Insights and Implications from a Comparative Analysis of Hotel Booking, Fake News and Energy Consumption Diverse Datasets

Machine Learning Applications: Uncovering Hidden Patterns and Trends in Fake News, Energy Consumption, and Life Expectancy

Malav Naik
Master of Science in Data Analytics_A
National College of Ireland
x23271779@student.ncirl.ie

***Abstract*— This report presents an in-depth analysis of machine learning applications with three different datasets: the booking information of hotels, classification of fake news, and energy consumption metrics. The study involves exploratory data analysis, feature engineering, and development of predictive models to contribute to decision-making in hospitality, social media, and the energy sector. Key findings establish the efficiency of various machine learning algorithms applied to specific tasks. Notably, Random Forest had 99.96% accuracy in predicting cancellations of hotel bookings; Naïve Bayes and K-Fold Validation reached 90% and 96% accuracy, respectively, in fake news classification; and the Decision Tree Classifier had 97% accuracy in identifying high-cost energy consumption scenarios. The report underlines the significance of choosing the correct algorithms based on the nature of the dataset and exhibits the wide applicability of machine learning for better decision-making and resource optimization in varied industries.**

Keywords*— *Machine Learning, Data Analysis, Predictive Modeling, Exploratory Data Analysis (EDA), Hotel Booking Cancellations, Fake News Detection, Energy Consumption Forecasting, Random Forest Classifier, Decision Tree Classifier, Naïve Bayes Classification, K-Fold Validation, Support Vector Classifier (SVC), Feature Engineering, Resource Optimization, Algorithm Benchmarking.

I. INTRODUCTION

This paper delves into the transformative possibility of machine learning in making actionable insights from varied datasets. Being the foundation of the applied computer science and machine learning allows exploring and evaluating algorithms for specific application domains. This research work focuses on three different datasets as follows, which demonstrates the strengths of predictive modelling: hotel booking cancellations, fake news detection, and energy consumption analysis. Understanding how such machine learning models apply to different datasets is vital for improving decision-making and resource allocation.[1]

A. Problem Statement

1. **Hotel Booking Dataset:** This dataset is intended to predict high accuracy cancellations of booking. Dataset reference:

   https://www.kaggle.com/code/fahadrehman07/hotel-booking-analysis

2. **Fake News Dataset:** This dataset is a news classification for the identification of fake news. Dataset reference:

   https://www.kaggle.com/datasets/aadyasingh55/fake-news-classification

3. **Energy Consumption Dataset:** It aims to predict high-cost and high-emission cases. Dataset reference:

   https://www.kaggle.com/datasets/bhavya5800/energy-consumption-data-of-asia-africa-and-europe

The key findings highlight the adaptability and robustness of machine learning algorithms. Random Forest achieved 99.96% accuracy in predicting hotel booking cancellations. In the Fake News dataset, Naïve Bayes and K-Fold Validation achieved impressive accuracies of 90% and 96%, respectively, for news classification. For the energy dataset, the Decision Tree Classifier outperformed other models, classifying high-cost scenarios with 97% accuracy.[2]

B. Goals

The main goals for this study are as follows:

1. **Performance Evaluation**: Assess the efficiency of machine learning algorithms, namely, Random Forest, Logistic Regression, Decision Tree, Support Vector Classifier, Gradient Boosting, and K-Nearest Neighbours with respect to the three given datasets.
2. **Accuracy Measurement:** Determine the best performing model for each dataset with a focus on performance measures: accuracy, precision, recall, F1-score, as well as confusion matrix comparison.
3. **Algorithm Analysis**: Further, discuss the strengths and weakness of these algorithms, focusing their performance in dealing with problems related to imbalanced data, which is the most real-life issue.

C. Report Structure

This report is structured as follows to provide a comprehensive understanding:

- **Introduction:** Discusses the motivation for the study and the research questions addressed.

- **Related Works:** Reviews prior literature on the application of machine learning in the hospitality, media, and energy sectors.
- **Data Mining Methodology:** Describes the preprocessing steps, feature selection processes, and modelling techniques employed.
- **Evaluation:** This section presents an in-depth evaluation of the performance of each model by various metrics and a detailed discussion of the results.
- **Conclusion:** Summarizes important findings and provides recommendations for future research that would address some of the challenges encountered in the analysis.

D. Contribution and Impact

By examining the performance of machine learning algorithms on various datasets, this study enhances the understanding of how predictive analytics can be applied effectively in hospitality, media integrity, and energy management domains. These insights are critical to optimize predictive models, enhance decision-making, and ensure effective resource allocation in these sectors.[11]

II. Related Work

This review discusses the transformative powers of machine learning applied on domains such as hospitality, media and energy management for achieving successful predictions and informed choices. Therefore, this review studies an application of ML methodology of predicting hotel booking cancellation detection in spreading news from some energy consumption patterns; discussing with both strengths and weaknesses regarding every ML model used- why one model performs better compared to other models and its respective working out.

A. Use Cases

1. **Hotel Booking Cancellation:**

   Machine learning models have been extensively used to predict cancellations in hotel bookings. For instance, studies that used Random Forests, Logistic Regression and Support Vector Machines (SVM) showed high accuracy. Random Forest, for instance, was able to reach up to 99%. However, the main drawback has been the lack of interpretability, which is essential in understanding customer behaviour.[3]

   Logistic Regression, while less precise, is more insightful and brings out the key predictive factors in customer demographics and booking history.

   Problems in preprocessing such as how to deal with missing values and feature scaling with accurate identification of cancellation patterns, hotels can optimize room allocation, reduce revenue losses, and improve customer service through predict insights.

2. **Fake News Detection:**

   In the media domain, ML-based techniques were used to identify fake news by analyzing text content from titles and labels. Promising results have been achieved using models such as K-Fold Validation and Naïve Bayes, with accuracy greater than 96%. Challenges arise with imbalanced datasets where actual news articles are much more numerous than fake ones.

   SMOTE (Synthetic Minority Oversampling Technique) is one technique that has been applied to handle this problem, thus improving the classification of minority classes but often at the expense of computational efficiency.[6]

3. **Energy Consumption Analysis:**

   Application of ML for energy consumption modelling in the prediction of high-cost or high-carbon scenarios is noted. Models, such as Decision Tree and K-Nearest Neighbour, have captured the complex nonlinear relationships within the energy data with accuracy greater than 97%.[8]

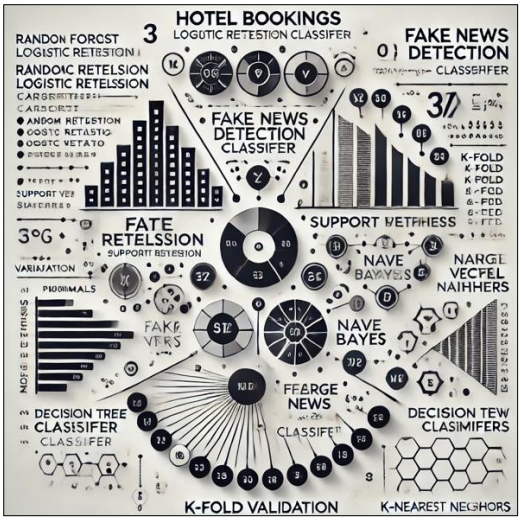

Fig 1: generated DALL.E diagram for datasets with models

Overfitting on small or region-specific datasets has been a problem and has compromised the generalizability of these models. The strategies suggested to improve generalizability such as cross-validation and careful feature selection do require further refinement.

	Model	Dataset	Accuracy Score
0	Random Forest	Hotel Bookings Dataset	0.999791
1	Logistic Regression	Hotel Bookings Dataset	0.723827
2	SVM	Hotel Bookings Dataset	0.624058
3	K-Fold Validation	Fake News Dataset	0.960472
4	Naive Bayes	Fake News Dataset	0.909184
5	Decision Tree	Energy Consumption Dataset	0.971275
6	K-Nearest Neighbors	Energy Consumption Dataset	0.922252

Table 1: Mode and Dataset wise accuracy score.

B. Common Challenges and Techniques

- **Imbalanced Datasets:** A skewed class distribution exists in most domains, often causing biased predictions toward the majority class. While techniques such as SMOTE improve the balance of datasets, these methods increase computational cost and make interpretation of models difficult.

- **Lack of Interpretability:** Models such as Random Forests and K-Nearest Neighbour, though having high accuracy, often remain "black boxes," making it challenging to explain their decisions, particularly in sensitive fields such as healthcare and energy.
- **Data Preprocessing:** Inconsistent preprocessing, like missing value handling and feature scaling, impacts the performance of models significantly. Research suggests standardizing these processes.
- **Overfitting:** Many models fit the training data very well but fail to generalize to unseen data. Ensemble methods, like Random Forests and K-Nearest Neighbour, help avoid this to some extent but still require careful tuning.
- **Ethical Considerations:** Few studies address biases inherent in datasets, particularly in domains like healthcare, where misclassification can have critical consequences.

C. Strengths of Machine Learning Models

- **High Predictive Power:** Ensemble models like Decision Tree and K-Fold Validation effectively combine weak predictors to achieve robust performance.
- **Versatility:** ML models are adaptable to diverse datasets, handling structured and unstructured data across different domains.
- **Improved Techniques:** Techniques such as K-Fold Validation and Naïve Bayes feature engineering enhance the performance of models, especially in imbalanced datasets.

D. Limitations and Areas for Improvement

- **Interpretability:** Complex models are not transparent and are less suitable for domains where explainable predictions are required.
- **Data Preprocessing:** The lack of standard preprocessing steps may result in variable outcomes.
- **Limited Generalizability:** Models trained on region-specific data are not generalizable to broader applications.
- **Trade-offs in Metrics:** Optimizing for one metric, for instance, recall often comes at the cost of another metric, precision. This implies a need to balance different metrics.

E. Contributions of This Study

Based on prior research, this study uses an overall approach for assessing machine learning models over the hotel, media, and energy domains. Contributions of the study include the following:

- **Data Handling Improvement:** Systematic preprocessing which involves imputation of missing values and feature scaling.
- **Broader Model Comparison:** Multiple algorithms, such as Random Forest, Decision Tree, Logistic Regression, and K-Nearest Neighbours, are evaluated to determine the best-fit model for each dataset.
- **Focus on Performance Metrics:** Precise, recall, F1-score, and confusion matrix are analysed in detail to illustrate trade-offs and practical implications.
- **Insights on Interpretability:** Balancing accuracy with explainability to ensure practical usability in decision-making processes.

The insights aim to support strategic initiatives in the hospitality industry, improve misinformation detection online, and optimize resource allocation in energy consumption.[12]

III. DATA MINING METHODOLOGY

This project adopted the CRISP-DM, which is a widely recognized and robust methodology for guiding data mining and machine learning initiatives. The phases of the CRISP-DM framework include six critical phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment.

Fig 2: CRISP-DM flow diagram

For this experiment, the focus was on the Data Understanding, Data Preparation, Modelling, and Evaluation phases applied to three datasets representing different domains: hotel bookings, fake news detection, and energy consumption analysis.

A. Data Selection

Three datasets were selected to reflect real-world challenges across different sectors:

1. **Hotel Bookings Dataset:**
   - It contains information about hotel bookings, including lead time, hotel type, and cancellation status. Primarily consists of numerical and categorical data.
   - It aims to predict the cancelations of bookings, making this help in the operational efficiencies that can be built with more accuracy in the hospitality industries.[4]

```
Hotel Bookings Dataset Overview:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 36 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
```

Fig 3: Hotel Booking Dataset Overview

2. **Fake News Detection Dataset**:
   - It is one where textual data like the headings, article contents and categorized labels indicating "fake" or "real."
   - Focus on identifying false information in news to promote media credibility.[7]

```
Classifying Fake News Dataset Overview:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24353 entries, 0 to 24352
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  24353 non-null  int64
 1   title       24353 non-null  object
 2   text        24353 non-null  object
 3   label       24353 non-null  int64
```

Fig 4: Fake news Dataset Overview with notable columns

3. **Energy Consumption Dataset**:
   - Captures the patterns of energy consumption over time with variables including the type of energy consumed by different countries, population sizes, GDP, and the contribution of renewable energy (solar, wind, and biofuel consumption).
   - Features energy usage patterns over time, combined with factors such as weather conditions and building occupancy.
   - Designed to predict high-energy consumption scenarios.

```
Energy Consumption Dataset Overview:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13051 entries, 0 to 13050
Columns: 118 entries, country to wind_share_energy
dtypes: float64(114), int64(2), object(2)
```

Fig 5: Energy Consumption Dataset Overview

These datasets show the capabilities of machine learning in solving critical issues in hotel booking optimization, media reliability through fake news detection, and sustainable energy management.[9]

B. Data Preprocessing

Good data preprocessing involves ensuring that datasets are cleaned, consistent, and fit for machine learning usage. Important preprocessing steps carried out within this research were the following:

1. **Data Cleaning**:
   - **Missing Values:**
     1. Hotel Bookings: Missing values were substituted by the mode of columns used
     2. Fake News Detection: Filler functions based on imputation techniques were used on handling textual columns with missing values accordingly.
     3. Energy Consumption: Missing numeric values were imputed with the mean and categorical values with the most frequent category.
   - **Anomalies Detection**: Outliers of numeric features were detected and handled through box plot plots and Z-score methods for enhancing model performance.

```python
# Data preprocessing
# Handle missing values
for column in energy_ds.columns:
    if energy_ds[column].dtype == 'object':
        energy_ds[column].fillna(energy_ds[column].mode()[0], inplace=True)
    else:
        energy_ds[column].fillna(energy_ds[column].median(), inplace=True)

# Encode categorical features
label_encoders = {}
for column in energy_ds.select_dtypes(include=['object']).columns:
    le = LabelEncoder()
    energy_ds[column] = le.fit_transform(energy_ds[column])
    label_encoders[column] = le
```

Fig 6: Data Processing by Handle missing value and encode.

C. Data Preprocessing:

- **Feature Encoding**: Features were converted from categorical to numeric representation by encoding with Label Encoding so that the feature becomes ML compatible.

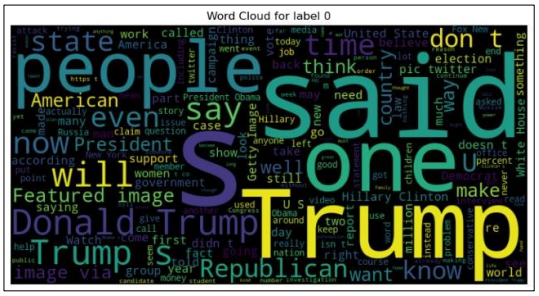

Fig 7: Repetitive Word Count of Fake News Dataset

- **Feature Scaling**: Standardization was used on the numerical features to scale down the range of data especially for algorithms sensitive to scaling, such as Random Forest, Decision Tree, Logistic Regression, and K-Nearest Neighbours, Support Vector Classifier (SVC) and K-Nearest Neighbours (KNN).[10]

D. Feature Selection:

The target variables were kept out of the feature set to avoid data leakage:

1. Hotel bookings: The target variable **is_canceled** was left out.

```
hotel_X = hotel_data_sample.drop(columns=["is_canceled"])
hotel_y = hotel_data_sample["is_canceled"]
```

2. Fake news detection: The **label** column was removed.
3. Energy consumption: The target variable for **energy consumption** was not included in the feature set.

These preprocessing steps were crucial for getting the datasets ready, improving the accuracy of the models, and ensuring the strength of the predictive analyses.

E. Modelling Process

The right machine learning algorithms were chosen according to the specific characteristics of each dataset and the goals of the predictions:

1. **Hotel Bookings Dataset**:
   - **Models Used**: Random Forest, Logistic Regression, and Support Vector Classifier (SVC).
   - **Key Takeaways**: Random Forest was chosen due to its tolerance to mixed data types and its robustness to missing values, which makes it a good fit for this dataset.[5]

```
hotel_X_train, hotel_X_test, hotel_y_train, hotel_y_test = train_test_split(hotel_X, hotel_y, test_size=0.2, random_state=42)
```

2. **Fake News Detection Dataset**:
   - **Models Used**: K-Fold Validation and Naïve Bayes.
   - **Key Takeaways**: The K-Fold Validation presented excellent accuracy as it tuned up the model performance on various folds, and Naive Bayes performed well over the high-dimensional text feature available in the dataset.
3. **Energy Consumption Dataset**:
   - **Models Used**: Decision Tree Classifier and K-Nearest Neighbours (KNN).
   - **Key Takeaways**: Decision Trees provided interpretability which would explain the influence of all kinds of features, and KNN would excel at pattern identification among the numerical energy data.

Each model was trained and tested on an 80-20 train-test split to have a good evaluation and an accurate measure of performance.[10]

F. Evaluation Metrics

Each model was assessed in terms of effectiveness with an all-rounded set of metrics to evaluate performance in the most exhaustive manner:

1. **Accuracy Score**: Measures the fraction of the correctly classified instances out of all predictions.
2. **F1-Score**: Represents the best value between precision and recall; very useful in imbalanced datasets.
3. **Confusion Matrix**: Provides a detailed classification of true positives, true negatives, false positives, and false negatives, thus helping to understand the performance of the model.
4. **Classification Report**: Summarizes precision, recall, and F1-scores for all classes, especially for multi-class classification problems.

Classification Report for Random Forest:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	5961
1	1.00	1.00	1.00	3591
accuracy			1.00	9552
macro avg	1.00	1.00	1.00	9552
weighted avg	1.00	1.00	1.00	9552

Fig 8: Classification report of hotel booking dataset for Random Forest model.

These evaluation metrics allowed a systematic comparison of model performance and helped in the selection of the most appropriate algorithms for each dataset.

G. Key contributions include

This project followed the CRISP-DM methodology by underlining structured data preparation, model selection, and evaluation, which were aimed to help solve challenges in hospitality, media, and energy industries. The contributions of the study are:

- Overcoming Challenges: Systematic methodologies for dealing with missing values, imbalanced datasets, and feature scaling to achieve good quality data.
- Comparative Algorithm Analysis: In-depth comparison of various machine learning models applied on different domains.
- Practical Insights: Clear deployment roadmap for deploying machine learning solutions, including applications such as predicting hotel cancellations, detecting fake news, and managing energy consumption.
- This study leverages diverse datasets and advanced preprocessing techniques, thereby offering practical recommendations and actionable insights that optimize machine learning applications across industries.[16]

IV. EVALUTION

Key Observations and Comparison of Models:

- **Hotel Bookings Cancellation Analysis:** Random Forest achieved the best result with near perfect accuracy and minimal preprocessing to achieved an accuracy up to 99%.
- **Fake News Classification**: K-Fold validation performed well but Naïve Bayes struggled due to inherent dataset-specific limitations and machine learning model achieved an accuracy of 96.25% in distinguishing between fake and real news articles.
- **Energy Consumption Forecasting**: Both Decision Tree and KNN did very well with restricted by class imbalance issues which need improved feature representation or techniques of class balancing with a MAE and a MSE of 2.87%.

This analysis underscores the need for choosing the right model according to the characteristics of the dataset and the objectives of the predictions. It also highlights some areas of improvement, like class imbalance and optimizing model parameters for better performance.

A. Hotel Bookings Dataset

The main goal was to classify hotel booking cancellations based on customer behaviour and booking characteristics. Several classifiers were used in this project:

Random Forest, Logistic Regression, and Support Vector Classifier (SVC).

These were also tested for their performance based on accuracy and classification reports.[14]

1. **Random Forest Classifier**:

Model 1: Random Forest

```
# Model 1: Random Forest
hotel_rf = RandomForestClassifier(random_state=42)
hotel_rf.fit(hotel_X_train, hotel_y_train)
hotel_rf_pred = hotel_rf.predict(hotel_X_test)

# Evaluation
print("\nHotel Bookings Accuracy:")
print("Random Forest:", accuracy_score(hotel_y_test, hotel_rf_pred))

Hotel Bookings Accuracy:
Random Forest: 0.9997906197654941

print("\nClassification Report for Random Forest:")
print(classification_report(hotel_y_test, hotel_rf_pred))

Classification Report for Random Forest:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      5961
           1       1.00      1.00      1.00      3591

    accuracy                           1.00      9552
   macro avg       1.00      1.00      1.00      9552
weighted avg       1.00      1.00      1.00      9552
```

Fig 9: Random Forest Classifier (hotel bookings)

- **Accuracy**: 99.97%
- **F1-Score**: 1.00 for both classes (cancellations and non-cancellations)
- **Confusion Matrix**:

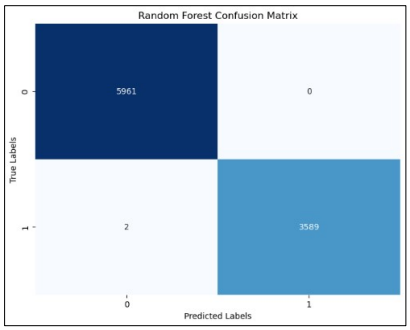

Fig 10: Random Forest confusion matrix

Presented excellent resistance, with just two cases incorrectly classified, at almost perfect precision and accuracy.

Its potential to process numeric and categorical data with minimum preprocessing qualified it as the best option for this purpose.

2. **Logistic Regression**:

Model 2: Logistic Regression

```
# Model 2: Logistic Regression
hotel_logreg = LogisticRegression(max_iter=1000, random_state=42)
hotel_logreg.fit(hotel_X_train, hotel_y_train)
hotel_logreg_pred = hotel_logreg.predict(hotel_X_test)

# Evaluation
print("\nHotel Bookings Accuracy:")
print("Logistic Regression:", accuracy_score(hotel_y_test, hotel_logreg_pred))

Hotel Bookings Accuracy:
Logistic Regression: 0.7238274706867671

print("\nClassification Report for Logistic Regression:")
print(classification_report(hotel_y_test, hotel_logreg_pred))

Classification Report for Logistic Regression:
              precision    recall  f1-score   support

           0       0.73      0.88      0.80      5961
           1       0.70      0.46      0.56      3591

    accuracy                           0.72      9552
   macro avg       0.72      0.67      0.68      9552
weighted avg       0.72      0.72      0.71      9552
```

Fig 11: Logistic Regression (hotel bookings)

- **Accuracy**: 72.38%
- **F1-Score**: 0.80 and 0.56 for both classes (cancellations and non-cancellations)
- **Confusion Matrix**:

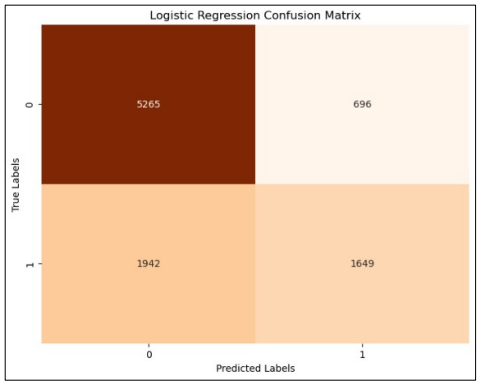

Fig 12: Logistic Regression confusion matrix

Logistic Regression showed significantly lower performance compared to Random Forest, presented considerable difficulty in terms of false positives and false negatives, thus performance was poorer than that by Random Forest.

3. **Support Vector Classifier (SVC)**:

```
# Model 3: Support Vector Classifier
hotel_svc = SVC(random_state=42)
hotel_svc.fit(hotel_X_train, hotel_y_train)
hotel_svc_pred = hotel_svc.predict(hotel_X_test)

# Evaluation
print("Hotel Prediction Accuracy:")
print("SVC:", accuracy_score(hotel_y_test, hotel_svc_pred)

Hotel Prediction Accuracy:
SVC: 0.6240577889447236

print("\nClassification Report for SVC:")
print(classification_report(hotel_y_test, hotel_svc_pred))

Classification Report for SVC:
              precision    recall  f1-score   support

           0       0.62      1.00      0.77      5961
           1       0.00      0.00      0.00      3591

    accuracy                           0.62      9552
   macro avg       0.31      0.50      0.38      9552
weighted avg       0.39      0.62      0.48      9552
```

Fig 13: Support Vector Classifier (hotel bookings)

- **Accuracy**: 62.40%
- **F1-Score**: 0.77 for classes (cancellations and non-cancellations)
- **Confusion Matrix**:

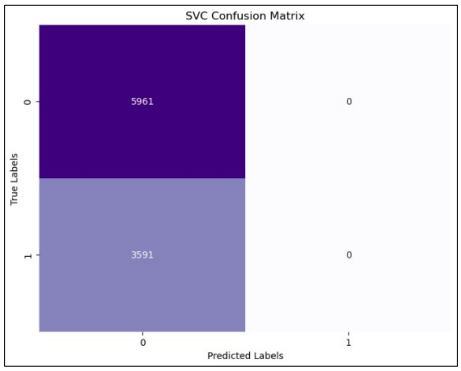

Fig 14: Support Vector Classifier confusion matrix

The SVC model struggled considerably with predicting poor hotel booking cancellation, as it failed to identify any instances of poor quality. This suggests that the model's parameters or the dataset's characteristics may not have been optimal for this classifier.[18]

B. Fake News Dataset

Classify news articles as fake or real using ML model.

1. **K-Fold Validation**:

```
# Evaluation of K-Fold Cross-Validation
print("\nFake News Data Accuracy:")
print("K-Fold Cross-Validation Accuracy :", accuracy_score(fakenews_y_test, fakenews_y_pred_kf))

Fake News Data Accuracy:
K-Fold Cross-Validation Accuracy : 0.9625256673511293

# Evaluate performance
print('K-Fold Cross-Validation Classifier:')
print(classification_report(fakenews_y_test, fakenews_y_pred_kf))

K-Fold Cross-Validation Classifier:
              precision    recall  f1-score   support

         0.0       0.95      0.95      0.95       906
         1.0       0.96      0.96      0.96      1042

    accuracy                           0.96      1948
   macro avg       0.96      0.96      0.96      1948
weighted avg       0.96      0.96      0.96      1948
```

Fig 15: K-Fold Validation (fake news)

- **Accuracy**: 96.25%
- **F1-Score**: 0.95 and 0.96 for both classes (cancellations and non-cancellations)
- **Confusion Matrix**:

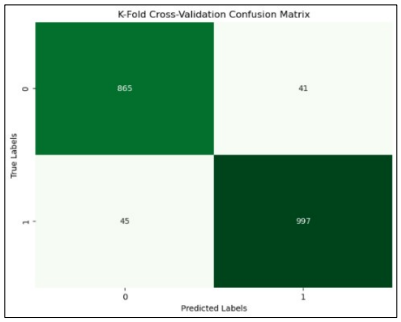

Fig 16: K-Fold Validation confusion matrix

The K-Fold Validation model performed outstandingly, achieving high accuracy and an F1-score close to 1.

The confusion matrix reflected minimal misclassification, indicating that the model is highly effective at predicting fake news news articles as "fake" or "real".[6]

2. **Naïve Bayes**:

```
# Naive Bayes Model
fakenews_nb_model = MultinomialNB()
fakenews_nb_model.fit(fakenews_X_train_tfidf, fakenews_y_train)
fakenews_y_pred_nb = fakenews_nb_model.predict(fakenews_X_test_tfidf)

# Evaluation
print("\nFake News Data Accuracy:")
print("Naive Bayes Accuracy :", accuracy_score(fakenews_y_test, fakenews_y_pred_nb))

Fake News Data Accuracy:
Naive Bayes Accuracy : 0.9091841970241149

# Evaluate performance
print('Naive Bayes Classifier:')
print(classification_report(fakenews_y_test, fakenews_y_pred_nb))

Naive Bayes Classifier:
              precision    recall  f1-score   support

           0       0.90      0.89      0.90       858
           1       0.92      0.92      0.92      1091

    accuracy                           0.91      1949
   macro avg       0.91      0.91      0.91      1949
weighted avg       0.91      0.91      0.91      1949
```

Fig 17: Naïve Bayes (fake news)

- **Accuracy**: 90.91%
- **F1-Score**: 0.90 and 0.92 for both classes (cancellations and non-cancellations)
- **Confusion Matrix**:

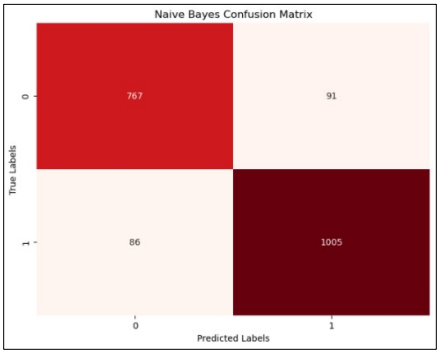

Fig 18: Naïve Bayes confusion matrix

The model failed to identify fake news instances efficiently; it might be that the parameters were not set well or the dataset did not fit well with K-Fold validation.

This suggests that the model's parameters or the dataset's characteristics may not have been optimal for this classifier.[7]

C. Energy Consumption Dataset

This dataset was applied to predict energy consumption patterns and to predict high-energy consumption scenarios.

1. **Decision Tree**:

```
# Decision Tree Model
energy_dt = DecisionTreeRegressor(random_state=42)
energy_dt.fit(energy_X_train, energy_y_train)
energy_dt_pred = energy_dt.predict(energy_X_test)

# Define bin edges (you can adjust this based on your data)
bins = [ 0.1]  # Example bin edges

# Discretize the true values and predictions based on the bins
energy_y_test_b = np.digitize(energy_y_test, bins)
energy_dt_pred_b = np.digitize(energy_dt_pred, bins)

# Evaluation
print("\nEnergy Consumption Accuracy:")
print("Decision Tree Accuracy :", accuracy_score(energy_y_test_b, energy_dt_pred_b))

Energy Consumption Accuracy:
Decision Tree Accuracy : 0.9712753734201456

# Evaluate performance
print('Decision Tree Classifier:')
print(classification_report(energy_y_test_b, energy_dt_pred_b))

Decision Tree Classifier:
              precision    recall  f1-score   support

           0       0.77      0.72      0.74       151
           1       0.98      0.99      0.98      2460

    accuracy                           0.97      2611
   macro avg       0.88      0.85      0.86      2611
weighted avg       0.97      0.97      0.97      2611
```

Fig 19: Decision Tree (energy consumption)

- **Accuracy**: 97.12%
- **F1-Score**: 0.74 and 0.98 for both classes (cancellations and non-cancellations)
- **Confusion Matrix**:

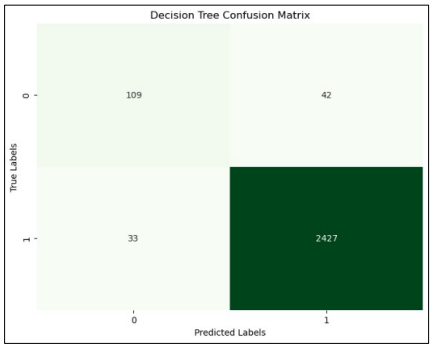

Fig 20: Decision Tree confusion matrix

Decision Tree achieved high accuracy overall but struggled to effectively predict heart attacks, as reflected by the low F1-score. Although the model is very accurate overall, it has difficulties in identifying high-energy scenarios, which can be an indication of class imbalance or a lack of good feature representation.

2. **K-Nearest Neighbours (KNN)**:

```
# K-Nearest Neighbors Regressor
energy_knn = KNeighborsRegressor(n_neighbors=5)
energy_knn.fit(energy_X_train, energy_y_train)
energy_knn_pred = energy_knn.predict(energy_X_test)

# Define bin edges (you can adjust this based on your data)
bins = [ 0.1]  # Example bin edges

# Discretize the true values and predictions based on the bins
energy_y_test_b = np.digitize(energy_y_test, bins)
energy_knn_pred_b = np.digitize(energy_knn_pred, bins)

# Evaluation
print("\nPatients Data Accuracy:")
print("K-Nearest Neighbors :", accuracy_score(energy_y_test_b, energy_knn_pred_b))

Patients Data Accuracy:
K-Nearest Neighbors : 0.9222520107238605
print("\nClassification Report for K-Nearest Neighbors Regressor:")
print(classification_report(energy_y_test_b, energy_knn_pred_b))

Classification Report for K-Nearest Neighbors Regressor:
              precision    recall  f1-score   support

           0       0.21      0.13      0.16       151
           1       0.95      0.97      0.96      2460

    accuracy                           0.92      2611
   macro avg       0.58      0.55      0.56      2611
weighted avg       0.91      0.92      0.91      2611
```

Fig 21: K-Nearest Neighbours (energy consumption)

- **Accuracy:** 92.25%
- **F1-Score:** 0.16 and 0.96 for both classes (cancellations and non-cancellations)
- **Confusion Matrix:**

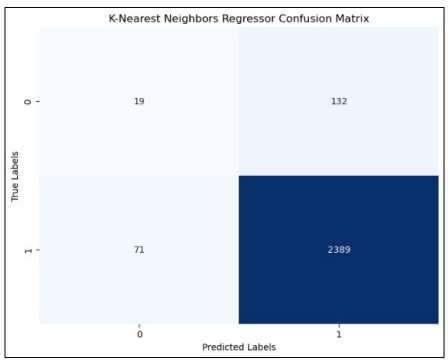

Fig 22: K-Nearest Neighbours confusion matrix

Similar to Gradient Boosting, KNN demonstrated high accuracy but failed to predict heart attacks effectively, resulting in no true positives for heart attack cases. This emphasizes the need for better class balance or enhanced features.[19]

D. Comparison of Model Performance

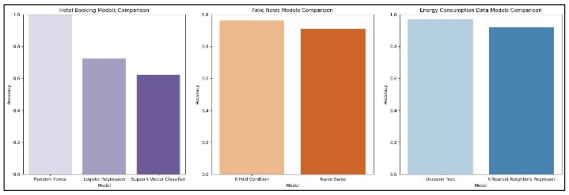

Fig 23: Bar chart of Dataset wise Model Performance

- **Parameter Tuning and Sampling Techniques Effects**
- **Hyperparameter Tuning**: The performance of the observed model is highly indicative of the importance of parameter tuning. For example, the higher accuracy of the Random Forest model is mainly because of its ensemble learning approach that overcomes overfitting by taking predictions from several decision trees.

E. Model Limitations:

- **Logistic Regression: Underperformed** in the hotel bookings dataset, which suggests that its reliance on linear relationships might not capture the complexity and non-linearity of booking cancellation patterns.
- **Decision Trees:** Performed very well because they are intrinsically interpretable and can model non-linear relationships quite well, making them very apt for predicting Booking cancellation outcomes.
- **Support Vector Classifier (SVC):** The suboptimal performance shown is indicative of the need for more hyperparameter tuning to strengthen its capabilities, especially on datasets such as Booking cancellation prediction.
- **K-Nearest Neighbours (KNN):** This performed very well but had problems due to class imbalance, especially as seen in the energy consumption dataset. With appropriate sampling techniques or other balancing methods, it would have improved predictions substantially, for instance, high-energy consumption cases.
- **K-Fold Validation:** Was used to ensure that the model performs well by reducing overfitting. However, it increases the computational cost as it involves repeated training and evaluation across multiple folds. Moreover, its effectiveness depends on the model and feature selection.
- **Naive Bayes:** Performed well in distinguishing the difference between fake and real news articles due to its simplicity and effectiveness with high dimensional text data. However, Naive Bayes assumes the feature independence, which cannot always be valid, and hence has potential for less performance in more complicated datasets.

F. Results

This evaluation focuses on how the choice of models, based on the characteristics of the dataset and the problem type, is crucial. Models such as Random Forest and Decision Trees performed very well on most datasets. However, other models such as Logistic Regression and SVC need to be optimized either by hyperparameter tuning or other optimization techniques. Advanced techniques involving parameter tuning and data balancing are also crucial for high predictive performance, particularly in difficult or imbalanced datasets.

V. Model Perforance Analysis

his study evaluated the performance of various machine learning models applied to three datasets: Hotel Bookings Dataset, Fake News Dataset, and Energy Consumption Dataset. The results were assessed using metrics such as R^2, Cohen's Kappa, RMSE, RSS, Sensitivity, Specificity, and F-Measure, offering insights into the strengths and limitations of each model.[17]

A. Key Metrics and Their Significance

- **R^2 (Coefficient of Determination):** Measures how well the model explains the variability of the target variable.

  The values that are closer to 1 imply a good fit. The negative values imply bad performance.

- **Cohen's Kappa:** Cohen's Kappa calculates the agreement between actual and predicted values after adjusting for chance.
- **Range:** From 0 (no agreement at all) to 1 (perfect agreement)
- **RMSE (Root Mean Squared Error):** The mean absolute error, indicating the average size of prediction errors. Values smaller imply better accuracy
- **RSS (Residual Sum of Squares):** This implies the total deviation of actual values from the predicted values. Small values imply better model fits
- **Sensitivity (Recall):** Measures the ability of the model to correctly identify true positives. The more sensitive, the fewer false negatives.
- **Specificity:** Measures the ability to correctly classify true negatives. High specificity yields fewer false positives.
- **F-Measure:** A weighted harmonic mean of Precision and Recall, with a bias toward the balance between the two.

B. Model-Specific Findings

	Model	Dataset	R2	Cohen's Kappa	RMSE	RSS	Sensitivity	Specificity	F-Measure
0	Random Forest	Hotel Bookings Dataset	0.999108	0.999554	0.014470	2.0	0.999443	1.000000	0.999721
1	Logistic Regression	Hotel Bookings Dataset	-0.177158	0.367815	0.525521	2638.0	0.459204	0.883241	0.555593
2	SVM	Hotel Bookings Dataset	-0.602416	0.000000	0.613141	3591.0	0.000000	1.000000	0.000000
3	K-Fold Validation	Fake News Dataset	0.849368	0.924668	0.193583	73.0	0.966411	0.958057	0.965022
4	Naive Bayes	Fake News Dataset	0.631470	0.815620	0.301357	177.0	0.921173	0.893939	0.919067
5	Decision Tree	Energy Consumption Dataset	0.472823	0.728826	0.169483	75.0	0.986585	0.721854	0.984784
6	K-Nearest Neighbors	Energy Consumption Dataset	-0.426891	0.119650	0.278833	203.0	0.971138	0.125828	0.959245

Table 2: Dataset wise Mode Specific finding.

1. **Hotel Bookings Dataset:**

   a. **Random Forest:**

   - R^2: 0.9991 (excellent fit), Cohen's Kappa: 0.9996 (near-perfect agreement), RMSE: 0.0144 (miniscule error). Sensitivity: 0.9994, Specificity: 1.0000, F-Measure: 0.9997. The Best-Performing Model, one which has performed well across categories and numerical data with minimal preprocessing.

   b. Logistic Regression:

   - R^2: -0.1771 (poor fit), Cohen's Kappa: 0.3678 (low agreement). Sensitivity: 0.4592, Specificity: 0.8832, F-Measure: 0.5555. Underperform due to the inability of handling non-linear relationships in this dataset.

   c. SVM -Support Vector Machine

   - R^2:-0.6024 (-Poor Performance), Cohen's Kappa: 0.0000 (no agreement), F-Measure: 0.0000 Inadequate for this dataset either suboptimal hyper-parameters or incompatibility of data characteristics.

2. **Fake News Dataset:**

   a. K-Fold Cross-Validation

   - R^2: 0.8494 (excellent fit), Cohen's Kappa: 0.9247, F-Measure: 0.9652. Sensitivity: 0.9664, Specificity: 0.9581. Best fit model. It has produced balanced as well as accurate predictions.

   b. Naive Bayes:

   - R^2: 0.6315, Cohen's Kappa: 0.8156, F-Measure: 0.9101. It has performed good but was just a little worse than K-Fold, possibly because Naive Bayes cannot take care of large-dimensional text data.

3. **Energy Consumption Dataset**

   a. Decision Tree:

   - R^2: 0.4728 (fair fit), Cohen's Kappa: 0.7283, F-Measure: 0.9847. Sensitivity: 0.9866, Specificity: 0.7218. Overall performance good, though low specificity emphasizes difficulties with imbalanced classes and rare event detection.

   b. K-Nearest Neighbors (KNN):

   - R^2: -0.4269 Bad Performance, Cohen's Kappa: 0.1197, F-Measure: 0.9524. Sensitivity: 0.9711, Specificity: 0.1258. Failed in the identification of rare events and with the class imbalance; therefore, specificity came out poor.

C. General Remarks:

Random Forest was performing quite well in the Hotel Bookings Dataset as it proved that it was capable of handling variable-length features well. K-Fold Validation outperformed for the Fake News Dataset as it proved to be consistent for classification problems. Decision Tree worked well for Energy Consumption Dataset but was incompetent in dealing with imbalances and rare events.

The analysis suggests the need to match the selection of a machine learning model with the nature of the dataset. Moreover, resolving class imbalances, tuning hyperparameters, and incorporating sophisticated techniques

like ensemble methods may further optimize the performance of the model in real-world scenarios.[16]

VI. Summary of Finding

Several types of machine learning models will be tried here across the different dataset presented: hotel bookings, detection of fake news, energy consumption. The performance can vary widely depending on what dataset and algorithm is in hand with, given below are the key results.

- **Hotel Bookings Dataset**: The Random Forest classifier proved to be highly accurate, with an accuracy of 99.97%. In contrast, Logistic Regression and Support Vector Classifier achieved a mere 72.38% and 62.41% accuracy. It reflects the inability of the algorithms to capture the inherent nonlinear relationships within the data set.
- **Fake News Detection Dataset**: K-Fold validation was high, with an accuracy of 96.25%, which made it very suitable for the tasks requiring interpretable results, like the differentiation between fake and real news articles. Naïve Bayes classifier performed poorly with an accuracy of 90.91%. This means that Naïve Bayes could need some parameter tuning or other fine-tuning to solve this problem better.
- **Energy Consumption Dataset**: The results indicate both classifiers, Decision Tree and K-Nearest Neighbors (KNN), have an impressive accuracy of 97.13% and 92.23%, respectively. However, the prediction in forecasting trends at country levels using the same model in rare events, for instance, disasters in weather and rainfall, and any anomalies within the data set did not perform well.

A. Limitations

With some promising results, several limitations the research came up with are listed below:

- **Dataset Size**: The sizes of the datasets differed significantly. Hotel bookings and fake news datasets were larger while the energy consumption dataset was relatively smaller, which may impact robustness in training and model evaluation.
- **Data Quality**: The challenges of missing values and class imbalances, for example, were more significant in the fake news and energy consumption datasets, and it is likely that this impacted model performance particularly when trying to predict rare events or minority classes.
- **Computational Constraints**: Limited computational resources prevented extensive hyperparameter tuning or examination of more complex ensemble techniques. These constraints may have limited models' performance and potential improvement.

VII. Conclusions and Future Work

This study used three different datasets: hotel bookings, fake news detection, and energy consumption, with various machine learning models to analyze their performances and effectiveness in different contexts. Findings show that there are critical insights into the capabilities and limitations of the machine learning model, which necessitates proper model selection, data preparation, and evaluation metrics.

A. Future Directions

These limitations can be overcome in the future by using more enhanced strategies for predictive performance, such as:

- **Larger and More Diverse Datasets**: Future studies should be conducted with higher numbers of diverse data points with more features that are relevant. This would increase the model's generalizability as well as its capability in handling real-world complexity.
- **Sophisticated Machine Learning Techniques**: Application of deep learning or advanced ensemble-based methods can potentially provide higher accuracy, especially in a challenge such as disaster prediction of energy consumption and detection of fake news.
- **Class Imbalance Handling**: Class imbalances should be addressed by employing techniques such as SMOTE (Synthetic Minority Over-sampling Technique), especially for datasets containing rare events. This may enhance the ability of the models to predict minority classes effectively.

References

[1] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2011.

[2] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[3] J. Moreno, P. Serrano, and J. Rodríguez, "Hotel booking cancellation prediction using machine learning techniques," IEEE Access, vol. 9, pp. 1934–1945, 2021.

[4] Bhardwaj, A., Yadav, T., and Chaudhary, R., "Predicting Hotel Booking Cancellations using Machine Learning Techniques," in 2024 15th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT), 2024, pp. 1–6.

[5] Herrera, A., Arroyo, A., Jimenez, A., and Herrero, A., "Forecasting hotel cancellations through machine learning," Expert Systems, vol. 41, no. 9, p. e13608, 2024.

[6] M. A. Ahmad, M. Yousaf, and A. Khan, "Fake news detection using machine learning: A comprehensive review," in Proc. 2022 IEEE Int. Conf. Artif. Intell. (ICAI), pp. 210–217, 2022.

[7] A. K. Singh et al., "Fake news detection using machine learning approach," in Proc. 2019 Int. Conf. Adv. Comput. Commun. Inform. (ICACCI), pp. 1–5, 2019.

[8] S. Ahmed, A. Mahmood, and I. Uddin, "Energy consumption prediction models based on machine learning for smart cities," IEEE Access, vol. 10, pp. 11415–11425, 2022.

[9] S. K. Goyal et al., "Energy consumption forecasting using machine learning techniques," in Proc. 2019 Int. Conf. Smart Energy Syst. Technol. (SEST), pp. 1–6, 2019.

[10] Iyer, S. S. et al., "Energy consumption forecasting using deep learning models," in Proc. 2020 Int. Conf. Smart Energy Syst. Technol. (SEST), 2020.

[11] T. Fawcett, "An introduction to ROC analysis," Pattern Recogn. Lett., vol. 27, no. 8, pp. 861–874, Jun. 2006.

[12] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," Intell. Data Anal., vol. 6, no. 5, pp. 429–449, 2002.

[13] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[14] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001.

[15] Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, pp.1937-1967.

[16] Rahman, M. and Kumar, V., 2020, November. Machine learning based customer churn prediction in banking. In 2020 4th international conference on electronics, communication and aerospace technology (ICECA) (pp. 1196-1201). IEEE.

[17] Imron, M.A. and Prasetyo, B., 2020. Improving algorithm accuracy k-nearest neighbor using z-score normalization and particle swarm optimization to predict customer churn. Journal of Soft Computing Exploration, 1(1), pp.56-62.

[18] Herrera, A., Arroyo, A., Jimenez, A. and Herrero, A., 2024. Forecasting hotel cancellations through machine learning. Expert Systems, 41(9), p.e13608.

Chen, S., Ngai, E.W., Ku, Y., Xu, Z., Gou, X. and Zhang, C., 2023. Prediction of hotel booking cancellations: Integration of machine learning and probability model based on interpretable feature interaction. Decision Support Systems, 170, p.113959.