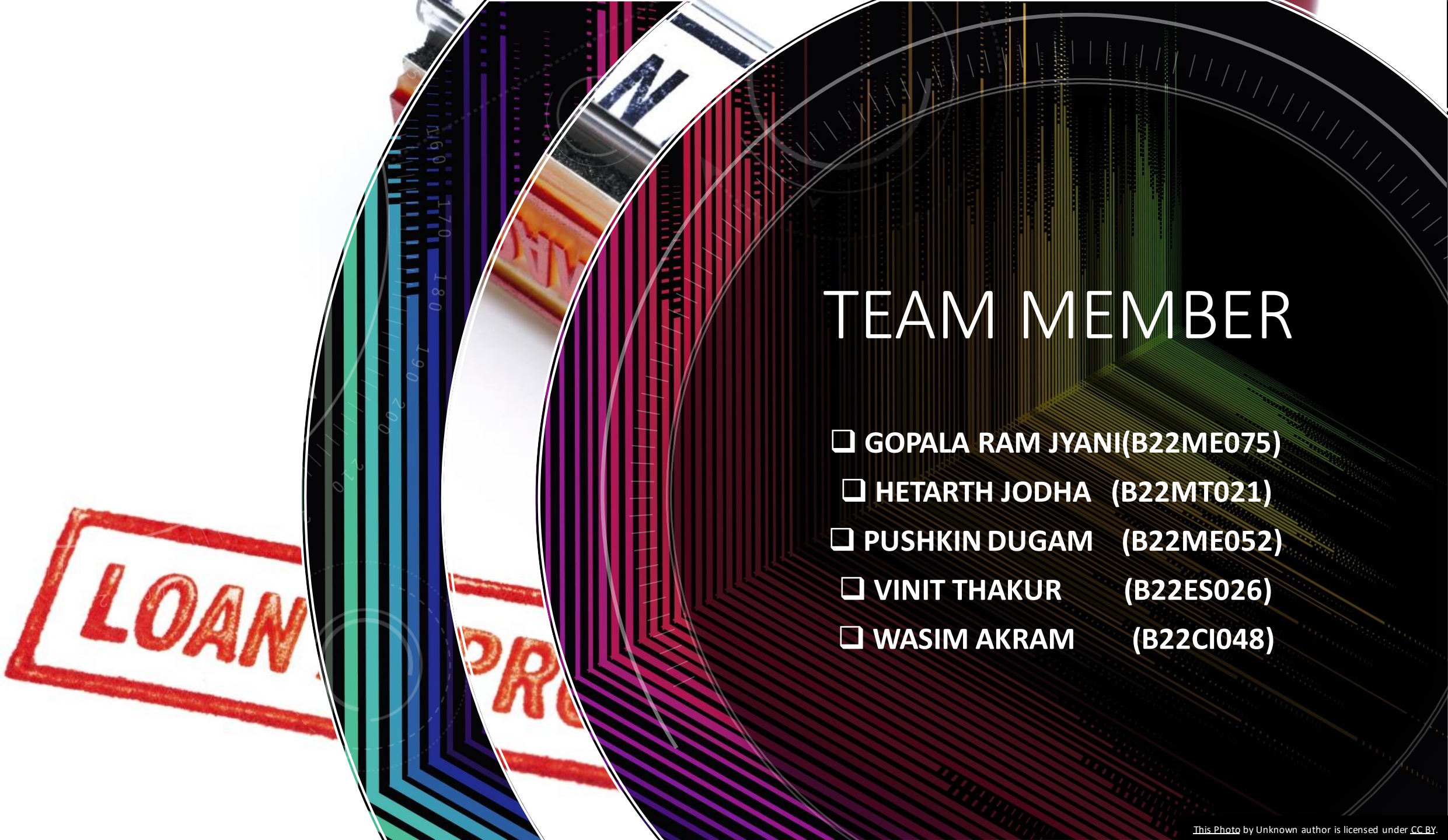
LOAN APPROVAL PREICTION

INTRODUCTION TO MACHINE LEARNING PROJECT



TEAM MEMBER

- GOPALA RAM JYANI(B22ME075)
- HETARTH JODHA (B22MT021)
- PUSHKIN DUGAM (B22ME052)
- VINIT THAKUR (B22ES026)
- WASIM AKRAM (B22CI048)

DATA PREPROCESSING

Scaling/Normalizing Features:-We are use in this dataset Min Max Scaling

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

DATA PREPROCESSING

In the minmax scaling all data lies between 0 to 1
In this given dataset income and cibil score are different range
Uniform Feature Scaling: Min Max scaling rescales the data to a common scale, ensuring that all features have the same range. This uniform scaling is particularly important when using machine learning algorithms that rely on distance-based calculations, such as k-nearest neighbors (KNN) and support vector machines (SVM). Without feature scaling, features with larger numeric ranges may disproportionately influence these algorithms
\
.

Encoding :- In this dataset we use label encoding

Label encoding :-Label encoding is a data preprocessing technique used in machine learning to transform categorical data into a numerical format. In this process, each unique category or label is assigned a unique integer value.

LOGISTIC REGRESSION

Logistic regression is a widely used statistical method for analyzing binary and categorical data. It is a type of regression analysis that is used to predict the probability of a binary outcome (e.g., success or failure, yes or no) based on one or more predictor variables.

Why we use logistic regression in this project:-

Binary Classification: Your loan approval task is inherently a binary classification problem - either approve or deny. Logistic regression is designed for precisely this type of problem and models the probability of loan approval effectively.

How to work logistic regression

$$Z = \beta 0 + \beta 1 * X1 + \beta 2 * X2 + \ldots + \beta n * Xn$$

$$P(Y=1|X) = 1 / (1 + e^{(-z)})$$

Logistic Function

The logistic function is a sigmoid function that maps any real-valued number to a value between 0 and 1. It is used in logistic regression to model the probability of a binary outcome.

Cost Function

The cost function in logistic regression is the negative log-likelihood function, which measures how well the model predicts the outcome labels. The goal of training the model is to minimize the cost function.

Gradient Descent

Gradient descent is an optimization algorithm used to minimize the cost function. It works by iteratively adjusting the model parameters in the opposite direction of the gradient of the cost function with respect to the parameters.

1] KNN

- K-Nearest Neighbors (KNN) is a simple and intuitive supervised machine learning algorithm used for classification and regression tasks. It is a non-parametric, instance-based algorithm, which means it makes predictions based on the similarity between a new data point and existing data points in the training dataset.
- KNN is non parametric algorithm ,which means it does not make any assumption on its own.
- It is used for regression as well as for classification.

HOW IT WORKS

- First step – first it starts with the data preparation. Starts with data preprocessing and all.
- Second step –starts with choosing the value of k ,which is the number of nearest neighbor to consider while making prediction ,k can be any integer but generally we take K as odd value .
- Third step -includes distance calculation when we have new unseen datapoint that you want to classify or predict calculate the distance between this data point and every data point in the training dataset .
- Fourth - neighbor selection ,in this we select k datapoints with the smallest distance to the new datapoint.

For classification tasks:

- Count the number of neighbors in each class among the K-nearest neighbors.
- Assign the class label that is most common among the neighbors to the new data point. This is known as a majority voting scheme.
- If there is a tie, you can resolve it using various techniques, such as weighted voting or considering a smaller K value.

For regression tasks:

- Calculate the average (or weighted average) of the target values of the K-nearest neighbors.
- Assign this average value as the prediction for the new data point.

FINAL -Return the class label (for classification) or the predicted value (for regression) as the output of the KNN algorithm.

DECISION TREE

- Decision tree is a supervised machine learning algorithm , it is also used for regression and classification tasks .
- It is a tree-like structure that helps make decisions or predictions by recursively splitting the dataset into subsets based on the values of input features.
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

HOW IT WORKS

- Root Node : The entire dataset is initially represented by the root node of the tree.
- Feature Selection : The Decision Tree algorithm selects a feature from the dataset to split the data. It chooses the feature that best separates the data into different classes or minimizes the variance (in the case of regression). This is determined using a splitting criterion, which could be Gini impurity, entropy, or mean squared error, depending on whether the problem is classification or regression.
- Splitting : The selected feature is used to split the dataset into subsets based on the feature's values. Each subset represents a branch or path in the tree, and the feature's value determines which path a data point should follow.
- Child Nodes : Each subset becomes a child node of the current node. These child nodes represent subgroups of the data that need further splitting.
- Recursion : The splitting process is repeated recursively for each child node, meaning that steps 2 to 4 are applied to each child node. This continues until certain stopping conditions are met, such as a maximum tree depth, a minimum number of data points in a node, or a node containing only one class (in the case of classification).
- Leaf Nodes : When the recursion ends for a particular branch, and no further splitting is possible, the final nodes are called leaf nodes. Each leaf node is associated with a class label (in classification) or a predicted value (in regression).
- Prediction : To make a prediction for a new data point, it starts at the root node and follows the path in the tree, navigating through the feature splits based on the values of the new data point. Eventually, it reaches a leaf node, and the class label or predicted value of that leaf node is the final prediction.

DATA PREPROCESSING

- Null Values Replacement
- Normalization (Min-Max Scaler)

5] SUPPORT VECTOR MACHINES

- Is a supervised Machine Learning Algorithm That can be used for finding the optimal hyperplane that separates the data points into 2 classes with the largest possible margins.
- Support Vectors are the Data points That are closest to the decision boundary which determine the position and orientation of the hyperplane .
- 3 Types of Kernel Function will be used Linear, Radial Basis Function, Polynomial.
- Grid search with cross validation will be used to find the optimal values of the Hyperparameters (C and Gamma)

RESULTS

At the end we plan to input a new custom dataset and predict the loan approval for all the 5 models along with its accuracy.

Finally we would give a score out
of 5 (1 point for approved in the loan status) for each model.

We also plan on adding xgboost , Random Forest , Voting Classifier ,etc in our model as when we learn it in our course and become thorough with its concepts.