**Transcript:**

Hello, everyone. My project name is the study of best fitted model to optimize machine learning accuracy and my name is Muhammad tamim ikbal and my ID is 230-66608.

First, you can see the table of content that containing machine learning objectives machine learning and the techniques explaining three models and how they work Implementing these three models on a particular data set Resultant analysis, conclusion and references.

So the first thing is the first thing Learning objectives. Then, we'll understand the basic machine learning Explore 3 machine learning techniques and their performance and then learn how this three model work And the operations evaluate these models and analyze the performance using a particular dataset.Learn feature engineering and preprocessing using this data set and then learn about the best for fitted model and the result of it.

So the first question is. What is machine learning?

So in one word, machine learning is a particular field of artificial intelligence that learn from data and try to make prediction or decision without being explicitly programmed. Here is four machine learning technique which is commonly used or supervised learning, unsupervised learning reinforcement learning and the semi-supervised learning.

So supervised learning. In this learning model is trained with level data.On the other hand, the unsupervised learning, it identifies the hidden patterns from unlabeled data.And in reinforcement learning, you try to learn from reward penalties and sometimes from feedback.And in semi-service learning, it mainly combines level data and unlabeled data for the training.

So now our three targeted machine learning techniques are decision tree, random forest and adaboost. We'll see the working procedure and we'll implement this tree model on a particular dataset later.

So the first one is decision tree. So, decision tree is mainly a supervised learning which is commonly used for classification and regression task so the classification mainly it shows the outcome in binary type of firm like Yes or no. For an example.The sender it can be female or male So the outcome will be like yes or no and the regression mainly it shows the outcomes in continuous form for examples my height Or my weight.

So this model mainly predicts the outcome is like a tree structure. That's why it's called this decision tree. It has mainly the root nodes, the branches and the leaf nodes.

And the leaf node mainly represents the outcome where The nodes mainly represent the features and the eades mainly represent the decision rule like which brands it should go the next.Here we can see the block diagram of decision tree where on the top we can see the root node here mainly the the features of the data set mainly split on the based on the criteria like Gini impurity or the information gain and it is split it into some branches .And it recursively splitting until it meets with the leaf node, here you can we can see in the yellow box there's the leaf node And this leaf nodes mainly shows the outcomes.This is how it mainly works.

So the second one is random forest So random forests mainly are ensemble machine learning method or technique that mainly uses like nth number of decision trees to predict outcome Here, all of this predict the Outcome separately then this random forest try to use the voting or aggregation for the outcomes we can see here is the block diagram random forest where the main data set is splitting into some training data And this training data mainly trained by The separate n number of decision tress and this comes to the voting box or aggregation section Well, for classification, measure to voting are the outcome of the prediction and for regression like averaging this outcomes from each of the decision tree is mainly the final predictions.So here we can see the steps like data sampling, feature selection training the trees and then aggregating and the final outcomes this is how the random twist mainly works And the third one is ada boost.

So AdaBoost is a short form of Adaptive boosting this is also enzyme technique combines with some weak liners And this weak learners are mainly decision trees is called stumps. So in addition 3, it tried to adjust the weight of every stump or the wheat liners and try to make it Try to make it stronger by adjusting the weights of the previous model's mistake.So here we can see the block diagram ,it is mainly a sequential model and here we can see first this model tried to initialize some words for every model and you try to adjust the next model from the previous one's mistake.So in here we can see from model one it tried to adjust model two from the weakness or from the mistake it made in model two and it goes on until the last model becomes stronger and shows us the predictions.

Now, we'll implement these three models on a particular data set that I have taken from kegel So this data set mainly consists of 1,259 records and 27 features.This data set mainly contains information about the mental health of some employees at their workplace and at their workplace how these factors are correlated to each other.So our task will be we'll see how many of these employees are receiving the treatment and how many of and not. And we'll implement three of this model and try to evaluate these models.

So in here we can like like first raws or records in here of this data set and 27th of column from this data set we can see there is like binary categorical and the numerical feature are available on this data set.

So now we'll implement this model for the particular three For this, we'll implement this particular three models on this data set and for this implementation we have to take some steps you can see here Like, first of all Two, the data processing and then class distributionAnd data splitting then Model training and evaluation, and then we'll visualize. So data preprocessing. In data preprocessing, perform data cleaning and cleaning data handling and data encoding.

So data cleaning i have seen here like there is some unwanted columns like time stamp comments and state this is and not related with our particular outcome so this way i have removed all this unwanted column then I have filled some missing values on this field

And I have… encoded this data to categorical to numerical.

And then class distribution. So our targeted variable or the feature is treatment and from counter plot, we can see the class distribution here, like almost 650 is uh no they're not getting the treatment and like a little bit more is a little bit more ,For yes they're getting the treatment And then I have performed data splitting data So in data splitting, I have taken 80% of this data for training and 20% data for the testing.

So after that, I have trained all of these three models And evaluate this model.

So from this model. I can see clearly in there the different trees classification report is here The accuracy rate is 62% and the recall precision and the F1 score is almost the same.

And then I have also trained the random forest and evaluate and their accuracy is 75 percent And similarly, I have trained and more evaluate the adaboost and their accursed rate is nearly same to random forest which is 74.2%. Now, result analysis.

So this project My aim to analyze this data set and optimize the accuracy of the employees which are receiving their mental health based on their work plus factor is receiving or not.

So, and find out the best model so from this uh from this from this uh from this three model i have seen like the addition tree model showed the moderate performance so over performance of this model is 62.7%.

On the other hand, the random forest is achieving the highest accuracy rate which is 75 person and And we all know that in dysentery there is some overfitting problem which is uh it shows like better performance is training but shows like lack of performance

The performance in testing it reduced in random forest you have seen like random for us to improve this overfitting problem and achieved the higher accuracy Because the arabos is performed almost well and it's nearly to the random forest but Its accuracy was 74.20%. So from all of this you can see that random forest achieve the highest accuracy rate.

So now the conclusion. So in conclusion I can say that random first is the best fitted model for this particular data set.

So for this type of particular data set we can see that like random forest shows the better performance than those other though here um Data Boost is performed well and it's like slightly lower than random forest but um The best model is random forest for this particular situation.

So this is all of the project and overall the random forest is showing the better performance So thank you everyone for listening. And here we can see the references what from where I have taken all of my information’s.

Thank you, everyone.