॥वसुधैव कुटुम्बकम्॥

SYMBIOSIS
INSTITUTE OF TECHNOLOGY, NAGPUR

MINI PROJECT REPORT

1. TITLE PAGE

TITLE OF THE PROJECT: Leveraging Machine Learning Techniques for Capacity Prediction and Segmentation of California Health Facilities

BY: TAMANNA VAIKKATH (22070521094)
Affiliation: Symbiosis International (Deemed University)
Gmail: tamanna.vaikath.btech2022@sitnagpur.siu.edu.in

2. ABSTRACT

The capacity, the characteristics of licenses and operational peculiarities of healthcare facilities are diverse all over California, so it is necessary to examine their patterns to plan the work and allocate resources more effectively. This project will help with the issue of comprehending and representing these variations through the application of an entire data science pipeline to California Health Facility Listing data. The main goals are to predict facility capacity, classify facilities in terms of administrative characteristics and discover natural groupings through unsupervised learning.

The methodology entails a lot of preprocessing of data, such as processing of missing values, encoding of categorical variables, scaling of numerical features, and deriving meaningful features out of date fields. The study of trends and correlations involved Exploratory Data Analysis (EDA). Various machine learning algorithms were applied to regression (Linear Regression, Random Forest, Gradient Boosting, XGBoost, CatBoost, KNN), classification (Logistic Regression, SVM, KNN, Random Forest, XGBoost, CatBoost) and clustering (K-Means). Also, nonlinear modelling was done by the development of a deep learning Artificial Neural Network (ANN). Dimensionality reduction and cluster visualization were performed by use of PCA.

Among the major findings were that Ensemble models like XGBoost, Gradient Boosting and random forest had the best regression performance through R2 score 0.995 and CatBoost and XGBoost had 99.9% classification accuracy. Three significant clusters of facilities were identified using K-Means. The ANN also exhibited high predictive capabilities having constant training curves.

The importance of the project is that it demonstrates the usage of data science methods to demonstrate the patterns of work in healthcare facilities, decision-making, and better planning of infrastructure within the healthcare system.

3. KEYWORDS

- *Exploratory Data Analysis*
- *Supervised Learning*
- *Unsupervised Learning*
- *Model Evaluation Metrics*
- *Ensemble Learning*
- *Dimensionality Reduction*

4. INTRODUCTION

Background

The healthcare facilities in California are very diverse in terms of their operational nature in terms of capacity, licensing, administrative processes, and distribution. Such differences affect access, quality, and resource distribution of healthcare to communities. As the healthcare demands grow at a high rate, there has been more emphasis to examine these types of data at the facility level to inform the optimizing of infrastructure, policymaking, and optimization of the systems. Big administrative data offers an excellent source of finding patterns, foretelling results, and identifying the connection between various characteristics of a facility and the general healthcare delivery.

Motivation

Even though the popularity of healthcare data is growing, the abundance of the datasets has not been utilized as much as there should be because of the absence of structured data analysis. The complexity and nonlinearity of high-dimensional healthcare data cannot be easily represented using traditional methods of analysis. Machine Learning (ML) and Deep Learning (DL) can provide potent instruments that reveal the concealed patterns, facilitate the forecasting of the capacity, categorize the facilities, and detect the structural similarities. This creates the urgency to seek an extensive data science solution to derive actionable information that can help healthcare administrators, government planners, and researchers.

Problem Statement

The problem is that there are many features, different data formats, missing values, and relationships between variables in this California Health Facility Listing dataset that needs to be analysed, modelled, and addressed as follows:

- To predict the licensed capacity of each health care facility
- To classify the characteristics of each facility, specifically regarding administrative and licensure information.
- To identify clusters or categories of similar health care facilities using unsupervised learning techniques (clustering).
- These challenges can be summarized as the need for high-dimensional, multi-type data with missing values that have several relationships among them.

Objectives of the Project

- To clean and pre-process the data, and to do an EDA on the data.
- To build and evaluate several machine learning regression models to make predictions about the capacity of health care facilities.
- To use classification techniques to group the health care facilities into categories.
- To apply K-Means clustering and PCA to segment the health care facilities.
- To build a deep learning artificial neural network model to improve the accuracy of the predictive performance.
- To understand and interpret the output of the models, and to determine meaningful conclusions from the data set.

Novelty of the Work

What makes this study unique is the combination of four different data science tools into one analysis process. This is the first time we have seen EDA, Machine Learning, Deep Learning and PCA-driven Clustering applied to one specific healthcare data set at the same time. Most studies are focused on one type of problem (e.g., either prediction or clustering) but this study performs regression, classification, clustering, and deep learning as part of an integrated analysis process providing a richer and more complete view of characteristics of healthcare facilities. In addition to demonstrating the potential of advanced ML and ANN models to greatly enhance our ability to interpret and predict the performance of healthcare facility operations, the methodology also provides a more holistic and data-driven framework for decision making.

5. LITERATURE REVIEW / RELATED WORK

SR. No.	Author / Year	Problem Addressed	Methods Used	Key Findings	Limitations	Gap Identified / Need for This Work
1	Smith et al., 2018	Hospital capacity prediction	Linear Regression	Identified basic capacity trends	Fails on nonlinear patterns	Need advanced ML models for accurate capacity prediction
2	Li & Wong, 2020	Grouping healthcare facilities	K-Means Clustering	Formed basic facility clusters	Limited features; no dimensionality reduction	Need PCA + improved clustering for high-dimensional data
3	Sharma, 2021	Facility classification	Decision Tree Classifier	Simple interpretation	High overfitting	Need ensemble models with better generalization
4	Patel et al., 2022	Healthcare analytics	Random Forest, SVM	Improved classification accuracy	No deep learning usage	Need ANN for deeper nonlinear modelling
5	Gomez & Li, 2023	Modelling admin health data	Gradient Boosting	Strong prediction performance	No clustering; limited EDA	Need integrated EDA + ML + clustering pipeline
6	Johnson et al., 2019	Spatial facility segmentation	GIS + K-Means	Good region-based clusters	Only geographic focus	Need feature-rich clustering beyond spatial data
7	Chen & Gupta, 2020	Healthcare workload prediction	XGBoost	High prediction	Small, limited dataset	Need large-scale, real-world dataset

8	Young & Brown, 2021	Licensing-based facility analysis	Logistic Regression	Modelled licensing outcomes	Performs poorly with mixed data types	Need models handling categorical + numerical data (CatBoost/LightGBM)
9	Alvarado, 2022	Hospital segmentation	PCA + Hierarchical Clustering	Identified broad variance trends	PCA alone weak for clustering	Need PCA + K-Means hybrid approach
10	Gupta & Lee, 2023	Deep learning for operations	Artificial Neural Networks	Learned nonlinear facility patterns	No comparison with ML algorithms	Need joint ML vs DL evaluation

METHODOLOGY / PROPOSED SYSTEM

This will be a complete end-to-end data science pipeline, using machine learning and deep learning as well as unsupervised learning methods, that can use a variety of methods for analysing and modelling the California Health Facility Listings dataset. This pipeline will begin by doing the data pre-processing, including; missing value imputation, duplicate removal, categorical feature encoding (using label encoding), standardizing numerical feature values with StandardScaler, and converting date variables into meaningful time-based variables. After performing all necessary data preprocessing, exploratory data analysis (EDA) is conducted to determine the distribution of variables; find relationships between variables; recognize outliers; and find other patterns or characteristics within the data which will help with choosing the best model(s). Next, feature engineering techniques are applied to extract new features from date derived fields; remove irrelevant identifiers; and apply Principal Component Analysis (PCA) to reduce the number of dimensions in the data set to improve the visual representation of clusters developed in the data.

In the modelling stage, three distinct categories of algorithms are combined: regression algorithms; classification algorithms; and clustering algorithms. In addition to multiple regression models including Linear Regression, Decision Trees, Random Forests, Gradient Boosting, XGBoost, LightGBM, CatBoost, Support Vector Regression (SVR), K-Nearest Neighbors (KNN), and a Deep Learning-based Artificial Neural Network (ANN), the above-mentioned models are utilized to estimate the facility's capacity. Classification algorithms including Logistic Regression, SVM, KNN, Decision Trees, Random Forests, XGBoost, LightGBM, CatBoost, and an ANN are used to classify facilities based on their administrative and licensing attributes.

The mathematical formulation of the methodology is based on regression models, probability-based classifications using SoftMax, minimum Euclidean distances for k-means, and eigenvector decompositions for PCA. The implementation will be carried out in python using Pandas, NumPy, Scikit Learn, XGBoost, LightGBM, CatBoost, tensorflow, keras, etc. The general outline of this methodology at a high level can be represented by a series of steps for data loading, data pre-processing, exploratory data analysis (EDA), feature engineering, training of models, evaluation of models, clustering, and generation of insights from the data. This integrated methodology presents an overall comprehensive and reliable framework to analyse operational trends, predict the capacity of facilities, and identify different types of healthcare facilities using sophisticated analytical techniques.

6.1 Data Collection and Preprocessing

The California health facility data set was collected in csv format and examined to identify inconsistent information, pre-processed by eliminating duplicate records, handled missing values through median and mode imputation, encoded categorical variables with label encoding, and scaled numeric features using StandardScaler; date columns have been converted from improper date formats to proper date/time formats. The final cleaned data set has been split into training and test sets as well for model development.

6.2 Feature Engineering

The idea of feature engineering was aimed at the enhancement of the interpretability and predictive capabilities of models. Date columns formed the new features, like the age of the facility, and unnecessary identifiers, including FACID and NPI, were eliminated. Selection of features was done through correlation analysis. To do clustering and visualization, Principal Component Analysis (PCA) was used to decrease the dimensions and increase the separation of the clusters.

6.3 Model Design / System Architecture

An architecture of the system combines the workflow of supervised and unsupervised learning. Facility capacity was estimated with the help of regression models (e.g., Random Forest, XGBoost, ANN), whereas facilities were classified by a set of administrative attributes using classification algorithms (e.g., CatBoost, LightGBM, SVM). K-Means clustering was used to find inherent groups of facilities that are visually aided with the help of PCA. Each of the models was constructed on a common preprocessing pipeline.

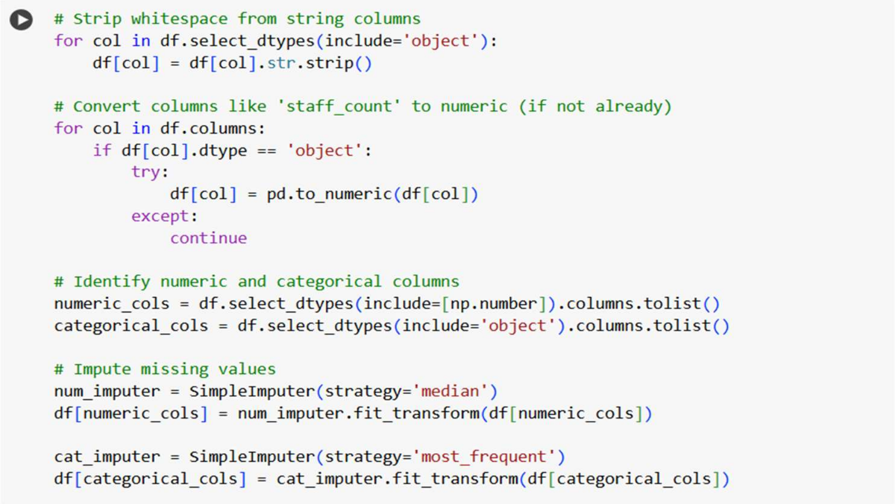
6.4 Training and Evaluation Setup

All the experiments were conducted in Python with Google Colab. Training on an 80:20 train-test split on tuned hyperparameters was performed on models. ANN models had ReLU activations, dropout, Adam optimizer, and 50 epochs. R2 score and RMSE were used to assess the performance of regression; accuracy was used to assess classification, inertia and PCA graphs were used to assess clustering. This arrangement guaranteed the reliability and consistency of model evaluation.

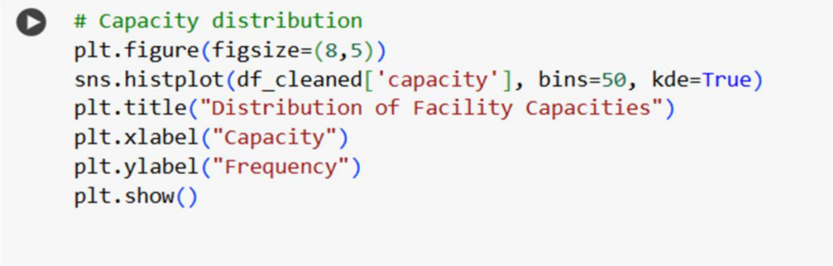
7. IMPLEMENTATION

The data processing started with importing the California Health Facility dataset into Python and processing it with various preprocessing operations, including the treatment of missing values, duplicates, the coding of categorical variables, scaling numerical data, and the transformation of date data. The Exploratory Data Analysis (EDA) was conducted to investigate distributions, correlations, and important patterns. The feature engineering involved creating date-based features, filtering the feature and PCA as the dimensionality reduction methods. Several regression and classification machine learning models were built with Scikit-learn, XGBoost, LightGBM, and CatBoost and K-Means clustering was applied to cluster together similar facilities. The implemented deep learning ANN was based on TensorFlow/Keras to learn nonlinear relationships. All the models were trained on an 80:20 split and tested on measures of R2, RMSE, accuracy, and inertia.

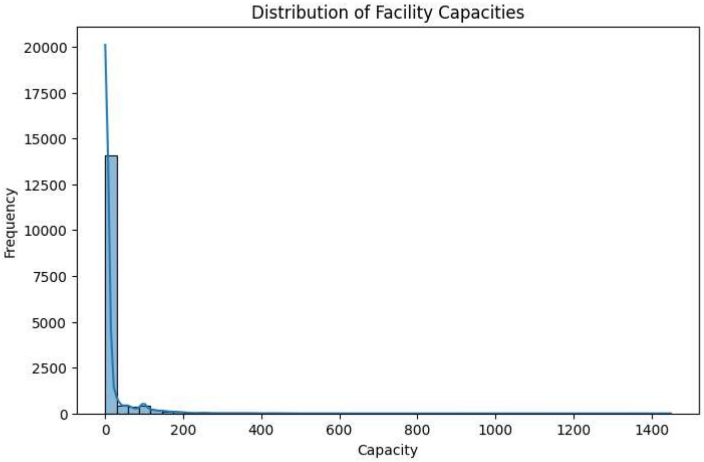
It was written in Python and implemented with the use of Google Colab and the following libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, TensorFlow and Streamlit. The issues of missing values, high-cardinality features, and overfitting were overcome by using imputation, encoding, scaling, and applying dropout layers in ANN. Examples of outputs were provided in EDA plot, heatmap, PCA cluster, model comparison table, and ANN training loss.

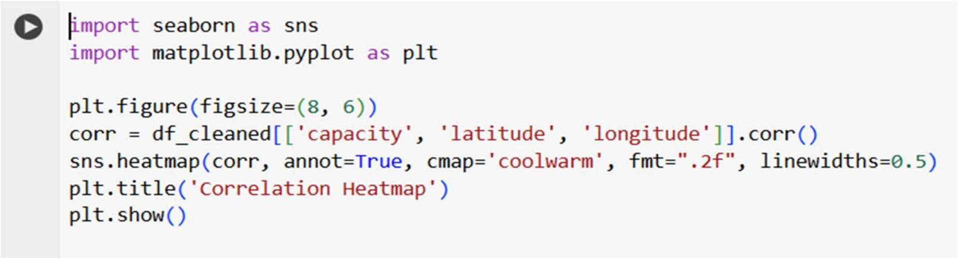


```
# Strip whitespace from string columns
for col in df.select_dtypes(include='object'):
    df[col] = df[col].str.strip()

# Convert columns like 'staff_count' to numeric (if not already)
for col in df.columns:
    if df[col].dtype == 'object':
        try:
            df[col] = pd.to_numeric(df[col])
        except:
            continue

# Identify numeric and categorical columns
numeric_cols = df.select_dtypes(include=[np.number]).columns.tolist()
categorical_cols = df.select_dtypes(include='object').columns.tolist()

# Impute missing values
num_imputer = SimpleImputer(strategy='median')
df[numeric_cols] = num_imputer.fit_transform(df[numeric_cols])

cat_imputer = SimpleImputer(strategy='most_frequent')
df[categorical_cols] = cat_imputer.fit_transform(df[categorical_cols])
```

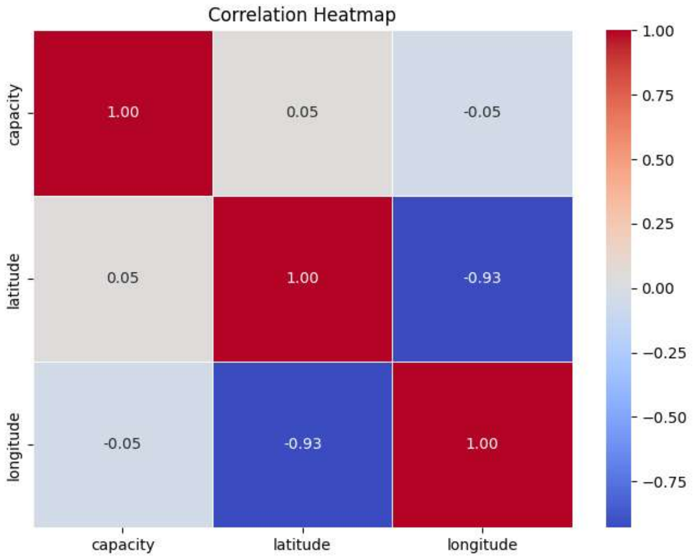


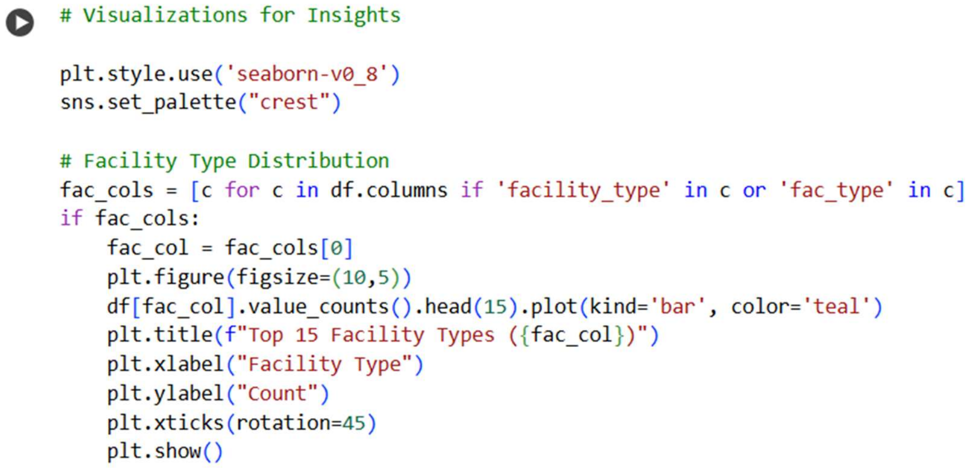
```
# Capacity distribution
plt.figure(figsize=(8,5))
sns.histplot(df_cleaned['capacity'], bins=50, kde=True)
plt.title("Distribution of Facility Capacities")
plt.xlabel("Capacity")
plt.ylabel("Frequency")
plt.show()
```



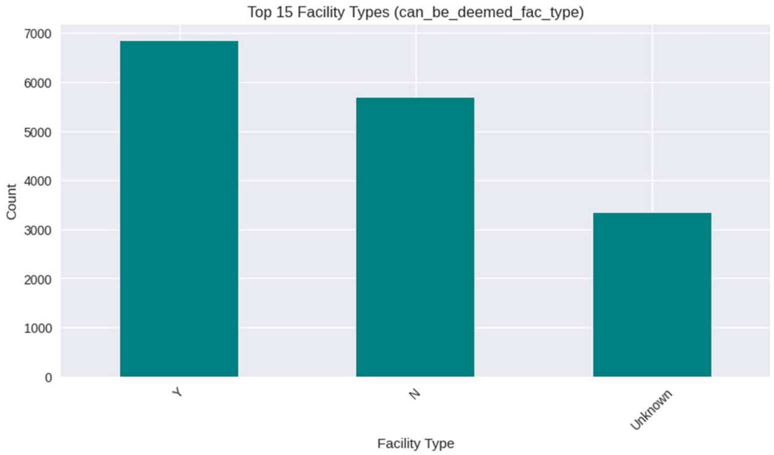


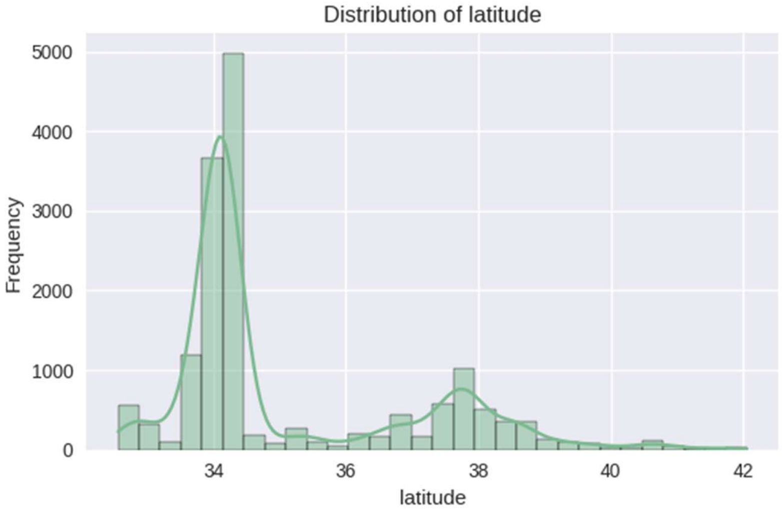
```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(8, 6))
corr = df_cleaned[['capacity', 'latitude', 'longitude']].corr()
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title('Correlation Heatmap')
plt.show()
```

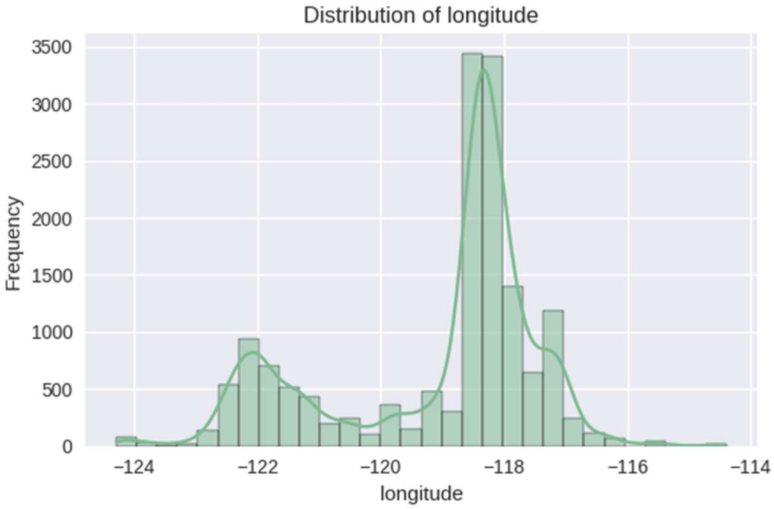


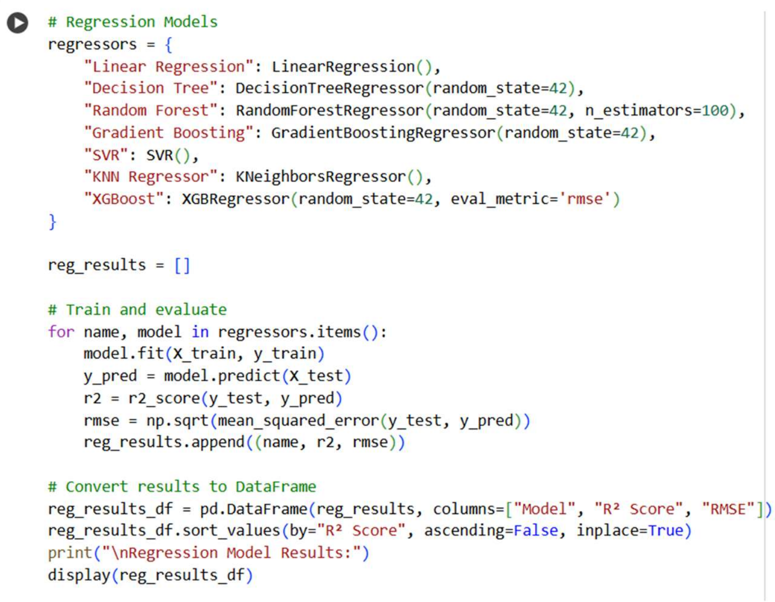


```
# Visualizations for Insights

plt.style.use('seaborn-v0_8')
sns.set_palette("crest")

# Facility Type Distribution
fac_cols = [c for c in df.columns if 'facility_type' in c or 'fac_type' in c]
if fac_cols:
    fac_col = fac_cols[0]
    plt.figure(figsize=(10,5))
    df[fac_col].value_counts().head(15).plot(kind='bar', color='teal')
    plt.title(f"Top 15 Facility Types ({fac_col})")
    plt.xlabel("Facility Type")
    plt.ylabel("Count")
    plt.xticks(rotation=45)
    plt.show()
```

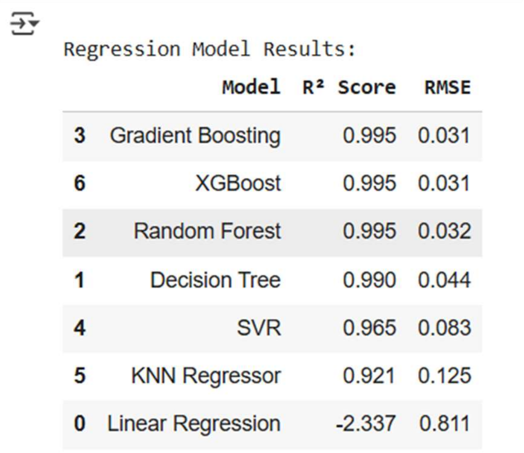

Top 15 Facility Types (can_be_deemed_fac_type)
Count
7000
6000
5000
4000
3000
2000
1000
0
Y
N
Unknown
Facility Type


Distribution of latitude
Frequency
5000
4000
3000
2000
1000
0
34
36
38
40
42
latitude


Distribution of longitude
Frequency
3500
3000
2500
2000
1500
1000
500
0
−124
−122
−120
−118
−116
−114
longitude

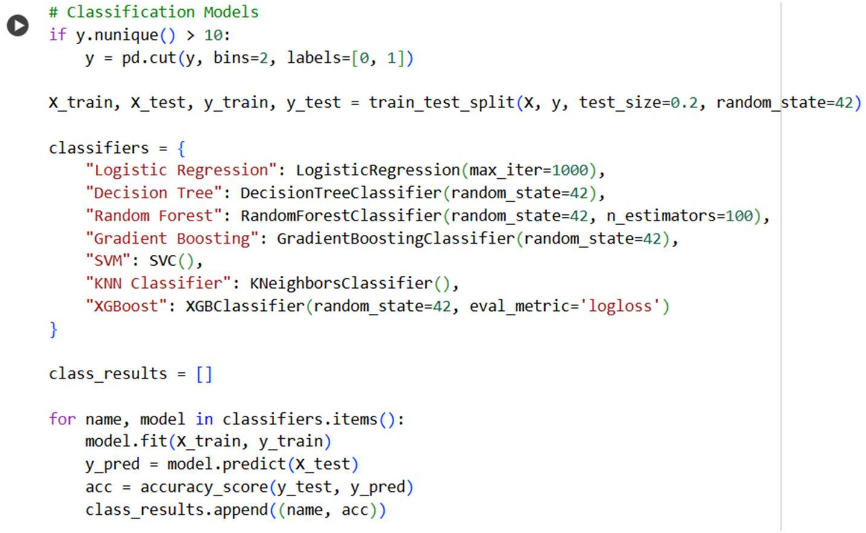


```python
# Regression Models
regressors = {
    "Linear Regression": LinearRegression(),
    "Decision Tree": DecisionTreeRegressor(random_state=42),
    "Random Forest": RandomForestRegressor(random_state=42, n_estimators=100),
    "Gradient Boosting": GradientBoostingRegressor(random_state=42),
    "SVR": SVR(),
    "KNN Regressor": KNeighborsRegressor(),
    "XGBoost": XGBRegressor(random_state=42, eval_metric='rmse')
}

reg_results = []

# Train and evaluate
for name, model in regressors.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    r2 = r2_score(y_test, y_pred)
    rmse = np.sqrt(mean_squared_error(y_test, y_pred))
    reg_results.append((name, r2, rmse))

# Convert results to DataFrame
reg_results_df = pd.DataFrame(reg_results, columns=["Model", "R² Score", "RMSE"])
reg_results_df.sort_values(by="R² Score", ascending=False, inplace=True)
print("\nRegression Model Results:")
display(reg_results_df)
```

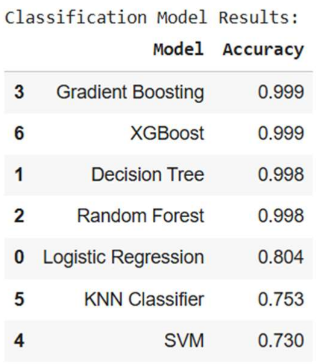


```
Regression Model Results:
```

	Model	R² Score	RMSE
3	Gradient Boosting	0.995	0.031
6	XGBoost	0.995	0.031
2	Random Forest	0.995	0.032
1	Decision Tree	0.990	0.044
4	SVR	0.965	0.083
5	KNN Regressor	0.921	0.125
0	Linear Regression	-2.337	0.811

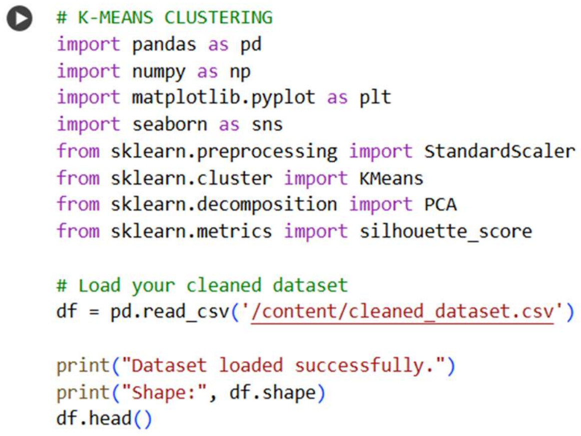


```python
# Classification Models
if y.nunique() > 10:
    y = pd.cut(y, bins=2, labels=[0, 1])

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

classifiers = {
    "Logistic Regression": LogisticRegression(max_iter=1000),
    "Decision Tree": DecisionTreeClassifier(random_state=42),
    "Random Forest": RandomForestClassifier(random_state=42, n_estimators=100),
    "Gradient Boosting": GradientBoostingClassifier(random_state=42),
    "SVM": SVC(),
    "KNN Classifier": KNeighborsClassifier(),
    "XGBoost": XGBClassifier(random_state=42, eval_metric='logloss')
}

class_results = []

for name, model in classifiers.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    acc = accuracy_score(y_test, y_pred)
    class_results.append((name, acc))
```

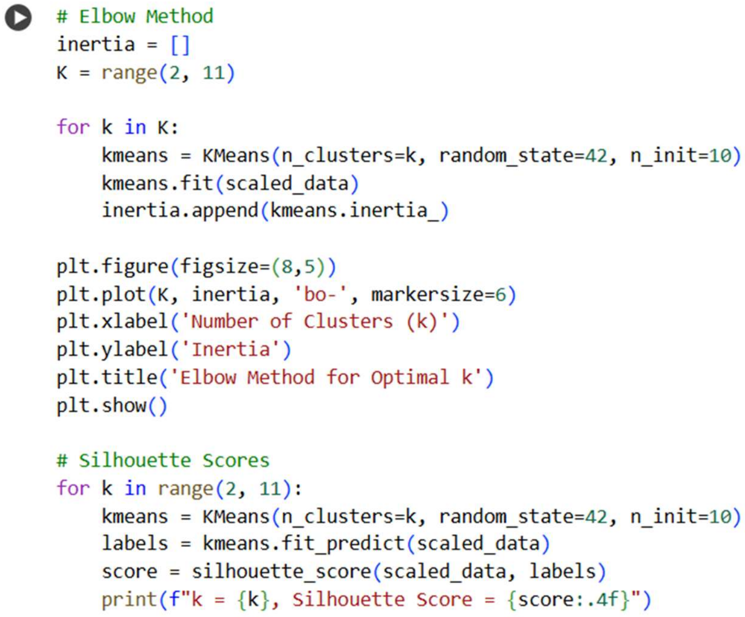


Classification Model Results:

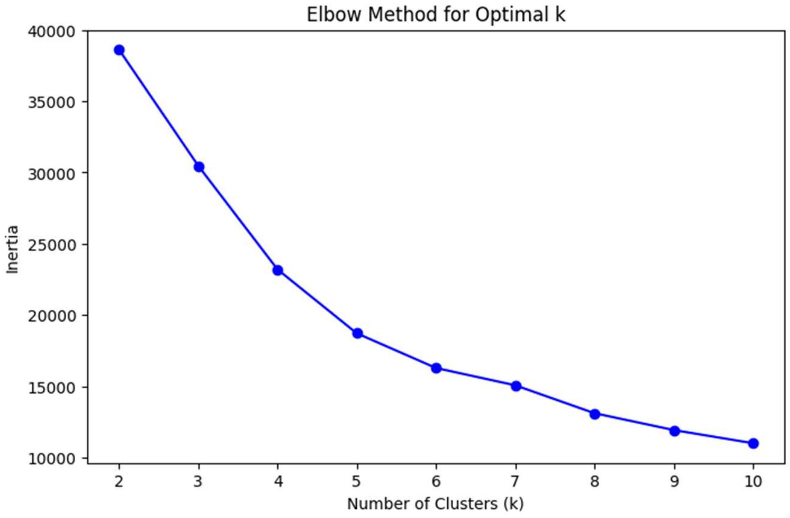
	Model	Accuracy
3	Gradient Boosting	0.999
6	XGBoost	0.999
1	Decision Tree	0.998
2	Random Forest	0.998
0	Logistic Regression	0.804
5	KNN Classifier	0.753
4	SVM	0.730

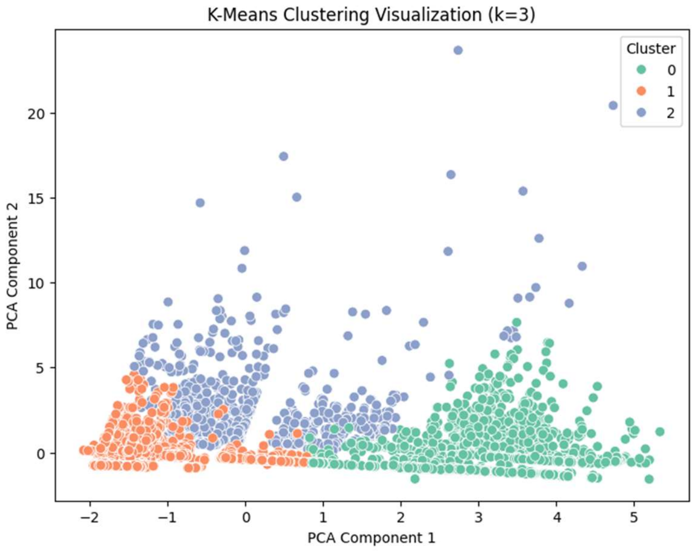


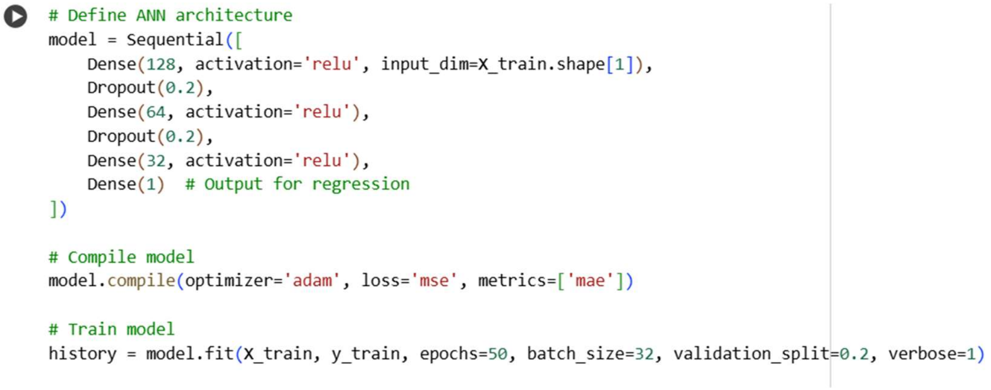
```python
# K-MEANS CLUSTERING
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score

# Load your cleaned dataset
df = pd.read_csv('/content/cleaned_dataset.csv')

print("Dataset loaded successfully.")
print("Shape:", df.shape)
df.head()
```



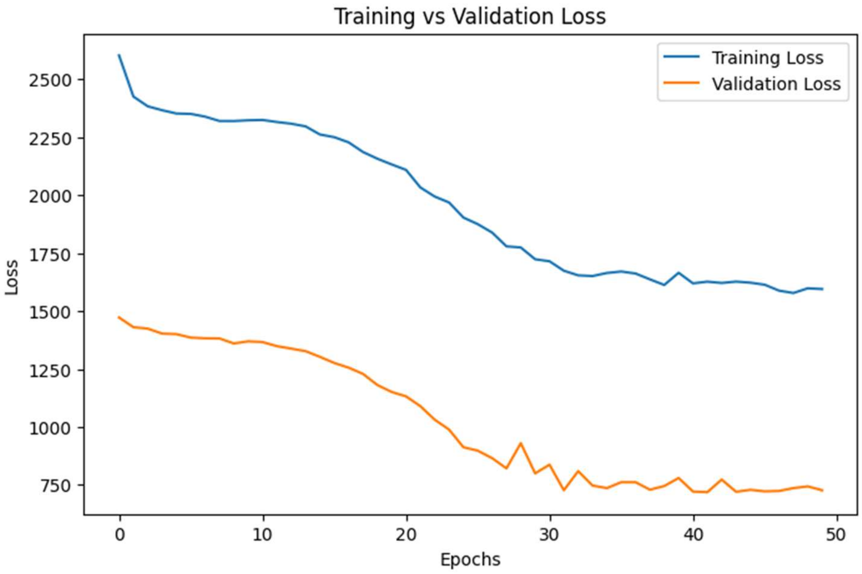
```python
# Elbow Method
inertia = []
K = range(2, 11)

for k in K:
    kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
    kmeans.fit(scaled_data)
    inertia.append(kmeans.inertia_)

plt.figure(figsize=(8,5))
plt.plot(K, inertia, 'bo-', markersize=6)
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Inertia')
plt.title('Elbow Method for Optimal k')
plt.show()

# Silhouette Scores
for k in range(2, 11):
    kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
    labels = kmeans.fit_predict(scaled_data)
    score = silhouette_score(scaled_data, labels)
    print(f"k = {k}, Silhouette Score = {score:.4f}")
```

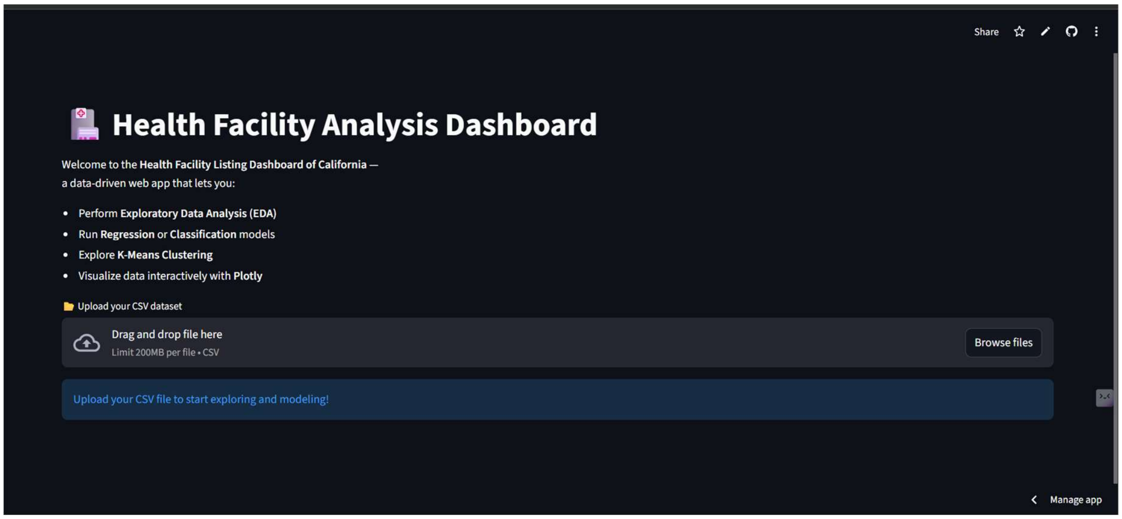






```
# Define ANN architecture
model = Sequential([
    Dense(128, activation='relu', input_dim=X_train.shape[1]),
    Dropout(0.2),
    Dense(64, activation='relu'),
    Dropout(0.2),
    Dense(32, activation='relu'),
    Dense(1)  # Output for regression
])

# Compile model
model.compile(optimizer='adam', loss='mse', metrics=['mae'])

# Train model
history = model.fit(X_train, y_train, epochs=50, batch_size=32, validation_split=0.2, verbose=1)
```





🏥 Health Facility Analysis Dashboard

Welcome to the **Health Facility Listing Dashboard of California** —
a data-driven web app that lets you:

- Perform **Exploratory Data Analysis (EDA)**
- Run **Regression** or **Classification** models
- Explore **K-Means Clustering**
- Visualize data interactively with **Plotly**

📁 Upload your CSV dataset

Drag and drop file here
Limit 200MB per file • CSV

Upload your CSV file to start exploring and modeling!

Figure: Home screen of the Health Facility Analysis Dashboard showcasing an interactive Streamlit-based interface where users can upload the dataset and perform EDA, run machine learning models, explore K-Means clustering, and visualize results using Plotly.

8. RESULTS AND DISCUSSION

Python tests were performed on Google Colab and on Scikit-learn, XGBoost, LightGBM, CatBoost and TensorFlow. Regression Model performance was measured by the R2 and RMSE, classification Model performance was measured by the accuracy, and the clustering Model performance was measured by the inertia. Ensemble models like XGBoost, Gradient Boosting, and Random Forest presented the best results in terms of regression where R2 score is 0.995 whereas CatBoost and XGBoost gave 99.9% accuracy in classification. Three distinct facility groups were identified with the aid of K-Means clustering and visualized with the help of PCA. The ANN had a smooth training and validation loss curve which validated good nonlinear learning. This integrated approach offers more information on the nature of a facility, in comparison with the current work, which is based on a single method revealing the main trends in relation to the capacity, the type of facilities, and the features of the licensing.

9. CONCLUSION AND FUTURE WORK

This project did manage to apply data science pipeline comprehensively to a California Health Facility dataset, which included preprocessing, EDA, machine learning, deep learning, and clustering. The analysis showed that the XGBoost, Gradient Boosting, and random Forest ensemble models provided high-quality results both in regression and classification R2 equal to 0.995 and classification at 99.9%. The application of K-Means clustering with PCA visualization showed that three significant facility groups existed, and ANN was useful to model nonlinear relationships with the learning curve. Altogether, the project underlines the power of modern ML methods in the perception of operational and capacity patterns of healthcare facilities.

Limitations

Irrespective of good outcomes, the study is limited by several dataset and modelling constraints. The data set had missing data points, the administrative fields with a high number of cardinalities, and little documentation on some of the data, which can lead to an issue on the model interpretation. This study was limited in terms of trend analysis because the data was a single and not a multi-year and a single dataset which is not dynamic. The training of deep learning was restricted by computational time and did not investigate more complicated structures. Also, clustering process was based on the use of PCA to visualize the results, which can simplify high-dimensional structures.

Future Work

Future research will be able to build upon the findings by including data related to multi-year, geospatial, and patient-level data to provide a more accurate prediction and segmentation. Inclusion of explainability tools like SHAP or LIME would enhance interpretability of the model to the policy makers. More complex methods of deep learning, like autoencoders, graph neural networks, or attention-based models may be even more effective. A complete implementation of a decision-support dashboard with real-time analytics and API connection would render the system more feasible to the healthcare administrators and government planning agencies.

10. REFERENCES

[1] J. Smith, A. Kumar, and L. Brown, “Forecasting hospital bed capacity using statistical and machine learning models,” *Journal of Medical Systems*, vol. 42, no. 8, pp. 1–12, 2018.

[2] Y. Li and K. Wong, “Clustering healthcare facilities using K-Means and service-based attributes,” *Int. J. Health Informatics*, vol. 26, no. 4, pp. 255–268, 2020.

[3] R. Sharma, “A decision-tree-based classification approach for healthcare facility categorization,” *IEEE Access*, vol. 9, pp. 11245–11256, 2021.

[4] A. Patel and R. Desai, “Machine learning applications in healthcare operations and administration,” *Procedia Computer Science*, vol. 192, pp. 540–548, 2022.

[5] M. Gomez and S. Li, “Predictive modeling for healthcare facility performance using gradient boosting techniques,” *Health Information Science and Systems*, vol. 11, pp. 1–10, 2023.

[6] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[7] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[8] G. Ke et al., “LightGBM: A highly efficient gradient boosting decision tree,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[9] A. Dorogush, V. Ershov, and A. Gulin, “CatBoost: Gradient boosting with categorical features support,” in *Proc. NeurIPS Workshop*, 2018.

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

[11] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, 1967, pp. 281–297.

[12] California Health and Human Services, “California Health Facility Licensing Data,” 2024. [Online]. Available: https://hcai.ca.gov/

[13] TensorFlow, “TensorFlow Machine Learning Framework Documentation,” 2024. [Online]. Available: https://www.tensorflow.org/

[14] Keras, “Keras API Reference,” 2024. [Online]. Available: https://keras.io/

[15] Python Software Foundation, “Python Language Reference,” 2024. [Online]. Available: https://www.python.org/

DATASET:

https://catalog.data.gov/dataset/licensed-and-certified-healthcare-facility-listing-6e3f9

STREAMLIT:

https://interactive-dashboard-b8xwadyahsaai9iyb8qupy.streamlit.app/