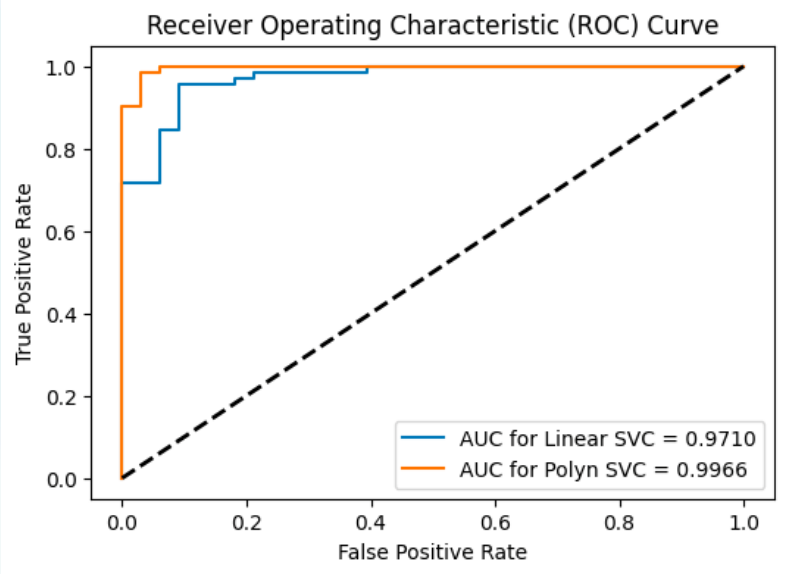
	precision	recall	f1-score	support
Negative	0.91	0.91	0.91	33
Positive	0.96	0.96	0.96	71
accuracy			0.94	104
macro avg	0.93	0.93	0.93	104
weighted avg	0.94	0.94	0.94	104

Classification Report for Polynomial SVM:

	precision	recall	f1-score	support
Negative	0.97	0.97	0.97	33
Positive	0.99	0.99	0.99	71
accuracy			0.98	104
macro avg	0.98	0.98	0.98	104
weighted avg	0.98	0.98	0.98	104

ROC_AUC: Logistic Regression vs. SVC

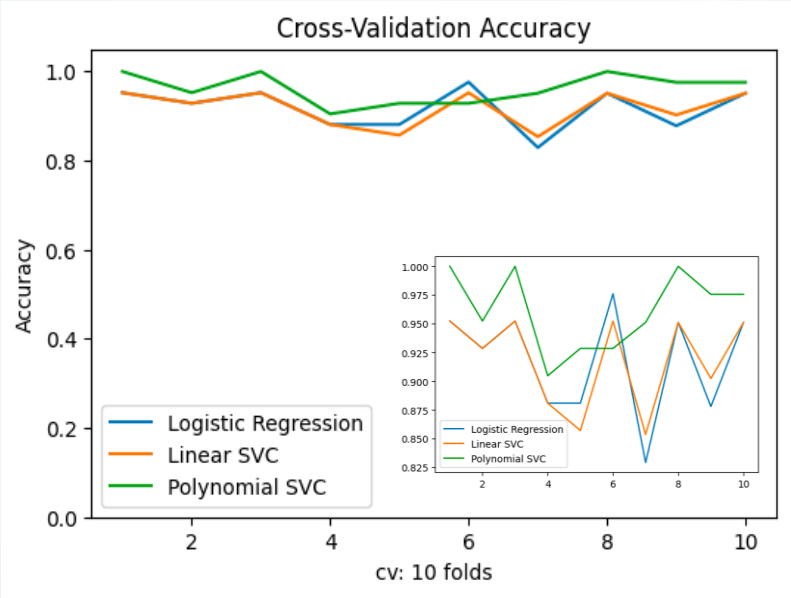




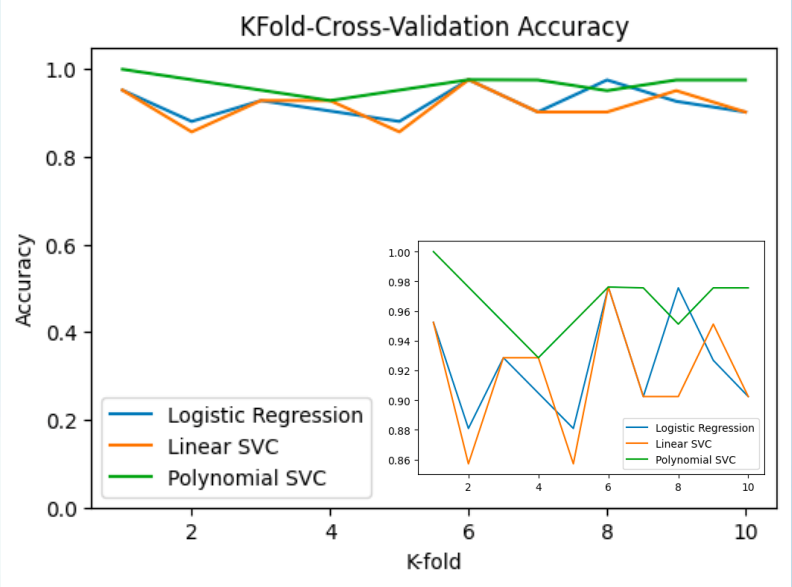
Margin of the decision boundary C: cost parameter
Linear SVC: C = 4, Polyn SVC: C = 1

Cross-Validation Evaluation

10-fold Cross-Validation



10-Fold Cross-Validation, shuffle data



Reference

[1] Shai Shalev-Shwartz and Shai Ben-David. 2014. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, New York, NY, USA.

[2] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[3] Travis E. Oliphant. 2006. A guide to NumPy, USA: Trelgol Publishing.

[4] Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference (SciPy 2010). 51 - 56.

[5] John D. Hunter. 2007. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering 9, 3 (2007), 90–95.

[6] Michael Waskom, Olga Botvinnik, Paul Hobson, Saulius Lukauskas, Emmanuelle Gouillart, Andreas Mueller, … Alistair Miles. (2017). seaborn: v0.8.1 (February 2017). Zenodo. http://doi.org/10.5281/zenodo.883859

[7] OpenAI. (2020). GPT-3.5. https://openai.com/blog/gpt-3-5-billion-parameters/

[8] David Newman, 2008. Bag of Words Data Set. UCI Machine Learning Repository, California, USA. https://archive.ics.uci.edu/ml/datasets/Bag+of+Words

Q&A?