

Career prediction using Decision tree algorithm and pass prediction using linear regression for higher secondary school students

Febin Babu
PG Scholar
Department of Computer Applications
Amal Jyothi College of Engineering
Kanjirappally, India
febinbabu2023a@mca.ajce.in

Ms. Meera Rose Mathew
Assistant Professor
Department of Computer Application
Amal Jyothi College of Engineering
Kanjirappally, India
meerarosemathew@amaljyothi.ac.in

Abstract— Career prediction is an essential issue that students face when deciding on their future education and career paths. In this seminar presentation, we will discuss the concept of career prediction using the decision tree algorithm, and also pass percentage prediction using linear regression algorithm, powerful tools for analyzing and predicting different data patterns. We will explain how this method can be used to predict a student's career path based on their academic performance, interests, and skills. In this presentation, we will start by introducing the basics of machine learning and decision tree algorithms, and how they can be applied to career prediction. We will then explore the various factors that are commonly used in career prediction models, such as academic performance, interests, skills, and personality traits. We will also discuss another algorithm for predicting the pass percentage of the students by using the linear regression algorithm. importance of data collection and analysis in building accurate career prediction models. we will provide some successful examples of career prediction models and discuss the limitations and challenges of using machine learning technologies for career prediction.

Keywords- *Decision Tree algorithm, Career prediction, Linear Regression, pass percentage prediction.*

I. INTRODUCTION

The idea of career prediction and pass% prediction utilizing the decision tree algorithm and linear regression algorithm will be covered in this section. Based on a student's academic achievement, hobbies, and skills, we may accurately anticipate their professional path with the use of these potent tools for analyzing and forecasting various data patterns. We will examine the many components, such as academic performance, hobbies, abilities, and personality traits, that are frequently included in career prediction models and pass prediction models, as well as the significance of data gathering and analysis in creating precise models. We will also examine the drawbacks and difficulties of applying machine learning technologies for career prediction, as well as successful instances of career prediction models.

II. LITERATURE REVIEW

Decision tree algorithm has emerged as a powerful machine

learning technique for decision making and prediction, particularly in the context of career prediction. The efficiency of decision tree algorithms in predicting career paths and selecting viable jobs for students based on a variety of characteristics, including academic achievement, interests, abilities, and personality traits, has been the subject of multiple studies in recent years.

Operational failures are closely related to many interest groups that exist both inside and outside of the companies. While continuing to operate, businesses may experience financial failure due to both internal and external economic and market conditions.

In these conditions [1], failure to manage the risks they face could result in bankruptcy. Because of this, businesses should be able to anticipate potential failures and take appropriate action by assessing their existing state. To estimate and categories the financial failures of organizations operating in diverse industries, a model utilizing artificial neural networks (ANN) and decision trees (DTs) is developed.

Cutting-edge in-vehicle technology [2] requires extensive study on risk estimation and cut-in features. These objectives are somewhat attained by 24 participants were selected to participate in the multi-driver simulation trials that were utilized in this study to collect data on risky driving. The threat posed by cut-ins was determined using the substitute measurements Time Exposure Time-to-Collision (TET) and Time Integrated Time-to-Collision (TIT). The risk was then divided into three groups using K-means clustering. Among the numerous candidate variables of two types, ten behavioral factors and seven driver trait variables were extracted. Three prediction models include the long-short-term memory (LSTM), gradient-boosting decision tree (GBDT), and decision tree (DT).

Transporting massive amounts [3] of freight across great distances with ocean shipping is an alternative logistics method. But the logistical planning for this mode is expensive due to the unknowns, such as climatic conditions, cargo

types, and port characteristics. As a result, determining how long ships will stay becomes a key goal for the waterway's planning and scheduling. The management of the port faces a dilemma in figuring out how long it can run the ship based on how long ships are typically expected to stay docked. To build a model for estimating the ship stay time using decision tree model-based algorithms, we gathered data on the major cargo flows in Brazilian ports in 2018 for the current study.

If thermal power reactors are to be operated independently [4], effective control of the startup and power-up procedures is necessary. An autonomous operating system's decision-making procedures must promptly foresee and suppress unwelcome power swings.

Medical Information Mart from the MIMIC-III [5] database for intensive care These were randomly split into two sets: one for training and the other for independent testing. The following information was gathered: demographics, admission details, vital signs, laboratory test results, critical illness scores, medications, comorbidities, and intervention methods. For early AKI occurrence prediction and feature extraction, ensemble models, random forest, LightGBM, XGBoost, and logistic regression were utilized. The SHAP analysis was used to determine how each feature's prediction would be affected.

We take into account how users' goals and background knowledge [6], which is available to them, can affect how they interpret and prefer the textual justifications created for the results predicted by decision trees (DTs). We create contrastive explanations that address potential conflicts between elements of DT predictions and plausible expectations authorized by background knowledge in order to examine the impact of background information. We identify four different forms of conflict, propose operational methods for identifying them, and outline explanatory models that handle each type. We use an interactive environment where, given a goal and an initial explanation for a predicted outcome, users select follow-up questions and evaluate the explanations that address these questions in order to examine the influence of users' goals.

Finding cluster structures and creating regression models [7] are two key problems that the fields of data analysis and machine learning frequently run into. Mixed-integer linear programming (MILP) has become a successful approach to solving these problems. Clustering and regression can be effectively solved using commercially available solvers if they are formulated as MILP problems. Bivariate cluster-wise linear regression (CLR) and piecewise linear regression (PWLR) are two modern methods that combine MILP into clustering and regression research.

Due to its simplicity and ease of use, [8] linear regression is a well-liked approach in machine learning for predictive

analysis. The model's success, however, is reliant on a number of presumptions, such as the linearity of the variables, homoscedasticity, normality, and measurement reliability, which may not always hold true in real-world data sets. Additionally, the performance of the model may be significantly impacted by inconsistent samples in the data set.

In summary, the literature review emphasizes how well decision tree algorithms predict career paths and suggest viable jobs for both students and workers. Recent research has shown that decision tree algorithms can be used to accurately forecast and recommend careers based on a variety of characteristics, including academic performance, personal preferences, industry trends, demographic data, and social network data.

III. MOTIVATION

Research on career prediction is crucial because it can offer professionals and students insightful information about their potential career trajectories. The decision tree method can be used to create precise models for career prediction by examining a variety of criteria, including academic achievement, hobbies, talents, and personality traits. Such models can help people make well-informed career decisions and can support businesses and educational institutions in guiding their employees and students. Policies and strategies for workforce development can be informed by career trends and patterns that are identified using the decision tree algorithm. In conclusion, by facilitating better professional decision-making, career prediction utilizing decision tree algorithms has enormous potential to benefit both individuals and society at large. Recent research has shown that decision tree algorithms offer a high degree of accuracy in predicting career pathways and selecting viable jobs for students. Decision tree algorithms are likely to continue playing a significant role in the field of career prediction, offering insightful information and suggestions to students and assisting them in making decisions about their future.

IV. METHODOLOGY

The methodology for predicting the career for students by using the decision tree algorithm involves the following steps:

A. Data Preparation: Academic information about students is gathered in a dataset. Preprocessing the data entails cleaning it, getting rid of outliers, and handling missing values. To assess the effectiveness of the decision tree method, the preprocessed data is divided into training and testing sets.

B. Model Training: The method builds a tree model from the training data that maps the input features to the associated output class labels, which stand in for several career routes. Recursively dividing the data into subgroups depending on the chosen features, the training procedure entails maximizing information gain or minimizing impurity with each split. After that, the tree model is trimmed to lessen overfitting and boost its ability to generalize to the testing data.

C. Model Evaluation: We use the testing data to evaluate the effectiveness of a decision tree algorithm for predicting career trajectories and the precision of the predictions made by the model. In order to do this, the projected class labels and the actual class labels must be compared, and metrics like accuracy score must be calculated.

D. Model Interpretation: In order to understand a decision tree method for career prediction, we look at the tree model to find the factors that have the biggest impact on the projected class labels. We examine the tree nodes' splitting criteria in order to identify the salient traits that help categories various career choices.

E. Prediction: Finally, we use the trained model to predict the students career using the decision tree algorithm for new data.

V. BUILD MODEL

For building a reliable and accurate machine learning model for the successful prediction of student performance we follow the following steps using linear regression algorithm.

Importing Libraries

```
from django.shortcuts import render, redirect
from django.contrib import messages
from .models import Student_Careerprediction
from sklearn import model_selection
from sklearn.ensemble import BaggingClassifier
from sklearn.tree import DecisionTreeClassifier
import pandas as pd
import numpy as np
import pickle
```

Perform Career Prediction

```
def careerprediction(request):
    if 'email' in request.session:
        email=request.session['email']
        data={'email':email}
        if request.method == 'POST':
            df = request.POST.get('df')
            ca = request.POST.get('ca')
            dcs = request.POST.get('dcs')
            cs = request.POST.get('cs')
            networking = request.POST.get('networking')
            sd = request.POST.get('sd')
            ps = request.POST.get('ps')
            pm = request.POST.get('pm')
            cff = request.POST.get('cff')
            tc = request.POST.get('tc')
            aiml = request.POST.get('aiml')
            se = request.POST.get('se')
            ba = request.POST.get('ba')
            cskills = request.POST.get('cskills')
            ds = request.POST.get('ds')
            td = request.POST.get('td')
            gd = request.POST.get('gd')
```

```
# Map values to numerical values
value_map = {"Not Interested": 1, "Poor": 2, "Beginner": 3, "Average": 5,
            "Intermediate": 6, "Excellent": 7, "Professional": 9}

df = value_map.get(df, 0)
ca = value_map.get(ca, 0)
dcs = value_map.get(dcs, 0)
cs = value_map.get(cs, 0)
networking = value_map.get(networking, 0)
sd = value_map.get(sd, 0)
ps = value_map.get(ps, 0)
pm = value_map.get(pm, 0)
cff = value_map.get(cff, 0)
tc = value_map.get(tc, 0)
aiml = value_map.get(aiml, 0)
se = value_map.get(se, 0)
ba = value_map.get(ba, 0)
cskills = value_map.get(cskills, 0)
ds = value_map.get(ds, 0)
td = value_map.get(td, 0)
gd = value_map.get(gd, 0)

Student_Careerprediction.objects.create(df=df, ca=ca, dcs=dcs, cs=cs, networking=networking, sd=sd, ps=ps, pm=pm,
                                       cff=cff, tc=tc, aiml=aiml, se=se,
                                       ba=ba, cskills=cskills, ds=ds, td=td, gd=gd).save()
```

```
# load the data
dataset = pd.read_csv('D:\Mainproj\T\SchoolProject\Schoolprjt\SchoolProject\Schoolapp\dataset9000.data', header=None)
X = np.array(dataset.iloc[:, 0:17])
Y = np.array(dataset.iloc[:, 17])
dataset.columns = ["Database Fundamentals", "Computer Architecture", "Distributed Computing Systems",
                  "Cyber-Security", "Networking", "Development", "Programming Skills", "Project Management",
                  "Computer Forensics Fundamentals", "Technical Communication", "AI ML", "Software Engineering", "Business Analysis",
                  "Communication skills", "Data Science", "Troubleshooting skills", "Graphics Designing", "Roles"]
dataset.dropna(inplace=True)

# initialize the base classifier
base_cls = DecisionTreeClassifier()

# no. of base classifier
num_trees = 50

# bagging classifier
model = BaggingClassifier(base_estimator=base_cls,
                          n_estimators=num_trees,
                          random_state=5)

# cross-validation
seed = 5
kfold = model_selection.KFold(n_splits=10, shuffle=True, random_state=seed)
results = model_selection.cross_val_score(model, X, Y, cv=kfold)
print("accuracy :", results.mean() * 100)
```

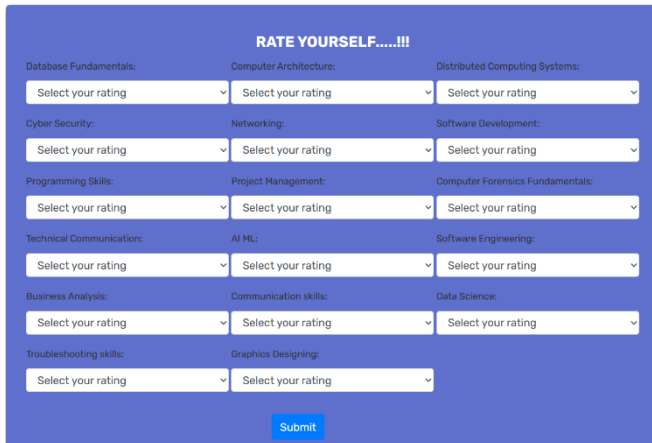
```
# fit the model to the training data
model.fit(X, Y)

# get input values from the Student_Careerprediction table
user_input = {}
user_input = [float(df), float(ca), float(dcs), float(cs), float(networking), float(sd),
                float(ps), float(pm), float(cff), float(tc), float(aiml), float(se), float(ba),
                float(cskills), float(ds), float(td), float(gd)]
print(user_input)

# predict the class using the trained model
predicted_class = model.predict([user_input])[0]
print("Predicted class: ", predicted_class)
temp=1

context={'predicted_class':predicted_class, 'temp':temp}
# add predicted class to the data dictionary and render the response
data['predicted_class'] = predicted_class
return render(request, "Student/student_careerprediction_cmptter.html", context)
else:
    temp=0
    context={'temp':temp}
    return render(request, "Student/student_careerprediction_cmptter.html", context)
temp=0
context={'temp':temp}
return render(request, "Student/student_careerprediction_cmptter.html", context)
```

VI. RESULT

A blue form titled "RATE YOURSELF.....!!!". It contains 12 dropdown menus for rating skills, arranged in a 4x3 grid. The skills listed are: Database Fundamentals, Computer Architecture, Distributed Computing Systems, Cyber Security, Networking, Software Development, Programming Skills, Project Management, Computer Forensics Fundamentals, Technical Communication, AI ML, Software Engineering, Business Analysis, Communication skills, Data Science, Troubleshooting skills, and Graphics Designing. A "Submit" button is located at the bottom right of the form.

VII.CONCLUSION

In conclusion, the decision tree algorithm is an effective tool that can assist companies in making defensible decisions based on massive amounts of data. Businesses can better comprehend the links between factors and forecast future events by building a decision tree based on pertinent features and traits.

Businesses can quickly and easily spot patterns and trends thanks to the decision tree algorithm's accurate and efficient analysis of large data sets. Businesses can optimize their strategy based on consumer and market demand by implementing decision tree analysis into their decision-making processes, which will enhance revenue and profitability.

The decision tree algorithm's capacity to handle both category and numerical input is one of its main advantages. As a result, it is a flexible tool that may be applied in a range of fields, such as marketing, healthcare, and finance. Businesses that need to swiftly and accurately analyse vast amounts of data are especially well-suited for decision tree analysis.

Overall, the decision tree algorithm is a useful tool that can assist companies in maintaining their competitiveness in a market that is continually changing. Businesses can benefit from useful insights into the interactions between variables by utilizing the power of decision tree analysis. They can then use this knowledge to make well-informed decisions regarding pricing, marketing, and other crucial business decisions.

REFERENCES

- [1] Tianyang Luo, Junhua Wang (2021). Risk prediction for cut-ins using multi-driver simulation data and machine learning algorithms: A comparison among decision tree, GBDT and LSTM
- [2] Levi R. Abreu, Ingrid S.F. Maciel , Joab S. Alves , Lucas C. Braga , Heráclito L.J. Pontes (2023). A decision tree model for the prediction of the stay time of ships in Brazilian ports
- [3] Mengqi Huang, Zhengyu Du, Yu Liu. (2023). Comparative study on peak power prediction methods during start-up and power-up of heat pipe reactor based on neural network and decision tree
- [4] Jian Zhou, Shuai Huang, Ming Tao, Manoj Kandelwal, Yong Dai, Mingsheng Zhao. (2023). Stability prediction of underground entry-type excavations based on particle swarm optimization and gradient boosting decision tree
- [5] Wenpeng Gao, Junsong Wang, Lang Zhou, Qingquan Luo, Yonghua Lao, Haijin Lyu , Shengwen Guo. (2022). Prediction of acute kidney injury in ICU with gradient boosting decision tree algorithms
- [6] Nezir Aydin, Nida Sahin, Muhammet Deveci, Dragan Pamucar, (2022). Prediction of financial distress of companies with artificial neural networks and decision trees models.
- [7] John Alasdair Warwicker , Steffen Rebennack (2023) A unified framework for bivariate clustering and regression problems via mixed-integer linear programming.
- [8] Rasyidah , Riswan Efendi, Nazri Mohd. Nawi, Mustafa Mat Derisf, S.M.Aqil Burney (2023) Cleansing of inconsistent sample in linear regression model based on rough sets theory.