Predicting Cervical Cancer Test Results

Jaedon Jensen

Data Science Intensive Capstone Project, April 2024 Cohort

The problem

- Cervical cancer is 100% preventable but causes 350,000 deaths every year

Who would care?

- Healthcare Professionals
- Patients

Data Information

- Survey conducted at hospital in Venezuela

https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors

Data Information

- Survey conducted at hospital in Venezuela
- 858 Records, 36 Features

https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors

Data Cleaning

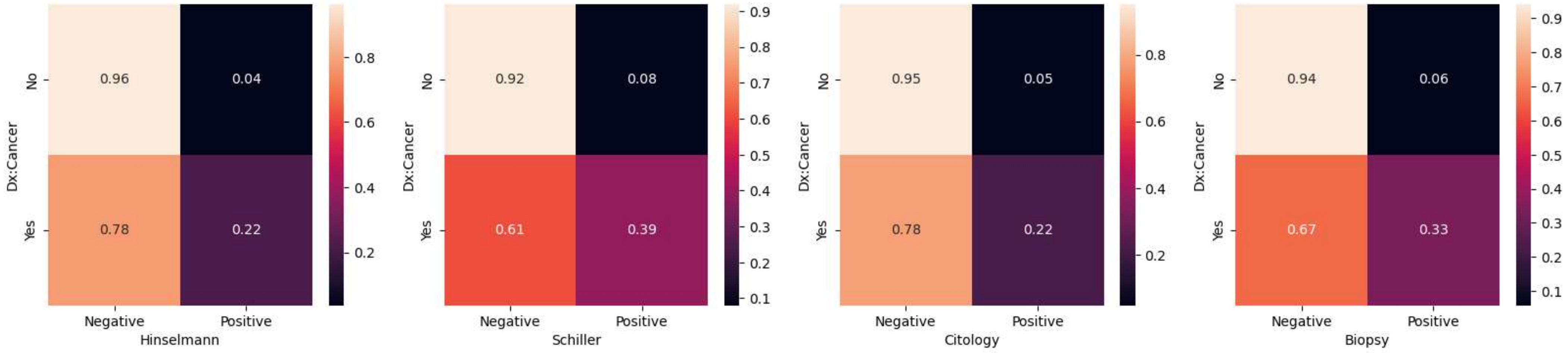
- Missing Values
- Dropped records with too many missing fields

Data Cleaning

- Missing Values
- Dropped records with too many missing fields
- Mean imputation and standardization for numerical values

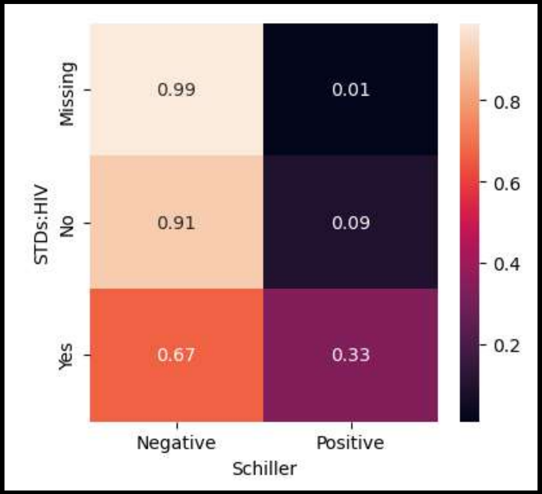
Data Exploration

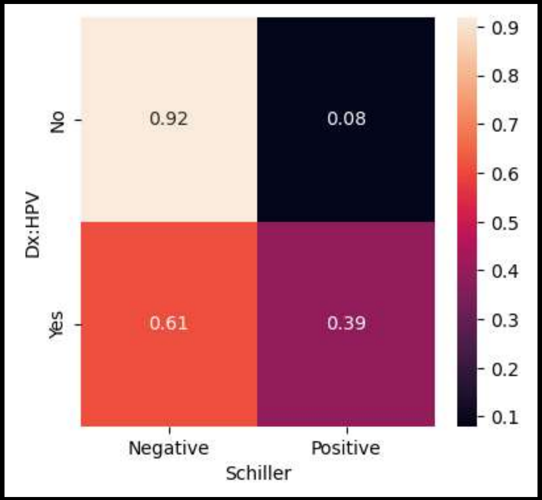
- Most significant test: Schiller

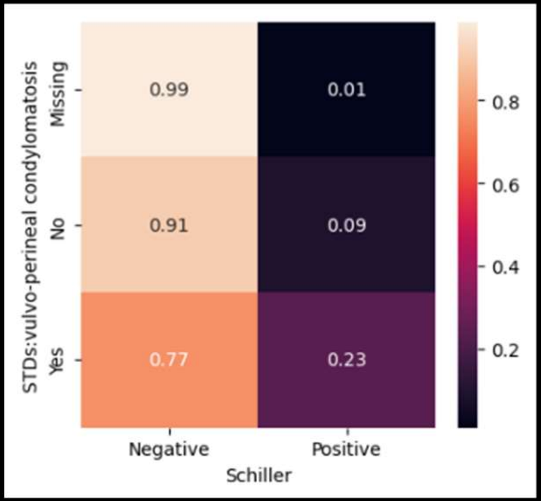


Data Exploration

- Important features:
  - HIV
  - HPV
  - Vulvo-perineal condylomatosis







Machine Learning Modeling

- Logistic Regression
- Random Forest
- SVM
- Metric: F-macro

Results

	Best Score	Train Time
Logistic Regression	0.851064	13.617378
Random Forest	0.888462	181.422026
SVM	0.872447	0.139997

Results

	Best Score	Train Time
Logistic Regression	0.851064	13.617378
Random Forest	0.888462	181.422026
SVM	0.872447	0.139997

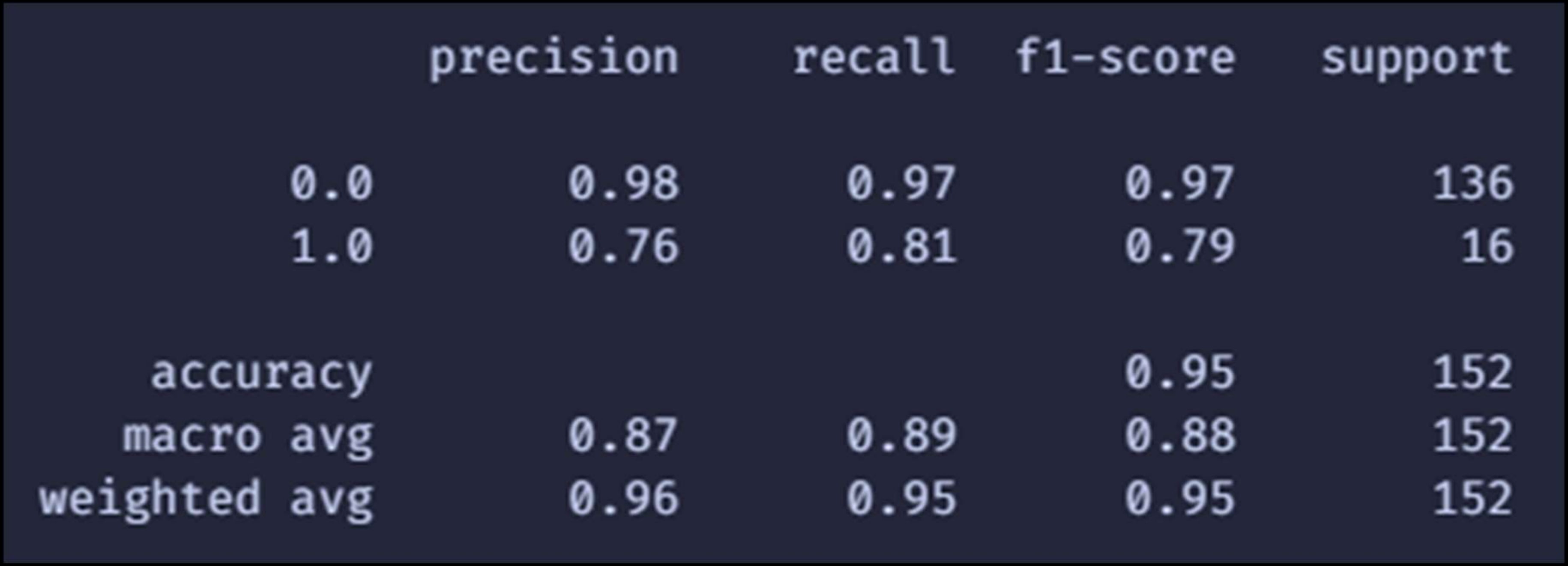
Results

	Best Score	Train Time
Logistic Regression	0.851064	13.617378
Random Forest	0.888462	181.422026
SVM	0.872447	0.139997

Results

	Best Score	Train Time
Logistic Regression	0.851064	13.617378
Random Forest	0.888462	181.422026
SVM	0.872447	0.139997

Logistic Regression

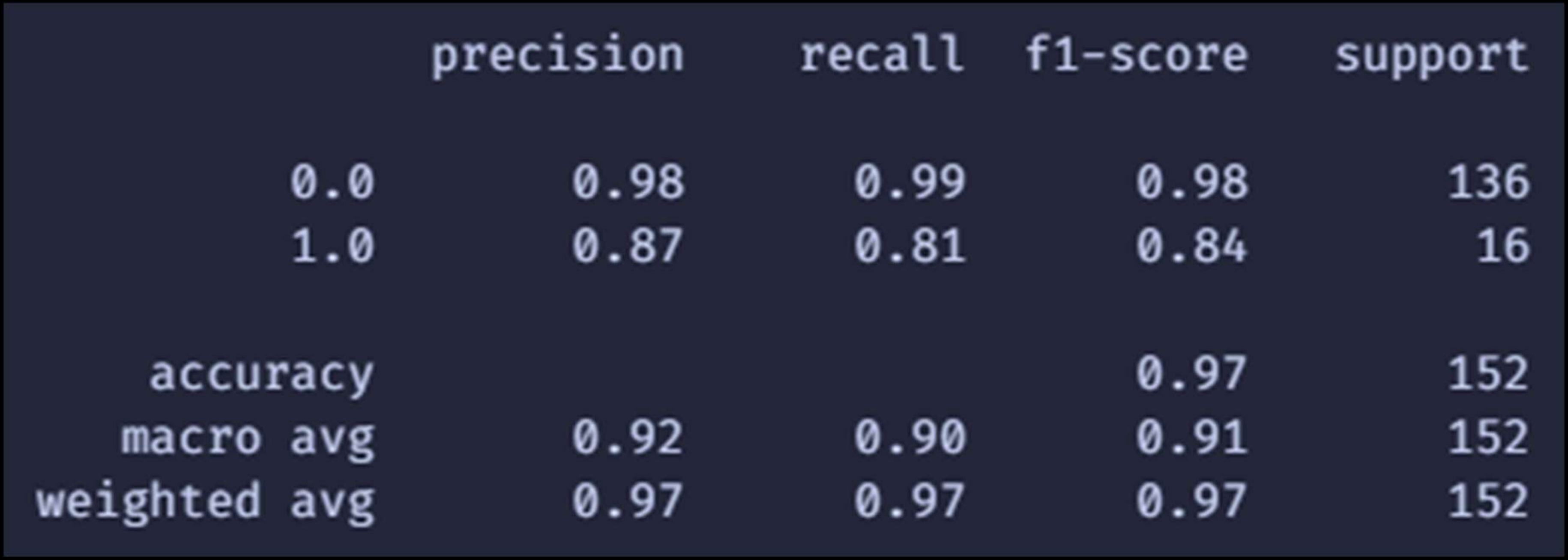


```
              precision    recall  f1-score   support

         0.0       0.98      0.97      0.97       136
         1.0       0.76      0.81      0.79        16

    accuracy                           0.95       152
   macro avg       0.87      0.89      0.88       152
weighted avg       0.96      0.95      0.95       152
```

Random Forest

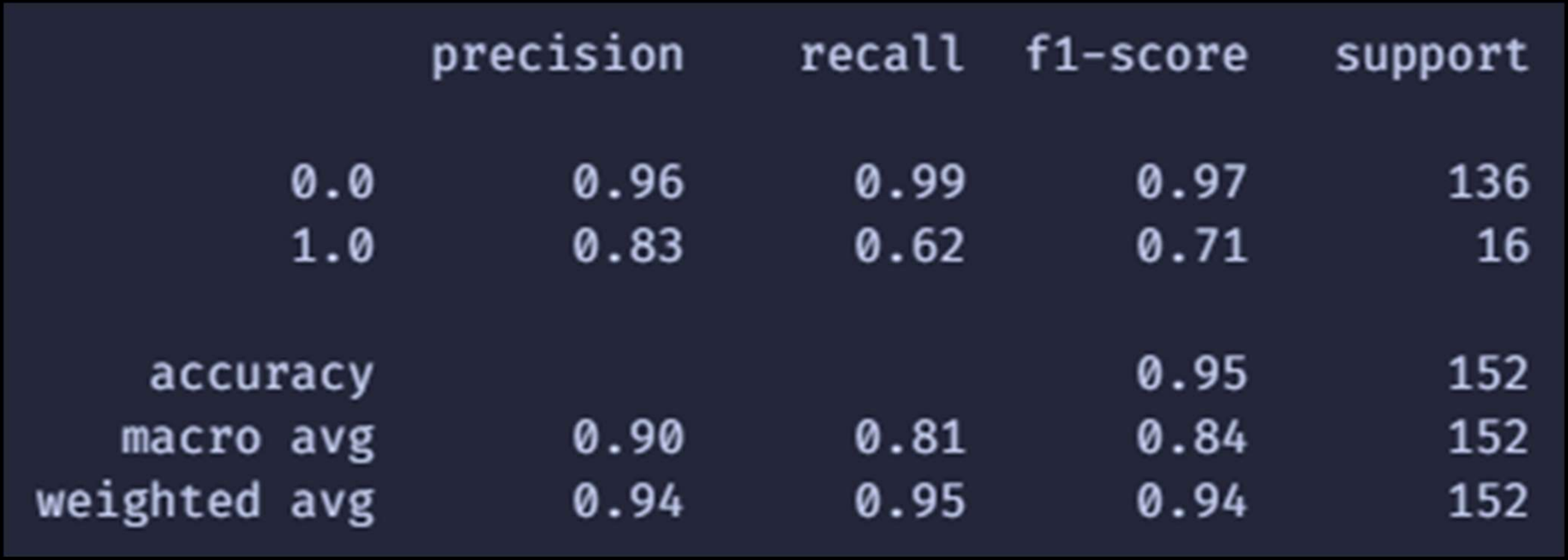


```
              precision    recall  f1-score   support

         0.0       0.98      0.99      0.98       136
         1.0       0.87      0.81      0.84        16

    accuracy                           0.97       152
   macro avg       0.92      0.90      0.91       152
weighted avg       0.97      0.97      0.97       152
```

SVM



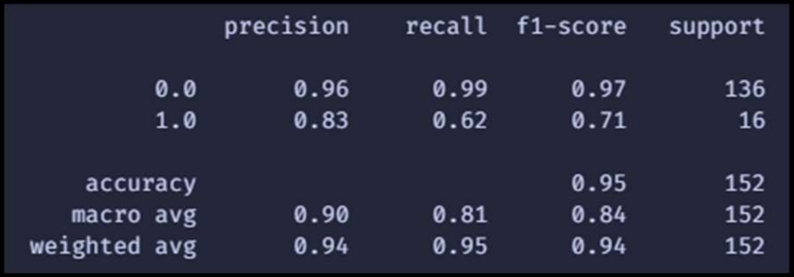
```
              precision    recall  f1-score   support

         0.0       0.96      0.99      0.97       136
         1.0       0.83      0.62      0.71        16

    accuracy                           0.95       152
   macro avg       0.90      0.81      0.84       152
weighted avg       0.94      0.95      0.94       152
```

Results

- Without Hinselmann test as a feature:

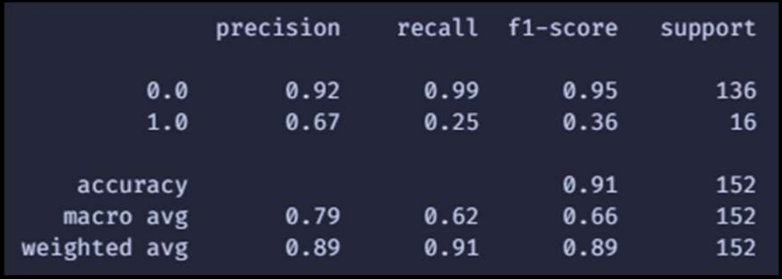


```
              precision    recall  f1-score   support

         0.0       0.96      0.99      0.97       136
         1.0       0.83      0.62      0.71        16

    accuracy                           0.95       152
   macro avg       0.90      0.81      0.84       152
weighted avg       0.94      0.95      0.94       152
```

- Without Hinselmann or Biopsy tests as features:



```
              precision    recall  f1-score   support

         0.0       0.92      0.99      0.95       136
         1.0       0.67      0.25      0.36        16

    accuracy                           0.91       152
   macro avg       0.79      0.62      0.66       152
weighted avg       0.89      0.91      0.89       152
```

Future Improvements

- Improved Feature Engineering
- Predict diagnosis instead

Conclusions

- SVM model the most ideal
- All models were inaccurate if no test results used as data features
- Options:
  - Work to improve AI models
  - Focus efforts elsewhere