**Machine Learning Based Residential Electricity Theft Detection**

**COMMUNITY SERVICE PROJECT**

***Submitted by***

**T.HARSHA VARDHAN (99210041292)**

**T.SHANTAN(9921004711)**

**U.SASI KUMAR(99210041646)**

**T.ROHIN(9921004717)**

***in partial fulfillment for the award of the degree***

***of***

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE AND ENGINEERING**

****

**SCHOOL OF COMPUTING**

**COMPUTER SCIENCE AND ENGINEERING**

**KALASALINGAM ACADEMY OF RESEARCH**

**AND EDUCATION**

**KRISHNANKOIL 626 126**

Academic Year 2023-2024

**DECLARATION**

We affirm that the project work titled **“Machine Learning Based Residential Electricity Theft Detection”** being submitted in partial fulfilment for the award of thedegree of **Bachelor of Technology in Computer Science and Engineering** is the original work carried out by us. It has not formed the part of any other project work submitted for award of any degree or diploma, either in this or any other University.

**T.HARSHA VARDHAN**

(99210041292)

**T.SHANTAN**

(9921004711)

**U.SASI KUMAR**

 (99210041646)

**T.ROHIN**

(9921004717)

A person wearing sunglasses

Description automatically generated with low confidence

**BONAFIDE CERTIFICATE**

 Certified that this project report **““Machine Learning Based Residential Electricity Theft Detection”** is the bonafide work of “**T.HARSHAVARDHAN, T.SHANTAN, U.SASI KUMAR, T.ROHIN”** who carried out the project work under my supervision.

**SUPERVISOR        HEAD OF THE DEPARTMENT**

**Mr.C.Sivamurugan**                                                 **Dr. N. Suresh Kumar**

Assistant Professor                                                               Professor & Head

Computer Science and Engineering                             Computer Science and Engineering Engineering Kalasalingam Academy of Research

Kalasalingam Academy of Research                            and Education

Education .   Krishnankoil 626126

Krishnankoil,626126

Submitted for the Project Viva-voce examination held on.......................................

\

**Supervisor                                          Faculty Advisor                                            External Examiner (s)**

**ACKNOWLEDGEMENT**

First and foremost, I wish to thank the **Almighty God** for his grace and benediction to complete this Project work successfully. I would like to convey my special thanks from the bottom of my heart to my dear **Parents** and affectionate **Family members** for their honest support for the completion of this Project work.

I express deep sense of gratitude to “Kalvivallal” Thiru. **T. Kalasalingam** B.com., Founder Chairman, “Ilayavallal” **Dr. K. Sridharan,** Ph.D., Chancellor, **Dr. S. Shasi Anand**, Ph.D., Vice President (Academic), **Mr. S. Arjun Kalasalingam** M.S., Vice President (Administration), **Dr. S. Narayanan**, Vice-Chancellor, **Dr. V. Vasudevan,** Ph.D., Registrar, **Dr. P. Deepalakshmi,** Ph.D., Dean (School of Computing), **Dr. N. Suresh Kumar**, Professor & Head, Department of CSE, Kalasalingam Academy of Research and Education for granting the permission and providing necessary facilities to carry out Project work.

I would like to express my special appreciation and profound thanks to my enthusiastic Project Supervisor **Mrs. J. Benita**, Assistant Professor/CSEof Kalasalingam Academy of Research and Education (KARE) for his inspiring guidance, constant encouragement with my work during all stages. I am extremely glad that I had a chance to do my Project under my Guide, who truly practices and appreciates deep thinking.

I will be forever indebted to my Faculty Advisor ‘Mr.C.Sivamurugan**’** for all the time. He gave me the moral support and the freedom I needed to move on.

Besides my Project guide, I would like to thank the committee members, all faculty members and non-teaching staff for their insightful comments and encouragement. Finally, but by no means least, thanks go to all my school and college teachers, well wishers, friends for almost unbelievable support.

****

**School of Computing**

**Department of Computer Science and Engineering**

**Project Summary**

|  |  |  |
| --- | --- | --- |
| Project Title | Machine Learning Based Residential Electricity Theft Detection | |
| Project Team Members (Name with Register No) | 1. T.HARSHA VARDHAN (99210041292) 2. T.SHANTAN (9921004711) 3. U.SASI KUMAR (99210041292) 4. T.ROHIN (9921004717) | |
| Guide Name/Designation | Mr.C.SIVAMURUGAN,Assistant Professor, Department of Computer Science and Engineering | |
| Program Concentration Area | Residential Theft Detection | |
| Technical Requirements | Streamlit is used by the developer to complete the project. | |
| Engineering standards and realistic constraints in these areas: (Refer Appendix in page 4 of this doc.) | | |
| **Area** | **Codes & Standards / Realistic Constraints** | **Tick ✓** |
| Economic |  |  |
| Environmental |  |  |
| Social |  |  |
| Ethical |  |  |
| Health and Safety |  |  |
| Manufacturability |  |  |
| Sustainability |  |  |

**ABSTRACT**

This study examined the indiscriminate theft of electricity, which is classified as a non-technical loss and affects both customers and electric distribution companies. It can have major repercussions, such as fires and blackouts. The goal of the study was to identify the most effective machine learning prediction model for electrical energy theft. A real-time dataset released by the E&D Corporation of India served as the source of statistics on the electricity use of 155 home clients. Feature extraction was the technique employed to enhance the detection of energy theft. There were eight machine learning models examined. Thus, 93.5% was the accuracy indicator for the SVM model, 93% for K-Nearest Neighbors, 96.7% for Random Forest, 96.7% for Logistic Regression, 93.4% for Naive Bayes, 93% for Decision Tree, 83.8% for Kmeans, and 93% for Gradient Booster. It is determined showed the Random Forest plus Logistic Regression model yields the highest results, with an accuracy of 96.7%.

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO** | **CONTENTS** | **PAGE NO** |
|  | **ABSTRACT** |  |
| 1) | **INTRODUCTION** |  |
| 2) | **LITERATURE REVIEW** |  |
| 3) | **PROBLEM DEFINITION** |  |
| 4) | **REQUIREMENTS** |  |
| 5) | **SYSTEM DESIGN** |  |
| 6) | **PROPOSED APPROACH** |  |
| 7) | **MODULE DESCRIPTION** |  |
| 8) | **IMPLEMENTATION AND RESULT**  8.1.Coding  8.2.Experiment Result  8.3.Output screenshot |  |
| 9) | **CONCLUSION AND FUTURE ENHANCEMENT**  9.1.Conclusion  9.2.Future Enhancement |  |
| 10) | **REFERENCES** |  |

**INTRODUCTION**

A variety of methods are used in machine learning-based home electricity theft detection to pinpoint instances of illicit electricity use. Usually, this is done to guard against energy theft, guarantee accurate invoicing, and preserve the integrity of the electrical infrastructure. Using smart meters, tracking voltage and current levels, evaluating usage trends, and applying data analytics to spot abnormalities are some techniques for identifying electricity theft. In addition to costing utility companies money, illegal electricity diversion puts community safety in danger and puts stress on the infrastructure supporting power distribution. The importance of effective energy management has increased due to the world's growing energy consumption and the requirement for sustainable energy resources. Utility companies encounter a significant obstacle in the form of domestic energy theft, a widespread problem that not only results in significant financial losses but also presents risks to public safety and inefficiencies in operations. The inaccuracies and delays in traditional theft detection systems make sophisticated technological interventions necessary. Within this framework, machine learning (ML) presents itself as a potent instrument for revolutionizing the field of electricity theft detection. Machine learning algorithms are a promising way to find abnormalities in electricity consumption that might be signs of theft because of their capacity to recognize complex patterns within big and complicated datasets. This study explores the field of machine learning (ML)-based household electricity theft detection with the goal of creating advanced models that can precisely identify and stop theft in real-time.This study analyzes the immense potential of predictive analytics by utilizing machine learning approaches, such as deep learning structures like neural networks and supervised learning algorithms like decision trees and support vector machines. Voltage swings, unusual usage patterns, and consumption patterns are just a few of the features that are thoroughly examined in this research of several elements taken from historical electricity usage data. Additionally, the study looks into how anomaly detection techniques might be included to identify minute abnormalities that could slip past traditional detection systems. The effective adoption of this measure would have significant ramifications not only for utility companies looking to reduce revenue losses and improve operational efficiency, but also for consumers who will benefit from more equitable billing procedures and a more secure electricity supply. This research endeavors to advance the field of electricity theft detection, contributing valuable insights that bridge the gap between theoretical advancements and practical applications, thereby paving the way for a more sustainable and equitable energy future.

**LITERATURE REVIEW**

1)The article Decision Tree based Electricity Theft Detection in Smart Grid focuses on the detection of electricity theft using machine learning algorithms. The authors implement decision tree, random forest, and gradient boosting methods on power consumption data collected from 114 single-family apartments. The goal is to detect non-technical loss, which refers to energy theft generated by various scenarios.

2) The article A practical feature-engineering framework for electricity theft detection in smart grids combines clustering and feature engineering technique to improve performance of fraud detection models. The study utilized demand data from over 4000 households and examined six different attack scenarios using five machine learning algorithms. The study compared the performance of different machine learning algorithms and found that GBM performed the best in detecting fraud.

3) In article Electricity theft detection using Empirical mode decomposition and K-nearest neighbor the significance of electricity in modern life is undeniable, with its crucial role in various sectors. Addressing electricity theft, particularly in countries like Pakistan, is essential for economic stability. Implementing smart grids, such as through smart meters, offers a promising solution to detect and reduce power losses from theft. Utilizing thirteen features like Mean, Standard Deviation, and others, the Fine KNN classifier achieved an accuracy of 91.0%.

4) The article "Electricity Theft Detecting Based on Density-Clustering Method" claims that unusual conduct by electricity users, especially electricity theft, has resulted in large financial losses for power companies globally. The methods based on clustering do not require unlabeled data. These methods take patterns from a vast amount of user features and use them to find outlier patterns. The confusion matrix divides the whole dataset into four categories: true positive (TP), false positive (FP), false negative (NP), and true negative (TN). TP, FP, FN, and TN.

5) In the article Real-time power theft monitoring and detection system with double connected data capture system the authors used a GSM (Global System for Mobile Communications) module for electricity theft detection can be a viable solution. To detect electricity theft, the GSM module needs to be connected to the smart meter that records energy consumption

**PROBLEM DEFINITION**

The challenge lies in developing a machine learning system to identify instances of domestic electricity theft, which is a serious problem that puts safety and profitability at risk. Using historical electricity usage statistics, customer information, and suspected tampering indicators, the objective is to create a highly accurate model. While balancing privacy issues, this system needs to be able to spot anomalous consumption patterns. Utility firms stand to gain from the project by reducing losses and increasing safety via real-time detection, which will ultimately increase the effectiveness of electricity distribution. Theft of electricity is a prevalent problem in many places. It can manifest itself in a variety of ways, including illegal connections, meter manipulation, and meter bypassing. The process of manually identifying this kind of theft is laborious and usually unsuccessful. Therefore, it's necessary to have an intelligent system that can identify odd usage patterns or other indicators of electricity theft.

**REQUIREMENTS**

1)Obtaining Data:

Access to historical data on electricity usage, client details, meter data, and maybe weather data.

2)Assurance of Data Quality:

Preparation and cleaning are required to ensure data accuracy and completeness.

3)Feature Creation:

Development of relevant characteristics to identify trends in power theft.

4)Model for Identifying Anomalies:

Selecting suitable machine learning methods for anomaly detection.

5)Model Guidance:

The selected model is trained using historical data.Model Evaluation

Evaluation criteria including the F1-score, AUC-ROC, accuracy, precision, and recall are used to assess the performance of the model.

6)Important Privacy Considerations:

Actions to stop theft and ease residential clients' privacy worries.

**SYSTEM DESIGN**

**PROPOSED APPROACH**

Our proposed system for residential electricity theft detection integrates cutting-edge machine learning techniques to create an efficient and accurate solution. The system consists of three main stages: data collection, feature engineering, and machine learning model implementation. Firstly, a comprehensive dataset of residential electricity usage patterns is gathered, ensuring diversity and representativeness. During feature engineering, relevant features such as consumption behavior, voltage irregularities, and temporal usage patterns are extracted to enhance the model's predictive power. Subsequently, various machine learning algorithms, including decision trees, random forests, Logistic Regression and K-Nearest neighbor are employed to build robust predictive models. These models are trained and validated using advanced techniques like cross-validation. Additionally, the system incorporates anomaly detection methods to identify suspicious activities indicative of electricity theft. By deploying this proposed system, utility providers can significantly enhance their capability to detect and prevent residential electricity theft, leading to reduced energy losses, fair billing practices, and overall improvements in operational efficiency.

**MODULE DESCRIPTION**

**1)Streamlit:**

Streamlit is the primary library that enables you to use Python to develop interactive web applications.

**2)Bringing numpy into NumPy as np:**

NumPy is used for numerical computations and array operations, which are commonly needed for data preprocessing and manipulation.

**3)pandas (please import as pd):**

In addition to other data manipulation tasks, Pandas is used to read and preprocess datasets.

**4)Bringing in matplotlib.pyplot from Matplotlib as a plt:**

Matplotlib is a popular tool for creating charts and data visualizations. It is used in your code to plot an accuracy comparison bar chart between the models.

**5) Scikit-Learn:**

Scikit-Learn is a powerful machine learning tool with a wealth of modules and classes for preprocessing data, training machine learning models, and evaluating their efficacy.

**6)**Specific modules that were used

* Use StandardScaler for feature scaling.Use the train\_test\_split function to split the data into training and test sets.
* The KNeighborsClassifier of the K-Nearest Neighbors model.Support Vector Computer is referred to as SVC.
* Use DecisionTreeClassifier for the Decision Tree model.Use RandomForestClassifier for the Random Forest model.
* Use LogisticRegression for the Logistic Regression model.For the Naive Bayes model, use GaussianNB.Use KMeans for K-Means clustering.Use GradientBoostingClassifier for the Gradient Boosting model.These libraries are required for many different project tasks, including handling data, creating user interfaces, building modelsInstalling these libraries typically requires using a package manager like conda or pip.

**IMPLEMENTATION AND RESULT**

**CODING:**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

df=pd.read\_csv("dataset.csv")

df

df.info()

df.drop(['date','id'],axis=1,inplace=True)

x=df.drop(['flag'],axis=1)

x

y=df['flag']

y

from sklearn.preprocessing import StandardScalerscaler=StandardScaler()

x\_scaled=scaler.fit\_transform(x

n\_clusters=2

from sklearn.model\_selection import train\_test\_split

from sklearn.neighbors import KNeighborsClassifier

from sklearn.metrics import accuracy\_score

X\_train, X\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.2, random\_state=0)

knn = KNeighborsClassifier(n\_neighbors=3)

knn.fit(X\_train, y\_train)

y\_pred = knn.predict(X\_test)

accuracy = accuracy\_score(y\_test, y\_pred)

print('Accuracy of KNN classifier:', accuracy)

from sklearn.svm import SVC

svm=SVC(kernel='linear')

svm.fit(X\_train,y\_train)

y\_pred=svm.predict(X\_test)

accuracy1=accuracy\_score(y\_test,y\_pred)

accuracy1

from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(n\_estimators=100)

rf.fit(X\_train, y\_train)

y\_pred1 = rf.predict(X\_test)

accuracy2 = accuracy\_score(y\_test, y\_pred1)

print('Accuracy:', accuracy2\*100)

from sklearn.tree import DecisionTreeClassifier

dt = DecisionTreeClassifier()

dt.fit(X\_train, y\_train)

y\_pred2 = dt.predict(X\_test)

accuracy3 = accuracy\_score(y\_test, y\_pred2)

print('Accuracy:', accuracy3)

from sklearn.linear\_model import LogisticRegression

lr = LogisticRegression()

lr.fit(X\_train, y\_train)

y\_pred3 = lr.predict(X\_test)

accuracy4 = accuracy\_score(y\_test, y\_pred3)

print('Accuracy:', accuracy4)

from sklearn.naive\_bayes import GaussianNB

NB = GaussianNB()

NB.fit(X\_train, y\_train)

y\_pred=NB.predict(X\_test)

accuracy5=accuracy\_score(y\_test,y\_pred)

print(accuracy5)

from sklearn.cluster import KMeans

kmeans = KMeans(n\_clusters=3)

kmeans.fit(X\_train,y\_train)

y\_pred=kmeans.predict(X\_test)

accuracy6=accuracy\_score(y\_test,y\_pred)

accuracy6

from sklearn.ensemble import GradientBoostingClassifier

gbc = GradientBoostingClassifier(n\_estimators=300,learning\_rate=0.05,random\_state=100,max\_features=5 )

gbc.fit(X\_train,y\_train)

y\_pred=gbc.predict(X\_test)

accuracy7=accuracy\_score(y\_test,y\_pred)

accuracy7

u=['KNN','SVM','RForest','DTree','Logistic','NaiveBias','kmeans','Gradientboster']

y=[accuracy,accuracy1,accuracy2,accuracy3,accuracy4,accuracy5,accuracy6,accuracy7]

plt.figure(figsize =(8.5, 5))

plt.xlabel('Algorithms')

plt.ylabel('accuracy')

plt.plot(u,y)

plt.figure(figsize=(9,5))

plt.bar(u,y)

lr.predict([[0.0995,0.137396,0.572,48,0.094495,6.595,0.065]])

lr.predict([[0.7575,0.778458,1.991,700,0.497389,51.366,0.201]])

y=gbc.predict([[0.485,0.432045,0.868,22,0.239146,9.505,0.072]])

if (y==0):

print('Faithfull')

else:

print('Unfaithfull')

import pickle

with open('svm.pkl', 'wb') as file:

pickle.dump(svm, file)

from IPython.display import FileLink

# Create a download link for the model.pkl file

FileLink(r'svm.pkl')

**Coding for Deployment:**

import streamlit as st

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler

from sklearn.model\_selection import train\_test\_split

from sklearn.neighbors import KNeighborsClassifier

from sklearn.metrics import accuracy\_score

from sklearn.svm import SVC

from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier

from sklearn.tree import DecisionTreeClassifier

from sklearn.linear\_model import LogisticRegression

from sklearn.naive\_bayes import GaussianNB

from sklearn.cluster import KMeans

st.title('Machine Learning Based Electricity Theft Detection')

# Load the dataset

df = pd.read\_csv("C:/Users/DELL/Desktop/New folder/csp.csv")

# Remove unwanted columns

df.drop(['date', 'id'], axis=1, inplace=True)

# Split the data into features (x) and target (y)

x = df.drop(['flag'], axis=1)

y = df['flag']

# Standardize the features

scaler = StandardScaler()

x\_scaled = scaler.fit\_transform(x)

# Split the data into train and test sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(x\_scaled, y, test\_size=0.2, random\_state=0)

# Create a dictionary for model names and model instances

models = {

'KNN': KNeighborsClassifier(n\_neighbors=3),

'SVM': SVC(kernel='linear'),

'DTree': DecisionTreeClassifier(),

'RForest': RandomForestClassifier(n\_estimators=100),

'Logistic': LogisticRegression(),

'NaiveBias': GaussianNB(),

'KMeans': KMeans(n\_clusters=2),

'Gradientboster': GradientBoostingClassifier(n\_estimators=300, learning\_rate=0.05, random\_state=100, max\_features=5)

}

# Train and evaluate each model

model\_accuracies = {}

for model\_name, model in models.items():

model.fit(X\_train, y\_train)

y\_pred = model.predict(X\_test)

accuracy = accuracy\_score(y\_test, y\_pred)

model\_accuracies[model\_name] = accuracy

# User input section

st.header("Predict Faithfulness of a Customer")

st.subheader("Enter customer features:")

feature\_names = x.columns.tolist()

user\_input = {}

for feature in feature\_names:

user\_input[feature] = st.number\_input(f"Enter {feature}:", min\_value=0.0)

# Predict using Logistic Regression

lr = models['SVM']

user\_input\_features = [user\_input[feature] for feature in feature\_names]

user\_input\_scaled = scaler.transform([user\_input\_features])

prediction = lr.predict(user\_input\_scaled)

st.subheader("Prediction:")

if prediction == 0:

st.write("Faithful")

else:

st.write("Unfaithful")

model\_names = list(model\_accuracies.keys())

accuracies = list(model\_accuracies.values())

st.write("Model Accuracies:")

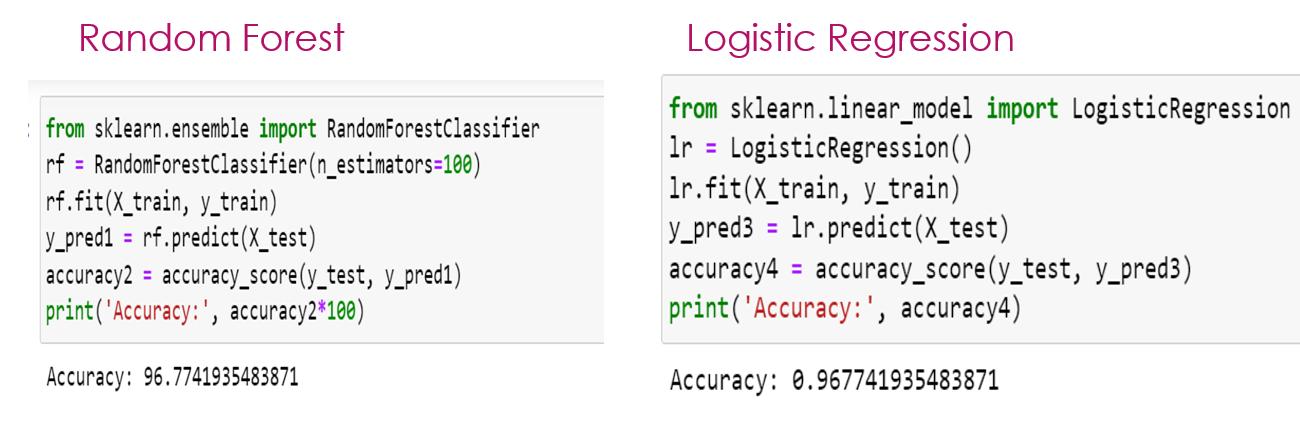
st.write(model\_accuracies)

# Create a bar plot of model accuracies

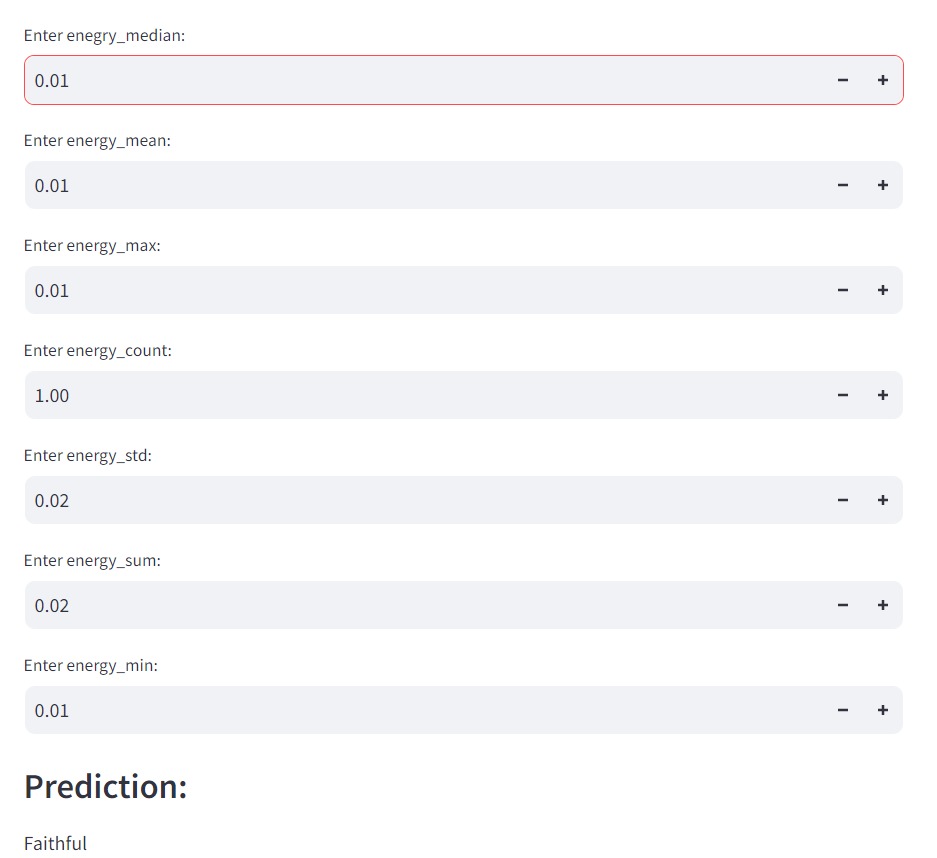
st.header("Classifier Model Comparision")

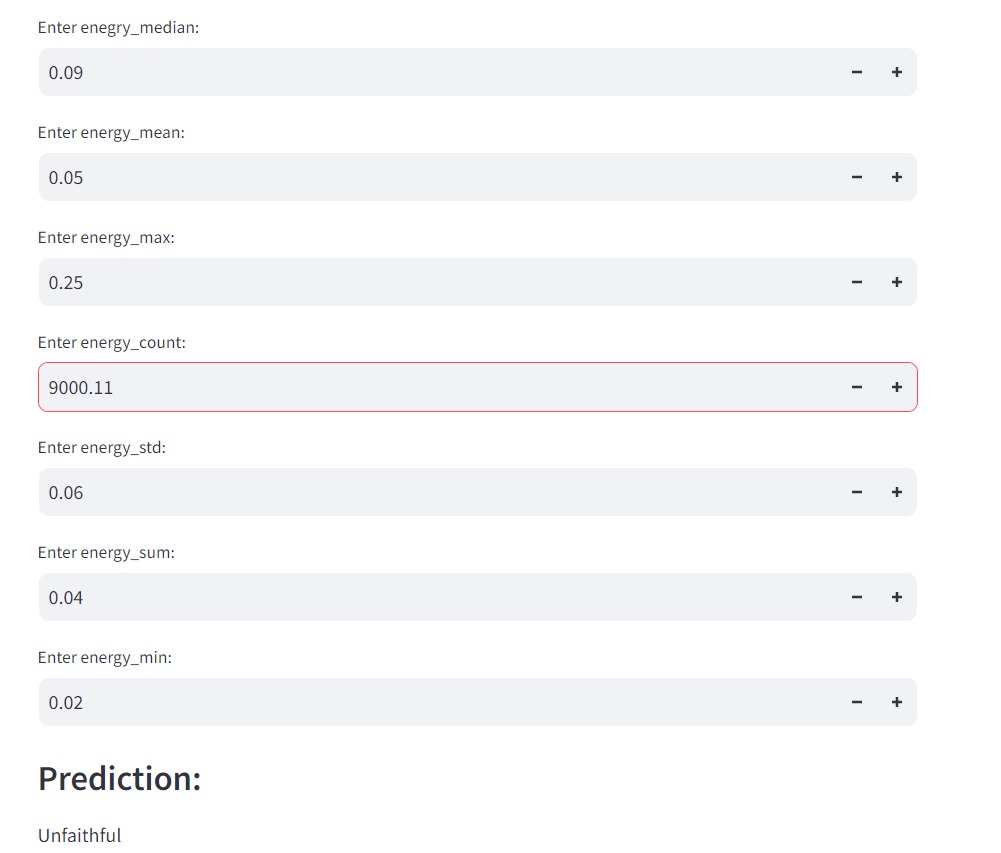
st.bar\_chart(accuracies)

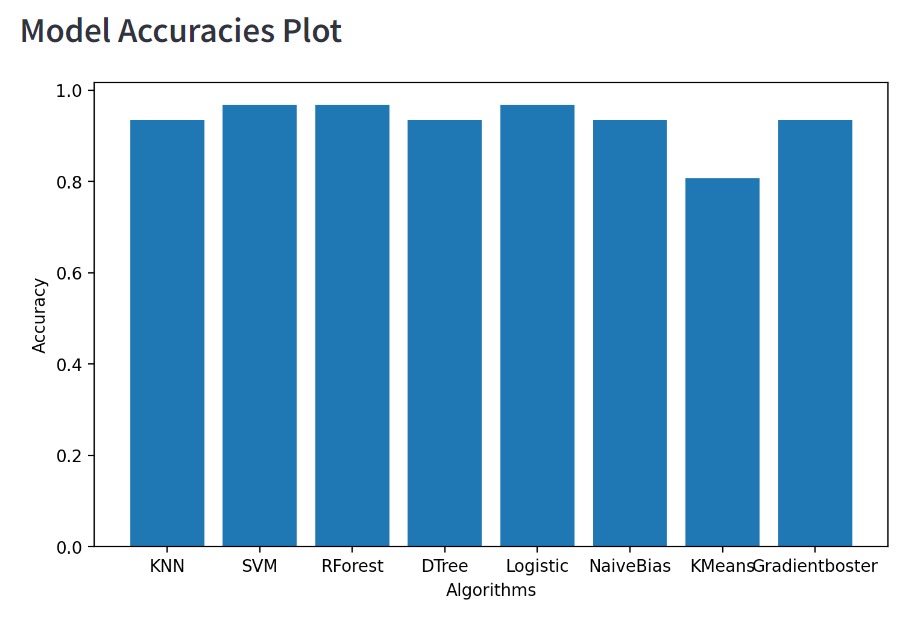
**Results**

****

**Output screenshot**

**1)** ****

2)****

3)****

**CONCLUSION AND FUTUTRE ENHANCEMENT**

**Conclusion:**

The Machine Learning-Based Residential Electricity Theft Detection project has successfully developed a system for identifying and preventing electricity theft in residential areas. Through the application of various machine learning models and real-time data analysis, the project has demonstrated its ability to enhance revenue collection for utility companies, promote safety, and ensure fair billing for customers. The project has a significant impact because it offers a preemptive solution to a persistent problem. However, deployment and stakeholder collaboration are the next steps for practical implementation in order to fully utilize this technology and its benefits.

**Future Enhancement**

1)Better Sources of Information:

Incorporate data from additional sources, such as smart meter data, meteorological data, and geographic information, to improve the accuracy of theft detection.

2)Deep Learning Models:

Analyze the use of deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), for more complex pattern recognition and prediction.

3)AI that is explicable:

Increase trust and transparency by using explainable AI techniques to provide clear, concise justifications for model predictions.

4)Continuous Learning:

Give the model the tools it needs to adapt to shifting consumer behaviors and theft trends over time.

5)Analysis Based on Maps:

Identify theft hotspots using geospatial analysis to prioritize maintenance and inspection efforts.

6)Cooperating with the Government:Form partnerships with law enforcement and regulatory agencies

**REFERENCES**

[1] Sumair Aziz Department of Electronics Engineering, University of Engineering and Technology, Taxila Taxila, Pakistan, sumair.aziz@uettaxila.edu.pk .

[2] Ivan Petrlik, Faculty of Industrial and Systems Engineering, National University Federico Villarreal, Lima, Perú .

[3] Soroush Omidvar Tehrani, Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran, omidvar@mail.um.ac.ir .

[4] Mohammad Hossein Yaghmaee Moghaddam Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran, [yaghmaee@ieee.org](mailto:yaghmaee@ieee.org).

[5] Mohsen Asadi Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran, mohsen.asadi@um.ac.ir .

[6] Rouzbeh Razavi Department of Management and Information Systems, Kent State University, USA, [rrazavi@kent.edu](mailto:rrazavi@kent.edu).

[7] Amin Gharipour School of Information and Communication Technology, Griffith University, Australia.

[8] Martin Fleury School of Computer Science and Electronic Engineering, University of Essex, UK .

[9] Pedro Lezama Faculty of Industrial and Systems Engineering, National University Federico Villarreal, Lima, Perú.

[10] Ciro Rodriguez Faculty of Electronic Engineering and Informatics, National University Federico Villarreal, Lima, Peru.

[11] Ricardo Inquilla Faculty of Engineering, National University of Cañete, Cañete, Peru.

[12] Julissa Elizabeth Reyna-González Faculty of Industrial and Systems Engineering, National University Hermilio Valdizán, Huánuco, Peru.