ESP IJAST
ESP International Journal of Advancements in Science & Technology
ISSN: 2583-9233/ Volume 4 Issue 2 April 2026
Paper Id: IJAST-V4I2P106/Page No: 39–47

Research Article

Diabetes Prediction Using Machine Learning Techniques: A Comparative Study

Ishu Pandey[1], Anuradha Misra[2]
[1,2] *Amity School of Engineering & Technology, Amity University Uttar Pradesh, India*

Received Date: 12 March 2026 Revised Date: 24 March 2026 Accepted Date: 11 April 2026

Abstract: *High blood glucose levels are a hallmark of diabetes, a chronic metabolic condition that can cause major side effects like heart disease, renal failure, nerve damage, and eyesight loss if undetected or untreated. Early identification and care are essential to lowering the risk and severity of problems connected to diabetes, which affects more than 537 million people worldwide. However, because of the disease's multifaceted nature and the complexity of patient data, fast and correct diagnosis is still difficult. In order to comprehend the dataset and get it ready for model training, data preparation and exploratory data analysis (EDA) were carried out. A variety of machine learning methods were examined, including Support Vector Machine (SVM), Random Forest, Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression, and Naive Bayes. SVM was chosen as the final model because it yielded the best results in the model comparison. The model's testing accuracy was 77.27% and its training accuracy was 78.66%. Additionally, a prediction algorithm was created that uses input health information to determine if a person has diabetes or not. The findings demonstrate that early diabetes prediction may be effectively supported by machine learning.*

Keywords: *Diabetes Prediction, Machine Learning (Ml), Support Vector Machine (Svm), Pima Indians Diabetes Dataset, Classification, Healthcare Analytics.*

I. INTRODUCTION

Millions of people worldwide suffer with diabetes mellitus, a metabolic condition that can have serious health consequences such renal failure, neuropathy, and cardiovascular disease. Early identification is essential. Despite their effectiveness, traditional diagnostic techniques sometimes include intrusive procedures and could miss the subtle risk factors that contribute to the establishment of diabetes. Healthcare professionals may increase the precision and effectiveness of diabetes prediction by utilizing machine learning, which makes it possible for earlier treatment and improved outcome for patients. Machine learning-based prediction systems' accessibility and potential influence in actual healthcare settings are further increased by their incorporation into digital platforms like web-based applications and applications for mobile devices. Recent developments in machine learning (ML) have made it possible to create prediction models that more accurately and efficiently identify people at risk for diabetes by analyzing complicated information. This paper focuses on building a diabetes prediction system using supervised machine learning algorithms. The study uses the Pima Indians Diabetes Dataset, which is a widely used dataset for diabetes-related prediction tasks. The project includes data preprocessing, Exploratory Data Analysis (EDA), model comparison, and final prediction. Different algorithms were tested, and the Support Vector Machine (SVM) model was selected because it gave the best performance among the compared models

II. LITRETURE REVIEW

Several studies have applied machine learning techniques for early diabetes prediction using the Pima Indians Diabetes Dataset, which is one of the most commonly used benchmark datasets in this area. Researchers have tested different supervised learning models such as Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, and Support Vector Machine to identify the most effective classifier for diabetes detection [1, 2].

Most previous work shows that model performance depends strongly on data preprocessing, feature selection, and evaluation methods. Some studies found Logistic Regression to be effective because of its simplicity and interpretability, while others reported Support Vector Machine and Random Forest as better-performing models for classification on structured medical data [2, 3]. Researchers have also highlighted that proper handling of missing or zero values in medical features can improve prediction quality [4].

Based on the existing literature, it is clear that no single model is always best for every diabetes dataset. However, comparative analysis of multiple machine learning models is a common and reliable approach for building an effective diabetes prediction system. In this project, a similar comparative approach was followed, and Support Vector Machine achieved the best performance among the tested models, making it the final model for diabetes prediction of multiple machine learning models is a common and reliable approach for building an effective diabetes prediction system. In this

project, a similar comparative approach was followed, and Support Vector Machine achieved the best performance among the tested models, making it the final model for diabetes prediction.

III. OBJECTIVE

This project's primary objective is to use machine learning techniques to develop a dependable and effective system that can forecast an individual's risk of developing diabetes. The project is to develop models that can assist in identifying individuals who are at risk of acquiring diabetes through the analysis of health data, allowing for earlier intervention and improved diabetes control.

To be precise, the project will help:

- To determine the best accurate technique for diabetes prediction, experiment with several machine learning algorithms.
- By properly organizing the data and choosing the most pertinent elements, you may increase the forecast accuracy.
- Create an intuitive tool or online application that enables patients or medical professionals to enter health data and obtain a diabetes risk assessment right away.
- Encourage the early detection and management of diabetes with the ultimate goal of lowering complications and enhancing patient outcomes.

By accomplishing these goals, the research hopes to show how machine learning can be extremely helpful in healthcare by improving the accuracy and accessibility of illness prediction.

IV. METHODOLOGY

The methodology of this project is based on building a diabetes prediction system using machine learning (Machine Learning) techniques. First, the Pima Indians Diabetes Dataset was used as the main dataset for the study. This dataset contains several medical attributes such as Glucose, Blood Pressure, Body Mass Index (BMI), Insulin, Age, and other health-related features that help in predicting whether a person is diabetic or not.

In the next step, the dataset was preprocessed to make it suitable for model training. This included checking the data, handling missing or invalid values if required, and separating the input features from the target output. After preprocessing, the dataset was divided into training data and testing data so that the models could be trained and then evaluated properly.

After data preparation, multiple machine learning (Machine Learning) classification algorithms were applied, such as Logistic Regression (Logistic Regression), K-Nearest Neighbors (K-Nearest Neighbors), Decision Tree (Decision Tree), Random Forest (Random Forest), and Support Vector Machine (Support Vector Machine). Each model was trained using the training dataset and tested on the testing dataset. Their performances were compared using accuracy as the main evaluation metric.

Finally, the model with the highest accuracy was selected as the best model for the project. In this study, Support Vector Machine (Support Vector Machine) gave the best performance and was chosen as the final model for diabetes prediction.

V. DATASET DESCRIPTION

The National Institute of Diabetes and Digestive and Kidney Diseases are the original source of this dataset. The dataset's goal is to use certain diagnostic parameters to provide a diagnostic prediction about a patient's likelihood of having diabetes. The selection of these examples from a broader database was subject to a number of restrictions. Specifically, all of the patients are Pima Indian women who are at least 21 years old.

The dataset contains 768 rows and 9 columns. Out of these 9 columns, 8 are input features and 1 is the target variable named Outcome. The Outcome column represents whether the person is diabetic or non-diabetic, where:

- 0 = Non-diabetic
- 1 = Diabetic

A. About the Data and the Features-

- Pregnancies: It is a numerical statistic that usually ranges from 0 to 17 and indicates how many times the patient has become pregnant.
- Glucose: Two hours following an oral glucose tolerance test, glucose tests the plasma glucose levels; the results range from 0 to 199.
- BloodPressure: It measures the diastolic blood pressure, which ranges from 0 to 122 mm Hg.
- SkinThickness: which ranges from 0 to 99 millimeters, is the thickness of the triceps skin fold.
- Insulin: With values ranging from 0 to 846, insulin represents the 2-hour serum insulin level (in mu U/ml).

- Body Mass Index: It is a number between 0 and 67.1 that is computed by dividing weight in kilograms by height in meters squared.
- Diabetes Pedigree Function: which varies from 0.078 to 2.42, is a computed number that predicts the risk of diabetes based on family history.
- Age: It just enters the patient's age in years, from 21 to 81.
- Outcome: The aim variable is outcome, where 0 denotes the absence of diabetes and 1 denotes its existence.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 1: Preview of the Pima Indians Diabetes Dataset

VI. DATA PREPROCESSING

Real-world data is typically unreliable, dirty, unprocessed, and erratic. It could contain missing values, incorrect structures, data input errors, etc. A crucial step in every Data Science project is doing an accurate and effective analysis. It ensures that the data quality is constant before using any kind of machine learning or data mining techniques.

A. Steps involved in preprocessing-

Data cleaning: It involves dealing with missing values in our data since they may provide unintended consequences or reduce the accuracy of our model.

We can deal with missing values in our data in a variety of ways-

- We can remove rows that include missing values, but this approach works best when there aren't many missing values else it affect our model.
- The mean, median, and mode can also be used to fill in the missing data.

```
diabetes_dataset.isnull().sum()
```

	0
Pregnancies	0
Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0
DiabetesPedigreeFunction	0
Age	0
Outcome	0

dtype: int64

Figure 2: Checking for missing values in the dataset

- Data Reductions: This is crucial since there may be a lot of unnecessary data in our dataset that doesn't help us reach our ultimate objective. As a result, we must eliminate this data and condense our dataset into more precise and understandable data that can be utilized for our project going forward.
- Data transformation: To achieve the desired outcomes, the data scientist will choose the most effective method for transforming the various data components in this section. This might involve structuring unstructured data, combining salient variables where appropriate, or selecting important ranges to concentrate on.

First, the dataset was tested for null or missing values. After checking the dataset, it was found that no missing values were present. Therefore, no additional missing value treatment was required.

Next, the dataset was checked for duplicate values, and it was found that no duplicate records were present.

After this, the data was separated into:

- X (input features) – all columns except Outcome
- Y (target variable) – only the Outcome column

Since the dataset contains features with different value ranges, data standardization was performed using StandardScaler from scikit-learn. Standardization is important because some algorithms, especially SVM, perform better when all features are scaled to a similar range.

Finally, the dataset was divided into:

- 80% training data
- 20% testing data

This train-test split was used to evaluate the performance of the model on unseen data.

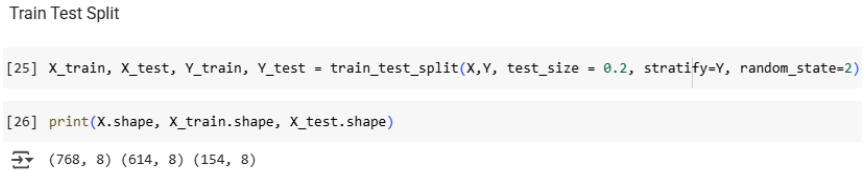

Figure 3: Data standardization and train-test split

VII. EXPLORATORY DATA ANALYSIS

In data science projects ('Exploratory data analysis') EDA plays a very important role. EDA is the method by which we study our data and identify the numerous kinds of variations and patterns that exist within it. Additionally, we can determine how many outliers are in our data and how various dataset attributes relate to one another. Plotting several graphs and charts for our data and identifying hidden patterns are also included in this.

We can identify a variety of patterns and trends in our dataset by conducting exploratory data analysis, which will enable us to make more precise predictions in the future. Analyzing historical information and researching previous trends are also important.

In this project, different types of visualizations were used, such as:

- Count plot for the Outcome variable
- Histogram plots for feature distributions
- Box plots to observe outliers
- Scatter plots to study relationships between features
- Pair plots for multivariable visualization
- Correlation heatmap for feature correlation analysis

The Outcome count analysis showed that the dataset contains more non-diabetic cases than diabetic cases. In the dataset:

- 500 patients are non-diabetic
- 268 patients are diabetic

This indicates that the dataset is slightly imbalanced, but still usable for classification.

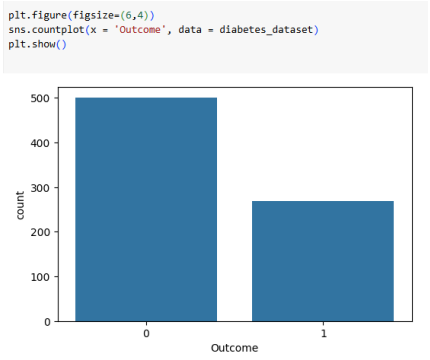

Figure 4: Distribution of diabetic and non-diabetic cases

The histogram plots helped in understanding the distribution of each feature, while box plots highlighted the presence of outliers in variables such as glucose and pregnancies. Scatter plots were used to visualize the relationship between pairs of features such as Age vs BMI and Glucose vs Blood Pressure.

Heatmap-A correlation heatmap displays a 2-dimensional correlation matrix comprising two dimensions by using colored pixels to represent data from a generally monochrome scale. The table row shows the values of first dimension and the table column shows the values of second dimension. The cell's color depends on the amount of measurements that match with the dimensional value.

EDA helped in identifying important patterns in the data and gave a better understanding before model training.

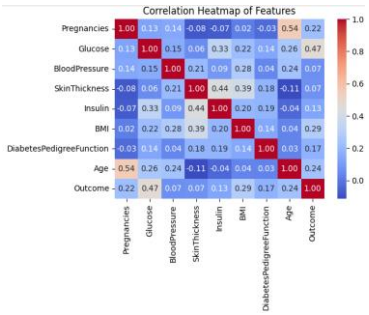

Figure 5: Correlation heatmap of dataset features

VIII. MODEL SELECTION

To determine the most suitable algorithm for diabetes prediction, multiple supervised machine learning models were compared in this work. The models considered in the project were:

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Decision Tree
- Random Forest
- Naive Bayes
- Support Vector Machine (SVM)

```
# Let's evaluate a few different classification algorithms
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score
from sklearn import svm # Import the svm module
```

Figure 6: Comparison of different machine learning models

The training dataset was used to train each model, and its performance was assessed. Following comparison, the Support Vector Machine (SVM) model was chosen as the final model as it demonstrated the highest accuracy of all the evaluated models.

```
Logistic Regression: 75.97%
SVM (Linear): 77.27%
Decision Tree: 67.53%
Random Forest: 74.68%
K-Nearest Neighbors: 72.08%
Naive Bayes: 77.27%
```

Figure 7: Accuracy comparison of different machine learning models

A. Support Vector Machines

A machine learning approach for classification problems is called Support Vector Classifier (SVC). Finding the appropriate hyper plane to divide data points of various classes in a feature space is the main objective of this particular implementation of the more general Support Vector Machine architecture. We need to maximize the margin, which is the distance between the hyper plane and the closest data points from each class, known as support vectors.

Increasing the size of this margin helps support vector classifier to increase the model's capacity to work with new and unknown data. The "kernel trick" refers to SVC's ability to apply kernel functions to convert data into a space with higher dimensions such that linear separation is possible there if the data is not linearly separable. SVC's ability to handle both linear and non-linear problems in classification makes it popular in domains including classification of text, picture recognition, and biotechnology. Many real-world applications choose SVC because of its adaptability and durability.

Kernel- In machine learning, a kernel is a function that implicitly transforms input into a higher-dimensional space, enabling algorithms—particularly Support Vector Machines to handle data that is not linearly separable. Through this procedure, the algorithm is able to determine that the higher-dimensional space's linear border corresponds to the original

space's non-linear border.

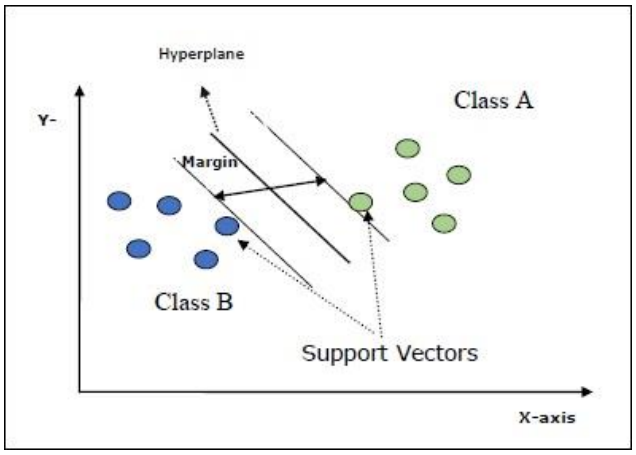

Figure 8: Support vector machine

IX. MODEL TRAINING AND EVALUATION

After selecting the Support Vector Machine (SVM) model, the model was trained using the training dataset. The standardized feature values were used as input to improve the performance of the classifier.

The trained SVM model was then evaluated on both the training and testing datasets.

A. Training Accuracy-

The model achieved a training accuracy of **78.66%.**

```
# accuracy score on the training data
X_train_prediction = classifier.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
```

```
print('Accuracy score of the training data : ', training_data_accuracy)
```

```
Accuracy score of the training data :  0.7866449511400652
```

Figure: 9 Training accuracy

B. Testing Accuracy-

```
# accuracy score on the test data
X_test_prediction = classifier.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
```

```
print('Accuracy score of the test data : ', test_data_accuracy)
```

```
Accuracy score of the test data :  0.7727272727272727
```

Figure 10 :Testing accuracy

C. The model achieved a testing accuracy of 77.27%.

The testing accuracy is slightly lower than the training accuracy, which is expected because the testing dataset contains unseen data. However, the difference between the two accuracies is not very large, which indicates that the model is learning reasonably well and is not heavily overfitting.

```
from sklearn.metrics import classification_report

# Classification report
print(classification_report(Y_test, X_test_prediction, target_names=['Non-Diabetic', 'Diabetic']))
```

```
              precision    recall  f1-score   support

Non-Diabetic       0.78      0.91      0.84       100
    Diabetic       0.76      0.52      0.62        54

    accuracy                           0.77       154
   macro avg       0.77      0.71      0.73       154
weighted avg       0.77      0.77      0.76       154
```

Figure 11: Classification report

The results indicate that the SVM model can provide a fairly reliable prediction for diabetes based on the selected health features. A classification report was also generated in the project to further analyze the performance of the model. This report provides more insight into how well the model performs for both diabetic and non-diabetic classes**.**

X. RESULTS AND DISCUSSION

The Pima Indians Diabetes Dataset was used to train and test various machine learning models. The findings revealed that while each algorithm was able to identify whether or not a person has diabetes, their performance varied somewhat. Logistic regression, K-nearest neighbors, decision trees, random forests, and support vector machines were among the models utilized in this research.

Out of all the examined models, Support Vector Machine provided the highest accuracy, according to the comparison. It was chosen as the final model for the diabetes prediction system as a result. Support Vector Machine performed more consistently on this dataset and generated superior overall prediction results, even if the other models also provided respectable results.

In order to determine whether a person is likely to have diabetes or not, the built predicting method uses critical medical input parameters such blood pressure, glucose, body mass index (BMI), insulin, and age. The trained model is then used to make this prediction. Because of this, the method may be used as an early-stage diabetes risk detection help tool. It can assist in providing a fast predictions based on data from patients, which might promote prompt medical response and awareness.

```
input_data = (5,166,72,19,175,25.8,0.587,51)

# changing the input_data to numpy array
input_data_as_numpy_array = np.asarray(input_data)

# reshape the array as we are predicting for one instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

# standardize the input data
std_data = scaler.transform(input_data_reshaped)
print(std_data)

prediction = classifier.predict(std_data)
print(prediction)

if (prediction[0] == 0):
  print('The person is not diabetic')
else:
  print('The person is diabetic')
```

```
[[ 0.3429808   1.41167241  0.14964075 -0.09637905  0.82661621 -0.78595734
   0.34768723  1.51108316]]
```

Figure 12: Predictive System

```
[1]
The person is diabetic
/usr/local/lib/python3.11/dist-packages/sklearn/utils/
  warnings.warn(
```

Figure: 13 Prediction output

After entering the values for each variable we got the final output as the person is diabetic.The above results shows the predictions for different input data, from this result we can say that our model is working properly. Similarly we can predict the future outcomes for different data. Our model can be further improved and enhanced by using different types of functions, algorithms and adding more data, by improving our model we will be able to predict the future values more precisely, which can help us to determine the changes in the body and it will help to improve medical diagnosis and early detection of diseases so that the person can take precautions and necessary treatments required to cure that disease.

XI. CONCLUSION

In conclusion we have built a diabetes prediction model using machine learning in python. This model analyses the past data of the patients and finds patterns and hidden trends in that data so we can predict diabetes in new patients as diabetes is a chronic metabolic condition that can cause major side effects like heart disease, renal failure, nerve damage, and eyesight loss if undetected or untreated. We have used various important libraries like pandas, numpy, matplotlib, seaborn etc.to preprocess the data, perform exploratory data analysis to find hidden patterns and trends, train the model and predict the disease.

Our model is working efficiently and shows a accuracy of 77.27% on testing data. We know that diabetes prediction is a difficult process and it affected by various numbers of factors like quality of dataset, algorithm used while training the model and many others, all these factors can affect the model's performance.

This project serves as an example of the capabilities and promise of machine learning and data science in diabetes prediction. The support vector classifier model showcases a dependable and effective tool for diabetes prediction, enhancing the capacity of health industries to.This project is a simple diabetes prediction system implementation, and the knowledge we acquire from it will enable us to create more intricate and advanced systems.

To improve the diabetes prediction system's accuracy, practicality we can use multiple ensemble techniques, such as gradient boosting, XGboost, and stacking or blending many models. These techniques can combine various algorithms to improve the prediction system by lowering the risk of overfitting. Further we can also use SMOTE(synthetic minority over-sampling technique) to address the class imbalance problems in the dataset which will help the model to correctly predict diabetes and non-diabetes patients.

XII. REFERENCES

[1] L. Xie, "Pima Indian Diabetes Database and Machine Learning Models for Diabetes Prediction," Highlights in Science, Engineering and Technology, vol. 66, pp. 400–406, 2024.

[2] Y. Tian, "Machine Learning Models for Diabetes Prediction: Logistic Regression, SVM, Random Forest, and Neural Networks," Theoretical and Natural Science, 2025.

[3] J. E. Priyatma and M. R. A. Sasmita, "Comparative Analysis of Random Forest and Support Vector Machine for Classifying Pima Indians Diabetes Dataset," United International Journal for Research & Technology (UIJRT), vol. 6, no. 9, pp. 117–126, 2025.

[4] A. A. Ali, G. R. Galal, and H. S. Hassan, "Diabetes Prediction on Pima Indians Dataset Using Machine Learning Techniques," International Journal of Environmental Sciences, 2025.

[5] Dhanashree S. Medhekar, Mayur P. Bote, Shruti D. Deshmukh, "Heart Disease Prediction System using Naive Bayes", INTERNATIONAL JOURNAL OF ENHANCED RESEARCH IN SCIENCE TECHNOLOGY & ENGINEERING, VOL. 2, ISSUE 3, MARCH.-2013, pp 1-5

[6] Mr. Sunil Navadia, Mr. Jobin Thomas, Mr. Pintukumar Yadav, Ms. Shakila Shaikh, "Weather Prediction: A novel approach for measuring and analyzing weather data", International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), (I-SMAC 2017), IEEE, pp 414-417

[7] Tasin, I., Nabil, T., Rahman, M. M., & Hossain, M. A. (2022). Diabetes prediction using machine learning and explainable AI. BioMed Research International, Article ID 2347326. https://pmc.ncbi.nlm.nih.gov/articles/PMC10107388/

[8] Introduction to Data Mining and Knowledge Discovery, Third Edition, Two Crowds Corporation, http://www.twocrows.com/introdm.pdf, accessed on 12 April 2009.

[9] Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.

[10] A. Gautam and P. Bedi, "MR-VSM: Map Reduce based vector SpaceModel for user profiling-an empirical study on News data," 015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, 2015, pp. 355-360.

[11] Evans, Hastings, and Peacock. "Statistical Distributions," 3rd Ed., John Wiley and Sons, 2000.

[12] Velleman, Paul and Hoaglin, David. "The ABC's of EDA: Applications, Basics, and Computing," 1981.

[13] Draper and Smith. "Applied Regression Analysis," 2nd ed., John Wiley and Sons, 1981.

[14] Efron and Gong. "A Leisurely Look at the Bootstrap, the Jackknife, and Cross Validation," The American Statistician, February 1983.