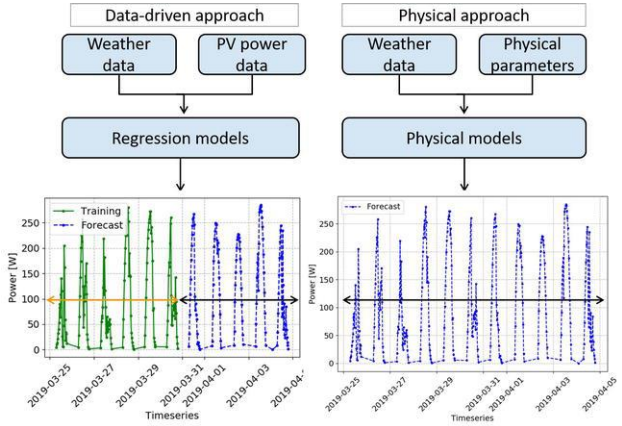
SOLAR POWER GENERATION PREDICTION MODEL



By : M JEEVAN
Data Scientist

INNODATATICS
Innovation • Data • Analytics

Project Leadership

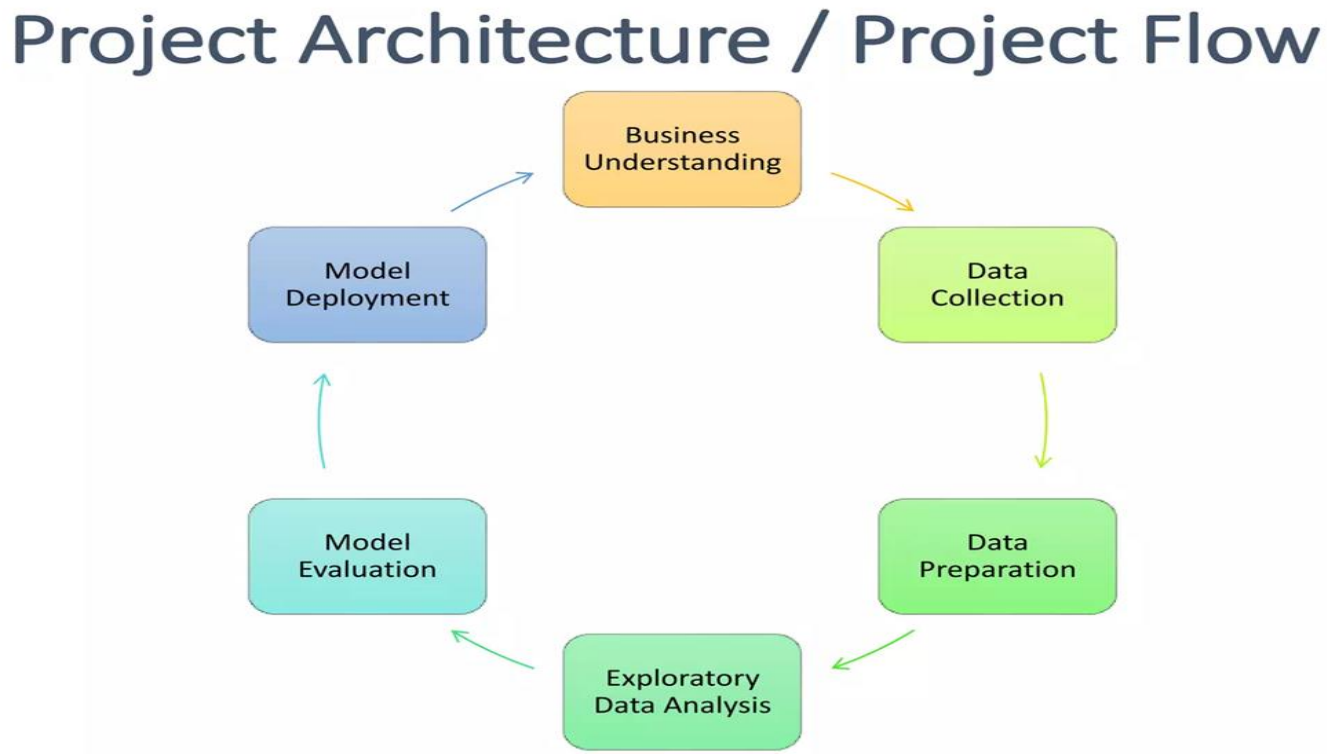
Sharat Manikonda
Director at Innodatatics and Sponsor
linkedin.com/in/sharat-chandra

Team Members

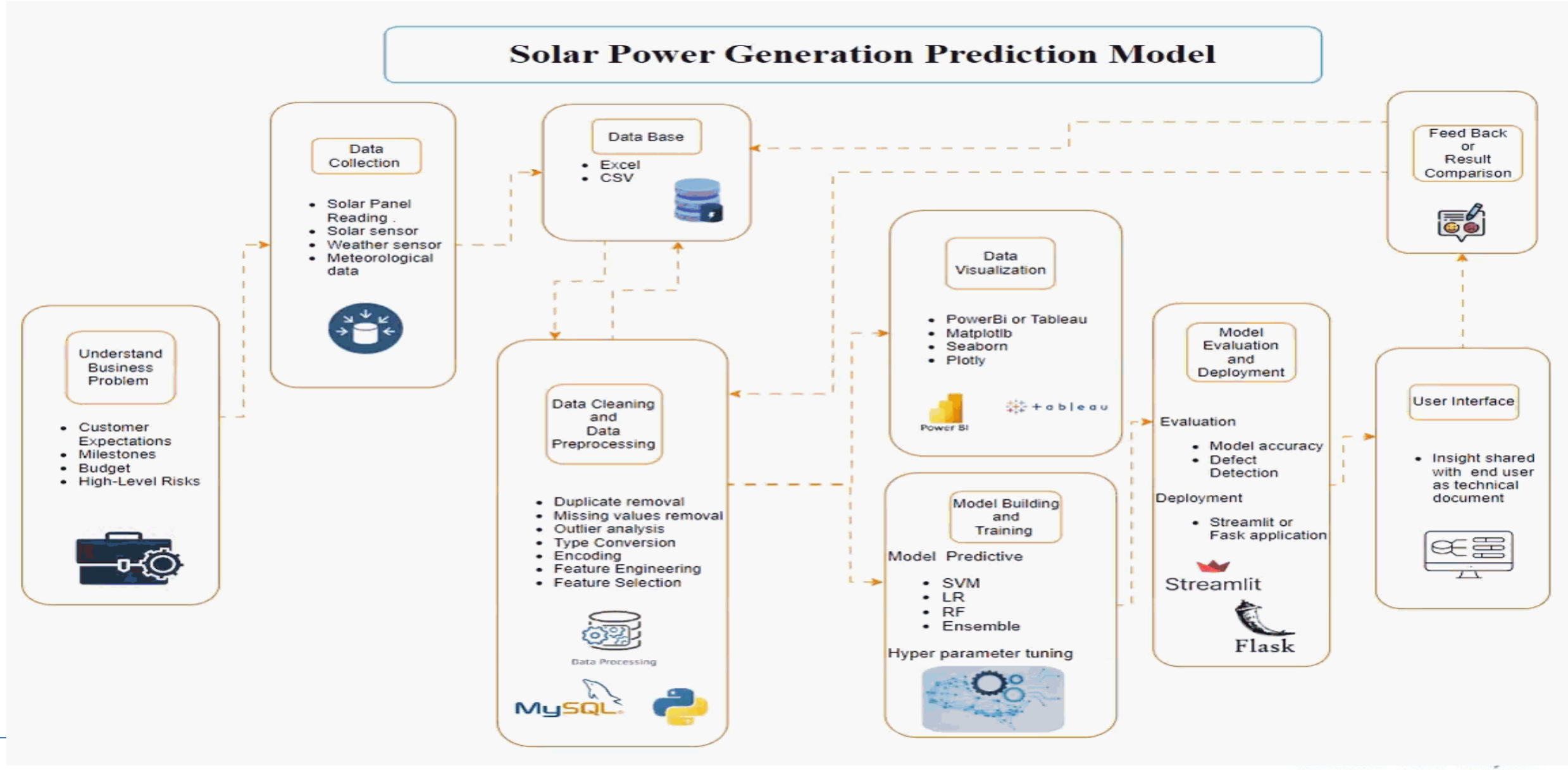
Name: M Jeevan
www.linkedin.com/in/m-jee8van78753

Contents

Project Overview and Scope



Project Architecture



Business Objective

Client:

One of the leading companies in solar power generation.

Business Problem:

Manual inspection limitations lead to undetected defects, decreasing energy production and systems.

Business Problem

Business Objective:

Maximise energy production and system reliability through efficient defect detection.

Business Constraint:

Minimize resource utilization.
Maximizing energy production and system reliability.

Success criteria:

Business success criteria:

Increase in energy production efficiently by at least 10% and improvement in system reliability with a 20% reduction in downtime.
Machine Learning success criteria: Achieve an accuracy of at least 95%.

Economic success criteria:

Reduction in maintenance costs by 15% and Increase in return on investment (ROI) by 20% through improved efficiency.

CRISP-ML(Q) Methodology

There are six stages of CRISP-ML(Q) Methodology

1.Business and data understanding

2.Data preparation

3.Model building

4.Model evaluation

5.Model deployment

6.Monitoring and maintenance

Technical Stacks

Programming Languages:
- Python

Data Manipulation and Analysis:
- Pandas
- NumPy

Data Visualization:
- matplotlib
- Seaborn

AutoEDA
- D-tale

Time Series Forecasting:
- Statsmodels

Optimization:
- SciPy

Machine Learning (for Continuous Improvement):
- scikit-learn

Database (for Data Storage and Retrieval):
- MySQL
- mysql.connector

Notebooks and Development Environment:
- Jupyter Notebook
- Visual Studio Code

Data Collection and Understanding

Data collection in this project involves gathering all the necessary information that will be used for analysis, modeling, and optimization.

Based on the secondary data source, the data types can be categorized as below:

CATEGORIZED	
Data	Data Type
Time: Time in seconds	Continuous, Ratio
Ipv: PV array current measurement	Continuous, Ratio
Vpv: PV array voltage measurement	Continuous, Ratio
Vdc: DC voltage measurement	Continuous, Ratio
ia, ib, ic: 3-Phase current measurements	Continuous, Ratio
va, vb, vc: 3-Phase voltage measurements	Continuous, Ratio
Iabc: Current magnitude	Continuous, Ratio
If: Current frequency	Continuous, Ratio
Vabc: Voltage magnitude	Continuous, Ratio
Vf: Voltage frequency	Continuous, Ratio
Defecitive/Non Defective	Categorical, Nominal

Data Information

1.Time (Continuous, Ratio):

- The time variable represents the timestamp of measurements.
- It's continuous as it can take any numerical value.
- Ratio scale as it has a true zero point (starting time).

2.Ipv (Continuous, Ratio):

- Represents the current generated by the photovoltaic (PV) array.
- Continuous and ratio scale as it represents real values with a true zero (no current).

3.Vpv (Continuous, Ratio):

- Represents the voltage generated by the PV array.
- Continuous and ratio scale, similar to Ipv.

4.Vdc (Continuous, Ratio):

- Represents the DC voltage measurement.
- Continuous and ratio scale as it represents real voltages with a true zero.

5.ia, ib, ic (Continuous, Ratio):

- Represent the three-phase current measurements for each phase.
- Continuous and ratio scale as they represent real current values with a true zero

Data Information

6.va, vb, vc (Continuous, Ratio):

- Represent the three-phase voltage measurements for each phase.
- Continuous and ratio scale, similar to the three-phase current measurements.

7.Iabc (Continuous, Ratio):

- Represents the current magnitude.
- Continuous and ratio scale as it represents real current values with a true zero.

8.If (Continuous, Ratio):

- Represents the current frequency.
- Continuous and ratio scale as it represents real frequencies with a true zero.

9.Vabc (Continuous, Ratio):

- Represents the voltage magnitude.
- Continuous and ratio scale as it represents real voltage values with a true zero.

10.Vf (Continuous, Ratio):

- Represents the voltage frequency.
- Continuous and ratio scale as it represents real frequencies with a true zero.

11.Defective/Non-Defective (Categorical, Nominal):

- Indicates whether the system is defective (1) or non-defective (0).
- Categorical and nominal scale as it represents categories with no inherent order.

Data Dictionary

DATA DICTIONARY -SOLAR POWER GENERATION PREDICATION MODEL [GPVS-FAULTS]						
SL NO	Field Name	Description	Data Type	Data Format	Data Size	Measuring Unit
1	Time	Time in seconds, average sampling T_s=9.9989 µs	Float	Floating-point	64-bit	Seconds(S)
2	Ipv	PV array current measurement	Float	Floating-point	64-bit	Amperes (A)
3	Vpv	PV array voltage measurement	Float	Floating-point	64-bit	Volts (V)
4	Vdc	DC voltage measurement	Float	Floating-point	64-bit	Volts (V)
5	ia, ib, ic	3-Phase voltage measurements (each phase)	Float	Floating-point	64-bit	Amperes (A)
6	va, vb, vc	3-Phase voltage measurements (each phase)	Float	Floating-point	64-bit	Volts (V)
7	Iabc	Current magnitude	Float	Floating-point	64-bit	Amperes (A)
8	If	Current frequency	Float	Floating-point	64-bit	Hertz (Hz)
9	Vabc	Voltage magnitude	Float	Floating-point	64-bit	Hertz (Hz)
10	Vf	Voltage frequency	Float	Floating-point	64-bit	Hertz (Hz)
11	Defecitive/Non Defective	Indicates whether the system is defective (1) or non-defective (0)	Integer	Integer (Binary)	8-bit	0 or 1

System Requirements

1. **Software Requirements:**

   - **Operating System:** Windows 10/11
   - **Python:** Python 3.11, packages (pandas, matplotlib, seaborn, scikit-learn, etc.)
   - **Database:** MySQL
   - **Data Visualization Tools:** Jupyter Notebook, D-tale,Autosweiz

2. **Collaboration Tools:**

   - Google Meet for communication and coordination within our project team.

3. **Backup and Recovery:**

   - Back up project data and code to prevent data loss using cloud storage.

Exploratory Data Analysis [EDA]-Business Insights

Statistical Insights									
SL.No	Column Name	Mean	Median	Mode	Standard Deviation	Range	Variance	Skewness	Kurtosis
1	**Time**	6.625781	6.501687	8.491664	3.932539	14.36905	15.46486	0.1148812	-1.11115
2	Ipv	1.757652	1.486237	1.544891	0.434874	2.343353	0.189115	0.7586745	-1.08309
3	**Vpv**	91.97654	101.1536	101.3	23.87105	105.188	569.8273	-3.278917	9.398946
4	Vdc	137.0768	143.8477	142.9688	37.59227	237.3047	1413.179	-2.680345	8.596569
5	**ia**	-0.02288	0.039428	0.381835	0.747003	17.63062	0.558014	-0.109045	36.66578
6	ib	0.025592	0.38269	-0.49011	0.770516	13.703	0.593694	0.0882433	35.45232
7	**ic**	-0.04881	-0.42884	0.094844	0.741227	19.04053	0.549417	0.0560101	40.94193
8	va	0.663288	-15.9239	-110.467	109.9207	319.3584	12082.56	0.0010671	-1.50279
9	**vb**	1.025666	-121.75	150.3912	109.9543	318.8641	12089.95	-0.004071	-1.50223
10	vc	0.747015	145.7061	-39.6792	109.4936	316.9595	11988.86	-0.000272	-1.4935
11	**Iabc**	0.618741	0.484271	0.465142	0.847299	6.460822	0.717915	6.1448486	37.7594
12	If	49.34148	50.07636	50.18584	4.97106	1.845093	24.71144	-8.327989	71.35928
13	**Vabc**	154.8871	155.8074	154.7514	6.953687	155.2683	48.35376	-21.57975	471.2706
14	Vf	49.99976	49.99888	49.99516	0.023231	0.969594	0.00054	-2.602877	101.6279

Exploratory Data Analysis [EDA]-Business Understanding

1.Time (Time in seconds): Average sampling time is 6.63s, with a standard deviation of 3.93s, indicating variability.

2.Ipv (PV array current): Average current is 1.76, with low variability (std dev = 0.43).

3.Vpv (PV array voltage): Average voltage is 91.98, with high variability (std dev = 23.87).

4.Vdc (DC voltage): Average voltage is 137.08, with considerable variability (std dev = 37.59).

5.ia, ib, ic (3-Phase currents): Represent three-phase current measurements, essential for system balance.

6.va, vb, vc (3-Phase voltages): Three-phase voltage measurements, crucial for assessing power supply balance.

7.Iabc (Current magnitude): Magnitude of current, important for identifying abnormalities or overloads.

8.If (Current frequency): Average frequency is 49.34, with deviations indicating potential issues.

9.Vabc (Voltage magnitude): Magnitude of voltage, crucial for identifying abnormalities.

10.Vf (Voltage frequency): Average frequency is 49.99, with low variability (std dev = 0.02), essential for stable power supply.

These insights are vital for optimizing renewable energy systems, power distribution, and ensuring reliable industrial processes.

Monitoring these parameters aids in performance optimization and issue identification.

Missing Values Observation

```
## Check Duplicates
data.duplicated().sum()

duplicates = data[data.duplicated()]
duplicates
```

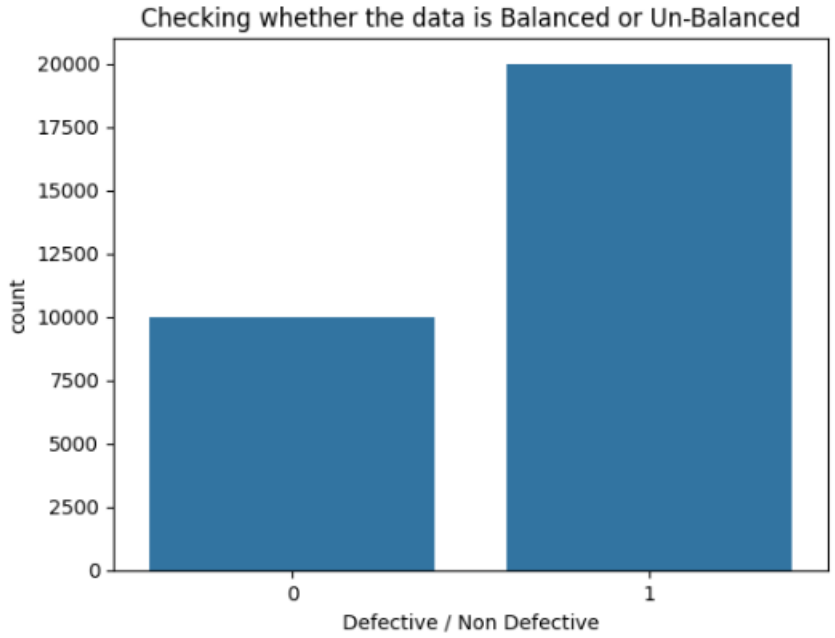
Time	Ipv	Vpv	Vdc	ia	ib	ic	va	vb	vc	Iabc	_If	Vabc	Vf	Defective / Non Defective

```
### Data checking

# Check for Missing values
data['Defective / Non Defective'].isnull().sum()
```

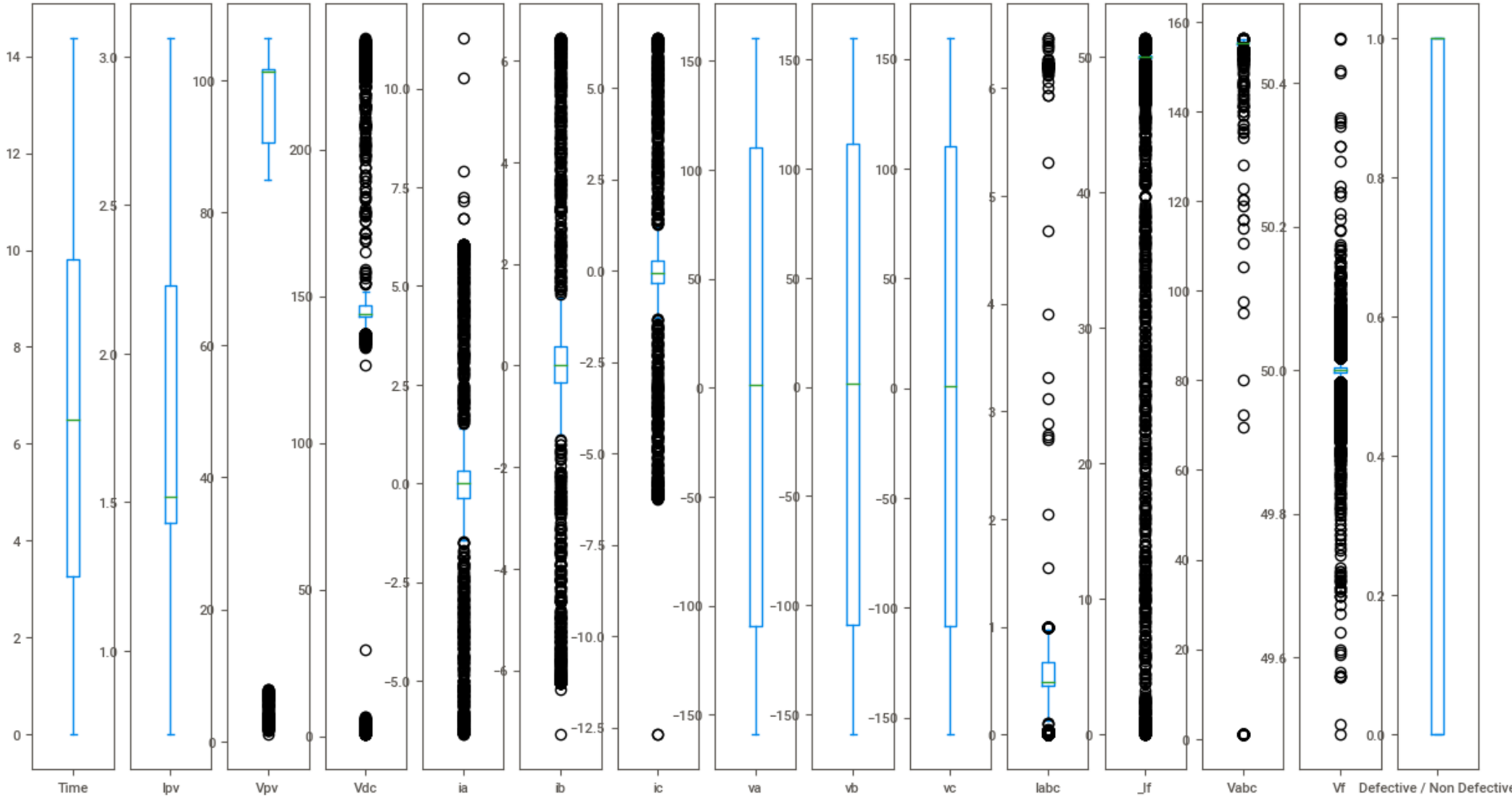
```
0
```

```
# Create the count plot
sns.countplot(x = x, data = data)

# tile for the plot
plt.title('Checking whether the data is Balanced or Un-Balanced')
```

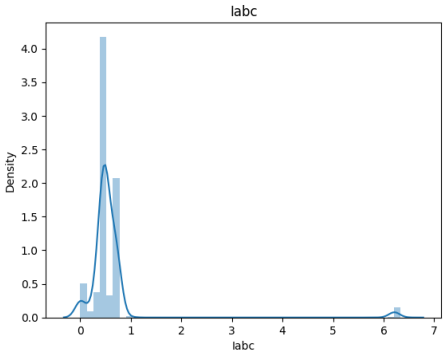
```
Text(0.5, 1.0, 'Checking whether the data is Balanced or Un-Balanced')
```

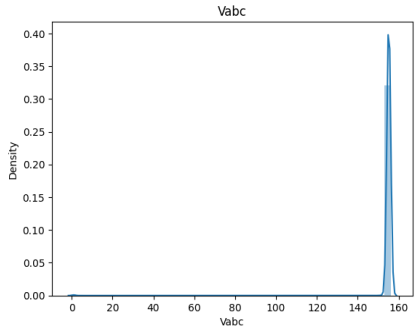


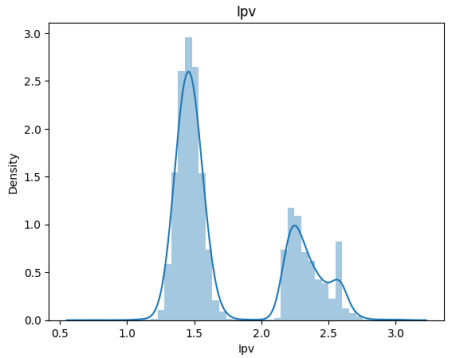
Data Preprocessing-Box plot _Outlier Analysis

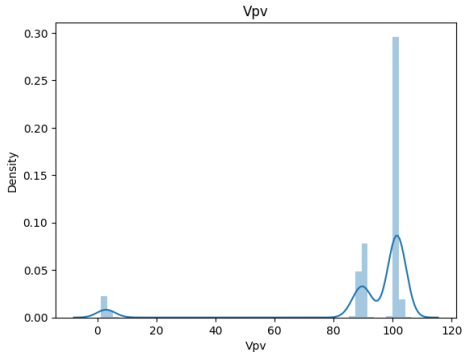


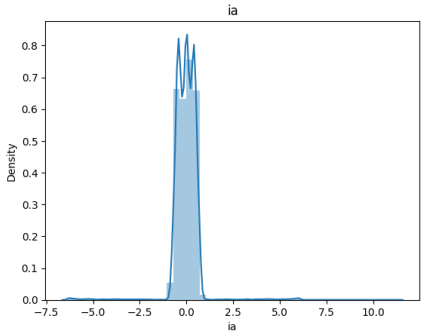
Data Preprocessing-Density Plot

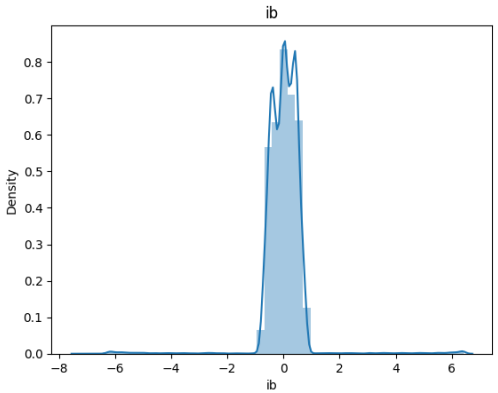


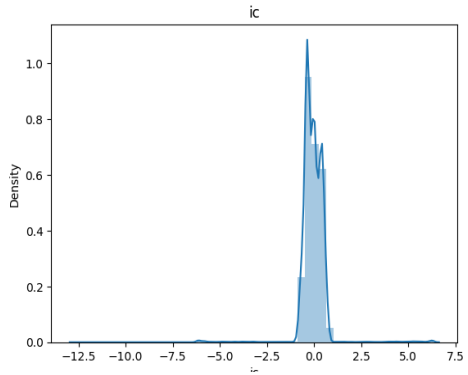


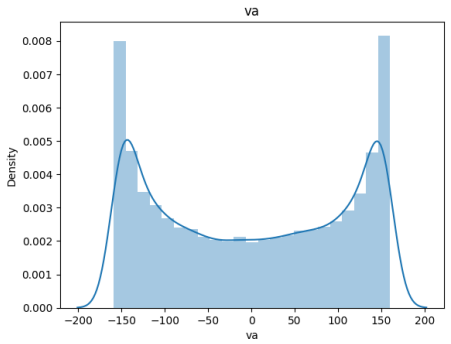


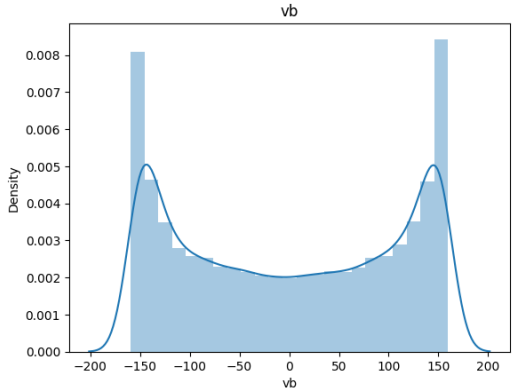


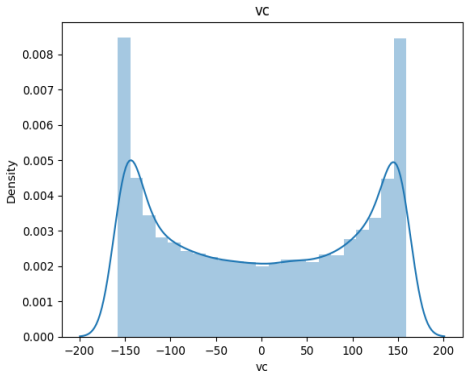


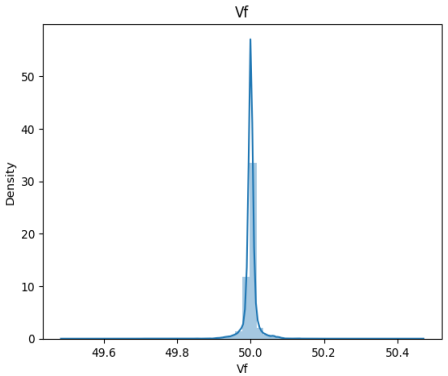


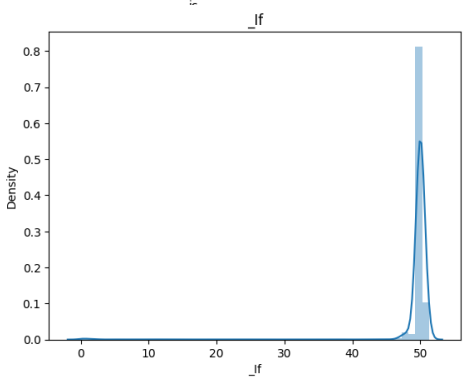






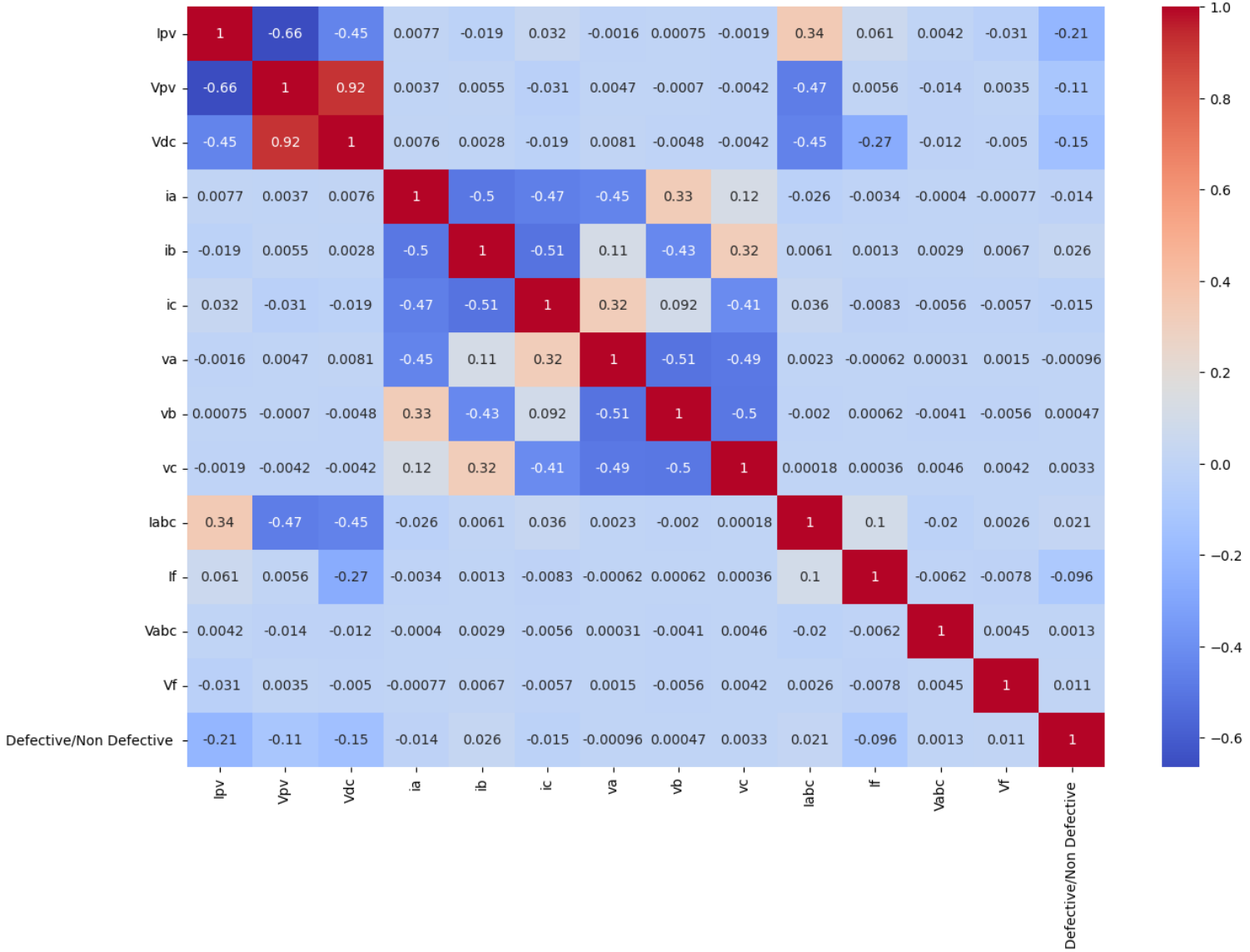






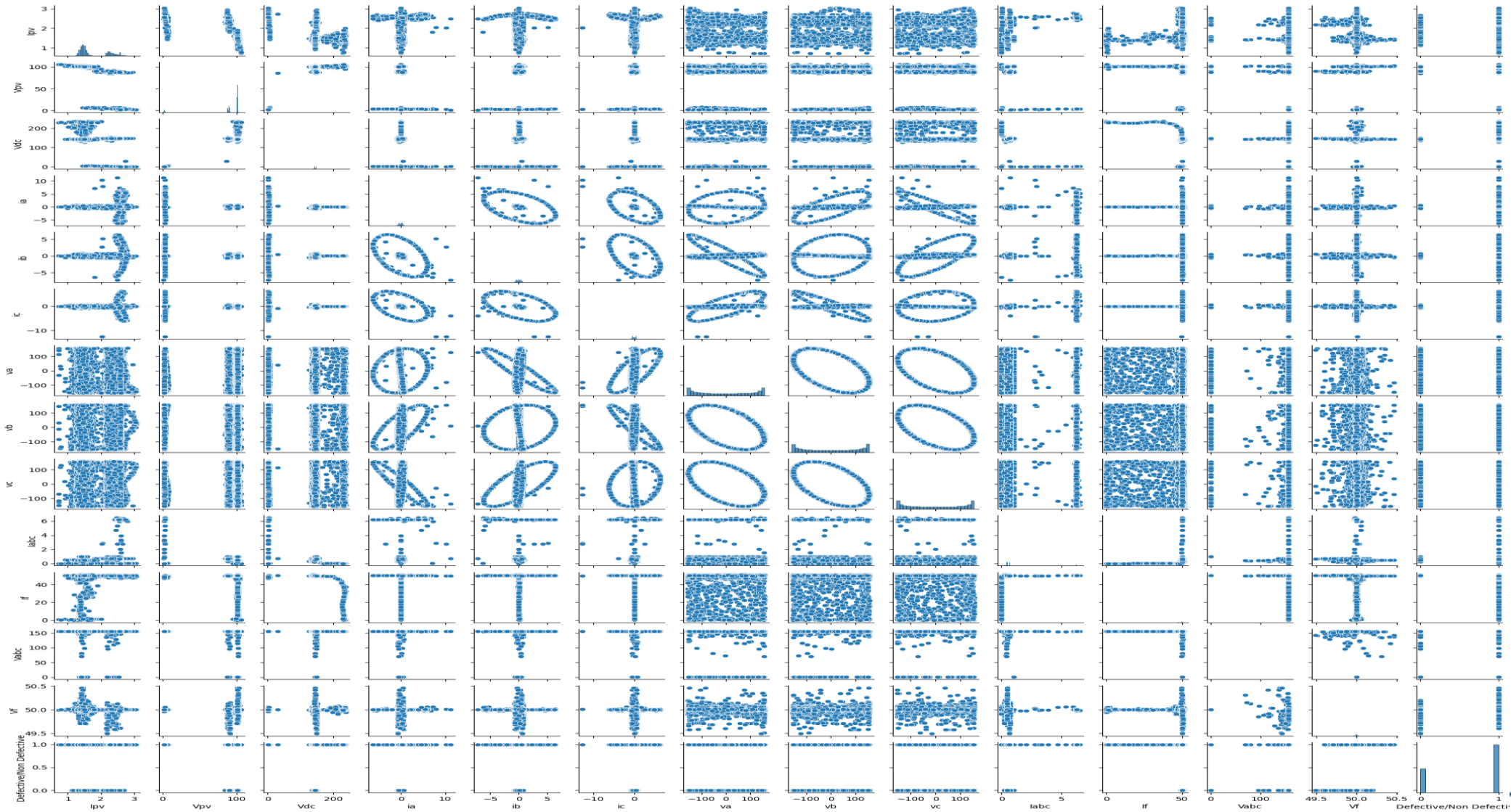
INNODATATICS
Innovation • Data • Analytics

Data Preprocessing-Correlation Coefficient



INNODATATICS
Innovation • Data • Analytics

Data Visualization -scatter plot



Auto EDA

SweetViz 2.3.1

Get updates, docs & report issues here

Created & maintained by Francois Bertrand

Graphic design by Jean-Francois Hains

DataFrame — NO COMPARISON TARGET

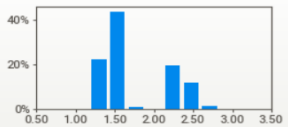
30000	ROWS
0	DUPLICATES
3.4 MB	RAM
14	FEATURES
1	CATEGORICAL
13	NUMERICAL
0	TEXT

ASSOCIATIONS

DataFrame

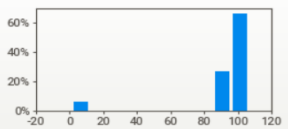
1 Ipv

VALUES:	30,000 (100%)	MAX	3.06	RANGE	2.34		
MISSING:	---	95%	2.57	IQR	0.800		
DISTINCT:	1,355 (5%)	Q3	2.23	STD	0.435		
ZEROES:	---	AVG	1.76	VAR	0.189		
		MEDIAN	1.52				
		Q1	1.43	KURT.	-1.08		
		5%	1.34	SKEW	0.759		
		MIN	0.72	SUM	52,730		



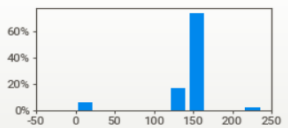
2 Vpv

VALUES:	30,000 (100%)	MAX	106	RANGE	105		
MISSING:	---	95%	102	IQR	11.1		
DISTINCT:	1,673 (6%)	Q3	102	STD	23.9		
ZEROES:	---	MEDIAN	101	VAR	570		
		AVG	92				
		Q1	90	KURT.	9.40		
		5%	3	SKEW	-3.28		
		MIN	1	SUM	2.8M		



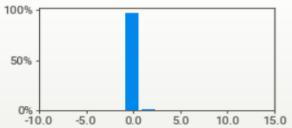
3 Vdc

VALUES:	30,000 (100%)	MAX	238	RANGE	237		
MISSING:	---	95%	149	IQR	3.81		
DISTINCT:	218 (<1%)	Q3	147	STD	37.6		
ZEROES:	---	MEDIAN	144	VAR	1,413		
		AVG	137				
		Q1	143	KURT.	8.60		
		5%	2	SKEW	-2.68		
		MIN	1	SUM	4.1M		



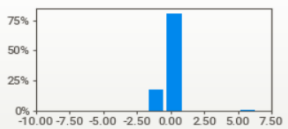
4 ia

VALUES:	30,000 (100%)	MAX	11.3	RANGE	17.6		
MISSING:	---	95%	0.6	IQR	0.718		
DISTINCT:	720 (2%)	Q3	0.3	STD	0.747		
ZEROES:	---	MEDIAN	-0.0	VAR	0.558		
		AVG	-0.0				
		Q1	-0.4	KURT.	36.7		
		5%	-0.7	SKEW	-0.109		
		MIN	-6.4	SUM	-686		



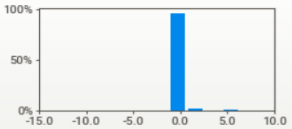
5 ib

VALUES:	30,000 (100%)	MAX	6.4	RANGE	13.7		
MISSING:	---	95%	0.7	IQR	0.698		
DISTINCT:	710 (2%)	Q3	0.4	STD	0.771		
ZEROES:	566 (2%)	AVG	0.0	VAR	0.594		
		MEDIAN	0.0				
		Q1	-0.3	KURT.	35.5		
		5%	-0.6	SKEW	0.088		
		MIN	-7.3	SUM	768		



6 ic

VALUES:	30,000 (100%)	MAX	6.4	RANGE	19.0		
MISSING:	---	95%	0.6	IQR	0.638		
DISTINCT:	726 (2%)	Q3	0.3	STD	0.741		
ZEROES:	---	AVG	-0.0	VAR	0.549		
		MEDIAN	-0.1				
		Q1	-0.3	KURT.	40.9		
		5%	-0.7	SKEW	0.056		
		MIN	-12.7	SUM	-1,464		



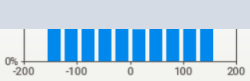
7 va

ZEROES: ---

Q1 -109 KURT. -1.50

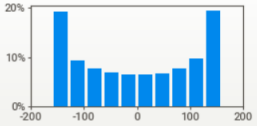
5% -152 SKEW 0.001

MIN -159 SUM 19,899



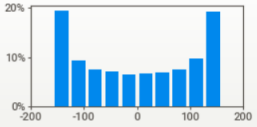
8 vb

VALUES:	30,000 (100%)	MAX	160	RANGE	319		
MISSING:	---	95%	154	IQR	220		
DISTINCT:	16,399 (55%)	Q3	111	STD	110		
ZEROES:	1 (<1%)	MEDIAN	1	VAR	12,090		
		AVG	1				
		Q1	-109	KURT.	-1.50		
		5%	-152	SKEW	-0.004		
		MIN	-159	SUM	30,770		



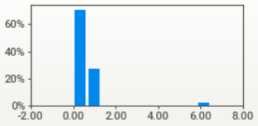
9 vc

VALUES:	30,000 (100%)	MAX	159	RANGE	317		
MISSING:	---	95%	154	IQR	218		
DISTINCT:	29,634 (99%)	Q3	110	STD	109		
ZEROES:	---	AVG	1	VAR	11,989		
		MEDIAN	1				
		Q1	-108	KURT.	-1.49		
		5%	-152	SKEW	-2.72e-4		
		MIN	-158	SUM	22,410		



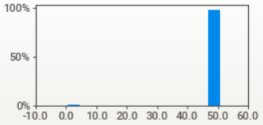
10 Iabc

VALUES:	30,000 (100%)	MAX	6.46	RANGE	6.46		
MISSING:	---	95%	0.73	IQR	0.218		
DISTINCT:	28,777 (96%)	Q3	0.67	STD	0.847		
ZEROES:	---	AVG	0.62	VAR	0.718		
		MEDIAN	0.48				
		Q1	0.45	KURT.	37.8		
		5%	0.01	SKEW	6.15		
		MIN	0.00	SUM	18,562		



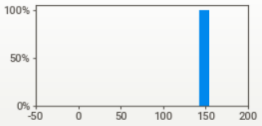
11 If

VALUES:	30,000 (100%)	MAX	51.3	RANGE	51.4		
MISSING:	---	95%	50.5	IQR	0.209		
DISTINCT:	28,767 (96%)	Q3	50.1	STD	4.97		
ZEROES:	---	MEDIAN	50.0	VAR	24.7		
		AVG	49.3				
		Q1	49.9	KURT.	71.4		
		5%	48.6	SKEW	-8.33		
		MIN	-0.0	SUM	1.5M		



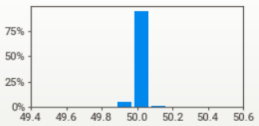
12 Vabc

VALUES:	30,000 (100%)	MAX	156	RANGE	155		
MISSING:	---	95%	156	IQR	0.467		
DISTINCT:	28,777 (96%)	Q3	155	STD	6.95		
ZEROES:	---	MEDIAN	155	VAR	48.4		
		AVG	155				
		Q1	155	KURT.	471		
		5%	155	SKEW	-21.6		
		MIN	1	SUM	4.6M		



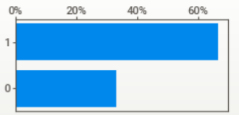
13 Vf

VALUES:	30,000 (100%)	MAX	50.462	RANGE	0.970		
MISSING:	---	95%	50.022	IQR	0.009		
DISTINCT:	28,768 (96%)	Q3	50.004	STD	0.023		
ZEROES:	---	MEDIAN	50.000	VAR	5.40e-4		
		AVG	50.000				
		Q1	49.996	KURT.	102		
		5%	49.977	SKEW	-2.60		
		MIN	49.492	SUM	1.5M		



14 Defective/Non Defective

VALUES:	30,000 (100%)		
MISSING:	---		
DISTINCT:	2 (<1%)		



Model Building

Classification Models

1. **Logistic Regression**: Simple linear model for binary classification.

2. **Decision Tree**: Non-linear model splitting data based on attributes.

3. **Random Forest**: Ensemble method combining multiple decision trees.

4. **SVM**: Finds the best hyperplane to separate classes.

5. **Naive Bayes**: Probabilistic classifier with strong independence assumptions.

6. **Ensemble Methods (Gradient Boosting)**: Combines weak learners sequentially for improved accuracy.

Model Accuracy Comparison

1.Logistic Regression

1. Accuracy: 71.83%
2. Train Accuracy: 71.21%
3. Test Accuracy: 71.83%

2.Decision Tree

1. Accuracy: 97.18%
2. Train Accuracy: 99.53%
3. Test Accuracy: 97.18%

3.Random Forest

1. Accuracy: 99.15%
2. Train Accuracy: 100%
3. Test Accuracy: 99.15%

4.Support Vector Machines (SVM)

1. Accuracy: 91.55%
2. Train Accuracy: 94.06%
3. Test Accuracy: 91.55%

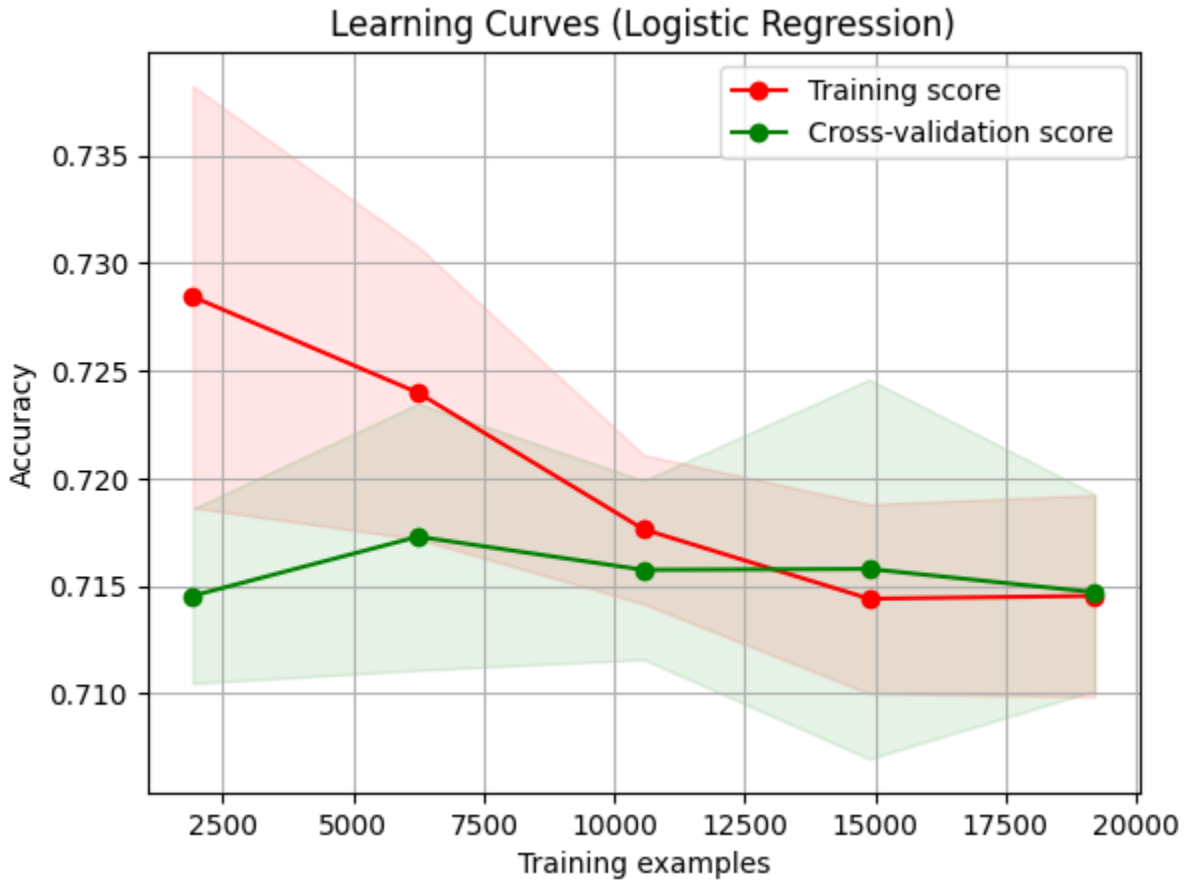
5.Naive Bayes

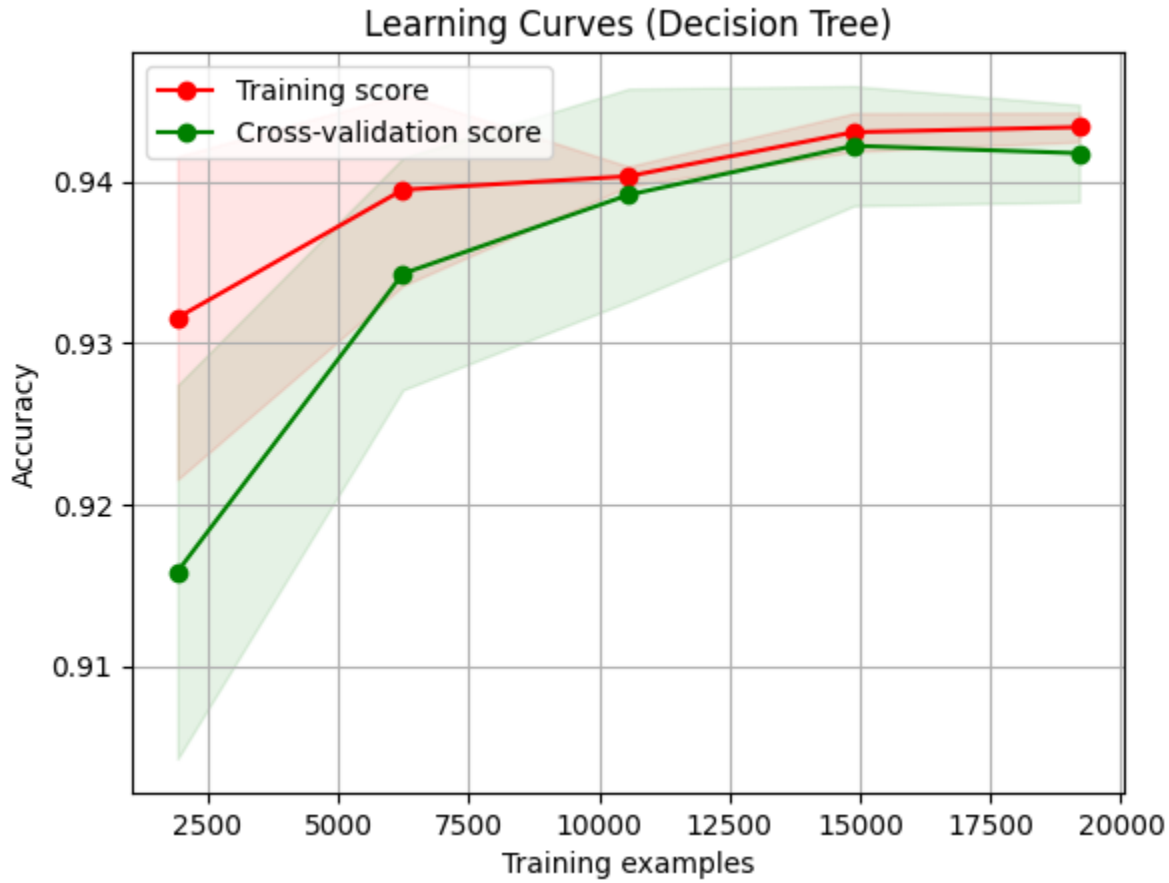
1. Accuracy: 45.22%

6.Ensemble Methods (Gradient Boosting)

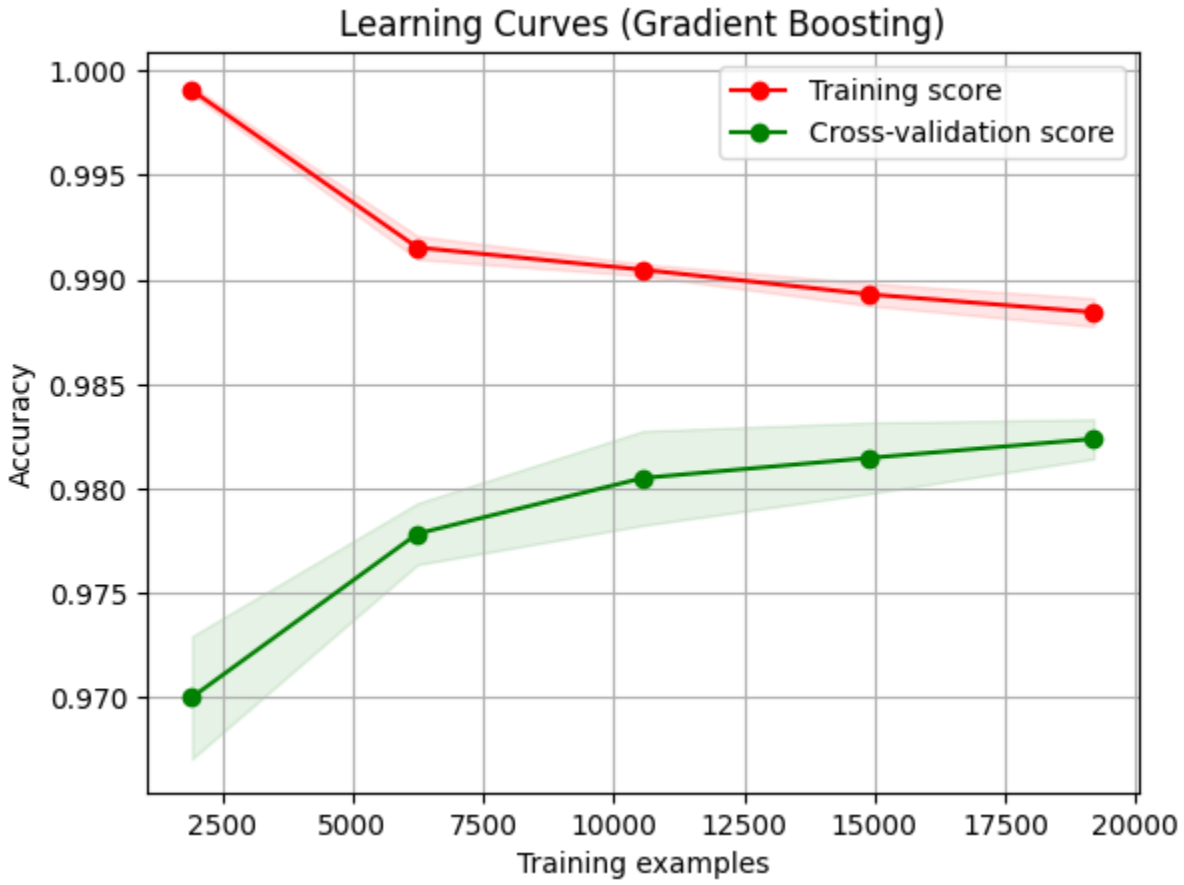
1. Accuracy: 99.02%
2. Train Accuracy: 99.66%
3. Test Accuracy: 99.02%

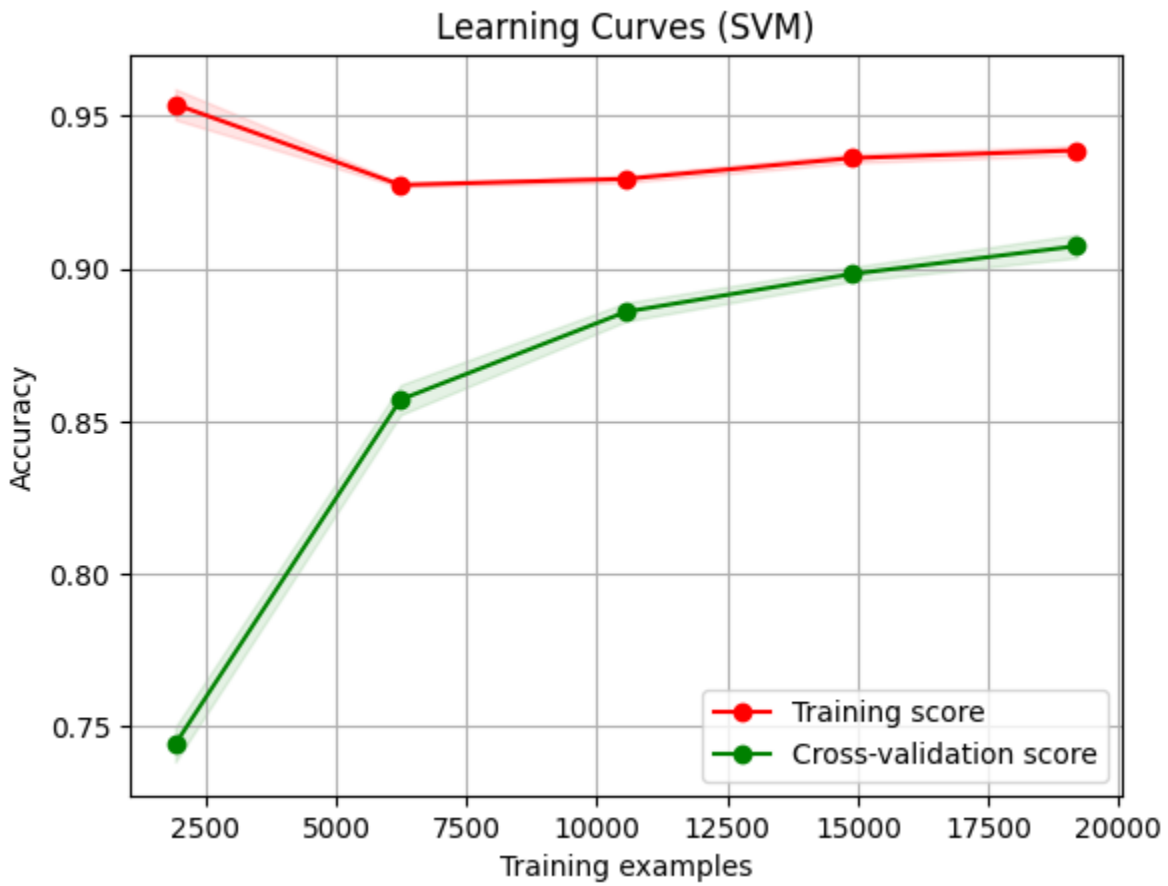
Model Accuracy Comparison



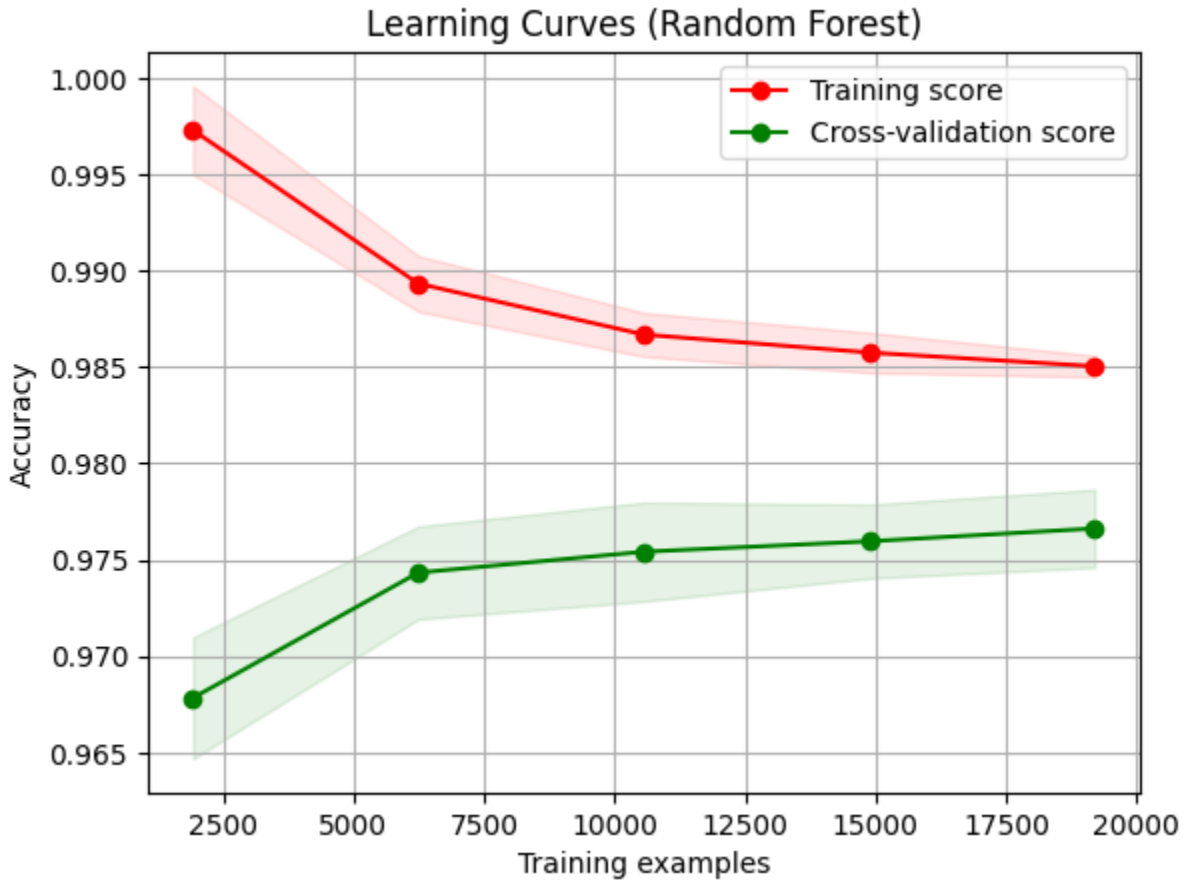


Model Accuracy Comparison



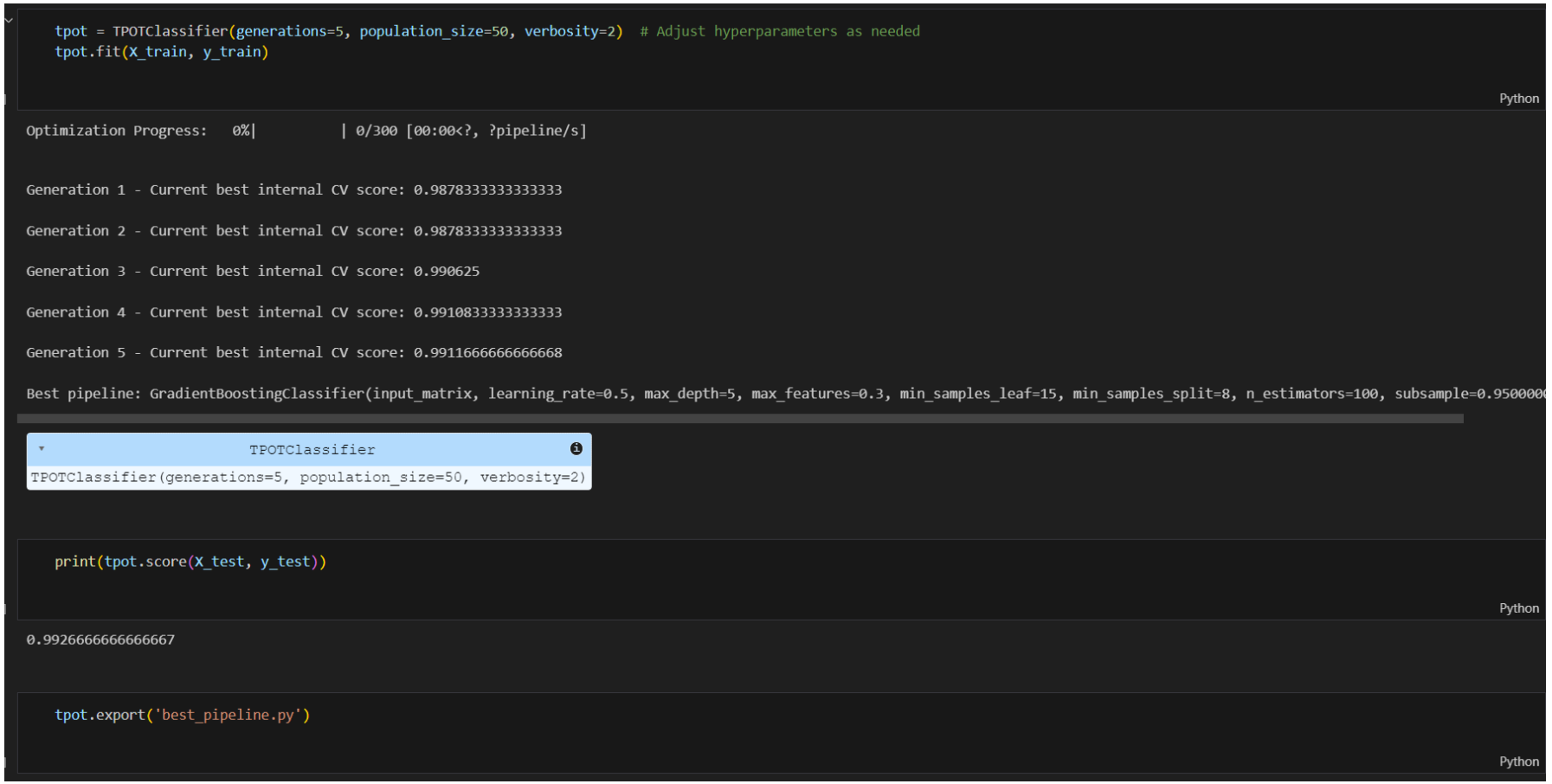


Best Model – Random Forest



- **Best Model:** Random Forest (Accuracy: 99.15%)
  - Random Forest exhibits the highest accuracy on both the training and testing datasets, indicating robust performance and generalization capability.
- **High Accuracy Model:** Gradient Boosting (Accuracy: 99.02%)
  - Gradient Boosting demonstrates a high accuracy level, slightly below Random Forest, making it another strong contender for classification tasks.

AutoML Model –TPOT



POT (Tree-based Pipeline Optimization Tool):

TPOT is a Python library that automatically creates and optimizes machine learning pipelines using genetic programming.

Auto ML _TPOT = 99.26%

Model Deployment - Strategy

Deployment Procedure:

Prepare Model:

Train your machine learning model and save it as a joblib file (model.joblib).

Create Streamlit or Flask App:

Write the code for your Streamlit or Flask application (app.py).

Load Model:

Load the trained model (model.joblib) within your application code.

Design User Interface:

Create a user interface where users can upload their dataset for prediction.

Data Handling:

Implement data preprocessing steps to prepare the uploaded dataset for model prediction.

Model Prediction:

Use the loaded model to make predictions on the uploaded dataset.

Display Results:

Show the prediction results to the user within the application interface.

Deployment:

Deploy your Streamlit or Flask application to a hosting platform or run it locally.

Testing:

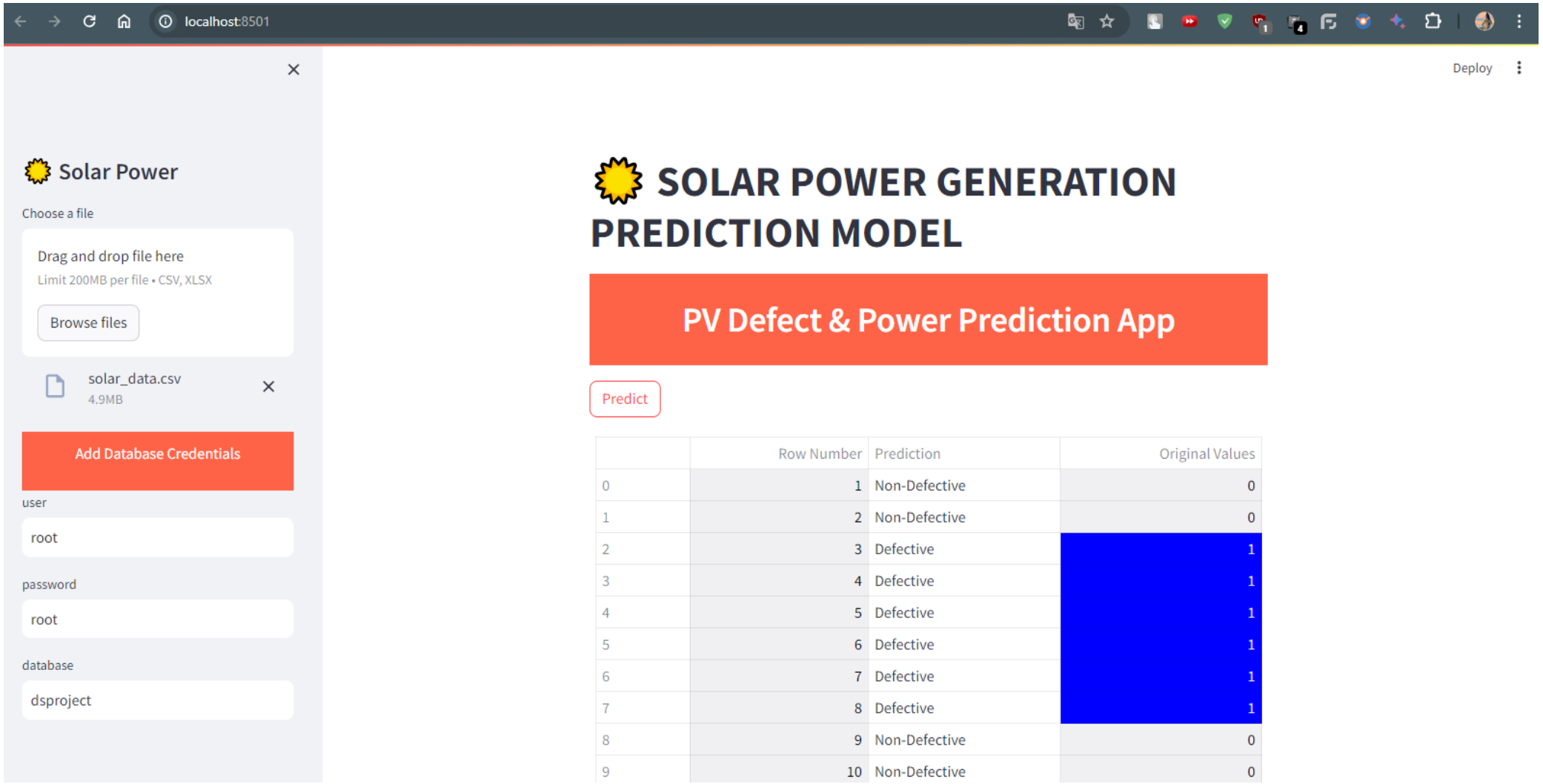
Test the deployed application to ensure it functions as expected.

Monitoring and Maintenance:

Monitor the application for performance issues and user feedback.

Update and maintain the application as needed.

Screen shot of output



	Row Number	Prediction	Original Values
0	1	Non-Defective	0
1	2	Non-Defective	0
2	3	Defective	1
3	4	Defective	1
4	5	Defective	1
5	6	Defective	1
6	7	Defective	1
7	8	Defective	1
8	9	Non-Defective	0
9	10	Non-Defective	0

Video of output

Refer to Project Artifacts

Challenges

1.Noisy Data:

- High-frequency measurements with noise and disturbances pose challenges for accurate fault detection.

2.Fault Severity Detection:

- Differentiating between severe and minor faults accurately is crucial for proactive maintenance.

3.Resource Constraints:

- Minimizing resource utilization while maximizing fault detection efficiency requires careful optimization.

4.Real-time Detection:

- Ensuring timely detection of faults before total failure, especially in dynamic operational environments.

Future Scopes

1.Advanced Modeling Techniques:

- Explore advanced machine learning techniques such as deep learning for more accurate fault detection.

2.Integration with IoT and Edge Computing:

- Utilize IoT devices and edge computing for distributed fault detection and real-time decision-making.

3.Continuous Model Improvement:

- Implement mechanisms for continuous model retraining and improvement based on new data and feedback.

4.Predictive Maintenance:

- Move towards predictive maintenance by leveraging historical data and predictive analytics to anticipate and prevent faults.

5.Collaborative Research:

- Collaborate with research institutions and industry partners to stay updated with the latest advancements in fault detection algorithms and methodologies.

Queries ?


Thanks