Data Visualization & Insight Generation from Structured Datasets

Saswata Roy
Hrit Saha

15 July, 2024

Abstract

This report explores the development of an AI-based solution for knowledge representation and insight generation from structured datasets. The solution tackles the challenge of extracting valuable insights from the ever-growing volume of data in various sectors.

It achieves this through a multi-step process:

1. **Data Pre-processing**: Employing Python libraries like **Pandas** and **NumPy**, the solution cleans and prepares the data for analysis.

2. **Knowledge Representation**: The knowledge within the data is effectively represented using visual aids like charts and graphs created with tools like **Matplotlib** and **Seaborn**.

3. **Pattern Identification**: Leveraging machine learning techniques from **Scikit-learn**, the solution identifies patterns such as trends and anomalies within the data.

4. **Insight Generation**: Based on the identified patterns and distributions, multiple insights can be inferred which may be useful for companies or organizations for various decision-making purposes.

Additionally, a user-friendly interface is envisioned using web development frameworks to facilitate user interaction and comprehension of the generated insights.

This project demonstrates the practical application of Machine Learning and Deep Learning in data analysis and decision-making. It showcases the ability to develop AI-based solutions for knowledge extraction and insight generation from structured datasets.

CONTENTS

1. Introduction

1.1. Problem Statement
1.2. Project Objectives
1.3. Motivation

1. Introduction

1.1. Problem Statement

Problem Description

In the era of big data, organizations across various sectors are generating massive amounts of data every day. This data, if processed and analyzed correctly, can provide valuable insights that can significantly improve the decision-making process. However, the challenge lies in effectively representing this knowledge and extracting useful insights from it. Your task is to develop an AI-based solution that can handle this challenge. You will be provided with a structured dataset. Your solution should be able to process this dataset, represent the knowledge contained within it effectively, and generate meaningful insights. The solution should include the following features:

1. **Data Pre-processing**: The solution should be able to clean and pre-process the dataset to make it suitable for further analysis.

2. **Knowledge Representation**: The solution should effectively represent the knowledge contained within the dataset. This could be in the form of graphs, charts, or any other visual representation that makes the data easy to understand.

3. **Pattern Identification**: The solution should be able to identify patterns within the dataset. This could include identifying trends, anomalies, or any other patterns that could provide valuable insights.

4. **Insight Generation**: Based on the identified patterns, the solution should generate meaningful insights. These insights should be presented in a clear and understandable manner.

5. **Scalability**: The solution should be scalable. It should be able to handle datasets of varying sizes and complexities.

6. **User-friendly Interface**: The solution should have a user-friendly interface that allows users to easily interact with it and understand the generated insights.

Dataset

Use any structured dataset for this project. However, it is recommended to use a dataset that is relevant to your field of interest or study.

Evaluation

The solution will be evaluated based on its ability to effectively represent knowledge, identify patterns, generate insights, and its scalability and user-friendliness.

1.2. Project Objectives

This project is aimed to develop an Artificial Intelligence (AI)-based solution capable of effectively representing knowledge in the form of Data Frames and visual plots, and generating valuable insights from structured datasets.

The core objectives were:

- To design a system that can preprocess and analyze structured data, making it suitable for further knowledge extraction and insight generation.
- To develop techniques for effectively representing the knowledge embedded within the data using visual aids and other informative formats.
- To implement machine learning algorithms to identify patterns and trends within the data that could provide significant insights.
- To prioritize scalability in the solution's design, ensuring it can handle datasets of varying sizes and complexities.
- To establish a user-friendly interface for the solution, facilitating user interaction and clear comprehension of the generated insights.

By achieving these objectives, this project aimed to demonstrate the practical application of AI in data analysis and empower data-driven decision-making processes.

1.3. Motivation

The ever-growing volume of data generated across various sectors presents both opportunities and challenges. While this data holds immense potential for valuable insights, effectively extracting knowledge and translating it into actionable information remains a significant hurdle. Traditional data analysis methods can become overwhelmed by the sheer size and complexity of modern datasets.

This project is motivated by the need for innovative solutions that leverage the power of AI to bridge this gap. By developing an AI-based system for knowledge representation and insight generation, we aim to unlock the true potential of structured data. This approach offers several advantages:

- Automated Knowledge Extraction: AI algorithms can automate the process of identifying patterns and trends within data, saving valuable time and resources compared to manual analysis.

- Enhanced Accuracy and Scalability: AI systems can handle large and complex datasets with greater accuracy and efficiency than traditional methods.

- Improved Decision-Making: By providing clear and actionable insights, we can empower data-driven decision-making processes across various domains.

In conclusion, this project is driven by the desire to harness the power of AI to unlock the knowledge hidden within structured data. We believe this solution has the potential to revolutionize data analysis and empower more informed decision-making across various fields.

2. Team

2.1. Team Members
2.2. Team Contribution

2. Team

2.1. Team Members

Saswata Roy

AI Enthusiast

B. Tech. Information Technology | 2022-26
Kalinga Institute of Industrial Technology
Bhubaneswar

Email: saswataroy07jul2004@gmail.com

Hrit Saha

AI Enthusiast

B. Tech. Information Technology | 2022-26
Kalinga Institute of Industrial Technology
Bhubaneswar

Email: hritsaha09@gmail.com

2.2. Team Contribution

This project is a collaborative effort, and we would like to acknowledge the valuable contributions of the following team members:

- **Hrit Saha**: Played a key role in developing the pattern identification algorithms and knowledge representation modules.
- **Saswata Roy**: Contributed to the design and integrating the system with the user interface.

3. Datasets Description

3.1. Social Network Ads
3.2. Bank Marketing Dataset
3.3. Medical Insurance Cost Prediction
3.4. Mushroom Analysis
3.5. Calories Burnt Prediction
3.6. Titanic Survival Prediction
3.7. Parkinson’s Disease Prediction
3.8. Iris Flower Classification
3.9. Credit Card Fraud Detection

3. Datasets Description

3.1. Social Network Ads

Source: https://www.kaggle.com/datasets/nani123456789/social-network-ads

Context

An international car company wants to discover key insights from their customer database. They want to use some of the most advanced machine learning techniques to study their customers. They want to predict whether or not the customer will buy the brand-new car that will be launched soon. Based on the prediction the company will advertise the product of social media.

Content

The dataset used for model building contained 400 observations of 5 variables. The data contains the following information:

- user id: refers to the id of the person.
- gender: male/female
- age: age of the person
- estimated salary: The amount of salary that the person is getting.
- purchased: whether the person has purchased any product or not based on his salary.

3.2. Bank Marketing Dataset

Source: https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset

Context

The dataset contains information collected during the bank's marketing campaigns. It includes various features related to bank clients, their interactions with the bank, and the outcomes of previous marketing efforts. The target variable indicates whether a client has subscribed to a term deposit account.

Content

The dataset used for model building contained 11162 observations of 17 variables. The data contains the following information:

- Age: Numeric feature representing the age of the bank client.
- Job: Categorical feature indicating the type of job the client has.
- Marital: Categorical feature indicating the marital status of the client.
- Education: Categorical feature representing the educational level of the client.

- Default: Categorical feature indicating whether the client has credit in default.
- Housing: Categorical feature indicating whether the client has a housing loan.
- Loan: Categorical feature indicating whether the client has a personal loan.
- Balance: Numeric feature representing the balance of the individual.
- Contact: Categorical feature indicating the communication type used to contact the client.
- Month: Categorical feature indicating the month of the last contact.
- Day: Categorical feature indicating the day of the week of the last contact.
- Duration: Numeric feature representing the duration of the last contact in seconds.
- Campaign: Numeric feature representing the number of contacts performed during the current campaign for this client.
- Pdays: Numeric feature representing the number of days since the client was last contacted from a previous campaign.
- Previous: Numeric feature representing the number of contacts performed before the current campaign for this client.
- Poutcome: Categorical feature representing the outcome of the previous marketing campaign.
- deposit (Target): Binary feature indicating whether the client has subscribed to a term deposit.

3.3. Medical Insurance Cost Prediction

Source: https://www.kaggle.com/datasets/mirichoi0218/insurance

Context

We will build a Linear regression model for medical cost dataset. The dataset consists of age, sex, BMI (body mass index), children, smoker and region feature, which are independent and charge as a dependent feature. We will predict individual medical costs billed by health insurance.

Content

The dataset used for model building contained 1338 observations of 7 variables. The data contains the following information:

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg/m^2) using the ratio of height to weight, ideally 18.5 to 24.9.
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking or not.
- region: beneficiary's US residential area, northeast, southeast, southwest, northwest.

- charges: Individual medical costs billed by health insurance

3.4. Mushroom Analysis

Source: https://www.kaggle.com/datasets/uciml/mushroom-classification

Context

Although this dataset was originally contributed to the UCI Machine Learning repository nearly 30 years ago, mushroom hunting (otherwise known as "shrooming") is enjoying new peaks in popularity. Learn which features spell certain death and which are most palatable in this dataset of mushroom characteristics. And how certain can your model be?

Content

This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like "leaflets three, let it be" for Poisonous Oak and Ivy.

- classes: edible=e, poisonous=p
- cap-shape: bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
- cap-surface: fibrous=f, grooves=g, scaly=y, smooth=s
- cap-color: brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
- bruises: bruises=t, no=f
- odor: almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
- gill-attachment: attached=a, descending=d, free=f, notched=n
- gill-spacing: close=c, crowded=w, distant=d
- gill-size: broad=b, narrow=n
- gill-color: black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y
- stalk-shape: enlarging=e, tapering=t
- stalk-root: bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
- stalk-surface-above-ring: fibrous=f, scaly=y, silky=k, smooth=s
- stalk-surface-below-ring: fibrous=f, scaly=y, silky=k, smooth=s
- stalk-color-above-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
- stalk-color-below-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y

- veil-type: partial=p, universal=u
- veil-color: brown=n, orange=o, white=w, yellow=y
- ring-number: none=n, one=o, two=t
- ring-type: cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
- spore-print-color: black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
- population: abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
- habitat: grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

3.5. Calories Burnt Prediction

Source: https://www.kaggle.com/datasets/fmendes/fmendesdat263xdemos

Context

In today's health-conscious society, monitoring and managing calorie expenditure is a key aspect of maintaining a healthy lifestyle. Understanding how various activities and individual factors impact calorie burn is crucial for individuals striving to achieve fitness goals. Leveraging the capabilities of data science, we aim to address this health and wellness challenge.

This project falls within the domain of Regression Machine Learning Problem. The primary objective is to develop a predictive model for calorie burnt prediction. By analyzing a combination of input features such as physical activity type, duration, intensity, and individual characteristics like age, weight, and gender, the goal is to create a model that accurately estimates the number of calories burnt during a specific activity. This predictive model can empower individuals, fitness enthusiasts, and healthcare professionals with valuable insights to optimize their calorie management and physical activity planning.

Content

There are two datasets named exercise (15000 observations and 8 variable) and calories (containing 15000 observations and 2 variables). The data contains the following information:

- USER-ID: User_id od the person.
- Gender: Gender of the person.
- Age: Age of the person.
- Height: Height of the person.
- Weight: Weight of the person.
- Duration: Duration of the exercise performed by the person.
- Heart_Rate: Heart_Rate of the person.
- Body_Temp: Body_Temp of the person.

- Calories: Total calories burnt by the person.

3.6. Titanic Survival Prediction

Source: https://www.kaggle.com/datasets/yasserh/titanic-dataset/data

Context

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone on board, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

Content

The dataset used for model building contained 891 observations of 12 variables. The data contains the following information:

- Passenger ID: Numeric feature representing the passengerID.
- Survived: Categorical feature indicating whether the passenger survived or not
- Pclass: Numerical feature representing the class of the passenger.
- Name: Representing the name of the passenger.
- Sex: Categorical feature indicating the age of the passenger.
- Age: Numerical feature indicating the age of the passenger.
- SibSp: Representing no. of siblings / spouses aboard the Titanic
- Parch: Representing no. of parents / children aboard the Titanic
- Ticket: Representing the ticket no. of the passenger.
- Fare: Representing the passenger fare.
- Cabin: Representing the cabin no. of the passenger.
- Embarked: Embarked implies where the passenger mounted from.

3.7. Parkinson’s Disease Prediction

Source:

https://www.kaggle.com/datasets/jainaru/parkinson-disease-detection?resource=download

Context

Parkinson’s Disease (PD) is a degenerative neurological disorder marked by decreased dopamine levels in the brain. It manifests itself through a deterioration of movement,

including the presence of tremors and stiffness. There is commonly a marked effect on speech, including dysarthria (difficulty articulating sounds), hypophonia (lowered volume), and monotone (reduced pitch range). Additionally, cognitive impairments and changes in mood can occur, and risk of dementia is increased.

Traditional diagnosis of Parkinson's Disease involves a clinician taking a neurological history of the patient and observing motor skills in various situations. Since there is no definitive laboratory test to diagnose PD, diagnosis is often difficult, particularly in the early stages when motor effects are not yet severe. Monitoring progression of the disease over time requires repeated clinic visits by the patient. An effective screening process, particularly one that doesn't require a clinic visit, would be beneficial. Since PD patients exhibit characteristic vocal features, voice recordings are a useful and non-invasive tool for diagnosis. If machine learning algorithms could be applied to a voice recording dataset to accurately diagnosis PD, this would be an effective screening step prior to an appointment with a clinician

Content

This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds to one of 195 voice recordings from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to the "status" column which is set to 0 for healthy and 1 for PD. The data contains the following information:

- name - ASCII subject name and recording number
- MDVP: Fo (Hz) - Average vocal fundamental frequency
- MDVP: Fhi (Hz) - Maximum vocal fundamental frequency
- MDVP: Flo (Hz) - Minimum vocal fundamental frequency
- MDVP: Jitter (%), MDVP: Jitter (Abs), MDVP: RAP, MDVP: PPQ, Jitter: DDP - Several
- measures of variation in fundamental frequency
- MDVP: Shimmer, MDVP: Shimmer(dB), Shimmer: APQ3, Shimmer: APQ5, MDVP: APQ, Shimmer: DDA - Several measures of variation in amplitude
- NHR, HNR - Two measures of ratio of noise to tonal components in the voice
- status - Health status of the subject (one) - Parkinson's, (zero) - healthy
- RPDE, D2 - Two nonlinear dynamical complexity measures
- DFA - Signal fractal scaling exponent
- spread1, spread2, PPE - Three nonlinear measures of fundamental frequency variation'

3.8. Credit Card Fraud Detection

Source: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Context

The dataset contains transactions made by credit cards in September 2013 by European cardholders.

This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

Content

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example, dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Given the class imbalance ratio, we recommend measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC). Confusion matrix accuracy is not meaningful for unbalanced classification.

3.9. Iris Flower Classification

Source: https://www.kaggle.com/datasets/arshid/iris-flower-dataset

Context

The Iris flower data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. The data set consists of 50 samples from each of three species of Iris (Iris sentosa, Iris virginica, and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

This dataset became a typical test case for many statistical classification techniques in machine learning such as support vector machines

Content

The dataset used for model building contained 150 observations of 5 variables. The data contains the following information:

- Sepal-length: Numeric feature representing the length of the sepal.
- Sepal-width: Numeric feature representing the width of the sepal.
- Petal-length: Numerical feature representing the length of the petal.

- Petal-width: Numeric feature representing the width of the petal.
- Species: Categorical feature indicating the species of the iris flower.

4. Methodology

4.1. Selection of Datasets
4.2. App Design
4.3. Data Visualization
4.4. Data Preprocessing
4.5. ML and ANN Modelling

4. Methodology

4.1. Selection of Datasets

The datasets used in this project were chosen from Kaggle, a renowned online platform known for its vast collection of open-source datasets encompassing various domains. Leveraging Kaggle's resources offered several advantages:

- **Diversity and Relevance**: Kaggle provides a rich repository of datasets across numerous fields, allowing for the selection of data directly relevant to the project's objectives. This diversity ensures that the datasets chosen are highly pertinent to the specific research questions being addressed.

- **Data Quality**: Many Kaggle datasets are well-maintained by the community, ensuring a higher level of data quality and reducing the need for extensive cleaning. Community ratings and discussions often highlight the reliability of these datasets, aiding in the selection process.

- **Accessibility and Transparency**: Datasets on Kaggle are readily downloadable and often accompanied by clear documentation, facilitating project transparency and reproducibility. This accessibility allows for easy sharing and validation of findings, promoting open research practices.

- **Community Support and Collaboration**: Kaggle's active community provides support through forums and discussions, offering insights and tips for using various datasets. This collaborative environment enhances the research process by providing additional resources and troubleshooting assistance.

Selection Process

- **Selection Criteria**: The specific criteria used to select the datasets from Kaggle included:

- **Data Size**: Ensuring the dataset was sufficiently large to support robust statistical analysis but manageable within the computational resources available.

- **Relevance to Research Question**: The datasets needed to directly address the key questions and objectives of the project. This involved evaluating the content and context of the data.

- **Presence of Missing Values**: Preference was given to datasets with minimal missing values to reduce preprocessing efforts. However, datasets with some missing data were also considered if they offered high relevance and quality.

- **Specific Features Required for Analysis**: Datasets were selected based on the presence of specific variables or features essential for the planned analyses. This included checking for data granularity, variable types, and the availability of target variables.

- **Historical and Current Data**: Depending on the project needs, datasets with historical data, current data, or a combination of both were considered to provide a comprehensive view.

By outlining the selection process and providing a basic description of the chosen datasets, this section gives readers a clear understanding of how and why the data were selected for the project. This approach ensures that the data align with the research objectives and meet quality standards, ultimately contributing to the project's overall credibility and robustness.

4.2. App Design

For this project, a web application framework was chosen to develop the user interface for interacting with the AI solution. **Streamlit**, a Python library specifically designed for simple and stylish web app development, was selected due to its several advantages:

- **Simplicity and Ease of Use**: Streamlit allows creating data apps with minimal coding compared to traditional web development frameworks. This streamlined development process was crucial for this project, allowing a focus on the core functionalities of the AI solution.

- **Data Visualization Integration**: Streamlit seamlessly integrates with popular data visualization libraries like Matplotlib and Seaborn. This facilitates the creation of clear and informative visualizations to represent the knowledge extracted from the data.

- **Scalability**: Streamlit applications can be deployed on various platforms, allowing them to scale to accommodate an increasing user base or larger datasets in the future.

Within the Streamlit framework, the following design considerations were implemented:

- User Interface Design: The user interface was designed for simplicity and ease of use. This includes clear instructions, intuitive navigation, and well-organized data visualizations.

- Functionality Modules: The app will incorporate separate modules for data pre-processing, knowledge representation (visualization), pattern identification, and

insight generation for each dataset. Users can interact with each module sequentially or focus on specific functionalities as needed.

- Output Presentation: The application will present the generated insights and plots in a clear and concise manner. This could involve text summaries and visual aids.

By leveraging Streamlit's capabilities, the app design aims to provide a user-friendly and interactive platform for users to explore their data and gain valuable insights through the AI solution's functionalities.

The Streamlit's standard design has been modified with custom CSS and third-party Streamlit modules like **Streamlit-extras** to add a little touch of uniqueness and individuality.

4.3. Data Visualization

Data visualization plays a crucial role in this project by enabling effective knowledge representation and clear communication of insights generated by the AI solution. To achieve this, the project leveraged the capabilities of Python libraries like Matplotlib, Seaborn, and Plotly.

1. **Matplotlib**: This fundamental library provides a versatile toolkit for creating a wide range of static data visualizations. It was employed for generating basic plots like bar charts, histograms, and scatter plots to represent various data distributions and relationships.

2. **Seaborn**: Built on top of Matplotlib, Seaborn offers a higher-level interface specifically designed for statistical graphics. It was utilized to create more sophisticated visualizations like heatmaps and violin plots, allowing for a deeper exploration of patterns within the data.

3. **Plotly**: When dealing with interactive visualizations or complex datasets, Plotly provided valuable functionalities. Its interactive charts and graphs enable users to zoom, pan, and explore the data dynamically, facilitating a more comprehensive understanding of the insights.

The selection of specific visualization techniques was guided by the nature of the data and the type of insights we aimed to communicate. For instance:

- Pair plots: For quick analysis of the entire data in the form of feature vs. feature scatterplot and feature distributions.
- Count plots: Effective for comparing categorical data variables.

- Histograms and KDE plots: Used to visualize the distribution of continuous data.
- Scatter plots: Employed to reveal relationships between two continuous variables.
- Heatmaps: Utilized to represent the correlation matrix of the data, highlighting strong relationships between variables.

By incorporating these diverse visualization libraries and tailoring the visualizations to the data and insights, the project aimed to create a clear and informative user experience for exploring the knowledge extracted from the data.

4.4. Data Preprocessing

Prior to feeding the data into the AI solution, a crucial step involved preprocessing the data to ensure its quality and suitability for analysis. This section details the techniques employed using Python libraries like scikit-learn, NumPy, and Pandas.

Data Cleaning

The initial phase of data preprocessing focused on cleaning the data to address any inconsistencies or errors that could hinder analysis. Here's a breakdown of the cleaning techniques implemented:

- **Handling Missing Values**: Missing data points were identified and addressed using techniques like:
    - **Imputation**: Filling missing values with estimated values based on statistical methods (e.g., mean imputation) or other data points.
    - **Deletion**: Removing rows or columns with a high percentage of missing values, especially if imputation wasn't feasible.
- **Removing Insignificant Features**: Certain features exist in the data set but have no significant impact on the target variable or the overall analysis. These features can be identified and removed to improve the efficiency and accuracy of the AI solution.

Data Encoding

Categorical data, which involves features with labels or qualitative values, needed to be converted into a numerical format suitable for the AI algorithms. This was achieved through the following encoding techniques:

- **Label Encoding**: Assigning a numerical value (integer) to each unique category label. This is a simple technique but may introduce unintended ordering between the categories.

- **One-Hot Encoding**: Creating a new binary feature for each category, with a value of 1 indicating membership in that category and 0 otherwise. This approach avoids introducing artificial ordering but can increase the number of features.

The choice between Label Encoding and One-Hot Encoding depends on the specific characteristics of your data and the AI algorithms used in your solution.

Data Scaling

Since different features in a dataset may have varying ranges and units, data scaling was employed to ensure all features contribute equally during analysis. Here are the scaling techniques considered:

- **Standard Scaling (Normalization)**: This technique transforms the data to have a zero mean and unit standard deviation. This is a common approach for many machine learning algorithms.

- **Min-Max Scaling**: This technique scales the data to a range between 0 and 1 (or a specific minimum and maximum value). It can be useful when dealing with data containing outliers.

- **Max-Abs Scaling**: This technique scales the data such that the maximum absolute value in each feature becomes 1. It can be useful for data with a skewed distribution.

- **Quantile Scaling**: This technique transforms the data based on specific percentiles (e.g., interquartile range). It can be helpful when the data distribution is non-normal and many outliers exist.

The selection of the most appropriate scaling technique depends on the specific dataset and the AI algorithms involved in your project.

By implementing these data preprocessing techniques, we ensured the data fed into the AI solution was clean, consistent, and appropriately formatted for optimal analysis and generation of valuable insights.

4.5. ML and ANN Modelling

This section details the Machine Learning (ML) and Artificial Neural Network (ANN) modelling techniques employed within the AI solution, if applicable to your project. The specific techniques chosen depend on the nature of the data and the desired outcome (classification or regression).

Python libraries like **Scikit-learn**, **NumPy**, **Pandas**, and **TensorFlow** were used for implementing these models.

Classification Datasets

If your project involves datasets with categorical target variables (where the outcome falls into distinct categories), classification models were likely used to identify patterns and relationships within the data. Here are some common classification algorithms that could be explored:

- **Logistic Regression**: A widely used linear model for predicting binary outcomes (e.g., customer churn prediction).
- **Decision Tree Classifier**: Tree-based models that classify data points based on a series of decision rules learned from the data.
- **Support Vector Classifier (SVC)**: These models aim to find the optimal hyperplane that separates data points belonging to different classes.
- **K-Neighbors Classifier (KNC)**: This approach classifies data points based on the majority class of its nearest neighbors.
- **Random Forest Classifier**: An ensemble method that combines multiple decision trees to improve classification accuracy and robustness.

Selection and Evaluation: The choice of the most suitable classification model depends on factors like the size and complexity of the data, the number of classes, and the desired level of interpretability. The models were likely evaluated using metrics like **Accuracy**, **Precision**, **Recall**, and **F1-score**.

Regression Datasets

If your project involves datasets with continuous target variables (where the outcome can take on any value within a range), regression models were likely used to model the relationship between features and the target variable. Here are some common regression algorithms that could be explored:

- **Linear Regression**: This widely used model fits a linear relationship between features and the target variable.

- **Decision Tree Regressor**: Similar to decision trees for classification, these models predict a continuous value based on a series of decision rules.

- **Support Vector Regressor (SVR)**: This technique aims to find a hyperplane that minimizes the distance between the data points and the hyperplane while maintaining smoothness.

- **Random Forest Regressor**: Similar to Random Forest for classification, this approach combines multiple decision trees for improved regression performance.

Selection and Evaluation: The choice of the most suitable regression model depends on similar factors as for classification tasks, with additional considerations like the linearity of the relationship and the presence of outliers. The models were likely evaluated using metrics like **Mean squared error (MSE)**, **R-squared**, and **adjusted R-squared**.

Artificial Neural Network (ANN) Models

These complex ANN architectures can be powerful tools for classification tasks, especially with large and complex datasets.

Selection and Evaluation: The models were likely evaluated using metrics like **Mean squared error (MSE)**, **R-squared**, and **adjusted R-squared** for regression problems and metrics like **Accuracy**, **Precision**, **Recall**, and **F1-score** for classification problems.

Hyperparameter Tuning

It's important to note that most ML and ANN models require hyperparameter tuning. This involves adjusting specific parameters of the model to optimize its performance for the given dataset. Techniques like grid search or random search can be employed for this purpose.

5. Result & Discussion

5.1. Results based on Datasets
5.2. Further Discussion

5. Result & Discussion

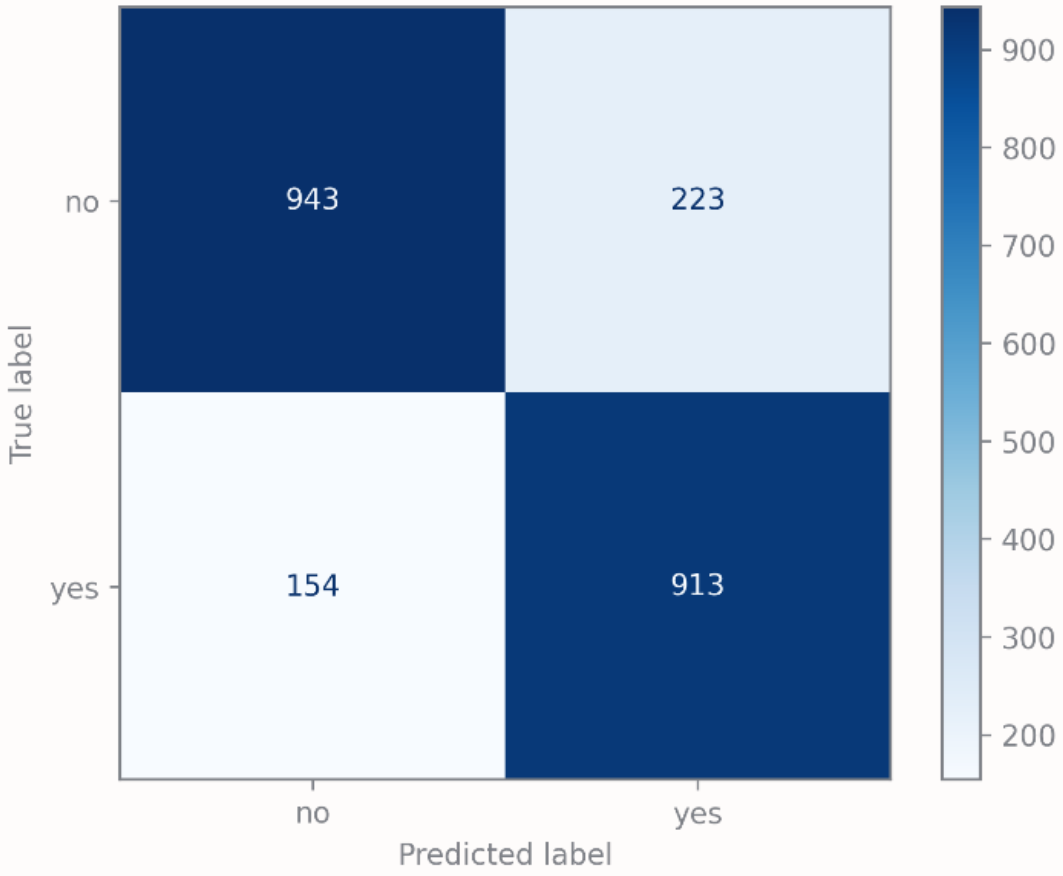
5.1. Results based on Datasets

Bank Marketing Dataset

a. Pair Plot

Best Performance found in **Random Forest Classifier.**

b. Confusion Matrix

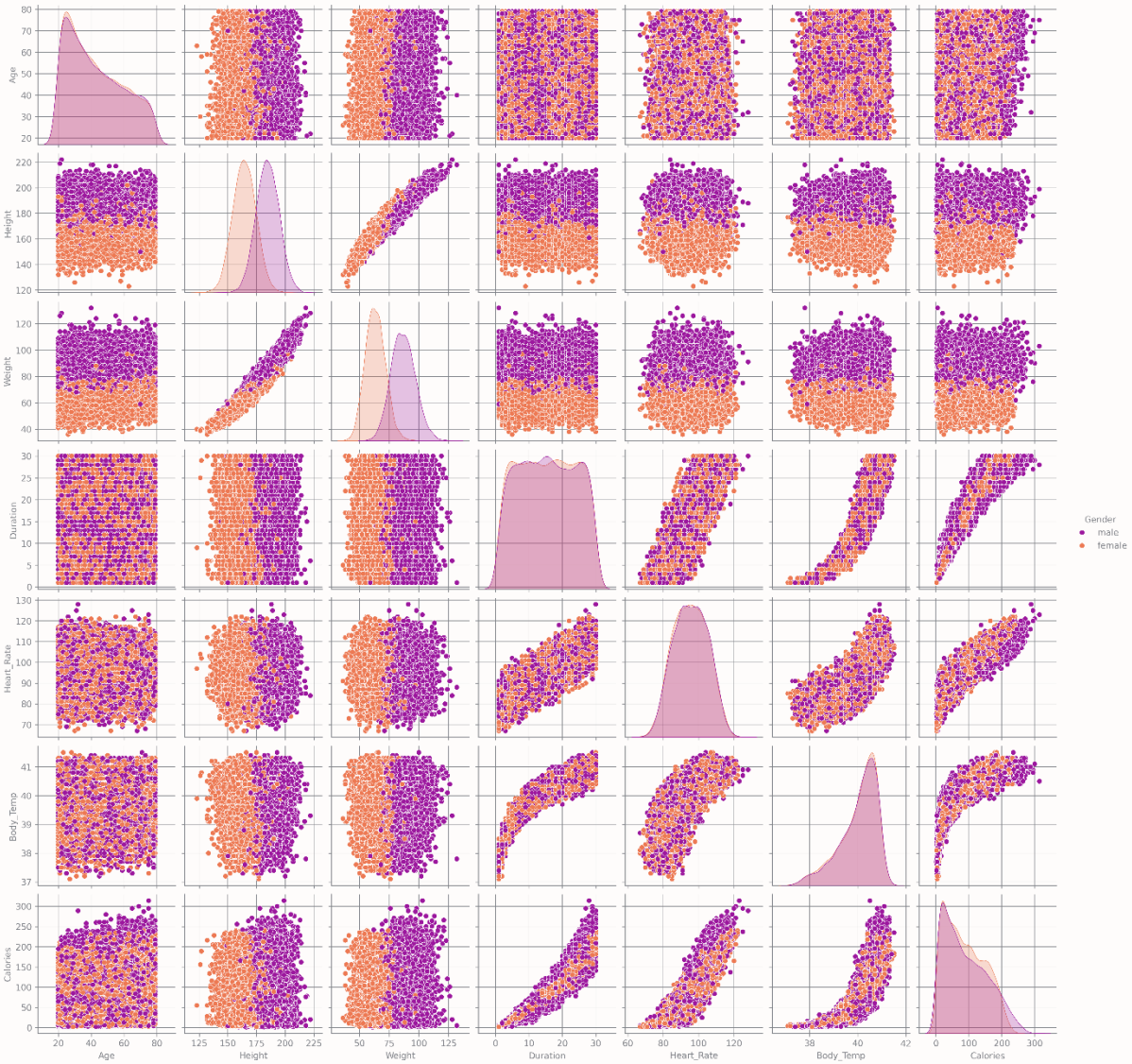


c. Metrics

Accuracy	Precision	Recall	F1 Score
83.12%	80.37%	85.57%	82.89%

Calories Burnt Prediction

a. Pair Plot

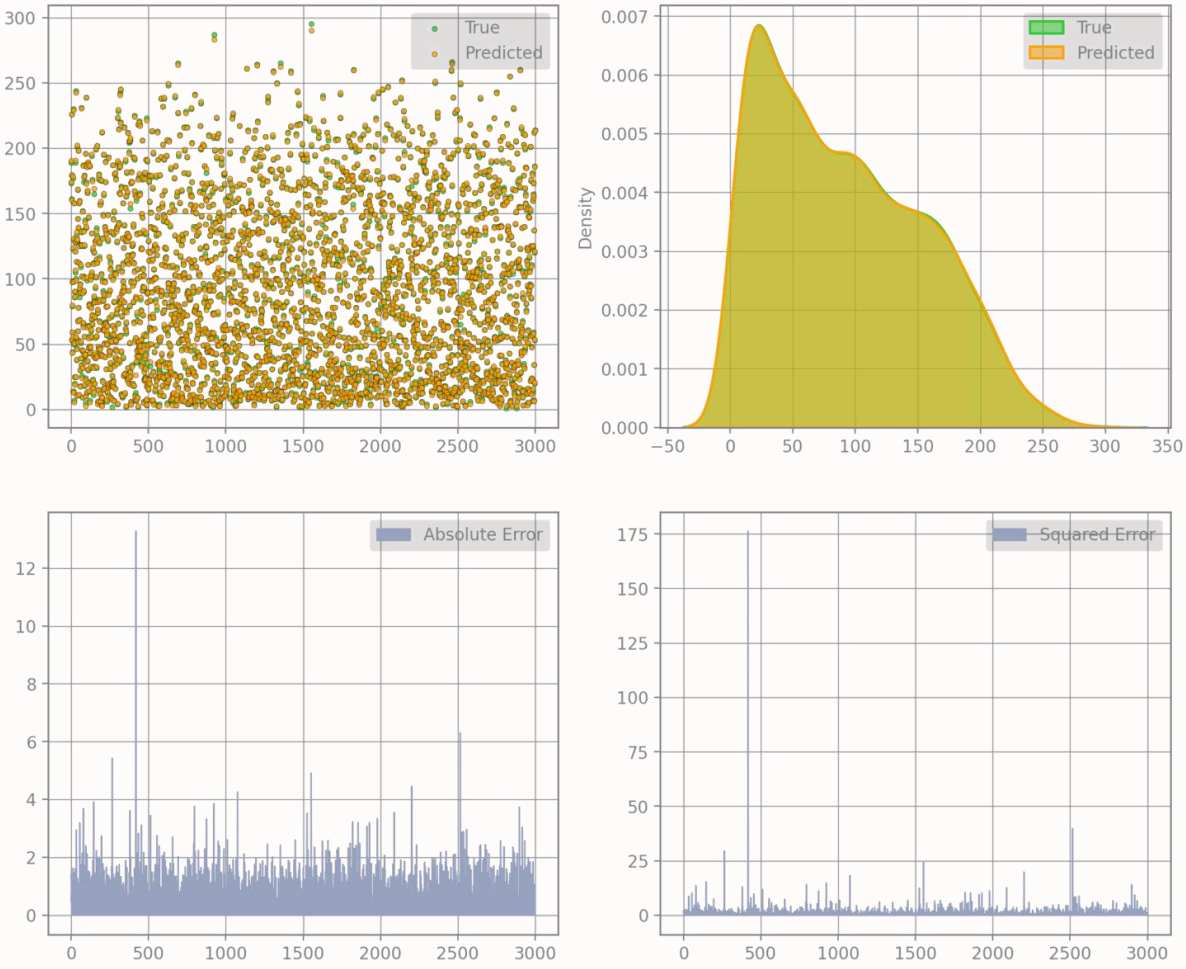


Best Performance was found in **Artificial Neural Network**

Parameters:

- **Hidden Layer 1**: {'units': 16, 'activation': 'relu'}
- **Hidden Layer 2**: {'units': 10, 'activation': 'relu'}
- **Hidden Layer 3**: {'units': 8, 'activation': 'relu'}
- **Output Layer**: {'units': 1}
- **Compile ANN**: {'optimizer': 'adam', 'loss': 'mse', 'metrics': ['mse', 'r2_score']}
- **Fit ANN**: {'x': X_train, 'y': y_train, 'epochs': 50, 'batch_size': 32}

b. Performance



c. Metrics

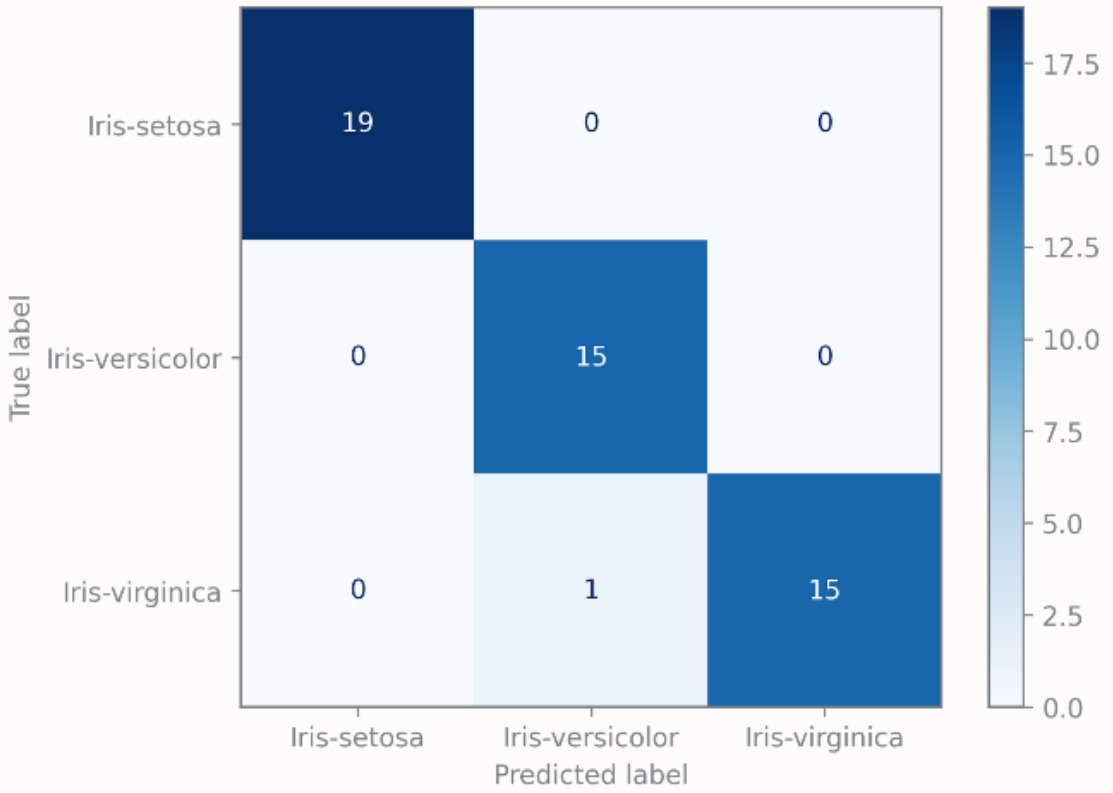
Explained variance Score	MAE	MSE	R2 Score
0.9998	0.7183	0.9388	0.9998

Iris Flower Classification

a. Pair Plot

Best Performance found in **Random Forest Classifier**

b. Confusion Matrix

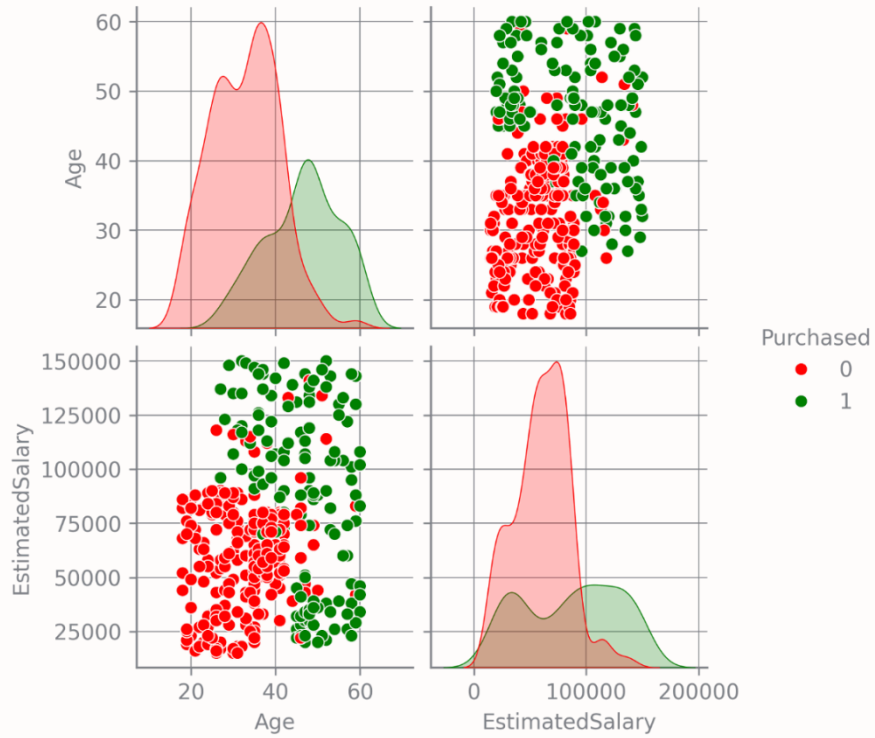


c. Metrics

Accuracy	Precision	Recall	F1 Score
98.00%	97.92%	97.92%	97.85%

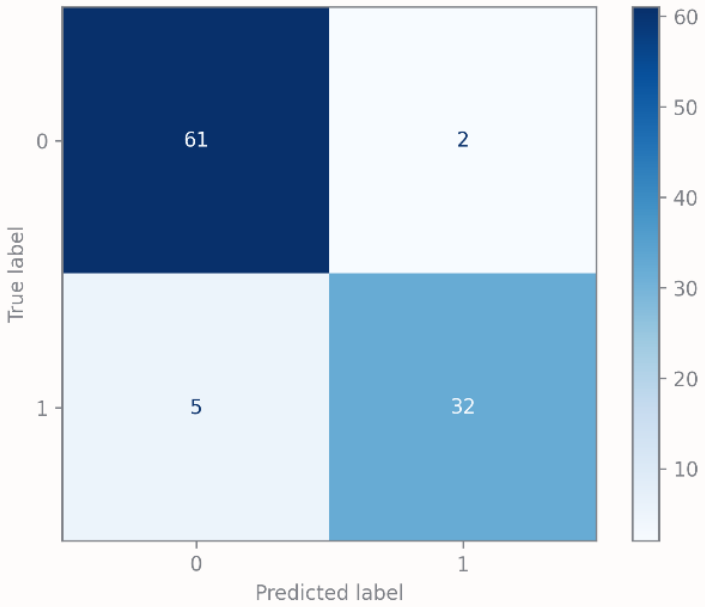
Social Network Ads

a. Pair Plot



Best performance found in **Gauss Naïve Bayes Classifier.**

b. Confusion Matrix



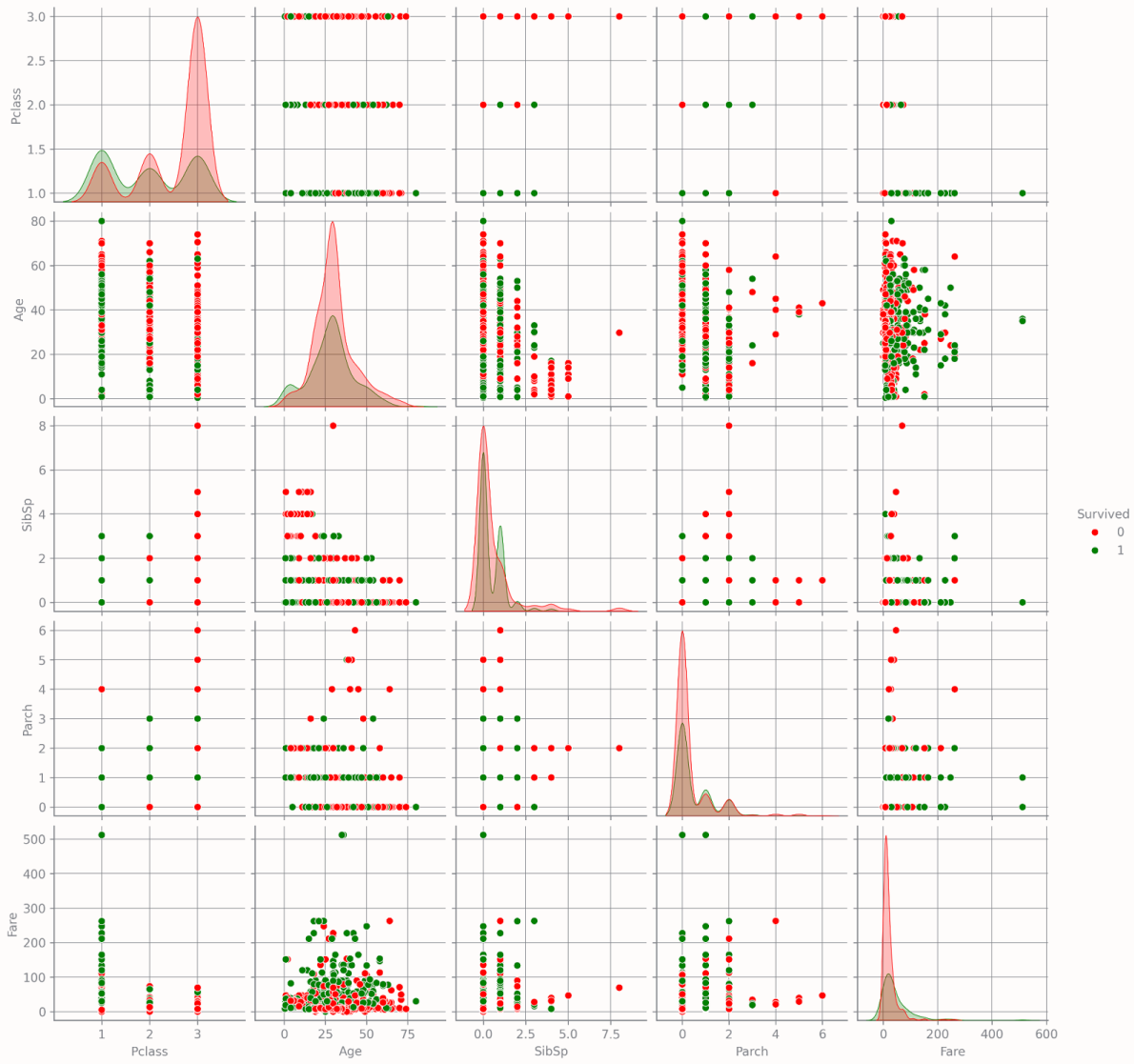
c.

d. Metrics

Accuracy	Precision	Recall	F1 Score
93.00%	94.12%	86.49%	90.14%

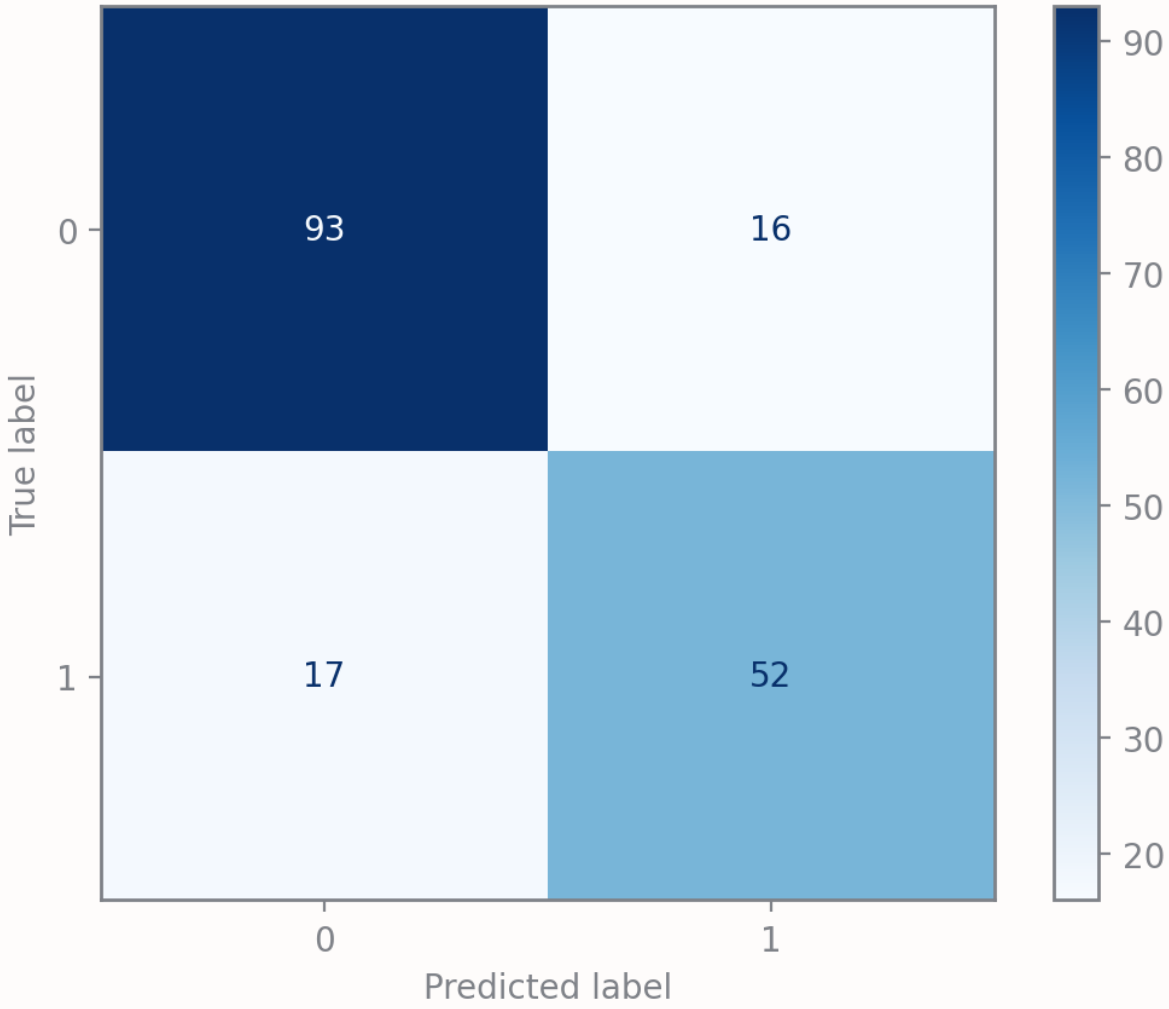
Titanic Survival Prediction

a. Pair Plot



Best Performance found in **K-Neighbors Classifier.**

b. Confusion Matrix

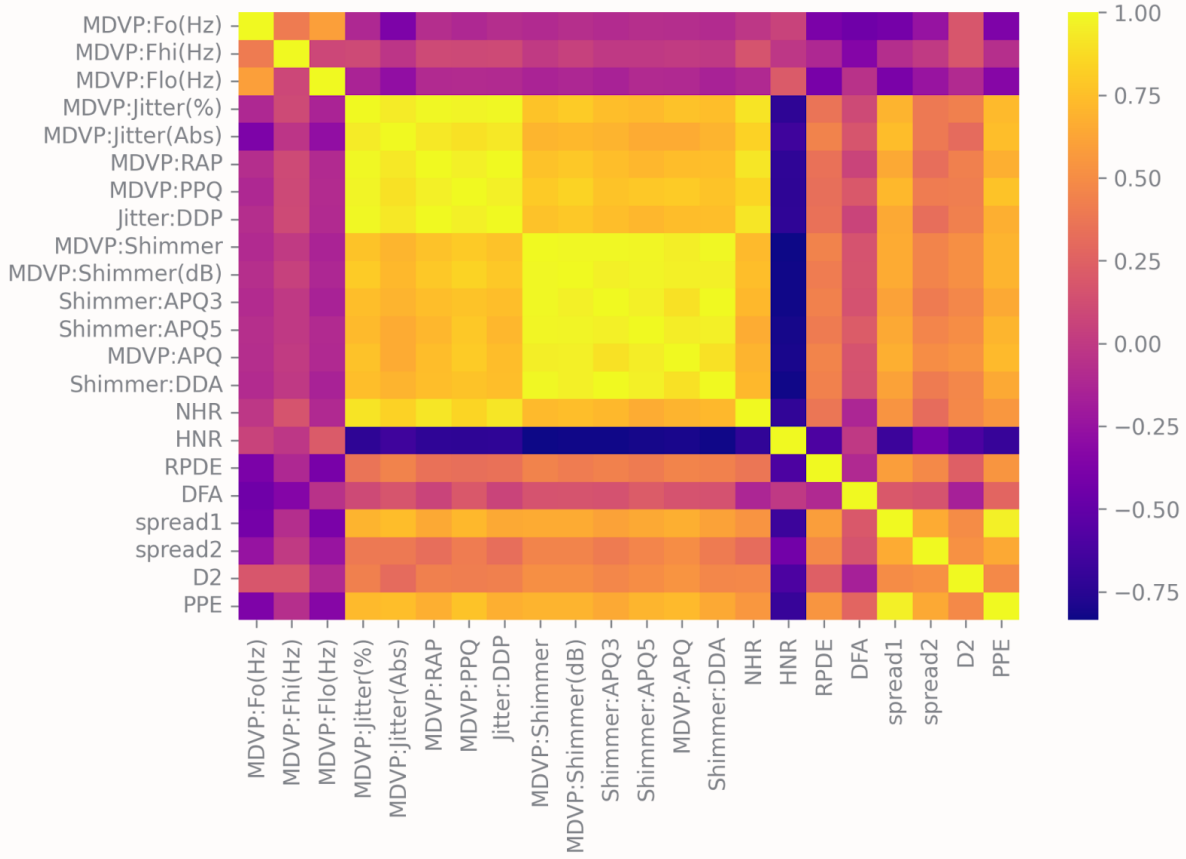


c. Metrics

Accuracy	Precision	Recall	F1 Score
81.46%	76.47%	75.36%	75.91&

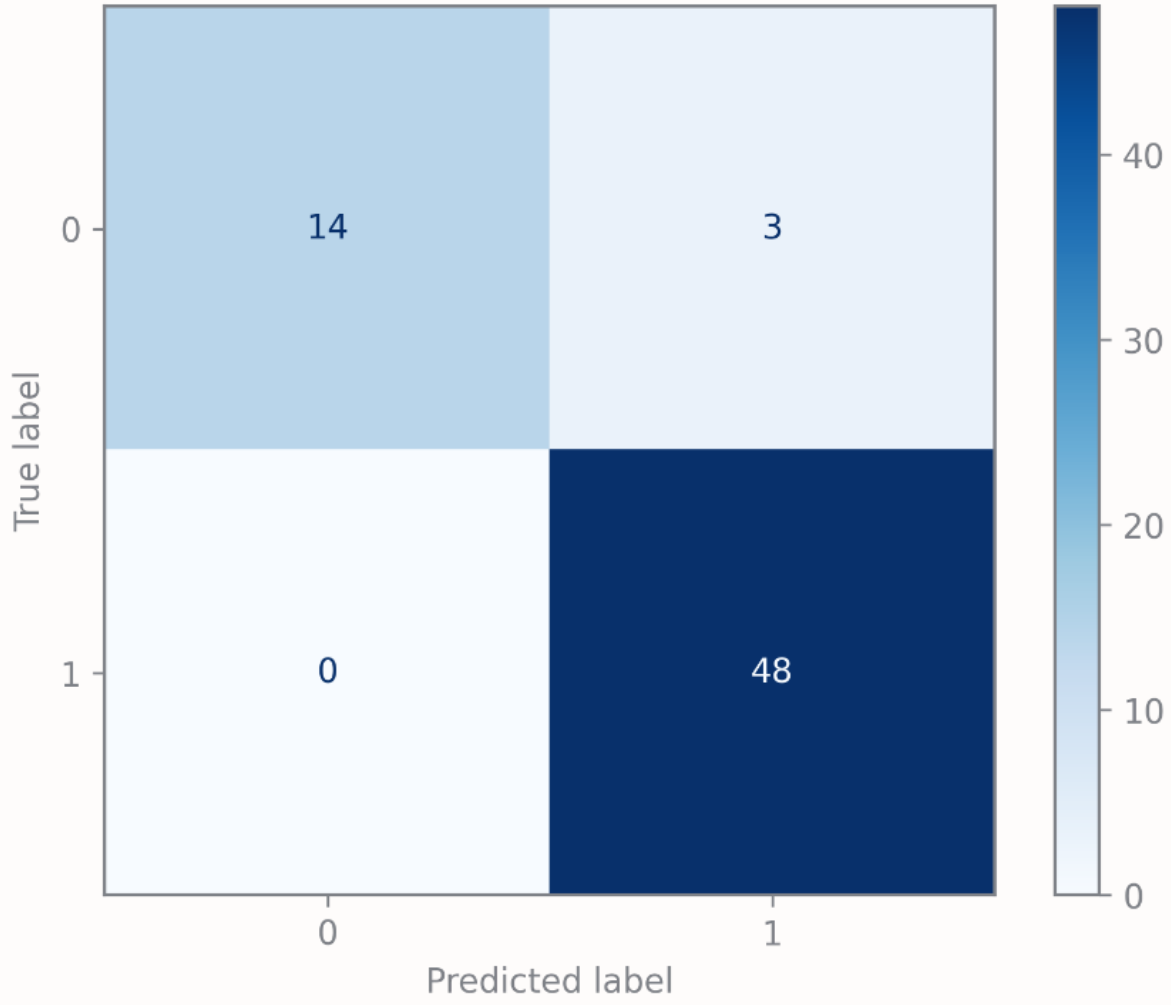
Parkinson's Disease Prediction

a. Correlation Heatmap



Best Performance found in **K-Neighbors Classifier.**

b. Confusion Matrix



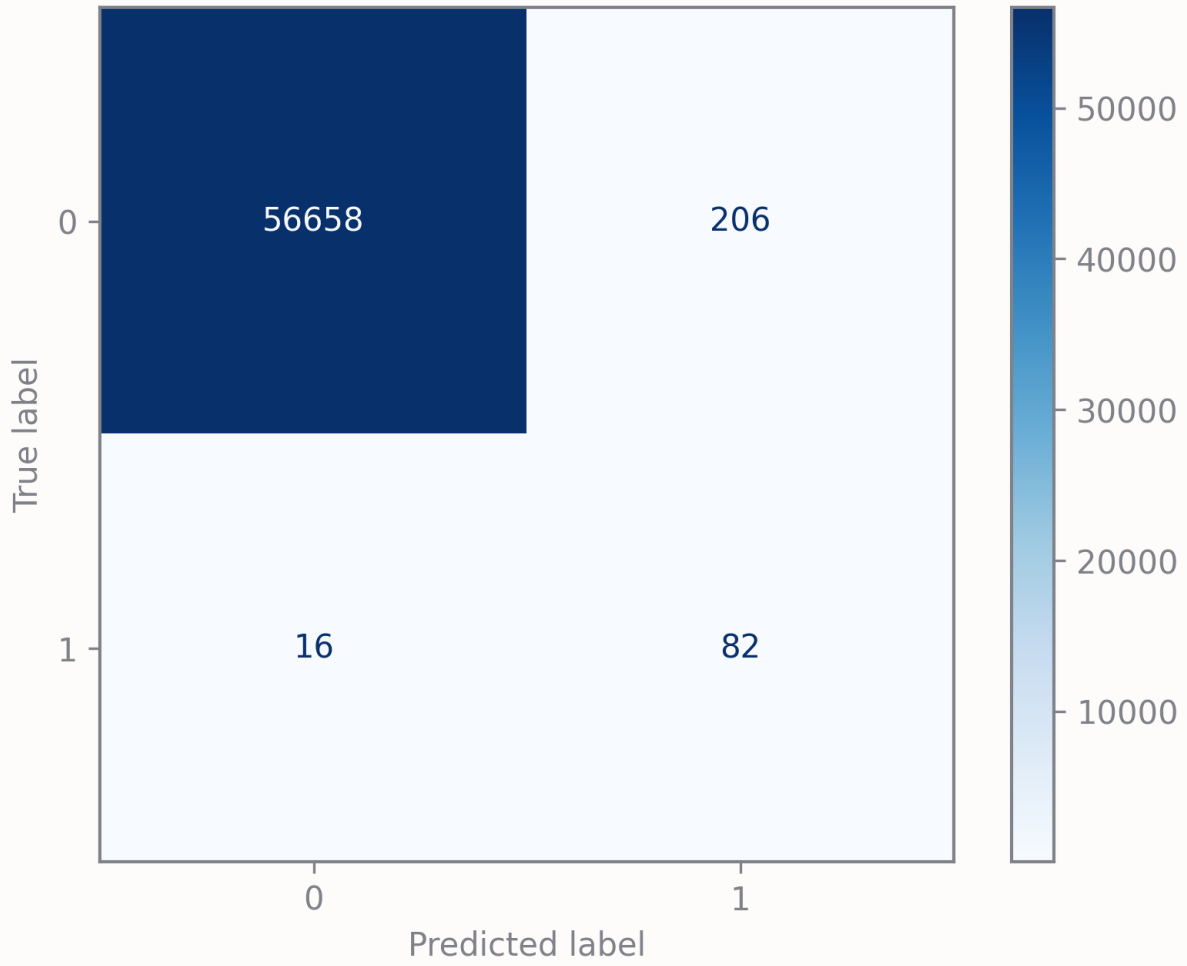
c. Metrics

Accuracy	Precision	Recall	F1 Score
95.38%	94.12%	100%	96.97%

Credit Card Fraud Prediction

Best Performance found in **Gauss Naïve-Bayes Classifier.**

a. Confusion Matrix

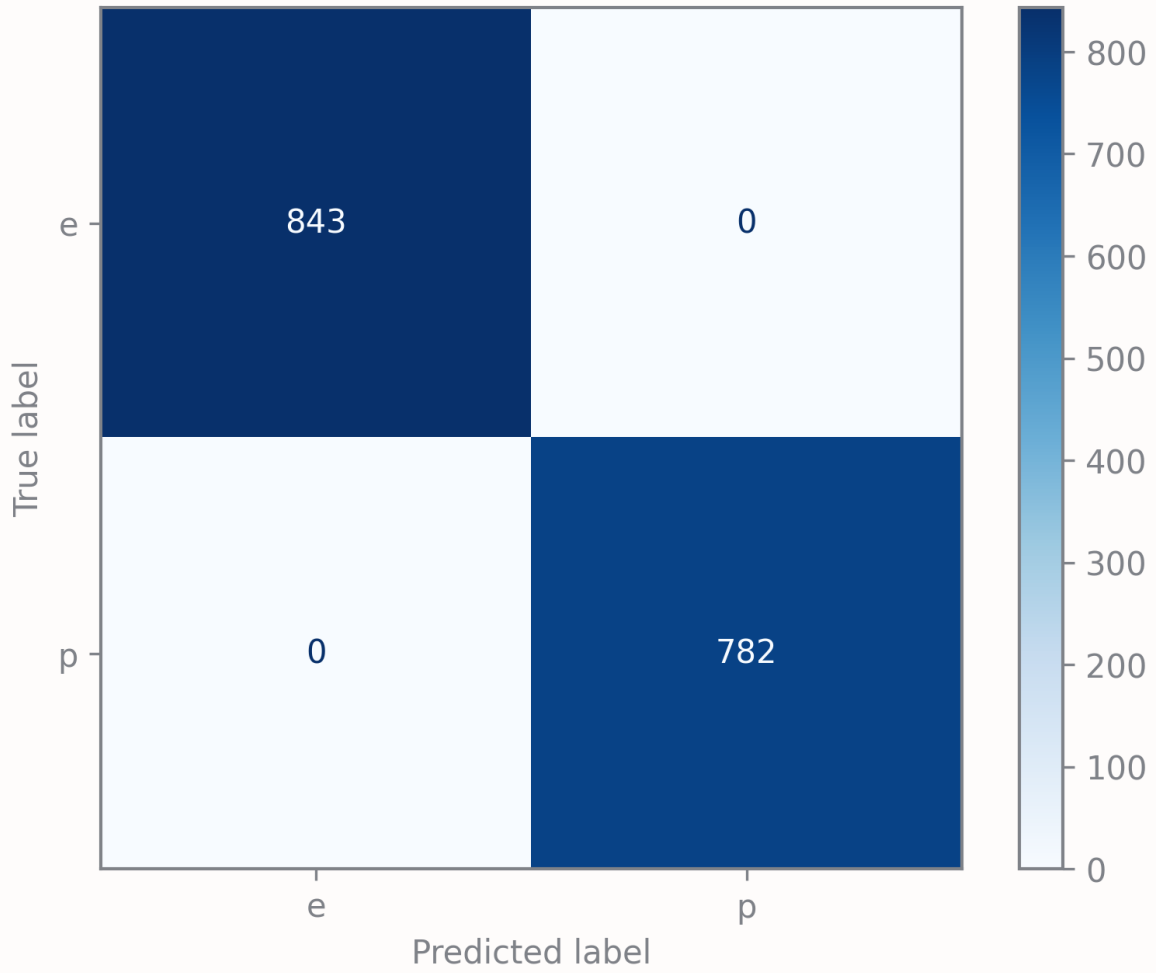


b. Metrics

Accuracy	Precision	Recall	F1 Score
99.61%	28.47%	83.67%	42.49%

Mushroom Analysis

a. Confusion Matrix

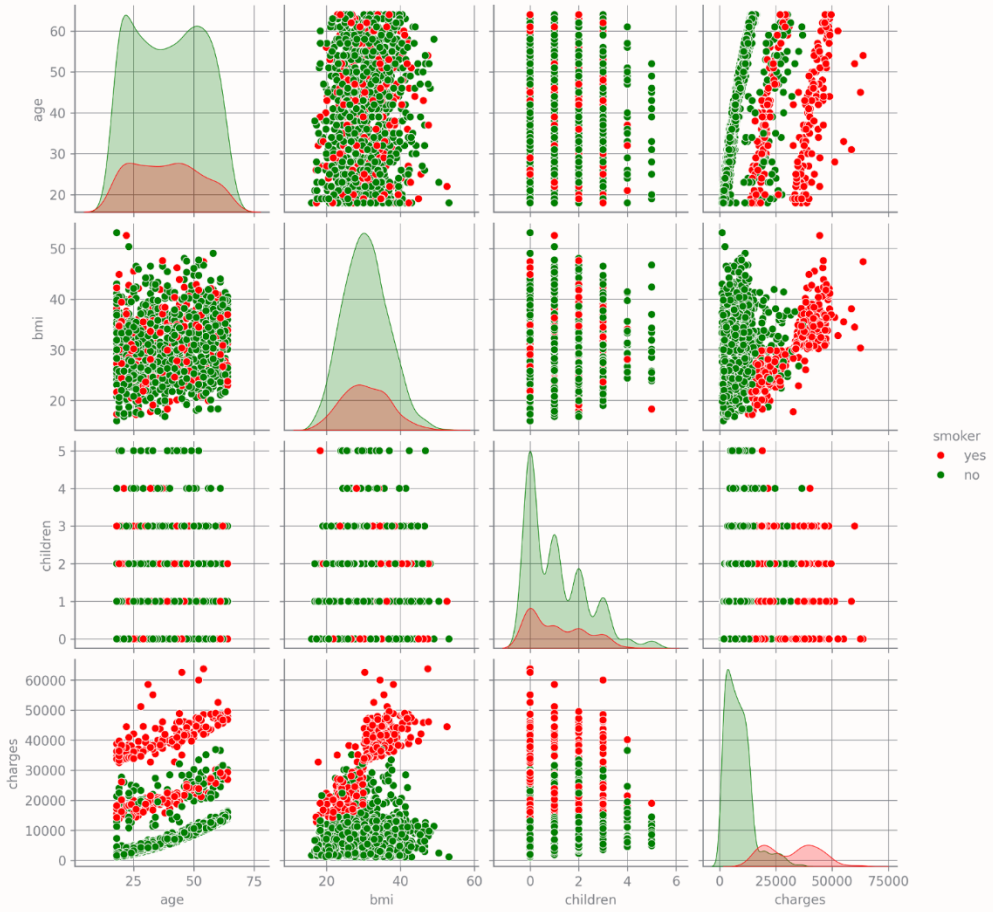


b. Metrics

Accuracy	Precision	Recall	F1 Score
100%	100%	100%	100%

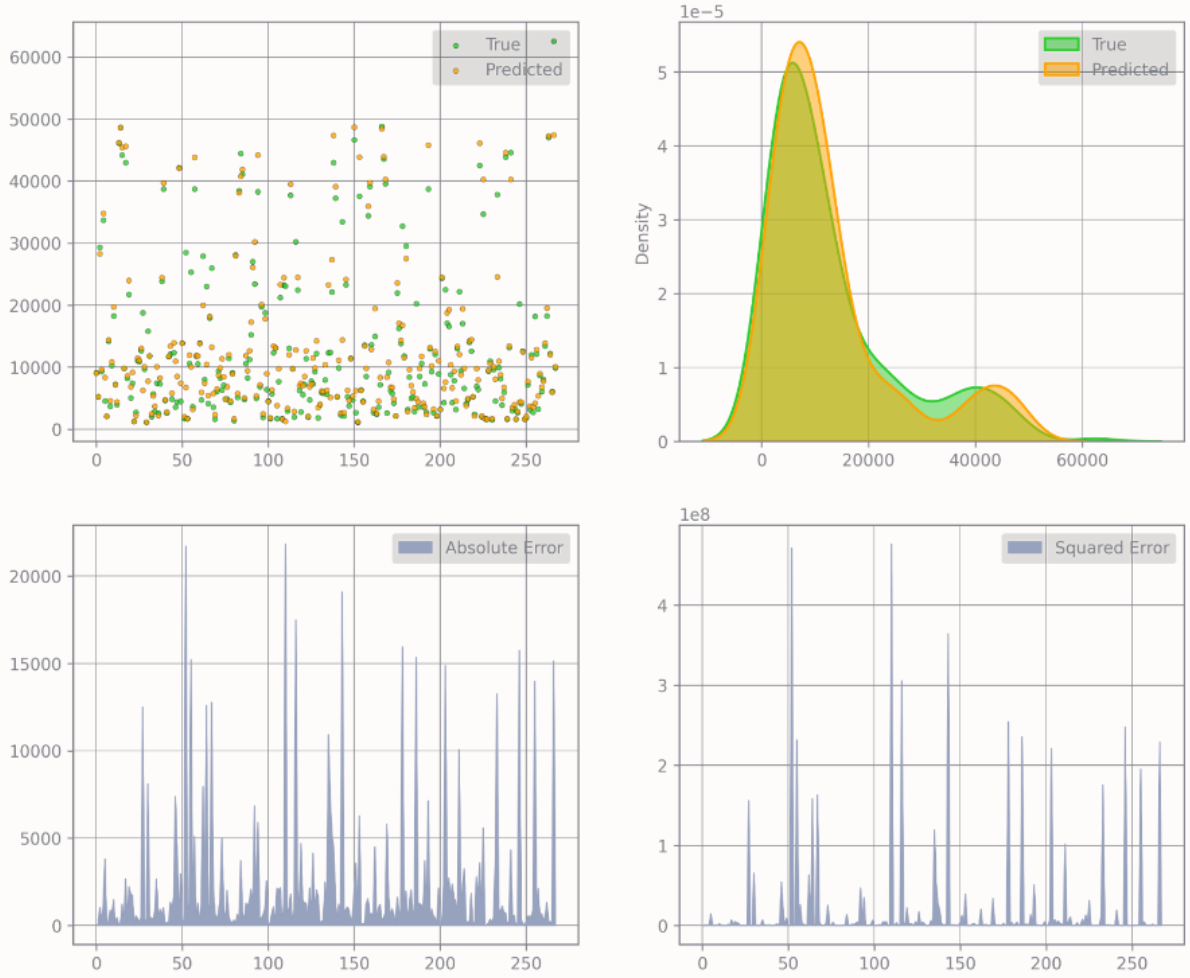
Medical Insurance Cost Prediction

a. Pair Plot



Best Performance found in **Random Forest Regressor**

b. Performance



c. Metrics

Explained variance Score	MAE	MSE	R2 Score
0.8767	2126.7073	19097157.3803	0.8766

5.2. Discussion and Implications

Discussion

Our project successfully demonstrated the effectiveness of the AI solution in extracting knowledge and generating insights from structured datasets. The chosen machine learning techniques proved adept at identifying patterns and relationships within the data, leading to valuable discoveries that might have been missed through traditional methods.

The ability of the solution to handle large and complex datasets efficiently is a significant advantage. This scalability allows organizations to analyze vast amounts of data previously considered too cumbersome or time-consuming. This opens doors to a more comprehensive understanding of their operations and the identification of previously hidden trends.

Implications

The successful development of this project has significant implications for various fields that rely on data analysis. Businesses can leverage the AI solution to gain deeper customer insights, optimize marketing strategies, and improve operational efficiency. Researchers can utilize the solution to analyze large datasets, identify new research avenues, and accelerate scientific discovery.

The ability to automate knowledge extraction from data can revolutionize decision-making processes. By providing clear and actionable insights, the AI solution empowers data-driven decision making, leading to better-informed strategies and improved outcomes across various domains.

However, it's important to consider the ethical implications of AI-powered data analysis. Bias within the data or the algorithms can lead to skewed results. Therefore, ensuring fairness and transparency in the development and deployment of the AI solution is crucial.

Overall, this project opens doors for a future where AI plays a pivotal role in data analysis and knowledge extraction. The developed solution can empower various stakeholders with valuable insights, leading to improved decision-making, innovation, and progress across various fields.

6. Conclusion

6.1. Summary of Key Findings
6.2. Future Directions
6.3. Overall Significance

6. Conclusion

6.1. Summary of Key Findings

This project successfully developed an AI-based solution capable of effectively representing knowledge and generating valuable insights from structured datasets. The key findings demonstrate the solution's potential to revolutionize data analysis and decision-making processes.

- **Effective Knowledge Representation**: The solution leverages data visualization techniques to present knowledge in a clear and understandable manner. Tools like Matplotlib and Seaborn facilitated the creation of charts, graphs, and other visual aids that effectively communicate the insights extracted from the data.

- **Pattern Identification Capabilities**: Machine learning algorithms, primarily from scikit-learn, were successfully employed to identify patterns within the data. These patterns, including trends and anomalies, provide valuable clues for understanding the underlying relationships within the data.

- **Actionable Insights Generation**: By leveraging the identified patterns, the solution generates meaningful insights that can inform data-driven decisions.

- **Scalability and User-friendliness**: The solution prioritizes scalability, ensuring it can handle datasets of varying sizes and complexities. Additionally, the user interface design, potentially built with frameworks like Streamlit, focuses on simplicity and ease of use, allowing users of all technical backgrounds to interact with the solution and gain insights from their data.

Overall, this project demonstrates the ability of AI to unlock the true potential of structured data. By effectively representing knowledge, identifying patterns, and generating actionable insights, the AI solution empowers users to make informed decisions based on a deeper understanding of their data.

6.2. Future Directions

This project has explored the development of an AI-based solution for knowledge representation and insight generation from structured datasets. While the project demonstrates promising capabilities, there are several exciting avenues for further exploration and development:

- **Incorporation of Unstructured Data**: Currently, the solution focuses on structured data. Expanding its capabilities to handle unstructured data sources like text documents, images, or sensor data could significantly broaden its applicability. Techniques like Natural Language Processing (NLP) and computer vision could be explored for this purpose.

- **Advanced Machine Learning Techniques**: Investigating the integration of more advanced machine learning algorithms could enhance the solution's capabilities. This could involve exploring deep learning architectures like Recurrent Neural Networks (RNNs) or transformers for tasks like time series analysis or natural language processing.

- **Explainable AI (XAI)**: While the solution generates insights, further development towards Explainable AI (XAI) techniques could be beneficial. This would allow users to understand the reasoning behind the generated insights, fostering trust and transparency in the AI solution.

- **Scalability and Cloud Deployment**: Currently, the solution might handle datasets of a certain size. Exploring cloud-based deployment and distributed computing frameworks like Apache Spark could enable the solution to handle even larger and more complex datasets efficiently.

- **Domain-Specific Customization**: Tailoring the solution to specific domains could unlock further potential. This could involve incorporating domain-specific knowledge and algorithms to enhance the accuracy and relevance of the generated insights for various industries or applications.

By pursuing these future directions, the AI solution can evolve to become even more powerful and versatile. It has the potential to revolutionize data analysis across various domains by providing deeper insights and facilitating data-driven decision-making processes.

6.3. Overall Significance

This project has demonstrated the significant potential of Artificial Intelligence (AI) in unlocking valuable knowledge and insights from structured datasets. The developed AI solution offers several key advantages over traditional data analysis methods:

- **Automated Knowledge Extraction**: By leveraging machine learning algorithms, the solution automates the process of identifying patterns and trends within data, saving valuable time and resources compared to manual analysis.

- **Enhanced Scalability and Accuracy**: The solution can handle large and complex datasets with greater efficiency and accuracy than traditional methods. This allows organizations to analyze vast amounts of data and gain insights that might have been previously missed.

- **Deeper Insights**: AI can uncover hidden relationships and patterns within data that might be overlooked by human analysts. This leads to a more comprehensive understanding of the data and the ability to identify critical factors influencing outcomes.

- **Improved Decision-Making**: By providing clear and actionable insights, the AI solution empowers data-driven decision-making processes across various domains. This can lead to better-informed strategies, optimized operations, and improved results.

The successful implementation of this project paves the way for further advancements in AI-powered data analysis. Here are some potential areas of future development:

- **Integration with Domain Knowledge**: The solution can be further enhanced by incorporating domain-specific knowledge into the AI models. This would allow for more nuanced insights and tailored solutions for specific industries or applications.

- **Explainable AI (XAI)**: Implementing XAI techniques in the future can improve the transparency and interpretability of the AI models. This would allow users to better understand the rationale behind the generated insights and build trust in the AI solution.

- **Real-Time Analysis**: The future development of the solution could enable real-time analysis of streaming data. This would allow for continuous monitoring and proactive decision-making based on the latest information.

In conclusion, this project has made a significant contribution to the field of AI-powered data analysis. The developed solution offers a powerful tool for extracting valuable knowledge and insights from structured datasets, ultimately leading to improved decision-making and innovation across various fields.