Airlines Passenger Satisfaction

Project Dataset:https://www.kaggle.com/code/frixinglife/airline-passenger-satisfaction-part-1

- The dataset contains information about airline passengers and their satisfaction with various aspects of their flight experience. The data includes over 129880 records with 24 features such as gender, age, flight distance, inflight wifi service, cleanliness, departure/arrival time convenient, etc.
- The target variable is "satisfaction" which indicates whether a passenger was satisfied or dissatisfied with their overall flight experience. The dataset also provides information on the type of travel (business or personal), the class of travel (economy, business, or eco plus), and the destination region.
- This dataset can be used to build classification models that can help airlines improve their customer experience by identifying the key factors that drive satisfaction or dissatisfaction among their passengers.

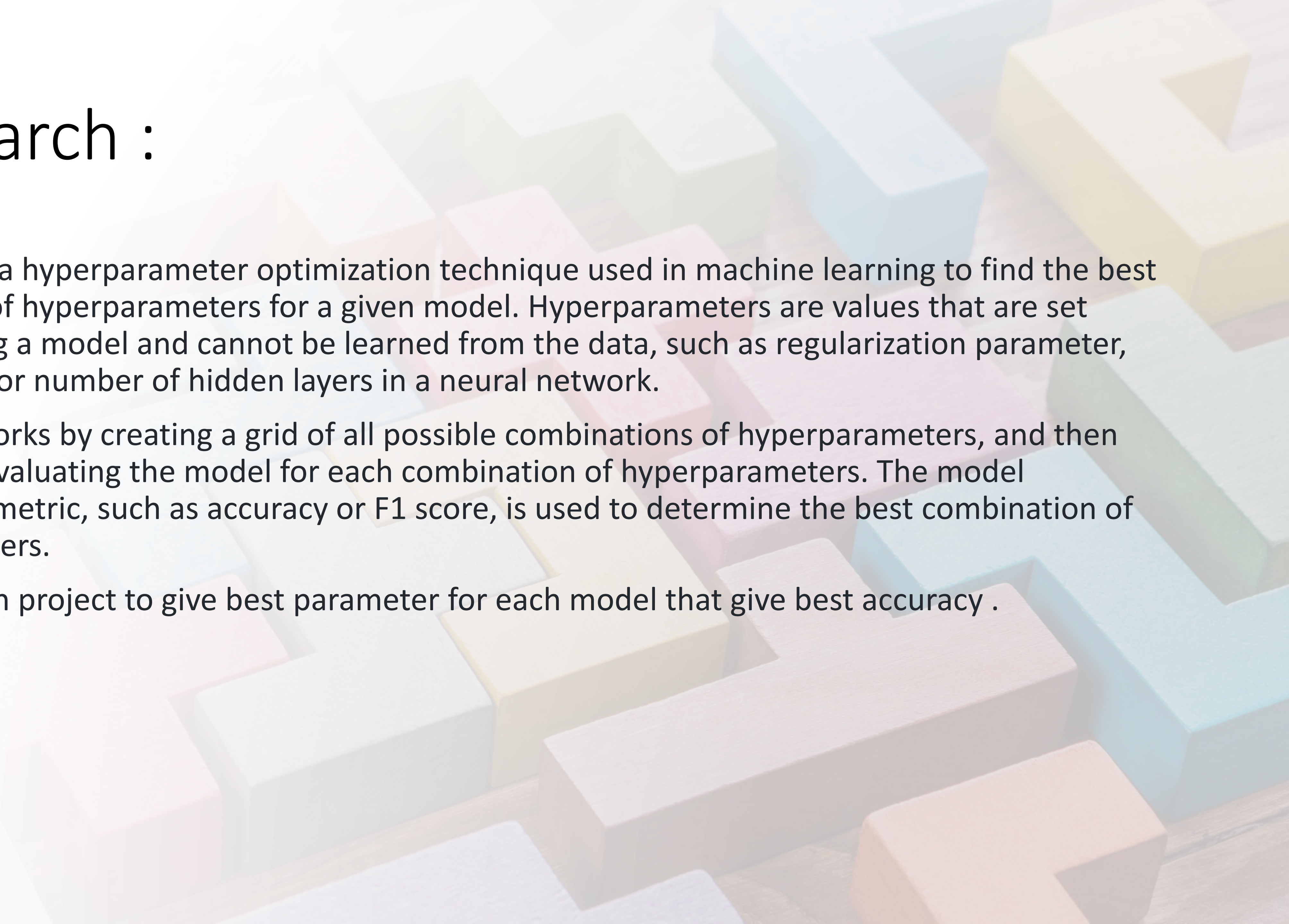
Data Preprocessing :

- Data Cleaning : we cleaned our data by Drop the unneeded rows and nan value, check if there any duplicated value and check if there is any outliers

- Data Encoding : we encoded our features with label encoding and one hot encoding

- Data Scaling

Grid Search :

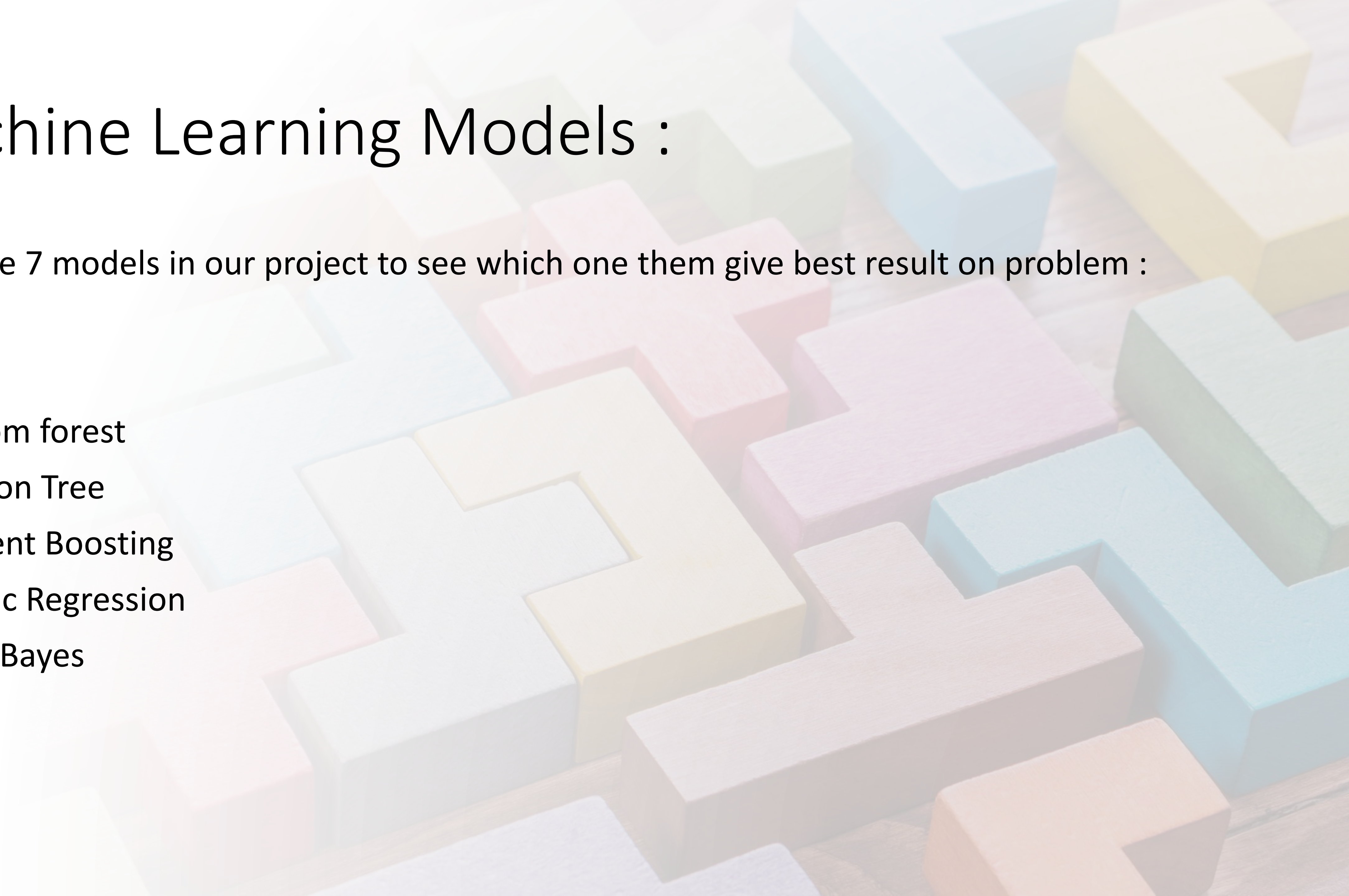


- Grid search is a hyperparameter optimization technique used in machine learning to find the best combination of hyperparameters for a given model. Hyperparameters are values that are set before training a model and cannot be learned from the data, such as regularization parameter, learning rate, or number of hidden layers in a neural network.
- Grid search works by creating a grid of all possible combinations of hyperparameters, and then training and evaluating the model for each combination of hyperparameters. The model performance metric, such as accuracy or F1 score, is used to determine the best combination of hyperparameters.
- So, we use it in project to give best parameter for each model that give best accuracy .

Machine Learning Models :

- We use 7 models in our project to see which one them give best result on problem :

- K-NN
- SVM
- Random forest
- Decision Tree
- Gradient Boosting
- Logistic Regression
- Naïve Bayes



K-NN :

K-Nearest Neighbors (KNN) is a machine learning algorithm used for both classification and regression problems. It is a non-parametric algorithm, meaning it makes no assumptions about the underlying distribution of the data. Instead, it simply compares a target instance to its k nearest neighbors in the training set and predicts the class or value based on the majority label or average value of those neighbors.

In other words, given a new input data point, KNN finds the k closest points in the training set and assigns the output value of the new point as the mode (for classification) or the mean (for regression) of the output values of the k closest points. The value of k is typically chosen through cross-validation techniques.

In our project we use knn for classification and use neighbors =5 ,p=1 after using grid search

this was the best value for parameters , knn model give an accuracy for training of 94.8%

and 92.7 % for validation .

SVM :

Support Vector Machine is a type of supervised learning algorithm that can be used for both classification and regression tasks. In the case of classification, SVM tries to find a hyperplane that best separates different classes of data points in the feature space.

SVC (Support Vector Classification) is one implementation of SVM for classification tasks. It works by finding the hyperplane that maximizes the margin between the closest data points of different classes. The data points closest to the hyperplane are called support vectors.

In our project we use svc ,svm model give an accuracy for training of 92.4% and 92.3 % for validation

Random Forest :

Random forest is a supervised machine learning algorithm used for classification, regression and other tasks. It is an ensemble learning method that combines multiple decision trees to create a more accurate and stable model. Random forest works by constructing a multitude of decision trees at training time and outputting the class (in case of classification)

Each decision tree in the random forest is constructed using a different subset of the training data and a different set of features, randomly selected from the total pool of features.

In our project we use Random forest classifier for classification and use max depth =12 after using grid search this was the best value for parameters , the model give an accuracy for training of 96% and 95.3 % for validation .

Decision Tree :

A decision tree is a popular machine learning algorithm used for both classification and regression analysis. It is a tree-like model where each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label or a numerical value that is the result of the classification or regression

The decision tree algorithm works by recursively partitioning the data into subsets based on the value of the attributes, with the goal of maximizing the homogeneity of the target variable within each subset. The process continues until a stopping criterion is met, such as a maximum depth of the tree or a minimum number of observations in a leaf node.

In our project we Decision Tree Classifier for classification and use criterion= “entropy ”, max_depth=15, splitter=“random ” after using grid search this was the best value for parameters, the model give an accuracy for training of 96.6% and 95.2 % for validation , and this is the best model

Gradient Boosting :

Gradient Boosting is a supervised machine learning algorithm that is used for both regression and classification problems. It is an ensemble method that combines multiple weak learners to create a strong learner. The idea behind Gradient Boosting is to leverage the strengths of decision trees while minimizing their weaknesses, such as overfitting.

The algorithm works by iteratively adding decision trees to the model, where each new tree corrects the errors made by the previous one. This is done by computing the gradient of the loss function with respect to the predictions of the current model, and using this information to fit a new tree that predicts the residual errors. The predictions from all the trees are then combined to obtain the final prediction.

In our project we use Gradient Boosting Classifier for classification and the model give an accuracy for training of 94.3% and 94.4% for validation .

Logistic Regression :

- Logistic Regression is a statistical method used to analyze and model the relationship between a dependent variable (also known as the response variable) and one or more independent variables (also known as predictor variables), where the dependent variable is categorical. The goal of logistic regression is to predict the probability of the dependent variable belonging to a particular category, given the values of the independent variables.
- It is called "logistic" regression because it uses the logistic function (also known as the sigmoid function) to transform a linear equation into a range from 0 to 1, representing the probability of the event occurring. The logistic function has an S-shaped curve that increases rapidly at first and then levels off as the probability approaches 1.

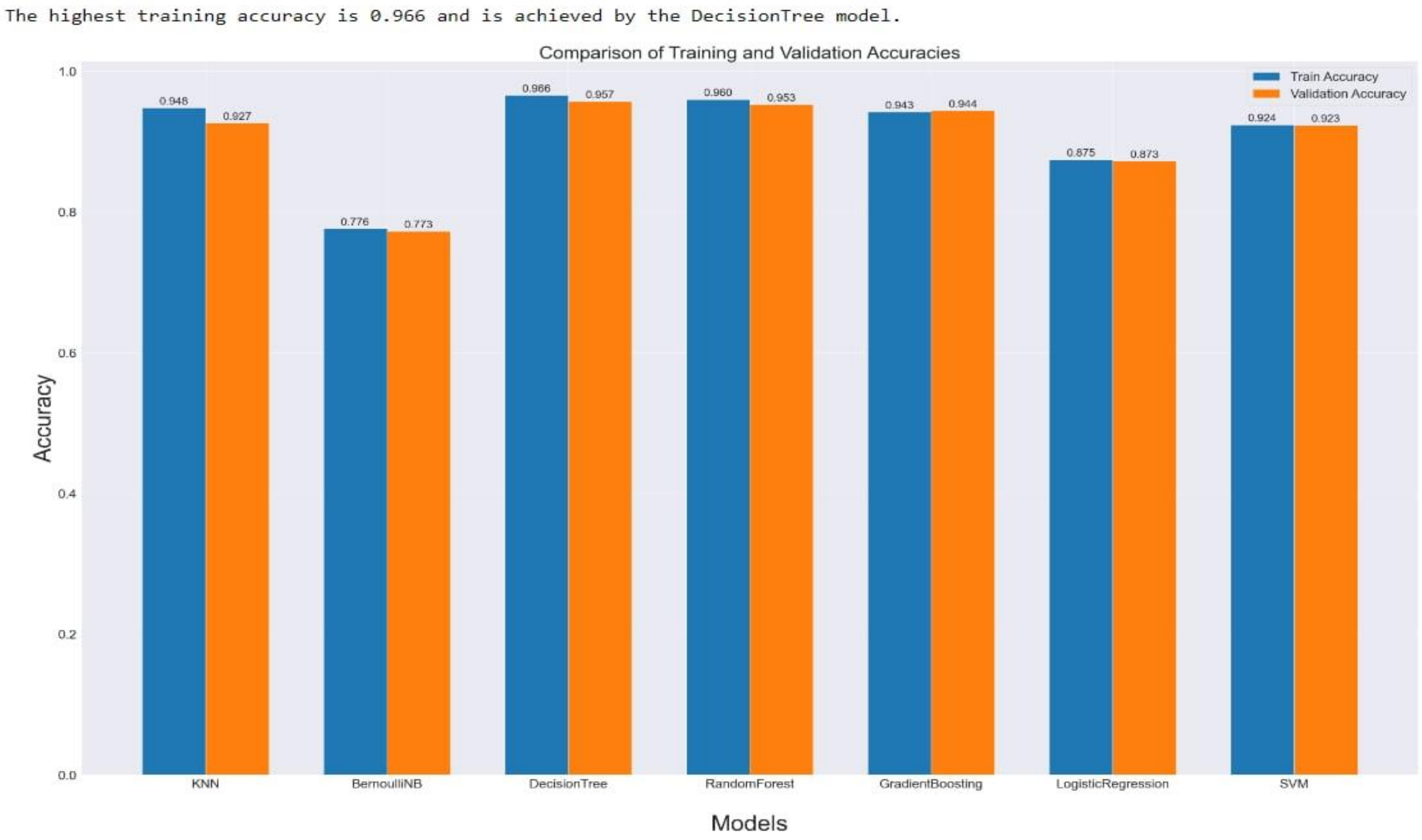
In our project we use Logistic Regression for classification and use multi_class=“multnomial” after using grid search this was the best value for parameters the model give an accuracy for training of 87.5% and 87.3% for validation .

Naïve Bayes:

- Naive Bayes is a popular machine learning algorithm for classification based on Bayes' Theorem, which describes the probability of an event occurring based on prior knowledge or information. In Naive Bayes, the assumption is made that the features (or attributes) used to classify instances are independent of each other, hence the name "naive".
- The algorithm works by calculating the probability of each class given a set of input features, using Bayes' Theorem, and then selecting the class with the highest probability as the predicted class. This is done by multiplying the conditional probabilities of each feature given the class, and then multiplying this value by the prior probability of the class.

In our project we use BernoulliNB for classification and use alpha =0.1 ,binarize=0.5 after using grid search this was the best value for parameters, the model give an accuracy for training of 77.6% and 77.3% for validation .

Comparison of training and validation accuracies



```
The highest training accuracy is 0.966 and is achieved by the DecisionTree model.
```

Team members

- عبدالرحمن أيمن دسوقي محمد
- أحمد ناصر امام
- احمد محمد مصطفي عبدالكريم
- ريهام ابراهيم عبدالدايم سليمان
- منة الله أحمد محمد صلاح الدين
- سهيلة ابراهيم رمضان

Thank You