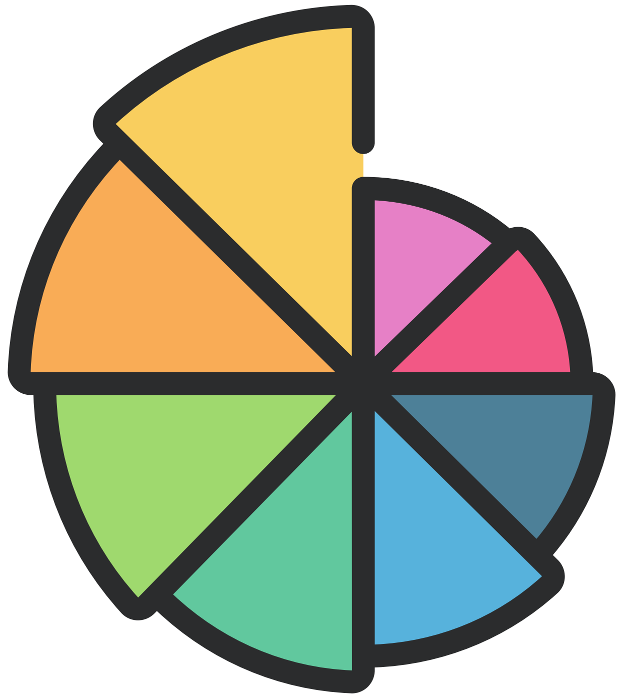
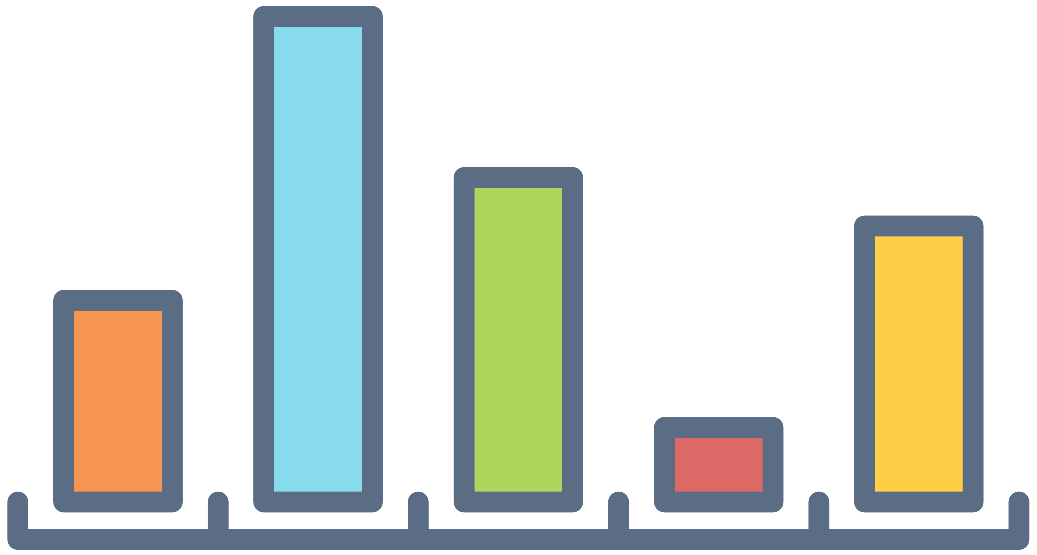
Loan Approval Prediction Analysis

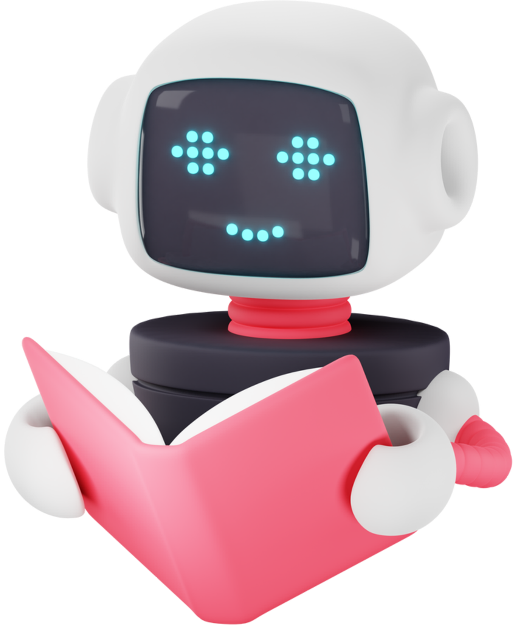
Exploring Data, Building Models, and Evaluating Performance



By Atharva & Chinmay

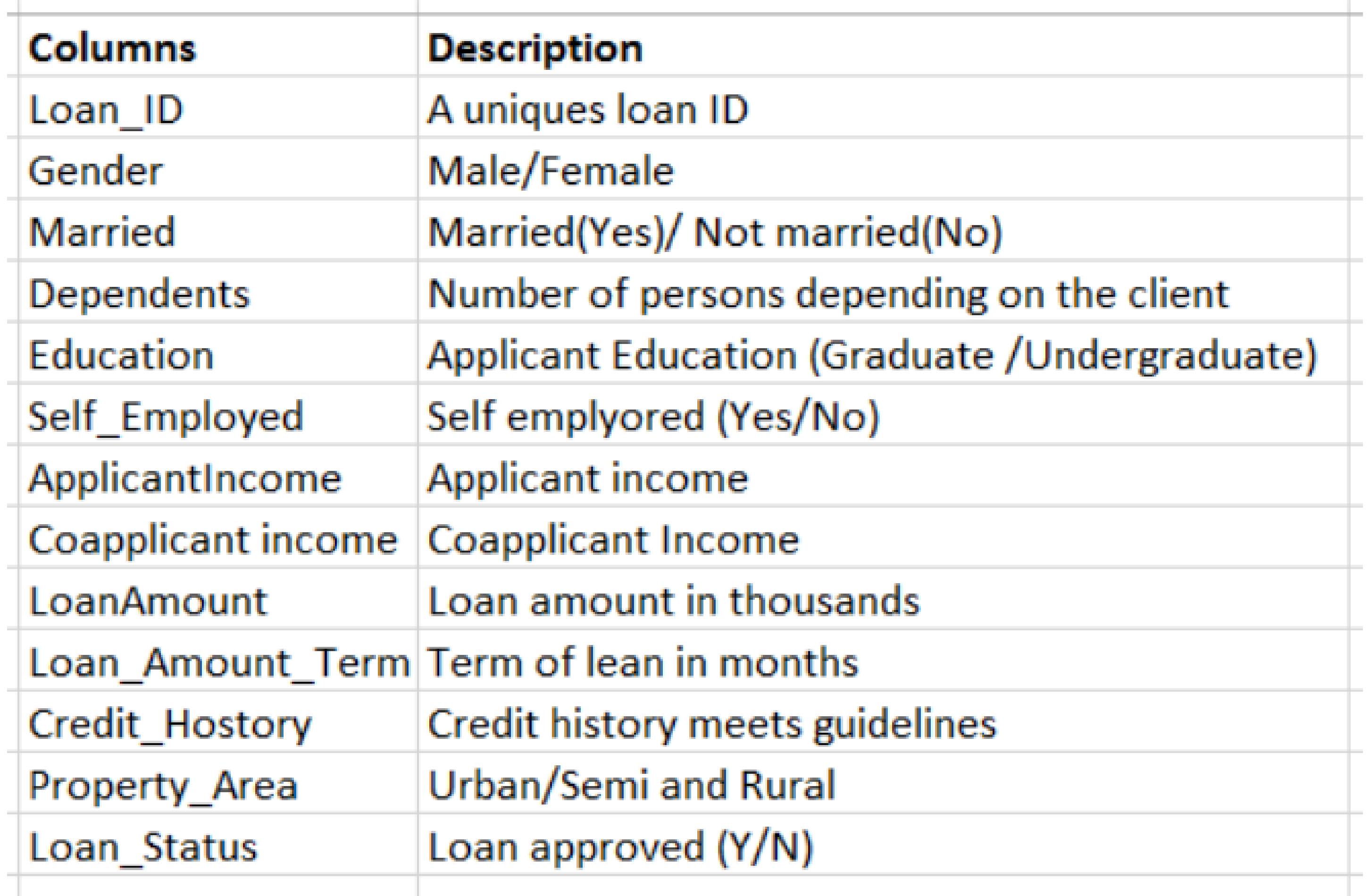


Introduction to the Problem:

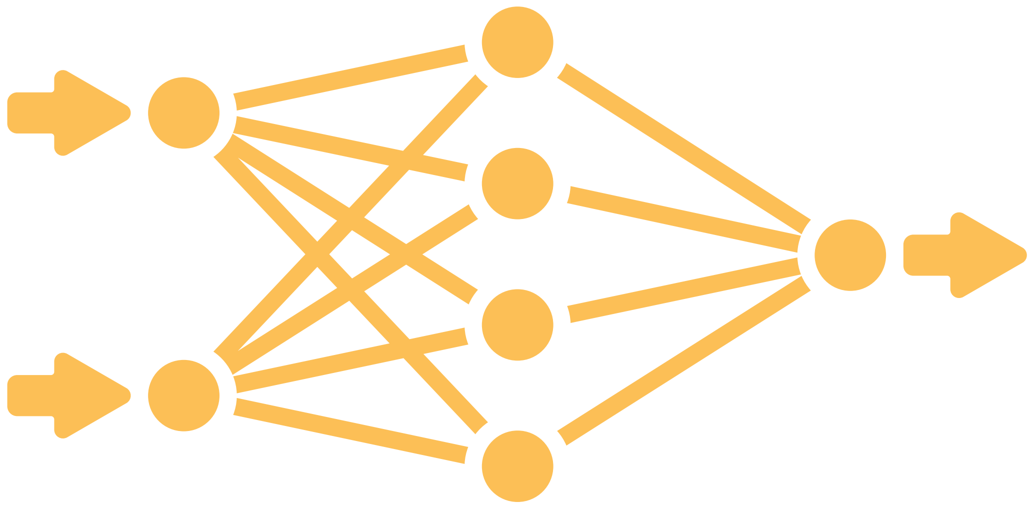


1. **Loan Approval Challenges:**
   - **Introduction to the problem statement reveals the intricate challenges in predicting loan approvals, navigating through factors like credit history, income, and debt-to-income ratios that influence lending decisions.**
2. **Critical Decision-Making:**
   - **Emphasizing the importance of accurate predictions underscores the significance of reliable loan approval forecasts, as they shape the financial well-being of both lenders and borrowers.**
3. **Modeling Approach Overview:**
   - **Providing an overview of our machine learning approach, we delve into the utilization of advanced algorithms to meticulously analyze diverse financial data, ensuring a robust model for accurate loan approval predictions.**

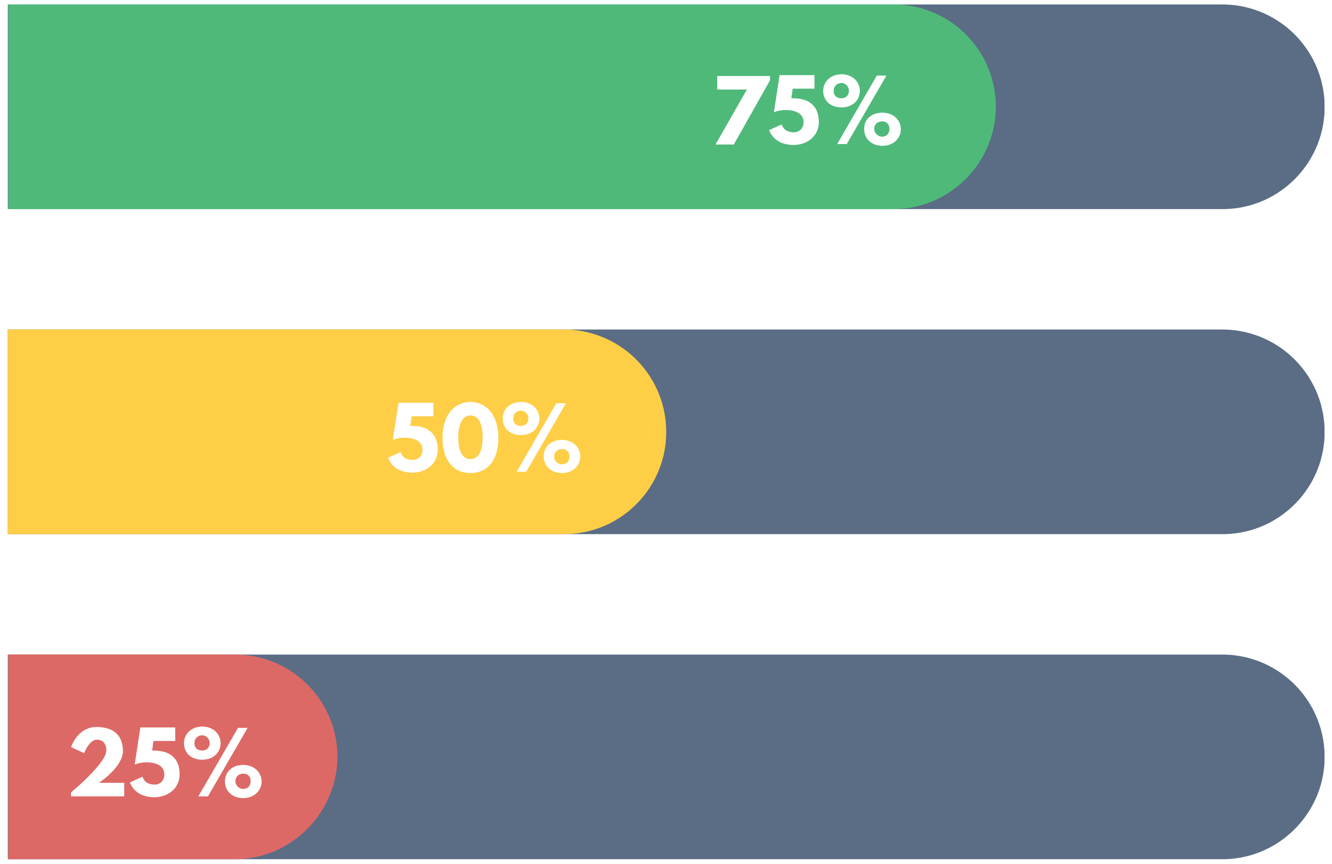
Loan Predictor



Columns	Description
Loan_ID	A uniques loan ID
Gender	Male/Female
Married	Married(Yes)/ Not married(No)
Dependents	Number of persons depending on the client
Education	Applicant Education (Graduate /Undergraduate)
Self_Employed	Self emplyored (Yes/No)
ApplicantIncome	Applicant income
Coapplicant income	Coapplicant Income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of lean in months
Credit_Hostory	Credit history meets guidelines
Property_Area	Urban/Semi and Rural
Loan_Status	Loan approved (Y/N)



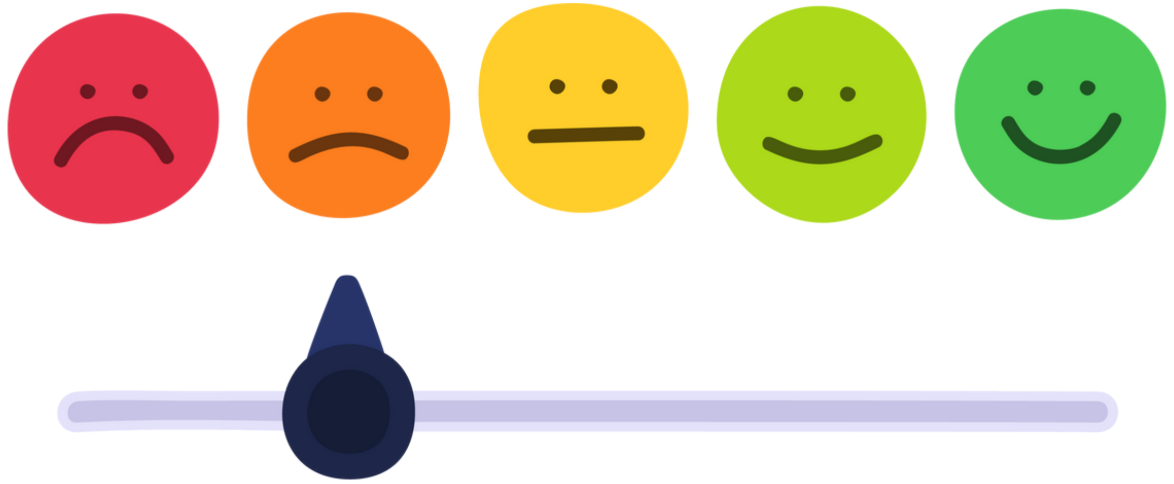
Project Focuses on



Performance Measures on Model

Data Visualization & Analysis

Performance Measures on Model

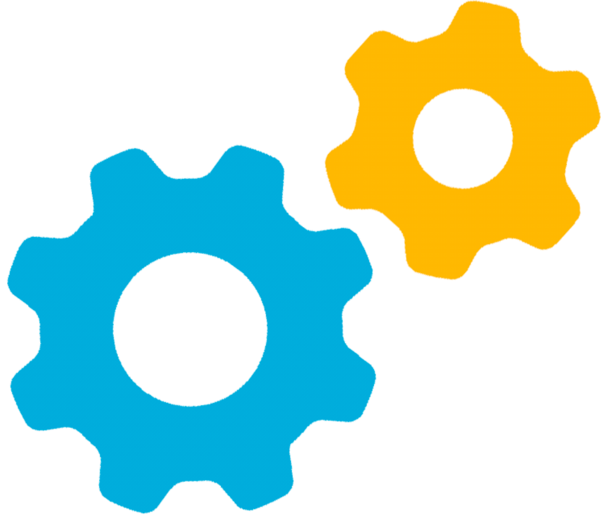


1. Accuracy Score:
   - Measures overall correctness.
   - Accuracy=Correct PredictionsTotal PredictionsAccuracy=Total PredictionsCorrect Predictions
   - High accuracy indicates good overall model performance.
2. F1 Score:
   - Balances precision and recall.
   - *F*1 = 2×Precision×RecallPrecision+RecallF1=2×Precision+RecallPrecision×Recall
   - Useful when false positives and false negatives have different impacts.
3. Precision:
   - Measures accuracy of positive predictions.
   - Precision=True PositivesTrue Positives+False PositivesPrecision=True Positives+False PositivesTrue Positives
   - Indicates reliability of positive predictions.

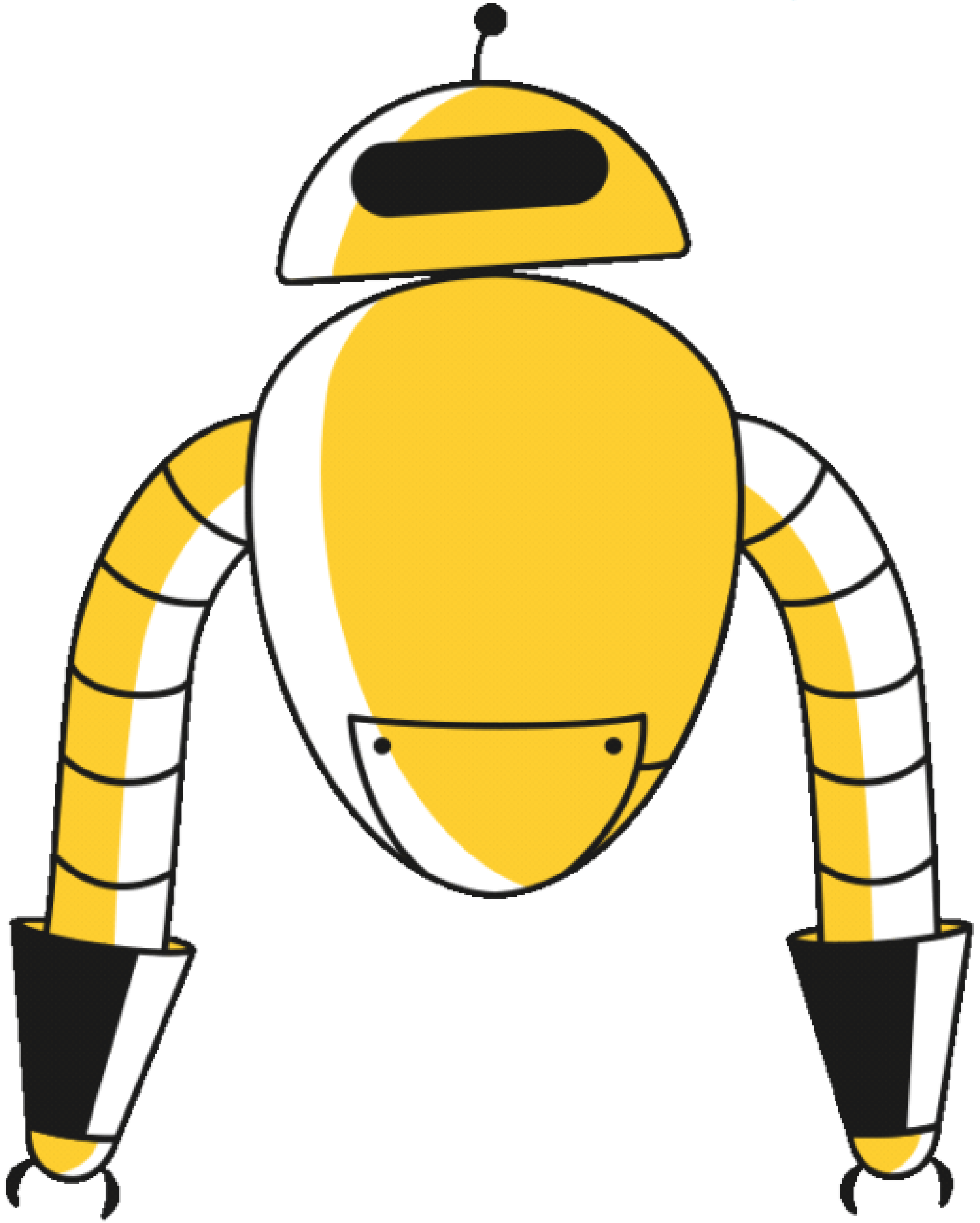
Data Visualization & Analysis

1. **Visual Insight:**
   - **Data visualization is crucial for understanding complex patterns, aiding in informed decision-making for loan approvals in our system.**
2. **Key Comparison Factors:**
   - **We focus on 11 vital factors, including income, marital status, age, and credit history, to discern trends influencing loan approval outcomes.**
3. **Matplotlib and Seaborn Tools:**
   - **Utilizing Matplotlib and Seaborn, we craft clear and informative visualizations, enhancing our ability to interpret and act upon the insights derived from the Loan Predictor system.**

Algorithms implemented on our model to seek performance measures.

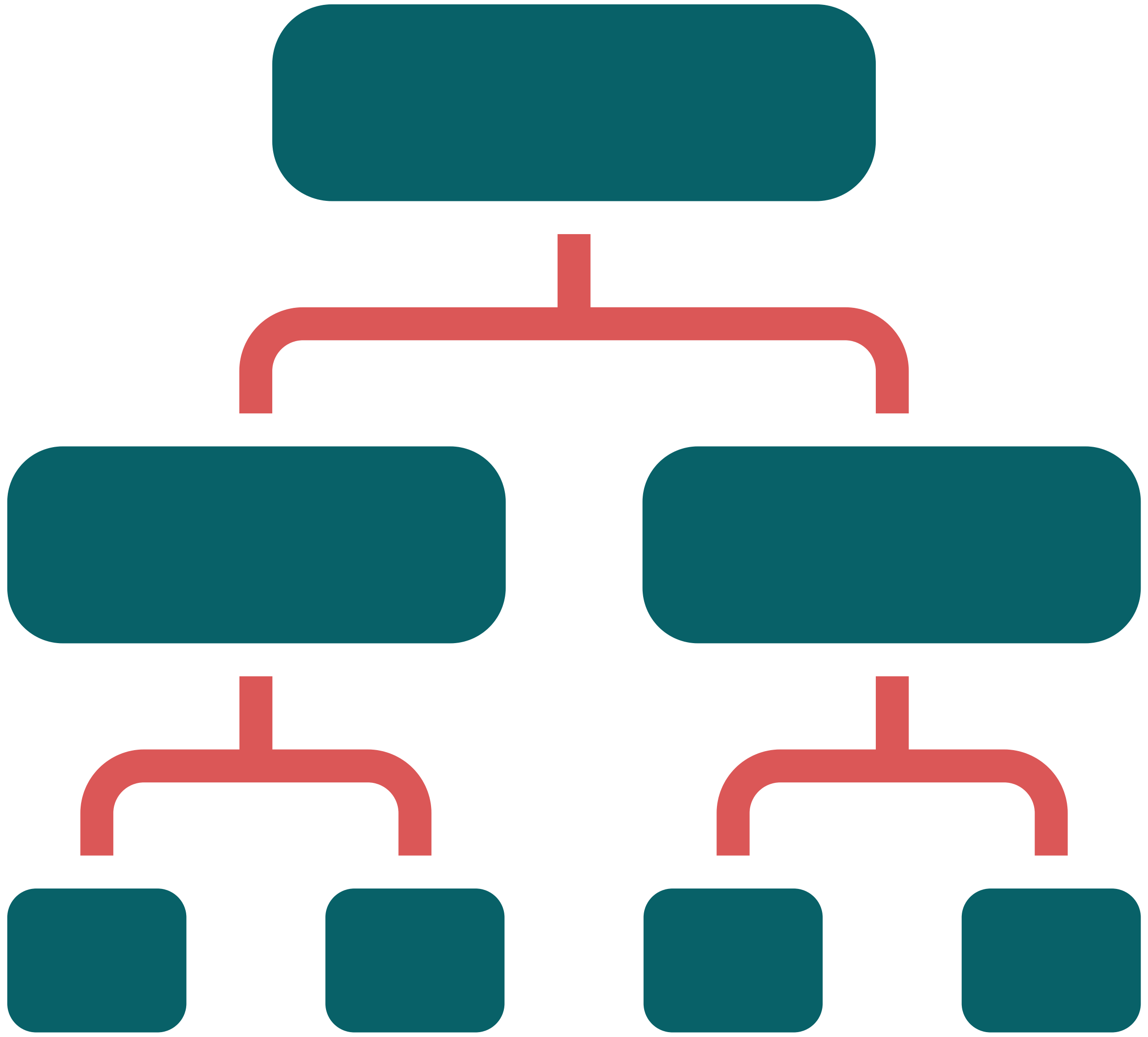


- Decision Tree
- Random Forest
- XGBoost
- Logistic Regression
- KNN
- SVM

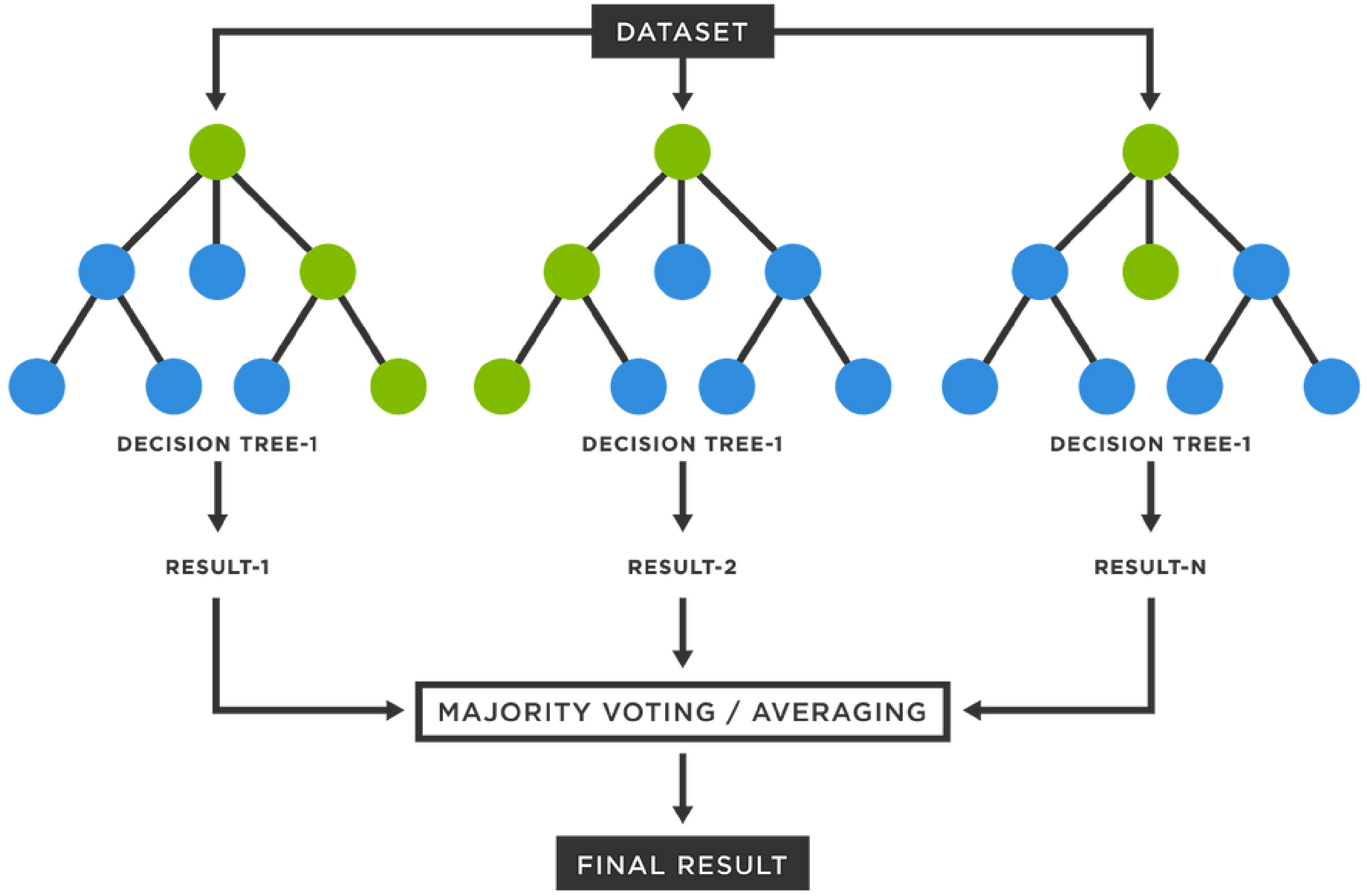


Decision Tree

1. **Tree-Based Structure:**
   - Decision Trees in ML utilize a tree-like structure, with nodes representing feature-based decisions, offering a clear and intuitive model.
2. **Real-world Applications:**
   - Widely employed in finance, healthcare, and marketing, Decision Trees find practical use in classifying and predicting outcomes based on distinct features.
3. **Advantageous Transparency:**
   - Decision Trees provide transparency by revealing feature importance, aiding in easy interpretation and understanding of the decision-making process.



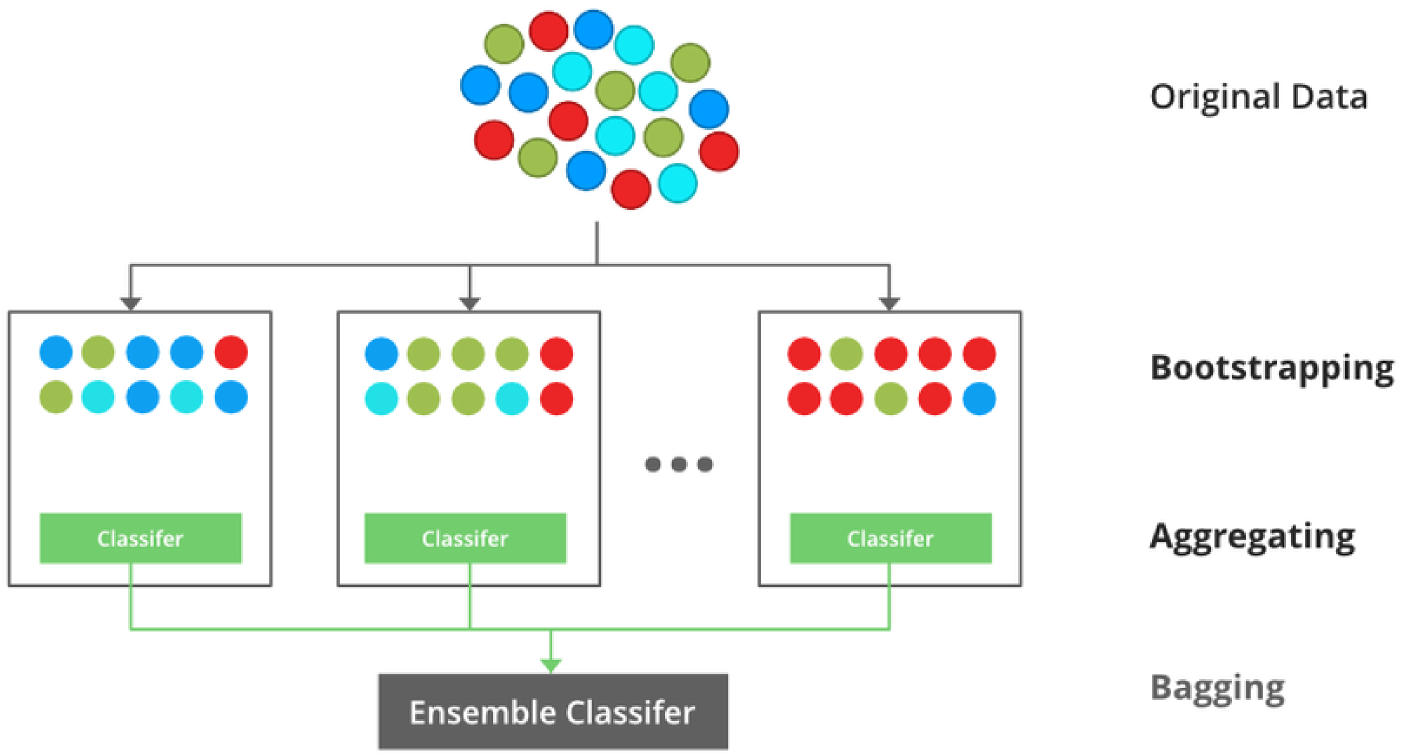
Random Forest



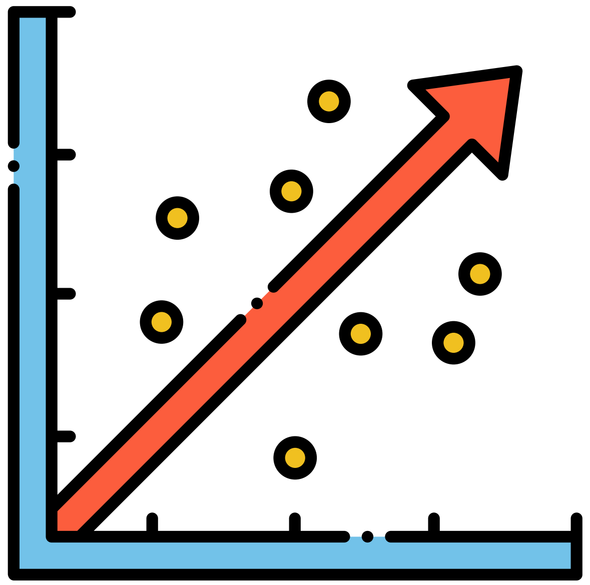
1. **Ensemble Power:**
   - **Random Forest, a potent ensemble method, combines multiple Decision Trees to enhance predictive accuracy.**
2. **Broad Applicability:**
   - **Applied in finance, healthcare, and diverse fields, Random Forest excels in real-world predictions, offering versatility across various datasets.**
3. **Robust Performance:**
   - **Its aggregated feature importance ensures stability, making Random Forest a robust choice for handling intricate data patterns with improved accuracy.**

XGBoost

1. **Boosted Ensemble Method:**
   - **XGBoost (eXtreme Gradient Boosting) is a boosted ensemble algorithm that sequentially combines weak learners, often decision trees, to create a robust and high-performance predictive model.**
2. **Efficiency in Real-world Scenarios:**
   - **Widely embraced in competitions like Kaggle and extensively used in industry, XGBoost demonstrates remarkable efficiency in real-world applications, showcasing its versatility and superior performance.**
3. **Optimized Training and Regularization:**
   - **XGBoost incorporates regularization techniques and advanced optimization algorithms, enabling faster training and mitigating overfitting, thereby enhancing model generalization and predictive accuracy.**



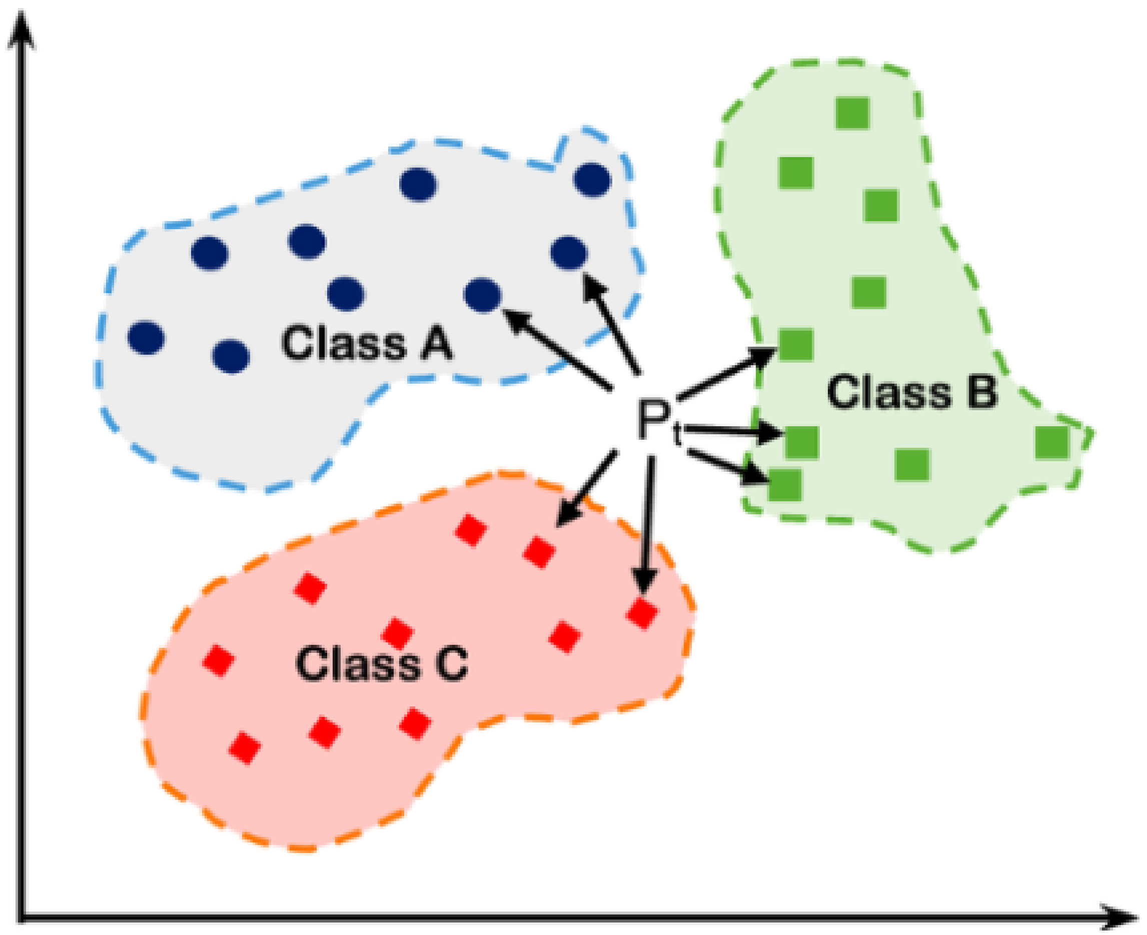
Logistic Regression



1. **Linear Classification:**
   - **Logistic Regression is a fundamental algorithm for binary classification, modeling the relationship between input features and the probability of belonging to a particular class through a logistic function.**
2. **Interpretability and Simplicity:**
   - **Known for its simplicity and interpretability, Logistic Regression is widely used in fields where understanding the impact of individual features on the outcome is crucial, such as in medical research or social sciences.**
3. **Probabilistic Predictions:**
   - **Logistic Regression provides probabilistic predictions, offering insights into the likelihood of an observation belonging to a specific class, making it valuable for decision-making and risk assessment in various domains.**

1. **Instance-Based Learning:**
   - **K-Nearest Neighbors (KNN) is an instance-based learning algorithm that classifies data points based on the majority class of their k-nearest neighbors, making it adaptable to various data distributions.**
2. **Simple and Intuitive:**
   - **Known for its simplicity and ease of implementation, KNN is particularly effective in scenarios where underlying patterns are not easily captured by parametric models, offering an intuitive approach to classification.**
3. **Versatility in Feature Spaces:**
   - **KNN is versatile in handling different types of feature spaces, making it applicable in diverse domains such as recommendation systems, image recognition, and anomaly detection, where distances between data points are crucial for decision-making.**

K Nearest Neighbors

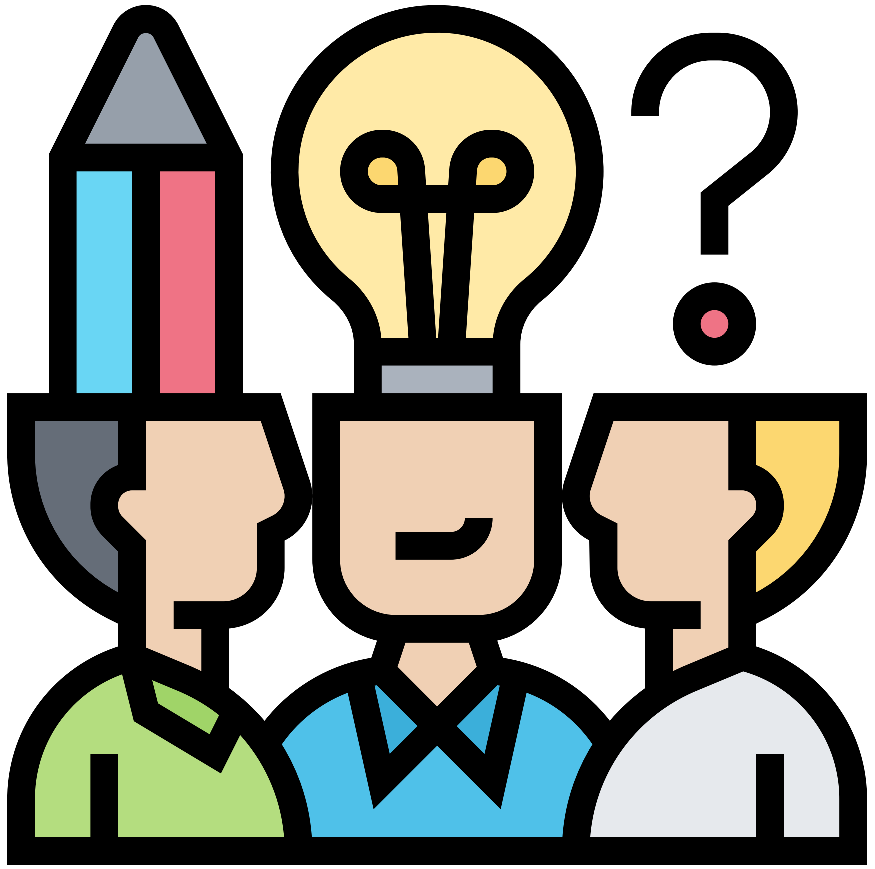


SVM

1. **Effective Hyperplane-based Classification:**
   - **Support Vector Machine (SVM) is a powerful algorithm for classification and regression tasks, utilizing hyperplanes to effectively separate data into distinct classes and maximize the margin between them.**
2. **Kernel Trick for Nonlinear Data:**
   - **SVM employs the kernel trick to handle nonlinear data, transforming input features into higher-dimensional spaces, enabling the algorithm to capture complex relationships and make accurate predictions in a variety of scenarios.**
3. **Robust Performance in High-Dimensional Spaces:**
   - **SVM exhibits robust performance in high-dimensional spaces, making it valuable for tasks such as image classification and text mining, where datasets have numerous features. The algorithm's ability to handle complex, multi-dimensional data enhances its applicability in real-world scenarios.**

Our Contributions

- Applying KNN & SVM on the existing model to analyze the performance metrics.
- Data Pre-processing, feature extraction, standardizing, and normalizing the data to increase the efficiency and accuracy of the models



Conclusion

1) Credit_History Importance:

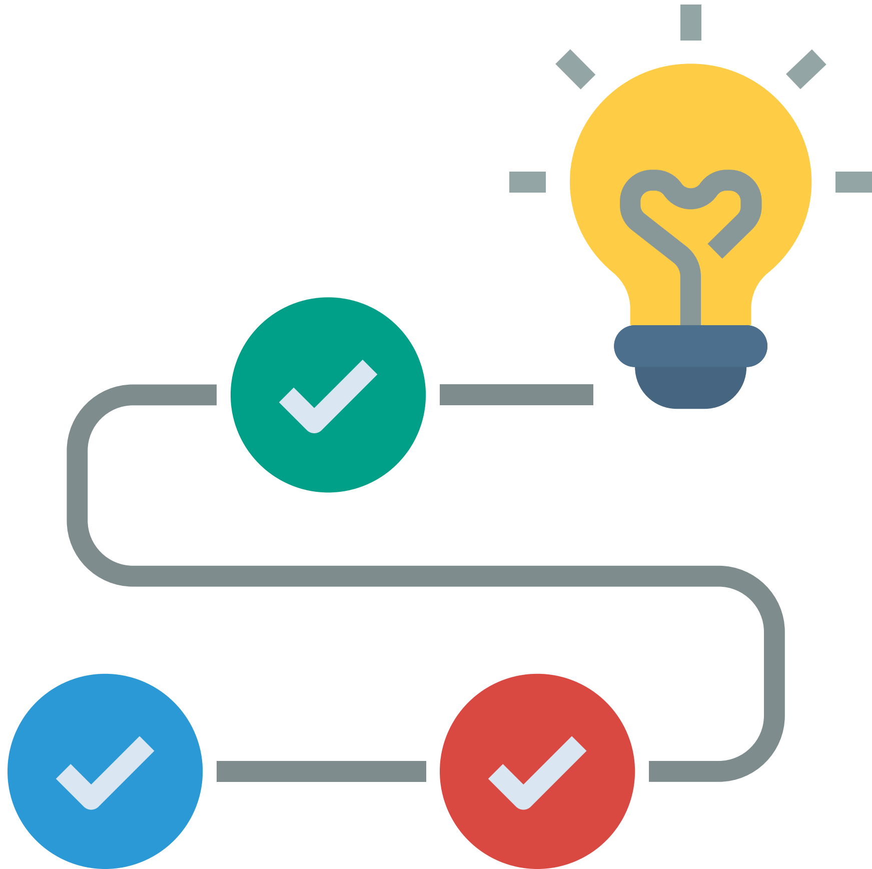
Identified Credit_History as a crucial variable due to its high correlation with Loan_Status. Noted its significant dependency on the loan approval outcome.

2) Model Performance:

Logistic Regression achieved the highest accuracy among the models, approximately 83%.

Listed the models in descending order of accuracy:

Logistic Regression (83.24%), XGBoost (80.54%), Random Forest (77.84%), Support Vector Machine (72.43%), Decision Tree (71.35%), K-Nearest Neighbors (62.70%).



Thank you!