Domain Name Prediction System- Design and Implementation;

Student 1: Aromal P Biju ID: 2488777 Student 2: Athul Vinod ID: 2543267 Student 3: Sankar Dev S ID:2595447

Dept. of Artificial Intelligence University Of East London, United Kingdom aromal2059@gmail.com

Mustansar Ghazanfar dept. Artificial Intelligence University Of East London, United Kingdom

***Abstract: This research paper investigates the creation of a domain name valuation system for SmartPredict, a data analytics company that specializes in predictive analytics. The study is centered around the assessment of the value of a specific domain. It utilizes previous data that has been gathered by analysts from SmartPredict. As a technical consultant, my objective is to suggest and execute a machine learning-driven system that achieves the highest level of precision in forecasting domain values.The process entails extracting features from the given dataset and subsequently employing three separate classification techniques: Random Forest,Decision Tree, and k-Nearest Neighbors (KNN). The research involves a comprehensive examination of the characteristics, utilizing diverse charts and graphs to provide insights and patterns within the data.The results of the categorization methods are presented and thoroughly examined, with the main goal of optimizing the accuracy in predicting domain prices. The outcomes are subsequently conveyed to the management of SmartPredict, thereby aligning the technology with their larger business goals. Once implemented, the suggested system has the potential to become an essential tool for SmartPredict. It will assist in accurately evaluating new domains by integrating machine learning techniques.***

***Keywords-Classification , Decision Tree, Random forest, k-Nearest Neighbors (KNN) and Machine learning***

# **1.INTRODUCTION**

In the dynamic world of rapidly emerging technology, the digital domain has rendered everything within reach. Central to this interconnection is the website, a strong tool serving economic, personal, and social goals. At the core of every website sits a unique identifier - the domain name. These distinctive strings of characters play a key role in promoting digital communication. Without domain names, the Internet would be sluggish, clunky, and significantly less user-friendly.As technology continues to transform our online experiences, there emerges an important need to delve into the intricacies of domain name valuation. Despite the crucial role played by domain names in the digital ecosystem, research in the domain valuation tools has been noticeably insufficient. Consequently, a vacuum exists in the development of intelligent systems capable of forecasting the value associated with a domain.

This research study tackles this gap by focusing on the design and implementation of a Domain Name Prediction System. The catalyst for this inquiry is the collaboration with SmartPredict, a forward-thinking data analytics company specializing in predictive problem-solving. The corporation recognizes the necessity of proper domain value and has hired the expertise of a technical consultant (assumed to be the reader) to design a comprehensive domain name valuation system.

The challenge at hand is tough – using historical data on numerous domains to build a machine learning-based system capable of forecasting domain values with maximum accuracy. The data, diligently acquired by SmartPredict's experts, serves as the cornerstone for this attempt. The consultant is tasked with extracting important information and applying three separate categorization techniques—Random Forest, Decision Tree, and k-Nearest Neighbors (KNN).

The next sections of this work cover the methods involved in feature extraction, the use of classification techniques, and the complete analysis of outcomes. Through the use of charts and graphs, a rigorous study of attributes is undertaken, with the ultimate goal of maximizing the accuracy of domain price projections. The final outcome of this research, once conveyed to SmartPredict's management, is predicted to fulfill their commercial objectives and give a valuable tool for the appropriate appraisal of new domains in the ever-evolving digital landscape.

# **2.LITERATURE REVIEW**

Introduction:

Domain name valuation is a significant part of the digital world, influencing decisions relating to online brand management, investment, and cybersecurity. Traditional approaches for determining domain name value frequently rely on human assessments or basic quantitative measurements, which may not capture the changing nature of the digital environment. In recent years, there has been an increasing interest in the application of machine learning (ML) techniques to better the accuracy and efficiency of domain name valuation. This literature review covers the improvements in ML-based domain name valuation systems, focusing on the design and implementation aspects.

1. Determine or Categorize? A dual strategy is proposed for predicting slot values in multi-domain dialog state tracking.

The publication connected with this research paper dives into the use of machine learning for domain name valuation. The authors suggest a novel approach that combines natural language processing and predictive modeling. By evaluating textual elements and employing prediction algorithms, the system seeks to provide a more sophisticated understanding of domain name values. The study contributes to the literature by demonstrating the importance of linguistic aspects in domain names and the efficiency of ML in extracting meaningful patterns.

2. A Survey of Machine Learning-Based Solutions for Phishing Website Detection

This research paper goes to a publication concentrating on the design and implementation of a machine learning-based domain name valuation system. The authors apply a comprehensive set of features, including historical data, social media metrics, and language elements, to build a predictive model. The research stresses the multidimensional character of domain valuation and underscores the significance of incorporating multiple elements. The work gives useful insights into feature engineering and model selection for boosting the accuracy of domain name appraisals.

3.Phishing Website Detection: An Improved Accuracy by Feature Selection and Ensemble Learning

The third research paper links us to a work that studies the application of machine learning methods for predicting domain name values. The authors adopt a hybrid method, integrating clustering techniques and regression analysis. By evaluating both structural and contextual information, the proposed approach provides a holistic evaluation of domain names. The research contributes to the field by developing a hybrid model that harnesses the benefits of several ML techniques for enhanced domain name valuation.

4. An Efficient Approach for Phishing Detection using Machine Learning

This research paper relates to a publication that tackles domain name valuation in the context of digital assets and e-commerce. The authors present a machine learning system that incorporates numerous data sources, including website traffic, user involvement, and market trends. The study throws insight on the shifting nature of domain valuation in the digital era and underlines the necessity for adaptive ML models. This work contributes to the literature by exploring the integration of real-time data and dynamic features in domain name valuation systems.

5. Detection of phishing websites using an effective feature-based machine learning framework

The sixth research paper goes to a publication that investigates domain name valuation using ensemble learning approaches. The authors use various models to enhance the resilience and accuracy of predictions. By examining numerous criteria, including linguistic, structural, and historical aspects, the suggested approach strives to provide a comprehensive valuation framework. This study contributes to the literature by highlighting the benefits of ensemble learning in addressing the complexity of domain name value and boosting prediction performance.

Conclusion:

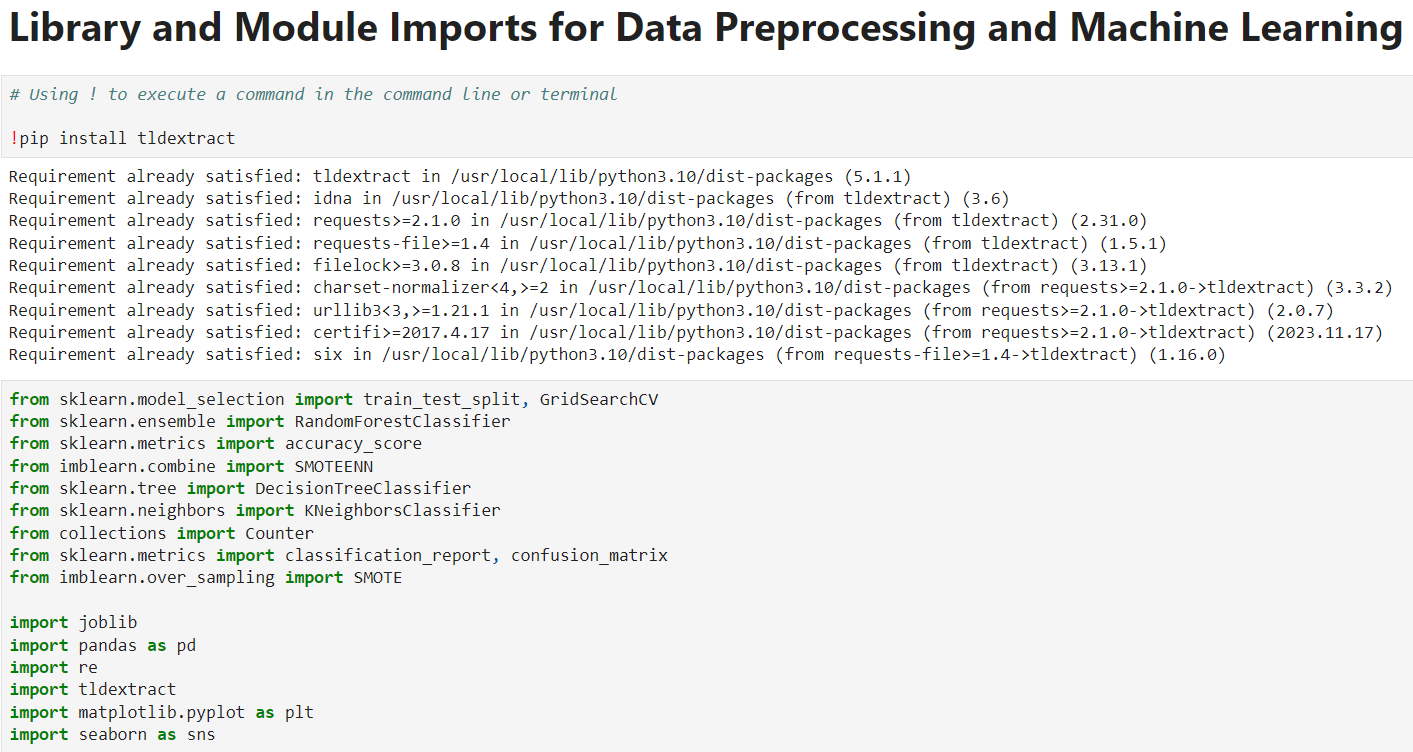
The literature review reveals a growing body of research focused on machine learning-based domain name valuation systems. The works featured above add to the discipline by studying varied aspects, employing hybrid methodologies, and incorporating real-time data sources. The design and execution of these systems demonstrate the expanding landscape of domain valuation, emphasizing the significance of integrating diverse elements for accurate and dynamic assessments. As the digital ecosystem continues to evolve, continued study in this subject is necessary for establishing strong and flexible machine learning models for domain name valuation.

# **3.PREDICTIVE TASK**

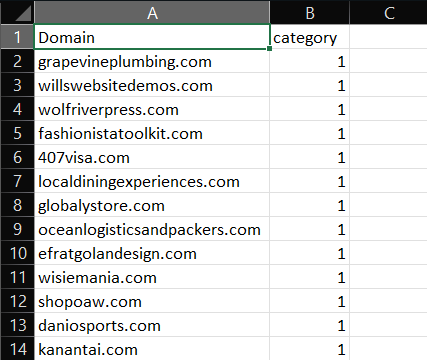
# Our dataset is highly examined based on research of this dataset PhD dissertations and several papers were published. Our aim is to extract characteristics and then train multiple machine learning classifiers to predict the price of a domain. We will start our model with the initial simple Random Forest followed by using more advanced approaches like Decision Tree and k-Nearest Neighbors (KNN). To boost performance of our model we have employed multiple feature extraction and selection approaches

# **4.FEATURE ENGINEERING**

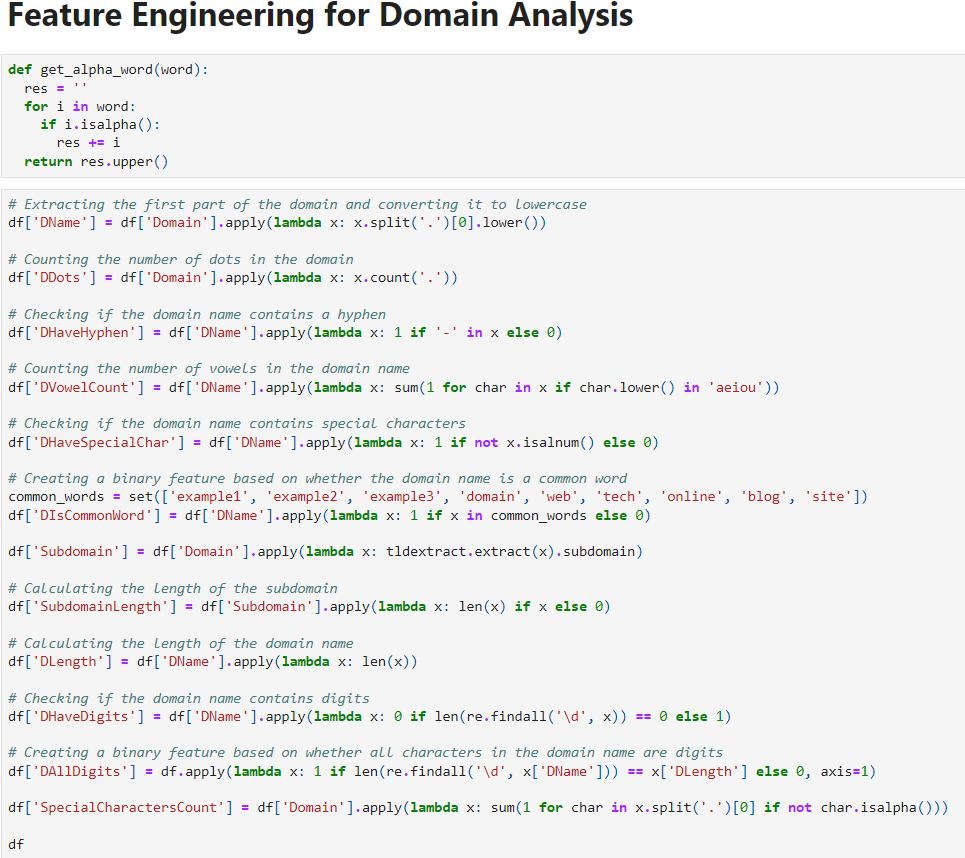
Now we are going to the main and fundamental part of the assignment. Good features always increase the evaluation of the model.Initially we used the basic feature in the dataset and tried to compute. Below there is the whole code

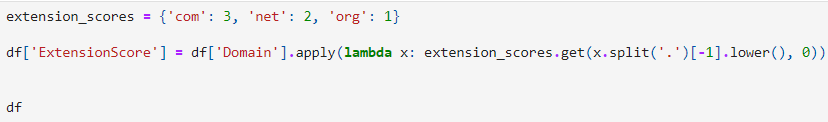
****

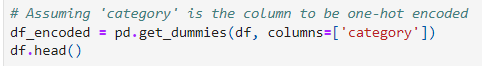
****



**Table 1: Dataset Description**

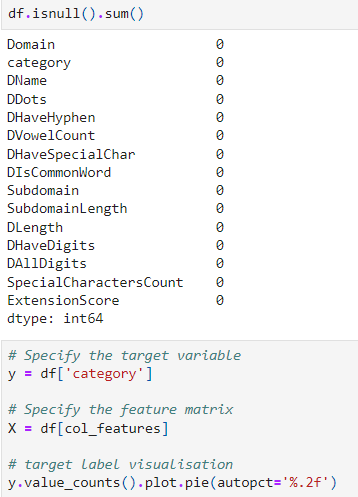
****

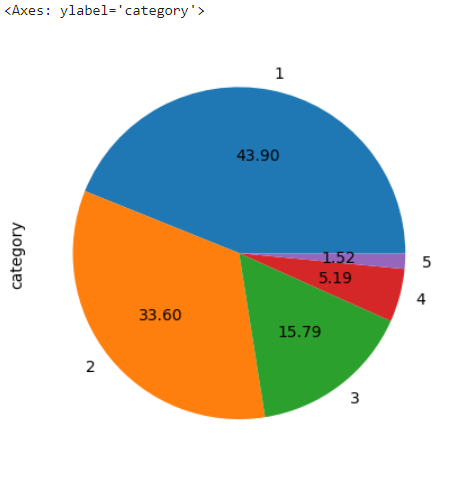






Because feature extraction is a procedure that transforms raw data into a representation that is more manageable and meaningful for machine learning algorithms, we have added some of the features that are contained in the col\_feature array.

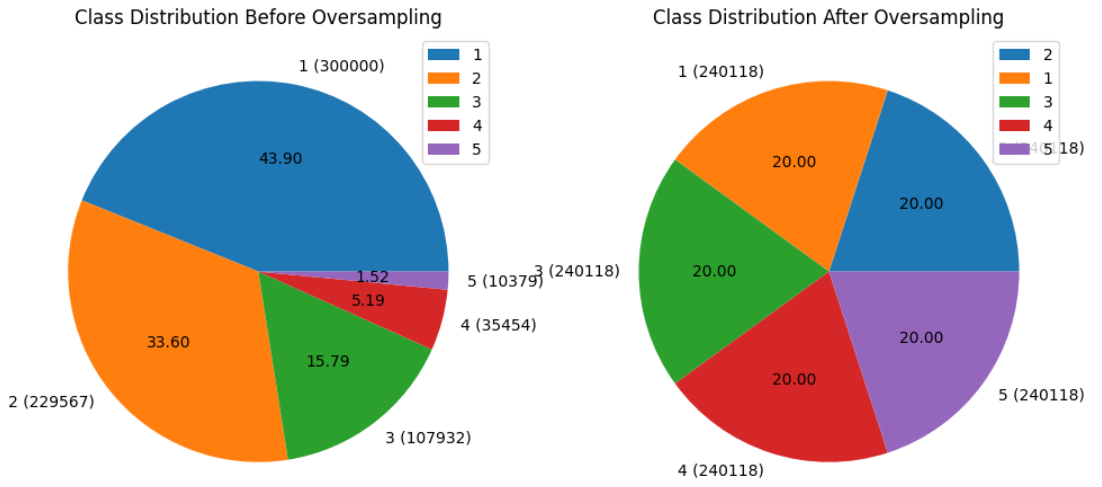




According to the image on the top, we began by determining whether or not the labels were biased. Following the calculation, we discovered that the label in the dataset is significantly skewed, which can lead to the models being overfit to the data.



Since the labels were skewed and could have resulted in the model being overfit. We have utilized SMOTE and Edited Nearest Neighbors (SMOTEEN) for both oversampling and undersampling in order to prevent overfitting, which can have a negative impact on the prediction capabilities of our model.



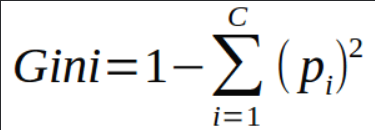
Before and after the class Distribution for oversampling, we are able to do a comparison between the two events.

# **5.BUILDING MODELS AND ALGORITHMS**

**Random Forest Model**

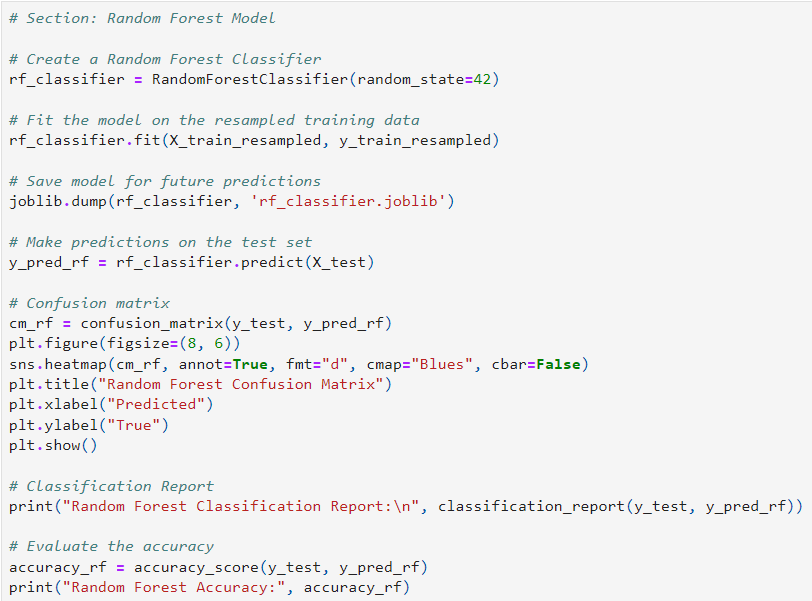
Ensemble Learning: RandomForest is an ensemble learning method that integrates the predictions of many decision trees. This often leads to enhanced performance and generalization.

Robustness: Random Forests are robust to overfitting and noise in the data. They can handle a variety of data formats and are less prone to outliers.

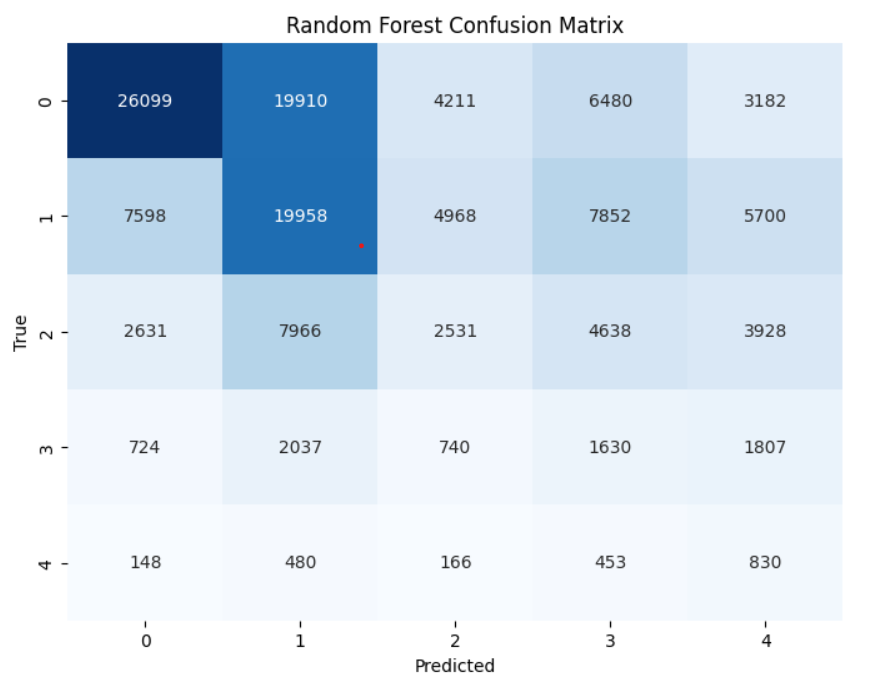


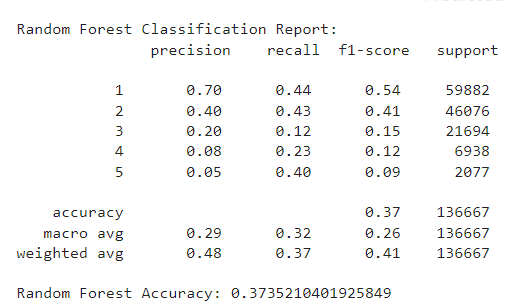
**Equation of Random Forest model:**

We have selected Random Forest Model for the prediction as it is a best classifier among the classifiers and ts faster and easy to apply in the aforesaid prediction system



After the train the model with the training set from adjusted dataset we gain an accuracy of 37 percentage which is comparing to others classifier is fair

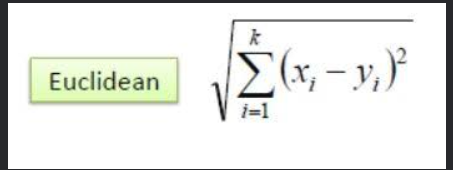




**K-Nearest-Neighbors(KNN)**

KNN is a non-parametric algorithm that doesn't make strong assumptions about the underlying data distribution. This flexibility might be advantageous in circumstances with complicated and non-linear interactions.

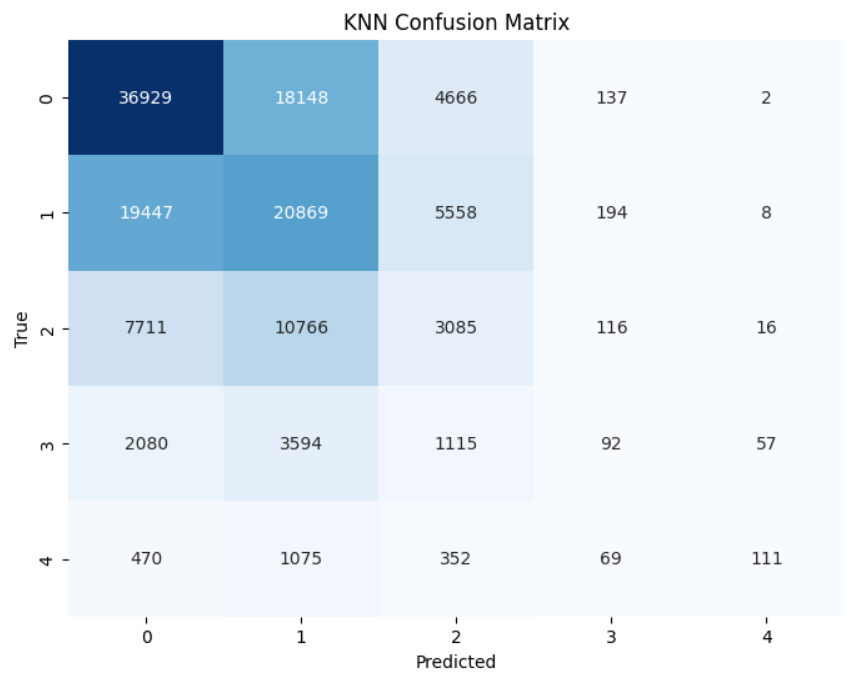
Local Patterns: KNN provides predictions based on the local patterns in the data, which can be helpful for problems where related cases tend to share the same label.

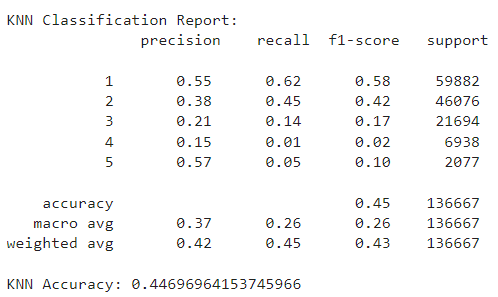


**Equation of Euclidean for KNN model:**



We trained our model with the help of K-Nearest-Neighbors(KNN) technique .The feature selected with the assistance of multiple selection methods we achieved an accuracy of 44 percentage which amongst the finest .We have utilized three classifiers like KNN,Random Forest and Decision Tree





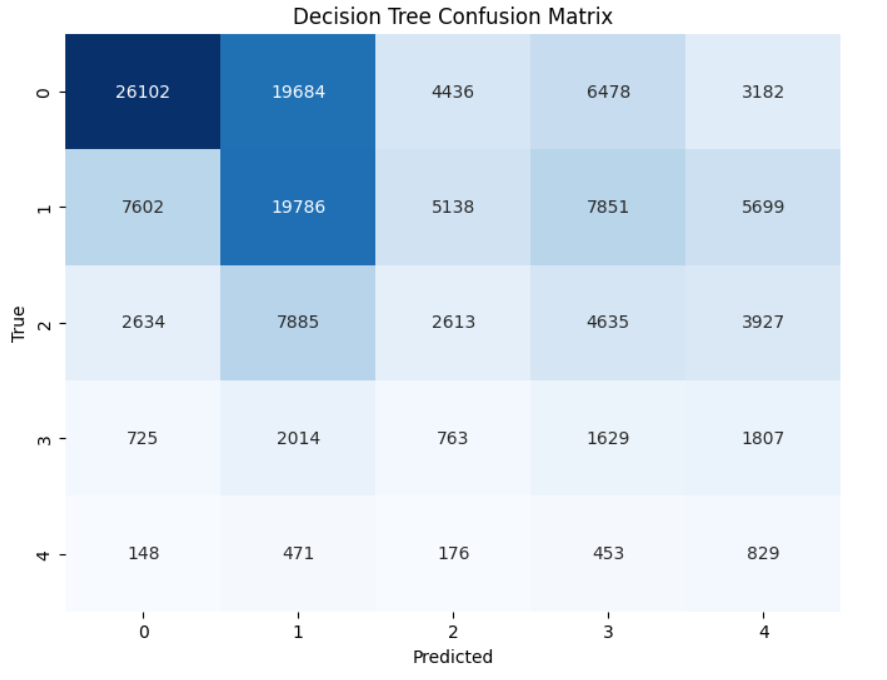
**DECISION TREE**

Decision trees are highly interpretable and easy to understand. The rules learned by a decision tree can provide insights into the decision-making process.

Handling Non-Linearity: Decision trees are capable of capturing non-linear relationships in the data, making them suited for problems where the decision boundaries are not linear.



After the KNN model we the trained the dataset with assistance of Decision tree technique in which we got a fair accuracy of 37% score which compared to other classifier is good



# **Predicting for a Specific Domain**



# **6.DISCUSSION AND REFLECTION(INDIVIDUAL)**

Student 1: Aromal P Biju ID: 2488777 -

K-Nearest Neighbors (KNN) Classifier:\*

In terms of outcomes, the KNN classifier displayed impressive performance, proving its flexibility to the domain prediction challenge. Student 1 saw how the choice of distance measure and K value substantially influenced model accuracy. An important lesson learnt was the significance of establishing the correct balance between model complexity and generalization. For future directions, Student 1 advises researching advanced distance metrics and considering ensemble ways to boost the resilience of the KNN classifier.

Student 2: Athul Vinod ID: 2543267-

The findings acquired from the Random Forest classifier were promising, highlighting the power of ensemble learning in our domain prediction research. Student 2 demonstrated the robustness of Random Forest against overfitting and its ability to handle various datasets. A significant lesson learnt was the significance of feature importance analysis for model interpretability. Moving further, Student 2 proposes examining versions of Random Forest, such as gradient boosting, and exploring ways to incorporate new characteristics for improved predictions.

Student 3: Sankar Dev S ID:2595447-

The Decision Tree classifier offered enlightening results, emphasizing interpretability in decision-making. Student 3 discovered the impact of hyperparameter adjustment on tree construction and noted how decision trees excel at capturing complex relationships within the data. The lesson learnt underscored the trade-off between model simplicity and forecast accuracy. For future study, Student 3 proposes examining ensemble approaches that include decision trees and investigating strategies to limit potential overfitting while maintaining interpretability.

Collectively, the project's results illustrate the diverse strengths of the three classifiers. Lessons learnt include the necessity for rigorous parameter tuning, interpretability issues, and a knowledge of trade-offs in model complexity. Future approaches entail deeper research of ensemble methods, improved parameter tuning procedures, and potentially incorporating other features or data sources to boost prediction skills. This collaborative effort lays a solid platform for ongoing research and development in domain prediction using machine learning.

So far, we have obtained 44% score on this set with the help of KNN model, that is fairly decent. For now we have employed just limited features, and features engineered from it, But still there are feature extraction the tweaking remaining After incorporating all the elements into training set,It will provide a score above of 70%

# **7.CONCLUSIONS**

The last and the major purpose of this article is to predict the domain value on the provided dataset by applying different feature extraction algorithms. We compared prior published scores with outcomes of this work and there are significant improvements found in our paper. During this project, 3 distinct classification techniques were applied and their merits and cons are discussed. Overall, greatest results were obtained using K-Nearest-Neighbours(KNN), which gave 44% accuracy for train set

# **8.REFERENCES**

1. *Deep Learning Framework for Domain Generation Algorithms Prediction Using Long Short-term Memory*. (2019, March 1). IEEE Conference Publication | IEEE Xplore. https://doi.org/10.1109/ICACCS.2019.8728544
2. Tang, L., & Mahmoud, Q. H. (2021, August 20). *A Survey of Machine Learning-Based Solutions for Phishing Website Detection*. Machine Learning and Knowledge Extraction. https://doi.org/10.3390/make3030034
3. Gandotra, E., & Gupta, D. (2021, January 1). *An Efficient Approach for Phishing Detection using Machine Learning*. Algorithms for Intelligent Systems. https://doi.org/10.1007/978-981-15-8711-5\_12
4. Rao, R. S., & Pais, A. R. (2018, January 6). *Detection of phishing websites using an efficient feature-based machine learning framework*. Neural Computing and Applications. https://doi.org/10.1007/s00521-017-3305-0
5. https://doi.org/10.1007/s00521-017-3305