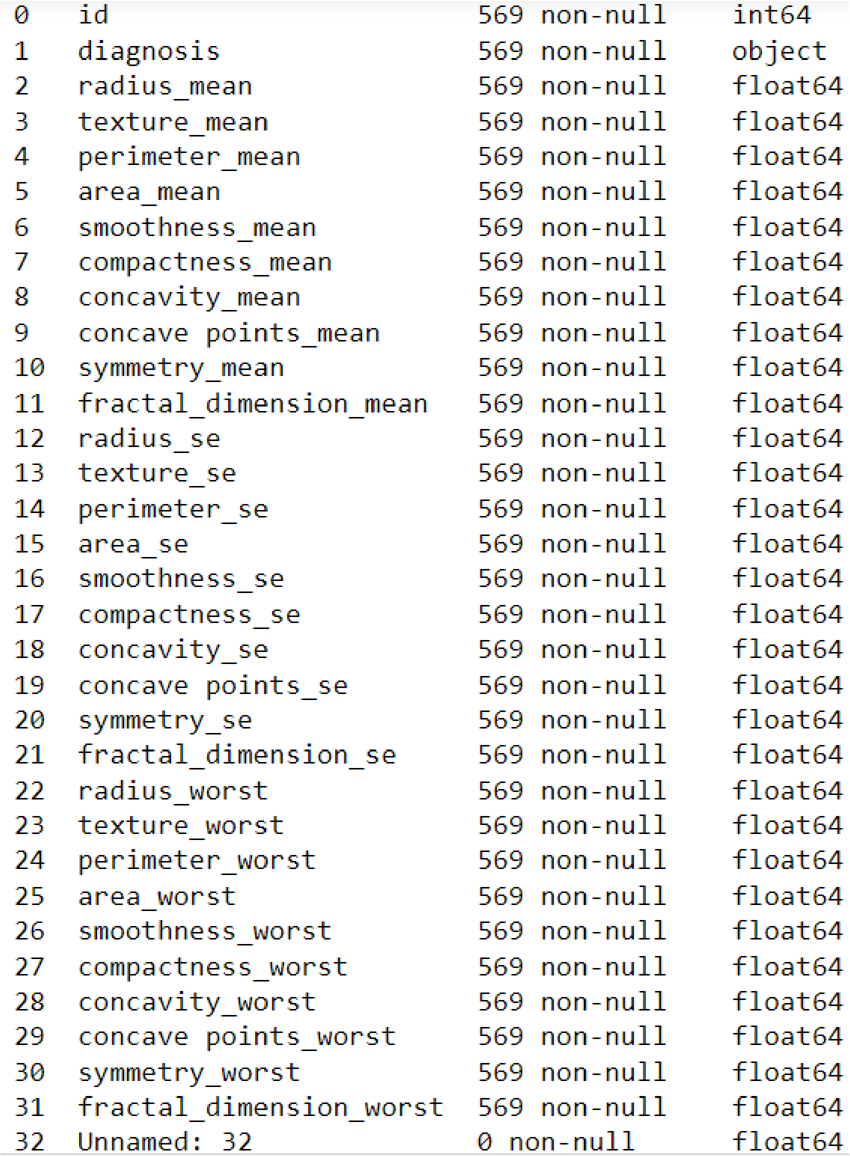
Analyzing Machine Learning Techniques to Detect Breast Cancer

By: Aroofa Mohammad(aam170007)

Breast Cancer Facts

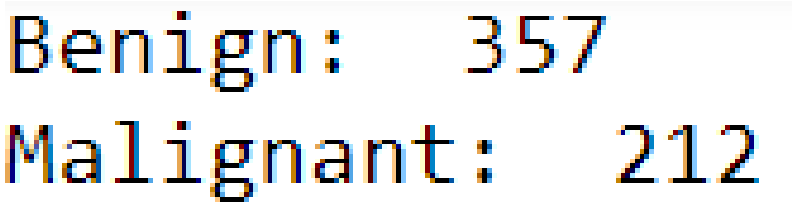
- 287,850 new cases of women in the United States in 2022 were discovered to breast cancer
- A woman that in the US has a 1/8 chance of developing breast cancer at some point
- Breast cancer is the 2nd largest cause of dead from cancer for women
- Roughly 1/39 woman dying from breast cancer

Data Cleanup and Target Defined



```
0   id                       569 non-null    int64
1   diagnosis                569 non-null    object
2   radius_mean              569 non-null    float64
3   texture_mean             569 non-null    float64
4   perimeter_mean           569 non-null    float64
5   area_mean                569 non-null    float64
6   smoothness_mean          569 non-null    float64
7   compactness_mean         569 non-null    float64
8   concavity_mean           569 non-null    float64
9   concave points_mean      569 non-null    float64
10  symmetry_mean            569 non-null    float64
11  fractal_dimension_mean   569 non-null    float64
12  radius_se                569 non-null    float64
13  texture_se               569 non-null    float64
14  perimeter_se             569 non-null    float64
15  area_se                  569 non-null    float64
16  smoothness_se            569 non-null    float64
17  compactness_se           569 non-null    float64
18  concavity_se             569 non-null    float64
19  concave points_se        569 non-null    float64
20  symmetry_se              569 non-null    float64
21  fractal_dimension_se     569 non-null    float64
22  radius_worst             569 non-null    float64
23  texture_worst            569 non-null    float64
24  perimeter_worst          569 non-null    float64
25  area_worst               569 non-null    float64
26  smoothness_worst         569 non-null    float64
27  compactness_worst        569 non-null    float64
28  concavity_worst          569 non-null    float64
29  concave points_worst     569 non-null    float64
30  symmetry_worst           569 non-null    float64
31  fractal_dimension_worst  569 non-null    float64
32  Unnamed: 32              0 non-null      float64
```

- The research was done on the Wisconsin Breast Cancer Dataset
- The feature are shown in the image on the left
- In the dataset there are two irreverent features that were dropped:
  - Id
  - Unnamed: 32
- The Target of the dataset is the diagnostic:
  - Benign: No Cancer
  - Malignant: Cancer
- The image shown below shows the cases of benign/malignant in the database

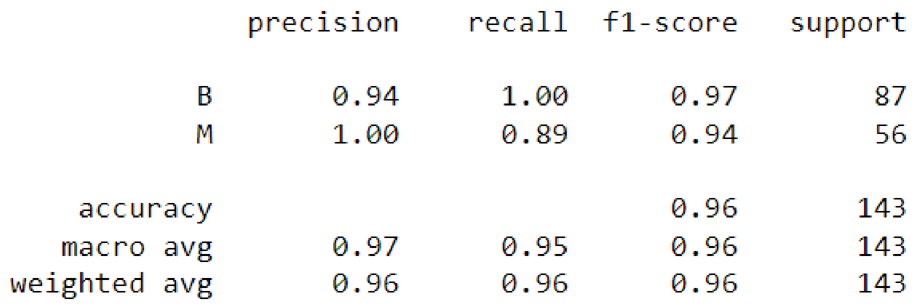


```
Benign:    357
Malignant:   212
```

Machine Learning Models Used to Tackle Breast Cancer Detection

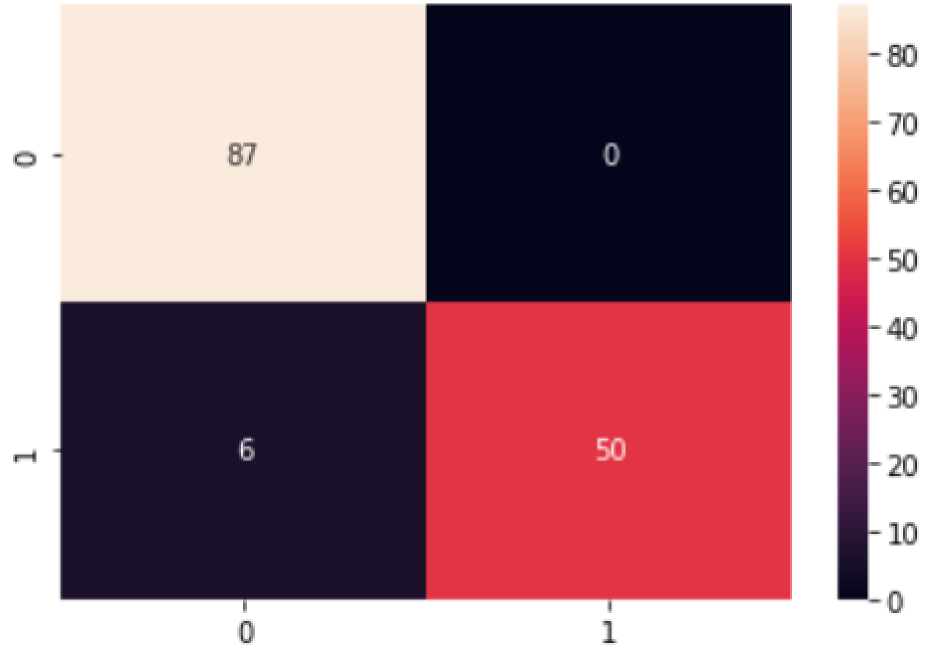
- The following techniques were used on the Wisconsin Breast Cancer Dataset:
    - Support Vector Machine (SVM)
    - Decision tree
    - K-Nearest Neighbors (KNN)
    - Logistic Regression

Support Vector Machine (SVM)



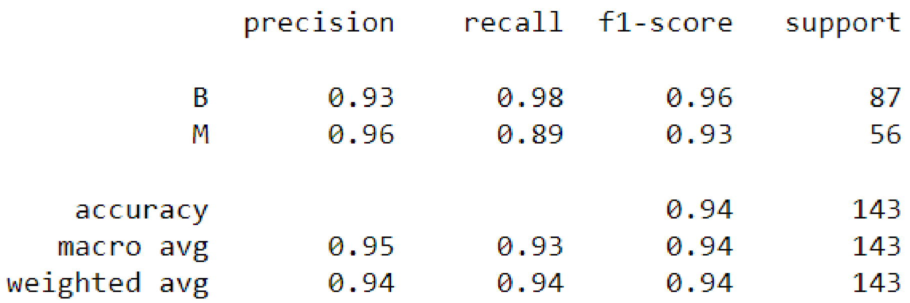
```
              precision    recall  f1-score   support

           B       0.94      1.00      0.97        87
           M       1.00      0.89      0.94        56

    accuracy                           0.96       143
   macro avg       0.97      0.95      0.96       143
weighted avg       0.96      0.96      0.96       143
```

- SVM categorizes data points by projecting them to a high-dimensional feature space even if not linearly separable
- After the separator is found between the categories, the data is transformed so that the separator can be drawn as a hyperplane

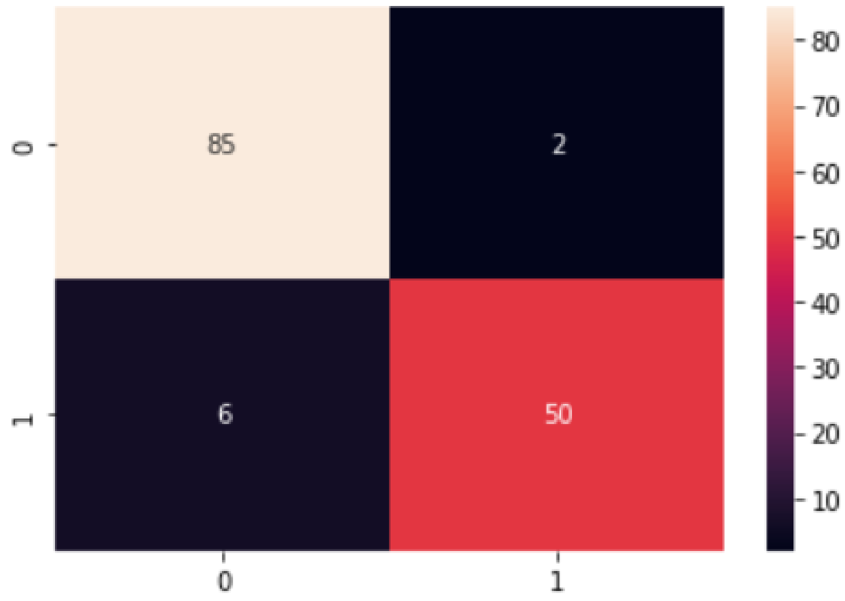


Decision Tree

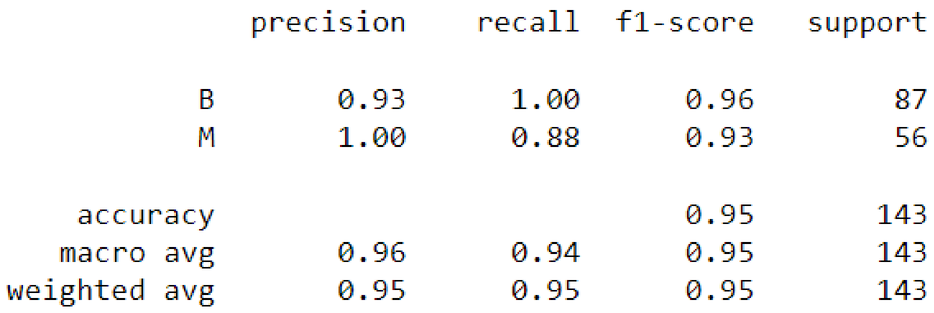
- A decision tree is a tree-like structure that serves as a decision-making aid by visually exhibiting actions and their probable outcomes, repercussions, and costs
- Overfitting is an issue that happens when the algorithm's depth is increased
  - Because number of nodes is larger than the size
- In this research the max you can raise the depth to is 3
  - Any binary classification issue can benefit from this optimum depth



```
              precision    recall  f1-score   support

           B       0.93      0.98      0.96        87
           M       0.96      0.89      0.93        56

    accuracy                           0.94       143
   macro avg       0.95      0.93      0.94       143
weighted avg       0.94      0.94      0.94       143
```

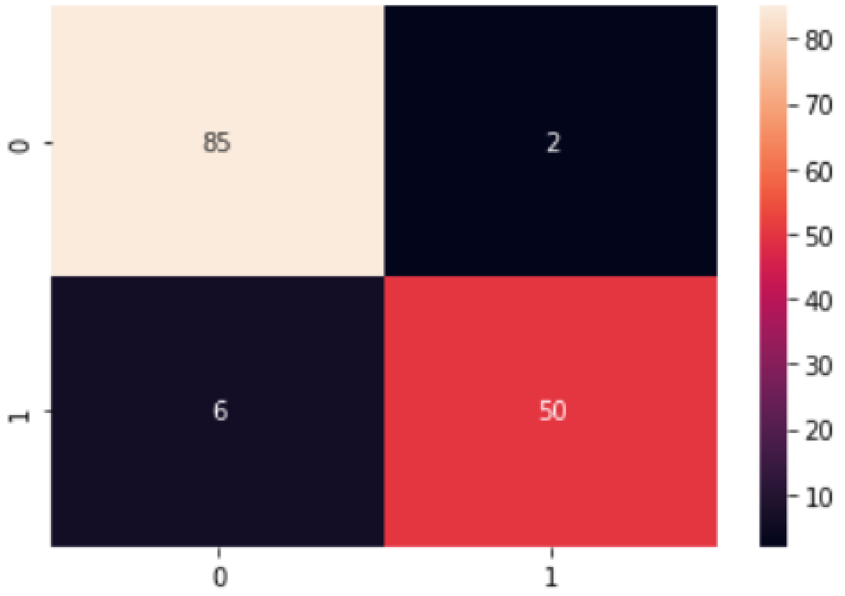


K-Nearest Neighbors (KNN)



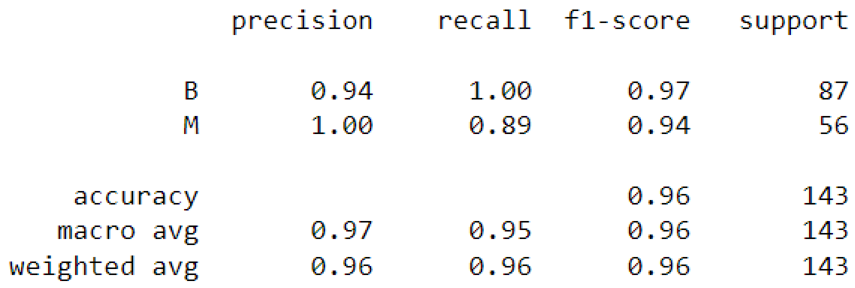
```
              precision    recall  f1-score   support

           B       0.93      1.00      0.96        87
           M       1.00      0.88      0.93        56

    accuracy                           0.95       143
   macro avg       0.96      0.94      0.95       143
weighted avg       0.95      0.95      0.95       143
```

- The KNN algorithm is a data classification approach that estimates the chance that a data point will belong to one group based on the data points closest to it
- K = number of nearest neighbors
- In this breast cancer study, the optimal value of k is 13
  - Higher number might cause underfitting

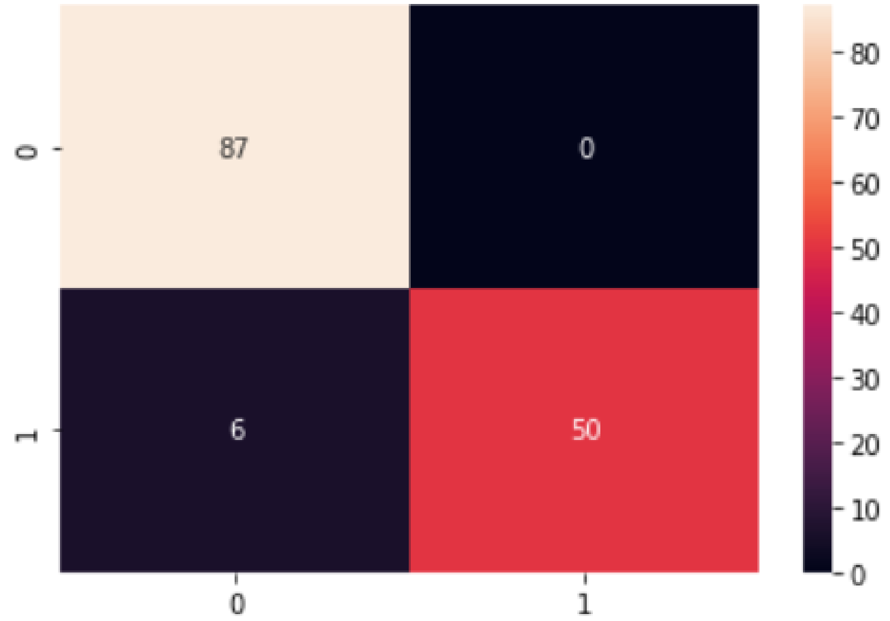


Logistic Regression

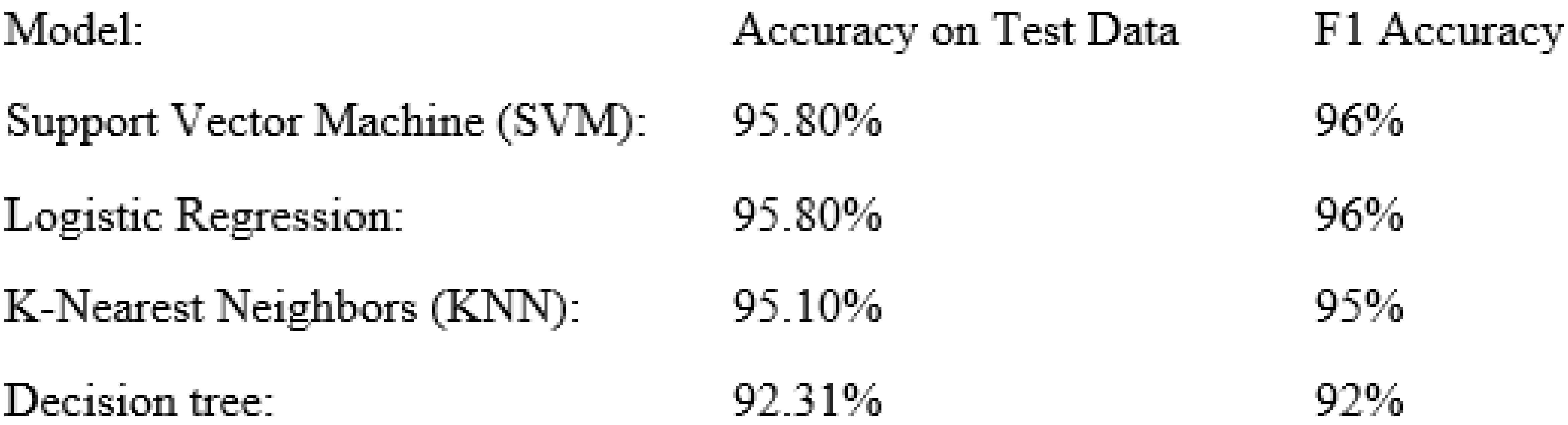
- Based on past observations of a data set, logistic regression is a statistical analytic approach for predicting a binary result, such as yes or no
  - Like ‘M‘ for malignant and ‘B’ for benign
- In the research, the training data is fit and forecasted using Logistic Regression with GridSearchCV
- Then we optimize the model performance by automatically tuning the hyperparameters



```
              precision    recall  f1-score   support

           B       0.94      1.00      0.97        87
           M       1.00      0.89      0.94        56

    accuracy                           0.96       143
   macro avg       0.97      0.95      0.96       143
weighted avg       0.96      0.96      0.96       143
```



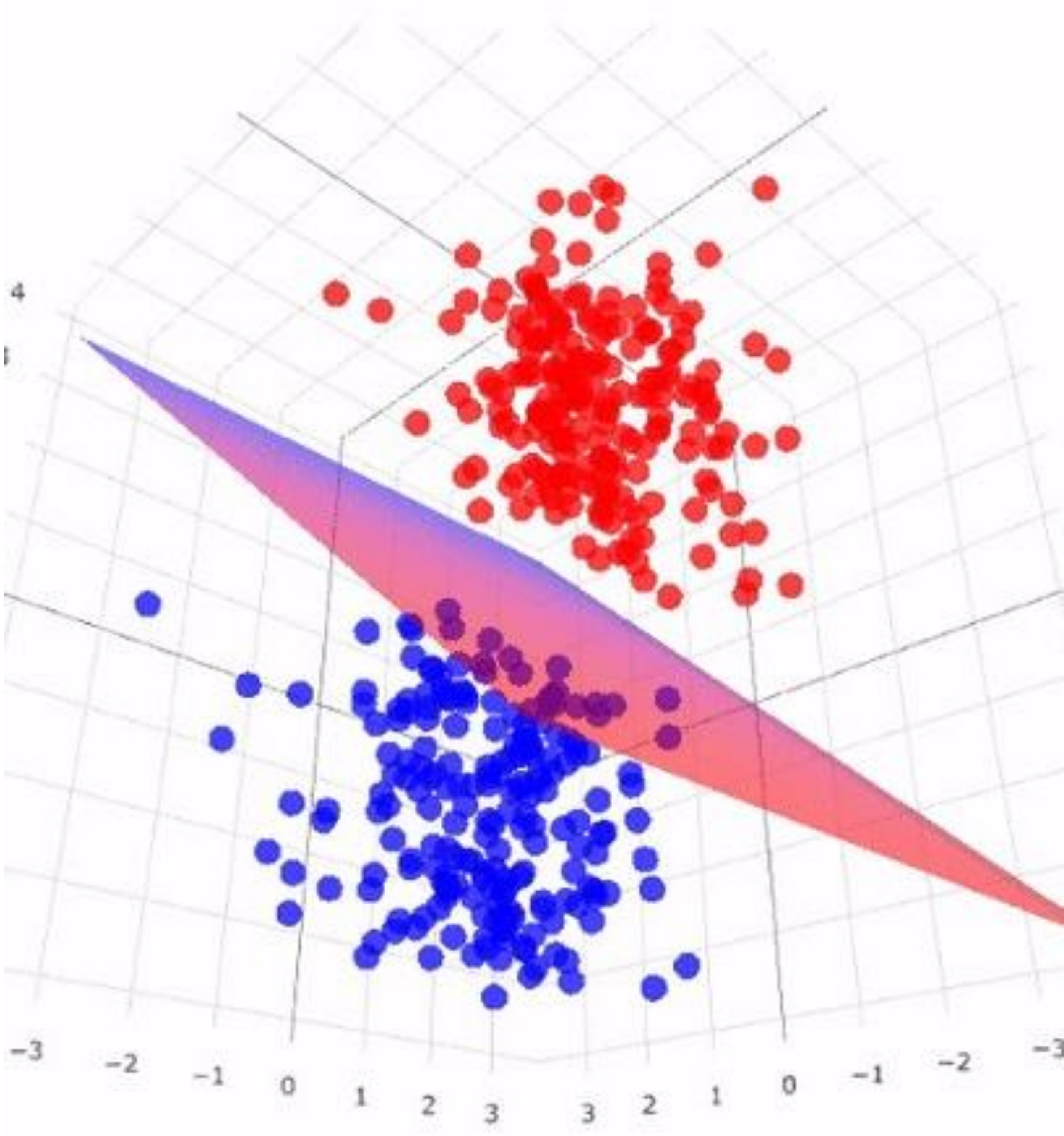
Result



Model:	Accuracy on Test Data	F1 Accuracy
Support Vector Machine (SVM):	95.80%	96%
Logistic Regression:	95.80%	96%
K-Nearest Neighbors (KNN):	95.10%	95%
Decision tree:	92.31%	92%

SVM is better

- SVM can be used for both classification and regression
- SVM tries to find the best distance between support vector and the line that separates the classes therefore lowering the risk of error of the data
- SVM has geometrical approach and is better with semi-structured data
- SVM is les vulnerable to overfitting compared to Logistic Regression


4
−3
−2
−1
0
1
2
3
3
2
1
0
−1
−2
−3