SUMMER INTERNSHIP PROJECT REPORT
ON
"Data Analysis and Study For Some UCI and Gene Sequences Datasets"

Aliah University

Public university in Kolkata, West Bengal

SESSION: (May 22, 2023 - Jul 22, 2023)

UNDER GUIDANCE

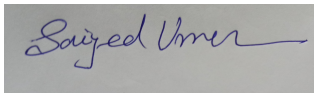
Dr. Saiyed Umer

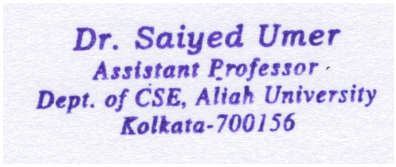
Assistant Professor

SUBMITTED BY

Pravin Kumar

B.Tech Student





Indian Institute of Engineering Science and Technology, Shibpur

Public university located at Shibpur, Howrah, West Bengal

REPORT - 1

Objective - 1: Training and Compute the accuracy of the model with all three datasets IRIS, TITANIC and mobile_price. {Analysis Of Three Various Datasets}

Solution:

1. Abstract
2. Introduction
3. Conclusion
4. References

ABSTRACT

A machine-learning model is similar to a software program. yet it is a clever program that learns from the given datasets. Learning indicates recognising new patterns in the datasets known as the training process. Suppose, a dataset contains pictures of animals.
And model is trained with this dataset. Once it is trained then give it new images of animals it has never seen before, and it will try to figure out what animals are in each image all by itself. Just like how a child learns from their teachers and gets better at solving new problems.

The training dataset should be clean and contain a specific category of data.
Cleaning the dataset indicates removing random rows or values that restrict the making of patterns. Such as outliers, nan values and so on.

This study focuses on the development of a model with titanic, iris and price datasets. A dataset was taken, and cleaned by using the preprocessing method. Then modified dataset was split into training and testing examples. Training examples were used in training the model and once it was trained, the testing examples were used to get the accuracy of the model. Finally, the accuracy of the model shows the overall performance of the model and tells us how well the model has learnt and doing to get things right.

The higher the accuracy, the better it is at solving new problems and making accurate predictions.

Training and Testing were taken in four ways:

1. Training data is 50% and testing data is 50%.
2. Training data is 60% and testing data is 40%.
3. Training data is 70% and testing data is 30%.
4. Training data is 80% and testing data is 20%.

The accuracy of the model in a specific dataset:

1. Titanic, the accuracy is 0.75, 0.756, 0.76, and 0.786.
2. Iris, the accuracy is 0.98, 1, 0.95, and 0.93.
3. Price, the accuracy is 0.65, 0.64, 0.57, and 0.54.

A support vector machine is used as a model.

INTRODUCTION

Note: following steps to run the program.

- Step-1: Open the given Colab file by clicking the link
- Step-2: Upload the mobile_price.csv dataset and remaining dataset access by the seaborn module
- Step-3: Run it
- Step-4: Get the result

Some Term Definition:

1. Machine Learning:

- Machine learning is a field of artificial intelligence that focuses on creating algorithms and models that enable computers to learn and make predictions or decisions without being explicitly programmed.
- It involves training a computer system using data, allowing it to automatically learn patterns, relationships, and insights to make accurate predictions or perform specific tasks.

2. Dataset:

- A dataset is a structured collection of data that represents a particular set of observations, measurements, or information.
- It can consist of various types of data, such as numbers, text, images, or other formats, organized into rows (instances) and columns (variables) for analysis or machine learning tasks.

3. Multiscaling value in a dataset:

- It typically refers to a situation where the dataset contains features or variables that have different scales or units of measurement. In other words, the values in the dataset are not uniformly distributed or measured in the same units.

- Here's a small sample of the dataset:

Area(sq.ft.)	price	Bedrooms
1500	200000	3
2000	300000	4
1200	180000	2

- In this example, we can see that the "Area" feature has values ranging from 1200 to t1500, the "Price" feature has values ranging from 180,000 to 300,000, and the "Number of Bedrooms" feature has discrete values of 2, 3, and 4.

- The multiscaling values in this dataset arise due to the different units and ranges of the variables. "Area" is measured in square feet, "Price" is measured in dollars, and "Number of Bedrooms" is a discrete variable with no specific unit.
- To address this issue, we can apply feature scaling techniques to bring all the variables to a similar scale. Let's use min-max scaling to transform the "Area" and "Price" features to a range between 0 and 1:

Area(sq.ft.)	price($)	Bedrooms
0.4167	0.2857	3
0.6667	0.4286	4
0.1667	0.1429	2

- Now, the "Area" and "Price" features are scaled between 0 and 1, allowing for a more standardized representation of the data. The "Number of Bedrooms" feature remains unchanged since it is a discrete variable.

4. Multi Data Types:

- A dataset with multiple data types refers to a collection of data that includes variables of different types, such as numerical, categorical, text, or date/time. Each data type represents a distinct kind of information and may require specific handling and analysis techniques. Understanding and appropriately managing these data types is crucial for accurate analysis and modelling of the data.

- Here’s a small sample of the dataset:

Age	Gender	Education Level	Income	Review	Date of purchase
30	Male	Bachelor’s	50k	“Excellent Product”	2022-5-10
45	Female	Master’s	80k	“Not satisfied with the quality”	2022-5-11
28	Female	High School	25k	“Great value for money”	2022-5-12
35	Male	Bachelor’s	60k	“Average Product”	2022-5-13
50	Male	Master’s	90k	“The best product”	2022-5-14

In this example, we have a dataset with multiple data types:

- “Age” and “Income” are numerical variables representing age and income, respectively.
- “Education Level” is an ordinal variable representing different levels of education.
- “Review” is a text variable containing customer reviews.
- “Date of Purchase” is a date/time variable representing the date of purchase.

5. **SVM Model:** Support Vector Machine (SVM) is a machine learning model that aims to find the best decision boundary or hyperplane to separate data points of different classes, maximizing the margin between the classes.

6. **Training data:** The portion of the dataset used to train or fit a machine learning model. It consists of input features and corresponding target labels. The model learns from this data to understand the underlying patterns and relationships in the data.

7. **Testing data**: The portion of the dataset used to evaluate the performance of the trained machine learning model. It is separate from the training data and contains input features, but the corresponding target labels are typically withheld. The model makes predictions on this data, and the predictions are compared to the actual labels to assess the model's accuracy or performance.

8. **UCI:** Its full form is “the University of California, Irvine”. This is a machine learning repository that is a collection of a large number of various datasets that are used in machine learning projects. Currently, It contains 624 datasets as a service to the machine learning community.
   The website link is https://archive.ics.uci.edu/

9. **Data Analysis:** This is a process including data preprocessing, extracting valuable information or pattern from data and data visualization with charts, graph and images so everyone can understand information in dataset and use dataset in their solving problems.
   data analysis is the key to unlocking the hidden knowledge within the data.
   The webpage link https://en.wikipedia.org/wiki/Data_analysis

When the dataset contains multiscaling and multi-data types values:

The steps for data analysis are,

- Missing value
- Balancing
- Normalization
- Splitting
- Classifier or model (using SVM model)

We worked on three datasets:

1. IRIS

- The Iris dataset is a famous dataset in machine learning and statistics.
- It consists of measurements of sepal length, sepal width, petal length, and petal width of three different species of iris flowers.
- The goal of using the Iris dataset is often to classify the iris flowers into their respective species based on these measurements.
- It is commonly used for tasks such as classification, clustering, and data visualization.

Case 1: Training data is 50% and testing data is 50%
The accuracy of the model is 98%.

Case 2: Training data is 60% and testing data is 40%
The accuracy of the model is 100%.

Case 3: Training data is 70% and testing data is 30%
The accuracy of the model is 95%.

Case 4: Training data is 80% and testing data is 20%
The accuracy of the model is 93%.

2. TITANIC

- The Titanic dataset is a dataset that contains information about the passengers onboard the Titanic ship.
- It includes attributes like age, gender, passenger class, fare, and survival status (whether the passenger survived or not).
- The dataset is often used for predictive modelling, aiming to predict whether a given passenger survived the Titanic disaster based on the available attributes.
- It is a popular dataset for learning data analysis and machine learning techniques, particularly for classification or survival prediction tasks.

Case 1: Training data is 50% and testing data is 50%
The accuracy of the model is 75%.

Case 2: Training data is 60% and testing data is 40%
The accuracy of the model is 75.36%.

Case 3: Training data is 70% and testing data is 30%
The accuracy of the model is 75.98%.

Case 4: Training data is 80% and testing data is 20%
The accuracy of the model is 78.67%.

3. PricePrediction

- The PP dataset contains features of the cell phone.
- It contains features such as Battery, Screen size, RAM, Storage and so on.
- The goal of this dataset is to train the model and model predict the price of cell phones by taking features as input.

Case 1: Training data is 50% and testing data is 50%
The accuracy of the model is 65%.

Case 2: Training data is 60% and testing data is 40%
The accuracy of the model is 64.6%.

Case 3: Training data is 70% and testing data is 30%
The accuracy of the model is 57%.

Case 4: Training data is 80% and testing data is 20%
The accuracy of the model is 54%.

Tech stack used:

- Numpy - It is a Python library. It is faster and has a better performance on large datasets.
- Python - It is a language for writing software programs.
- Pandas - It is also a Python library. It is more user-friendly and has a lot more options for handling missing data.
- Important modules for machine learning models.
- Important modules for data preprocessing method.

CONCLUSION

- Nature of features present in the database.
- How Classifier learns on those features of the dataset.
- Considering different cases, the model accuracy variation is seen.
- Only at 60-40 split in the iris dataset provides 100% accuracy.
- Learned to deal with random examples and make-to-model training examples.
- Learned to use various Python libraries.

REFERENCES

Dataset

- Titanic - https://www.kaggle.com/competitions/titanic
- Iris - https://www.kaggle.com/datasets/uciml/iris
- Price - https://docs.google.com/spreadsheets/d/1rv8jH0Dg8jUIkU_Dxfe587vGiTtKXffsFLcnHVn1gEg/edit#gid=1154984736

Numpy

- https://numpy.org/doc/

SVM Model

- https://scikit-learn.org/stable/modules/svm.html

Data Preprocessing method

- https://en.wikipedia.org/wiki/Data_pre-processing
- https://github.com/alod83/data-science/tree/master/Preprocessing

UCI Machine Learning Repository

- https://archive.ics.uci.edu/

REPORT - 2

Objective 2: Compute features of given gene sequence dataset and accuracy of the model. {Study Of Gene Sequences}

Solution:

1. Abstract
2. Introduction
3. Conclusion
4. References

ABSTRACT

Gene sequence is also known as genetic code. In other words the arrangement of the nucleotides of a particular gene. For example, “ATCGTACGTA” is a part of the gene sequence where A, C, G, and T are only four nucleotides. adenine (A), cytosine (C), guanine (G), and thymine (T). It always makes pairs in gene sequence (A-T and C-G). simply, Genetic sequence represents the DNA of a living object.
Each person's DNA contributes to differences in appearance and health. People who are closely related have more similar DNA.
Human DNA is 99.9% identical from person to person. Although a 0.1% difference doesn’t sound like a lot, it actually represents millions of different locations within the genome where variation can occur. And differentiate the persons from each other.

This study focuses on the development of a model with gene sequence and exploring the given gene sequence dataset.
Taken three gene sequence datasets, each dataset represents one kind of living object. Then loaded the dataset, extracted features using Shanon entropy, hurst exponent, and run length encoding features method and after getting features, cleaned the features dataset using preprocessing method, now modified dataset splits into training and testing examples. Training examples were used to train the model and testing examples were used to compute the accuracy of the model. The accuracy of the model shows the how well model has learned.

Four ways we have taken features as training dataset:

1. Used Shanon entropy features
2. Used Hurst exponent features
3. Used Run length encoding features
4. Used Shanon entropy and Hurst exponent as multi-features

Accuracy of the model in the four above cases:

1. 0.38
2. 0.43
3. 0.61
4. 0.47

The SVM model is used.

INTRODUCTION

Note:

- Step-1: Open the given Colab file of the gene sequence
- Step-2: Load the AR, DOM, and PR folder (contains gene sequence data) in Colab
- Step-3: Run it
- Step-4: The feature will show in result with the accuracy of the SVM model as well.

Some Terms Definition:

- **Gene Sequence**: The **sequence** tells scientists the kind of **genetic** information that is carried in a particular **DNA** segment. It contains a huge number of characters in one sequence and there are only four characters A, C, G, and T that represent amino acids.

  Each gene sequence is composed of a series of nucleotides (A, T, G, C), and the distribution of these nucleotides within the sequence can provide insights into the sequence's characteristics.

  For example,
  A small DNA sequence: ATGGCTCCAACTCAGGATCCCAACAGTGT

- **Feature:** A feature refers to an individual measurable property or characteristic of a data sample. Features are typically represented as columns in a dataset, and they provide information that can be used to make predictions or perform analyses.

- **Label:** Set the label to each sample in entire datasets, so we can categorize them.

  Each given gene sequence folder AR, DOM, and PR has its own category.
  For example,

  Feature — class
  AR ---- 0,
  DOM —-- 1,
  PR ---- 2

- **SVM model:** Support Vector Machine (SVM) is a machine learning model that aims to find the best decision boundary or hyperplane to separate data points of different classes, maximizing the margin between the classes.

- **Testing dataset:** The portion of the dataset used to train or fit a machine learning model. It consists of input features and corresponding target labels. The model learns from this data to understand the underlying patterns and relationships in the data.

- **Training dataset:** The portion of the dataset used to evaluate the performance of the trained machine learning model. It is separate from the training data and contains input features, but the corresponding target labels are typically withheld. The model makes predictions on this data, and the predictions are compared to the actual labels to assess the model's accuracy or performance.

For calculating the features of gene sequence we perform the following concept:

1. **Shannon Entropy Feature:** Shannon entropy, also known as information entropy, is a measure of the uncertainty or randomness in a dataset or information source. It quantifies the average amount of information or surprise contained in each data point.

   For Example,
   Original sequence: "ATCGATCGATCGATCG"
   Numeric array: [0, 3, 1, 2, 0, 3, 1, 2, 0, 3, 1, 2, 0, 3, 1, 2]
   Compute_SE_feature(Numeric array): 1.13

   Case 1: Training data = 50% and Testing data = 50%
   Model Accuracy: 38%

2. **Hurst Exponent Feature:** The Hurst exponent is a statistical measure that quantifies the long-term memory or persistence of a time series data. It is named after Harold Edwin Hurst, a British hydrologist who first introduced this concept.
   The Hurst exponent is a measure of the long-term memory or self-similarity of a time series data. It can also be applied to analyze DNA sequences to capture certain characteristics or patterns in the data.

   For Example,

   Original sequence: "ATCGATCGATCGATCG"
   Numeric array: [0, 3, 1, 2, 0, 3, 1, 2, 0, 3, 1, 2, 0, 3, 1, 2]
   Compute_HE_feature(Numeric array): 0.03 (Import hurst module to use this function)

   Case 1: Training data = 50% and Testing data = 50%
   Model Accuracy: 43%

3. **Run Length Encoding Feature:** Run-length encoding (RLE) is a simple compression technique that can also be used as a feature extraction method. It involves representing consecutive repeated elements in a sequence with a count of the element and the element itself.
   Example,

   sequence = "AATTTCCGGGGGAA"
   Features = [(2, 'A'), (3, 'T'), (2, 'C'), (5, 'G'), (2, 'A')]
   Numeric_features = [2,3,2,5,2]

For computing statistic feature, we have to find following terms:

- **Energy:** Energy measures the total power or magnitude of a dataset. It is calculated as the sum of squared values.

- **Skewness:** Skewness measures the asymmetry of a dataset's distribution. A positive skewness indicates that the distribution has a longer tail on the right side, while a negative skewness indicates a longer tail on the left side. The 'skew' function from Scipy calculates the skewness of the dataset.
  Gene Z: [1,2,3,4,5,20]
  The skewness of Gene Z will be positive, as it has a longer tail on the right side due to the outlier value 20.

- **Kurtosis:** Kurtosis measures the "tailedness" of a probability distribution. It indicates whether the distribution has heavier or lighter tails compared to a normal distribution.

  For example,
  in Leptokurtic Distribution (High Kurtosis),
  Suppose we have a dataset representing the number of goals scored by a football team in 10 matches: [0, 1, 2, 4, 6, 8, 10, 12, 14, 16].
  in Platykurtic Distribution (Low Kurtosis),
  consider another dataset representing the temperature in degrees Celsius on 10 consecutive days: [20, 20.5, 21, 21.5, 22, 22.5, 23, 23.5, 24, 24.5]

- **Entropy:** Entropy measures the amount of uncertainty or information content in a dataset. It is higher for more diverse datasets and lower for more uniform datasets.

  Sequence 1: "ATCGATCGATCG" (more diverse)
  Sequence 2: "AAAAAAAAAAAA" (more uniform)
  The entropy of Sequence 1 will be higher than that of Sequence 2.

- **Correlation:** Correlation measures the linear relationship between two variables. A correlation of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

Gene A: [1, 2, 3, 4, 5]
Gene B: [2, 4, 6, 8, 10]
The correlation between Gene A and Gene B will be 1, as they show a perfect positive linear relationship with each other

- **Mean:** The mean is the average value of a dataset, calculated by summing all values and dividing by the number of data points.

- **Std_dev:** The standard deviation measures the spread or dispersion of data points from the mean. It quantifies how much the values deviate from the average value.

  Gene X: [10, 15, 12, 18, 20]
  Mean = 15.0
  Deviations: [-5, 0, -3, 3, 5]
  Squared Deviations: [25, 0, 9, 9, 25]
  Variance = (25 + 0 + 9 + 9 + 25) / 5 = 13.6
  Standard Deviation = $\sqrt{(13.6)} \approx 3.68$

Hence,
Compute_statical_information(numeric_features) = [Kurtosis, Entropy, Correlation, Mean, Std_dev, energy, skewness] = [0.2,0.4,0.6,0.23,0.11,0.44,0.5]

Case 1: Training data = 50% and Testing data = 50%
Model Accuracy: 61%

4. **Shanon entropy and Hurst exponent used together:**
   Export a CSV file with both features as two columns and label as 3rd column.

Case 1: Training data = 50% and Testing data = 50%
Model Accuracy: 47%

Used Tech Stack:

- Numpy - It is a Python library. It is faster and has a better performance on large datasets.
- Python - It is a language for writing software programs.
- Pandas - It is also a Python library. It is more user-friendly and has a lot more options for handling missing data.
- Important modules for machine learning models.
- Important modules for data preprocessing method

CONCLUSION

- Shannon entropy features and Hurst exponent features train the models in a similar way. But the Run length encoding features is applied in a different way.
- Run length encoding provides more accuracy than all features.
- After taking two features Shanon entropy and Hurst exponent as training datasets,
  Then model accuracy is increased a bit.
- Learned preprocessing method to clean the datasets.
- Learned various methods to find features.
- Model accuracy is between 40% and 60%.

Result of the accuracy of the model with all features given below in the tabular format:

Features method	**Training data**	**Testing data**	**Accuracy of model**
Shanon Entropy	50%	50%	38%
Hurst Exponent	50%	50%	43%
Run Length Encoding	50%	50%	61%
Shanon Entropy and Hurst Exponent	50%	50%	47%

REFERENCES

Dataset:
AR gene sequence - https://drive.google.com/drive/u/2/folders/1mBmxXefJ5j47atTM9ftSdir1bN9zSfI5

DOM gene sequence -
https://drive.google.com/drive/u/2/folders/1E3hm_S2fPNfrMKkNM4c0GlMxg3-HBa6U

DR gene sequence -
https://drive.google.com/drive/u/2/folders/1jrkB4idjmRAufsa0pJYftyExqDeefaEo

Numpy Python - https://numpy.org/doc/

DNA sequence - https://en.wikipedia.org/wiki/DNA_sequencing

SVM Model

- https://scikit-learn.org/stable/modules/svm.html

Data Preprocessing method

- https://en.wikipedia.org/wiki/Data_pre-processing
- https://github.com/alod83/data-science/tree/master/Preprocessing