Prediction and Its Impact on Its Attributes While Biasing MachineLearning Training Data

J.Kavitha
BVRIT HYDERABAD College of Engineering for Women, Hyderabad, Telangana
j.kavitha5555@gmail.com

J.Sasi Kiran
Lords Institute of Engineering and Technology , Himayath Sagar, Hyderabad, Telangana
sasikiranjangala@gmail.com

Srisailapu D Vara Prasad
GITAM Deemed to be University, Hyderabad, Telangana
sdvprasad554@gmail.com

Krushima Soma
Malla Reddy University, Hyderabad, India
krushima.s@gmail.com

G Charles Babu
Gokaraju Rangaraju Institute of Engineering and Technology Hyderabad, India.
charlesbabu26@gmail.com

S Sivakumar
Lakshmi Bangaru Arts and Science College
Melmaruvathur, India.
sivakumarmca@gmail.com

2022 Third International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE) | 978-1-6654-5664-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSTCEE56972.2022.10100010

***Abstract-* Machine learning models are built utilizing biased training data that comes from human experience. The data gathered reflects the cognitive bias that human's display in their actions and thought processes. The most effective machine learning models are considered to resemble human cognition; as a result, these models are biased. For improved explainable models, bias detection and evaluation are crucial. This study identifies bias in learning models in relation to cognitive bias in humans and suggests a cutting-edge method for identifying and evaluating machine learning bias. It also tries to identify bias in the dataset. The potentially skewed qualities are seen in the deployed dataset. Prior to employing the notion of alternation function to swap the values of PBAs and analyze the influence on prediction using KL Divergence, we first choose a few common biased characteristics. Compare the KL divergence values from the various models used to train the dataset and forecast the output in this section.**

Keywords: KL divergence, machine learning algorithm, Bias, project bias detection, potentially biased attributes.

1. INTRODUCTION

Machine learning models are built utilizing biased training data that comes from human experience. The data gathered reflects the cognitive bias that human's display in their actions and thought processes. Bias is the degree to which a model's prediction deviates from the target value as measured against the training data. Technology bias is a well-known issue. The code review process, which tends to rely more on past participation than anything else and can be a big barrier to people starting their careers or joining a new organization, is one potential source of ongoing prejudice. To provide an extensive insight into the technical and real time work aspects of the project- Bias Detection (bias in dataset). The bias in the data set is detected by finding the Potentially Biased Attributes (PBA's) and the impact of these attributes on the prediction. The divergence is measured using the KL Divergence value. However, this data bias (KL Divergence value) includes the machine learning model bias, it is still considered as a challenging problem. To detect whether an attribute impacts the prediction if we alternate the values of that attribute. If a predictor is dependent on one or more PBA given the class label, it is considered biased. so, we try to detect whether an attribute is PBA and its impact on prediction. The application of machine learning models in the actual world has risen dramatically as technology has advanced. Prediction has become a crucial task in a variety of scientific and academic fields. We use datasets from many sources for this, and the data is acquired by humans, resulting in human cognitive bias. As a result, the predictions of the machine learning model are skewed. For that, we employ an alternation function to detect biasin the data set. We do this by using various machine learning models to forecast the class label, which are then fed into the Sequential model, which then uses the model's prediction to determine the Potentially BiasedAttributes.

The proposed system presents an approach for detecting the data bias (or Historical Data Bias) in the dataset and evaluating this with the help of KL Divergence. The average predicted wage of a group of attributes before and after the alternation function is plotted. The proposed system also plots the graph showing the Bias evaluation between different attribute values using KL-divergence.

The KL divergence value does not truly represent the Data bias because machine learning models' predictions include the model's prediction bias. So, to detect bias in the data set, we employed two approaches and compared the findings of these two approaches. We would conclude whether the data set is biased towards which category of

978-1-6654-5664-7/22/$31.00©2022 IEEE

Attributes based on the outcomes of these two methodologies. Historical data bias develops when socio-cultural assumptions and viewpoints are repeated in methodical processes. When machine learning models are trained on data from historically biased sources, this becomes more challenging. So, for mitigation of the bias, we have to detect whether the bias is induced in the dataset or not. We also try to evaluate that using KL Divergence function. The KL Divergence value represents how divergent the prediction values of the model before and after alternation. If we notice that the data set is biased towards a certain set of attributes, we can strive to mitigate bias by adding more useful and representative data for protected groups to guarantee that the algorithms treat them equally. Data resampling, data augmentation, and the gathering of additional data are all options for generating incremental data.

Fair ML

To detect bias in a machine learning model, Fair ML uses the technique of determining the relative significance of the characteristics included in the model. The model is said to be unduly reliant on a characteristic if it is one of the protected attributes and it is determined to have a high significance. This indicates that the feature is crucial to the model's prediction. As a result, the model could be considered biased. The working of Fair ML is shown in Figure 1.

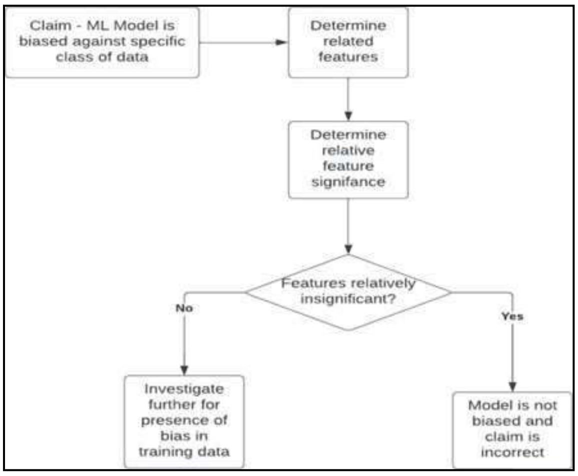

Claim - ML Model is biased against specific class of data
Determine related features
Determine relative feature signifance
Features relatively insignificant?
No
Yes
Investigate further for presence of bias in training data
Model is not biased and claim is incorrect

Figure 1: Flowchart of Fair ML for Bias Detection

Introduction to the problem domain

When a data set used to train an AI system incorporates human prejudices or discriminations towards any subset in the population, data bias arises. It's common to think that bias is directed towards a protected group, yet bias against any group is unacceptable. Bias can emerge at any point during the dataset lifespan, including creation, sampling, collection, and processing. The point in the dataset lifecycle when bias appears is linked to the bias's root cause. Data bias occurs in the phase of data creation. Data bias has generated serious concerns regarding the accuracy and trustworthiness of research forecasts in the burgeoning research industry. Consider DNA testing, which has the potential to provide crucial information. Many DNA test companies heavily rely on artificial intelligence (AI) algorithms to expedite the research process and decrease reliance on human specialists, which causes these problems. Although historical data samples are used to train the algorithms, it's possible that these samples don't accurately reflect the genetic make-up of a customer. Businesses that rely on AI must manage data bias and the related ethical, legal, and financial issues in a variety of industries. This calls for a deep comprehension of the root causes of data bias. Understanding the origins of bias can help businesses design AI systems responsibly as well as detect and mitigate it.

2. LITERATURE REVIEW

Algorithms for pedestrian detection exhibit age and gender bias. Writers: Martim According to Brandao Martim[2], who examined several pedestrian recognition algorithms, algorithmic bias made it so that certain types of walkers were more likely to have missed detections. Martim focuses on age and gender bias evaluation in particular and comes to the conclusion that algorithms clearly perform worse on kids.

The author evaluated prejudice quantitatively by comparing male/female and child/adult miss rates, as well as objectively by computing average performance differences and Wilcoxon rank-sum test p-values. The rank-sum test is used to determine whether or not the performance distribution for male/female and child/adult consistently holds across several algorithms.

Simultaneous labels for males and females or children and adults are not permitted by the labelling system. If multiple measurements have the same value, the Wilcoxon test may yield false findings. The test is weakened when several values are comparable since their relative ranks are also similar. Large data sets are not tested using the rank sum test.

LOGAN: Clustering to Detect Local Group Bias Authors: Kai-Wei Chang and Jieyu Zhao

The authors claim that understanding how biases are embedded in a model requires more than just bias analysis at the corpus level. A model that claims equivalent performance for two groups in a corpus may behave differently for these two groups in a local area. Authors suggested LOGAN, a local group bias detection method, to find biases in local areas in order to uncover local group prejudice. A clustering technique is modified by LOGAN to group instances according to their features.

The goal of LOGAN is to arrange instances from the test corpus into clusters where each cluster exhibits the local group bias from the trained model. LOGAN is utilised in this to identify local group bias in texts. LOGAN has the drawback of only taking into account binary properties. Model biases that were previously hidden from the global bias measurements can be found with the aid of LOGAN. The effectiveness of synthetic data was demonstrated by Adam Kortylewski, Bernhard Egger, and others[4] in their investigation of the detrimental impacts of data set bias on deep face recognition systems. The impact of various types of bias on the generalisation capacities of neural network topologies are examined using the face recognition rate as a function. The data set bias is reduced once the neural network model has been fine-tuned with real-world data and pre-trained with synthetic data. A basic problem with face recognition systems is the difficulty in analysing the effects of data set bias on generalisation performance. Utilizing synthetic face images made with a parametric 3D Morphable Face Model, the suggested method addresses this problem.

3. PROPOSED MODEL

The proposed system works as shown in Figure 2. Synthetic images of various facial identities are created and then transformed along the nuisance transformation axes. The training data is biased, for example, first separating the synthetic data into a training and test set, then removing particular facial expressions. The generalization of the DCNNs to the unbiased test data is then assessed after numerous DCNN architectures have been trained on the biased training data. It is possible to assess the identification rate as a function of the biased nuisance transformation since the synthetic data is fully parametric. A huge sample of artificial images is produced in the proposed approach to lessen the effect of dataset bias on the generalization capacity of neural networks.

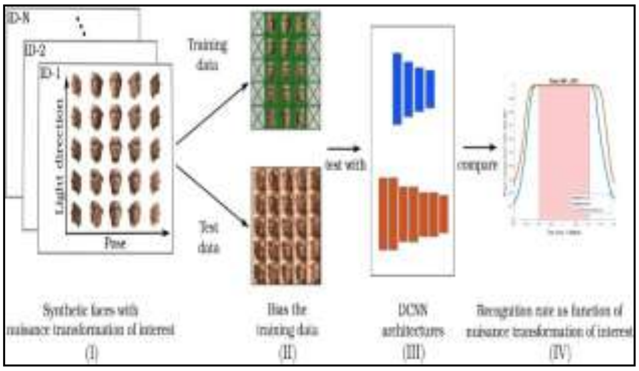


Figure 2: Experimental Setup for analysis of biased training dataset

The design of our proposed system is explained using a block diagram of our model and module description along with the foundation of the algorithm.

Block Diagram

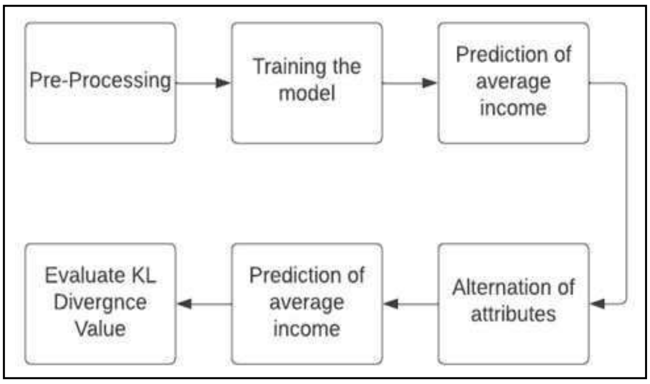

Pre-Processing
Training the model
Prediction of average income
Alternation of attributes
Prediction of average income
Evaluate KL Divergnce Value

Figure 3: Proposed system model

The flow of the proposed system

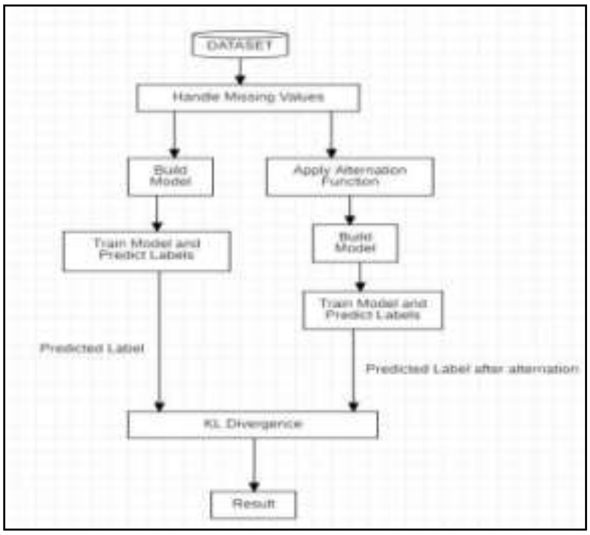

DATASET
Handle Missing Values
Build Model
Apply Alternation Function
Train Model and Predict Labels
Build Model
Train Model and Predict Labels
Predicted Label
Predicted Label after alternation
KL Divergence
Result

Figure 4: Basic flowchart of Proposed System

Module Description

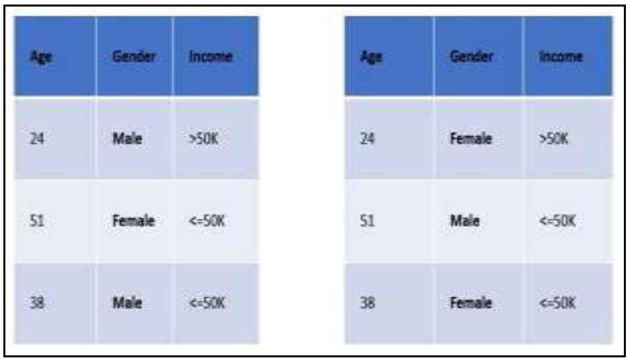
The modules of the proposed system are discussed in this chapter. The proposed system can be enclosed in two modules i.e, Prediction module and Alternation module.

Prediction Module

This module includes the majority of the tasks of our proposed system. This module starts with the task of pre - processing. Pre - processing involves removing the rows with null values, replacing the invalid characters ‘ ?’ with the mode of that column and balancing the data frame using SMOTE. K-Fold cross validation is then used to split the data frame into training and testing sets. The class label is then predicted using the testing set after a model has been trained on the training set.

Alternation Module

This module takes the data frame and the column to be alternated is given. For example, the column to be alternated is sex and the values to be alternated be Male and Female.



Age	Gender	Income
24	Male	>50K
51	Female	<=50K
38	Male	<=50K

Age	Gender	Income
24	Female	>50K
51	Male	<=50K
38	Female	<=50K

Figure 5: Working of Alternation Function

In the above figure, the table on the right side is the dataset after applying the alternation function on the dataset on the left side.

Algorithms

The algorithms used for the proposed system. We proposed two different methods for our system. They are as follows:

a. Layered Method

b. Averaging Method

Layered Method

In the Layered Model, we used four different classifiers for prediction. They are Random Forest Classifier, Decision Tree Classifier, KNN Classifier and the Sequential model. In this model first we take the predictions of the ml models and form a data frame from these predictions and this data frame is given as input to the sequential model.

Sequential Model

A sequential paradigm is better suited for a straightforward stack of layers. The sequential stacking of Keras layers is the fundamental concept behind Sequential API, which is also how the name Sequential model was developed. In the majority of ANNs, which comprise layers that are arranged in a sequential order, data flows from one layer to the next in the predetermined order until it reaches the output layer. The activation functions we used are RELU, LINEAR. For the proposed system, we used 4 dense layers with different number of nodes in neural networks. The number of nodes in the layers is 12, 8, 4 and 1 nodes respectively. The loss function is Binary Cross entropy and the optimizer is Adam Optimizer.

Decision Tree Classifier

It is a tree-structured classifier, with core nodes expressing dataset attributes, leaves reflecting the outcome, and branches denoting decision rules. Decision nodes and leaf nodes are the two different sorts of nodes. Decision nodes are used to make decisions and can contain multiple branches. The outputs of those decisions are called leaf nodes, and they don't have any further branches.

Random Forest Classifier

A method for supervised learning is called Random Forest. It is based on the idea of ensemble learning, which is a method for resolving a challenging issue and enhancing the performance of the model by combining various classifiers. In order to increase the predictive accuracy of a dataset, a classifier called Random Forest averages the outcomes of many Decision Trees on various subsets of the dataset. KNN Classifier

A method for supervised learning is K Nearest Neighbor. It places the new instance in the category that is most similar to the existing cases, presuming that the new case/data and the old cases are equivalent. All of the available data is kept, and new data points are categorized based on how much they resemble the current data. During the training phase, the KNN algorithm stores the dataset, and when it receives new data, it classifies it into a category that is quite similar to the new data.

Averaging Method

In the Layered Model, we used three different classifiers for prediction. They are Random Forest Classifier, Decision Tree Classifier and KNN Classifier. In this model first we take the predictions of the three models and then find the mean of these predictions and interpret them as results i.e., average income. The basic design of this method.

4. RESULTS AND DISCUSSIONS

It will enlist and displays all the outputs and results obtained from the training and execution of the model. The below figure 6, shows the average predicted wage for male before applying the alternation function.

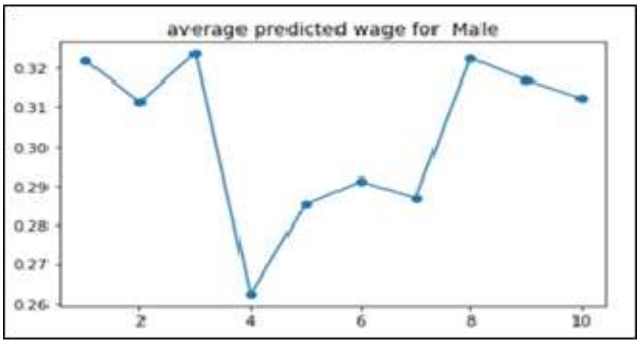


Figure 6: Average predicted wage for Male before alternation

Figure 6, depicts the average predicted wage for male after applying the alternation function (Male/Female alternation).

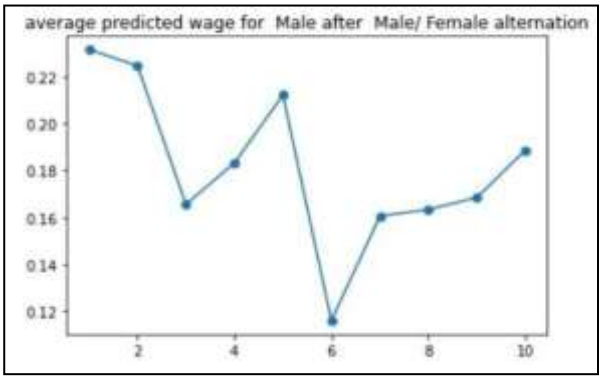


Figure 7: Average predicted wage for Male after Male/Female alternation

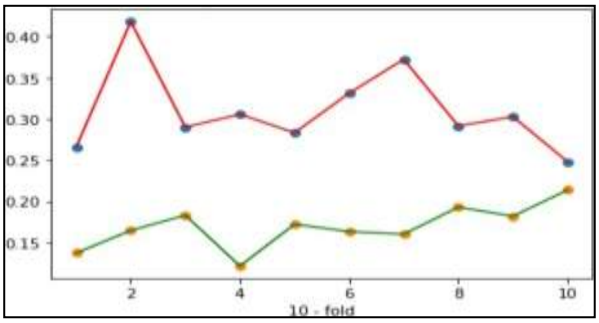


Figure 8: Average predicted wage for male.

The above graph Figure 8 shows the average predicted wage for males before and after alternation. The Red Line indicates the average predicted wage of male before alternation. Green Line indicates average predicted wage of male after alternation. From that, we can say that average predicted wage for males decreases after applying alternation.

The below figure shows the average predicted wage for females before applying the alternation function

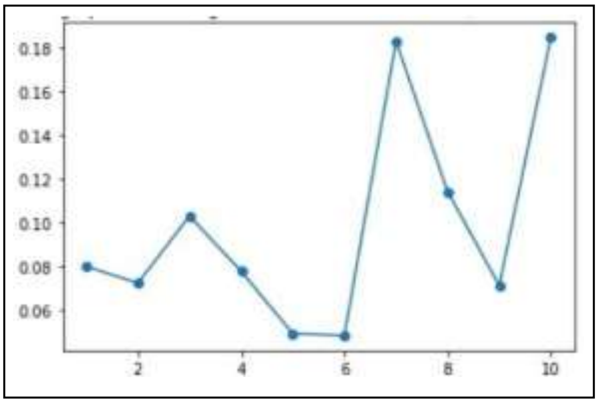


Figure 9: Average predicted wage for Female before alternation

The below figure, depicts the average predicted wage for females after applying the alternation function (Female/Male alternation).

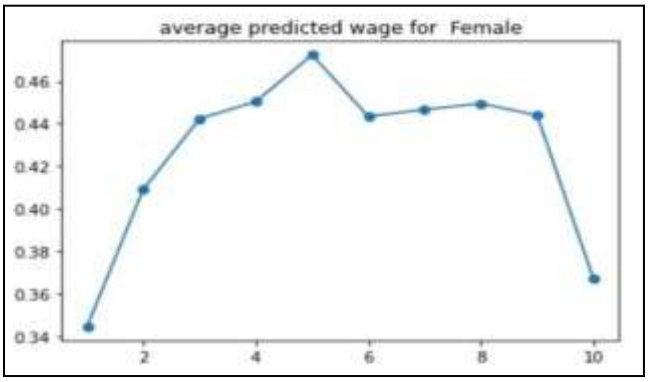


Figure 10: Average predicted wage for Female after Male/Female alternation

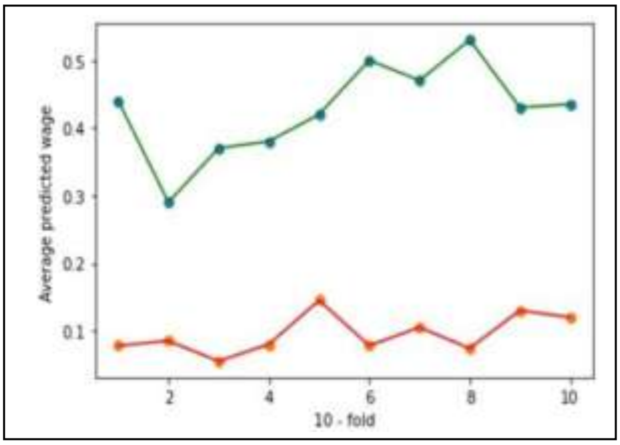


Figure 11: Average predicted wage for Female.

The above graph, shows the average predicted wage for females before and after alternation. The Red Line indicates the average predicted wage of female before alternation. Green Line indicates average predicted wage of female after alternation. From that, we can say that average predicted wage for females increases after applying alternation.

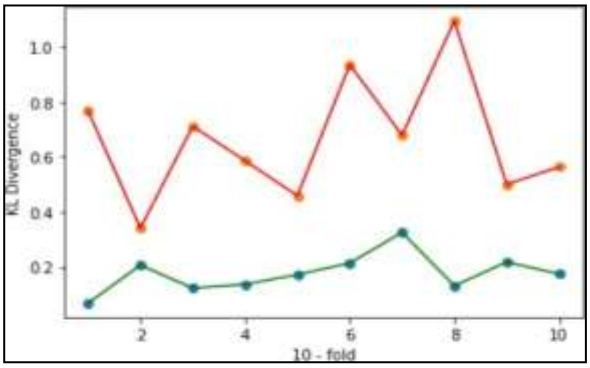


Figure 12: KL divergence of male and female.

The KL divergence between the original dataset's male-predicted wage and the female-predicted wage is depicted by the green line in the aforementioned graph, Fig.13. The red line shows the KL divergence between the original dataset's anticipated wage for women and men and the expected wage for women. The graph above indicates that there is more bias against women than against men.

It was discovered that the average KL-divergence across all folds was 0.486. The model's accuracy is observed to be 82 percent.

Averaging Method

The below Figure, depicts the average predicted wage for male after applying the alternation function (Male/Female alternation).

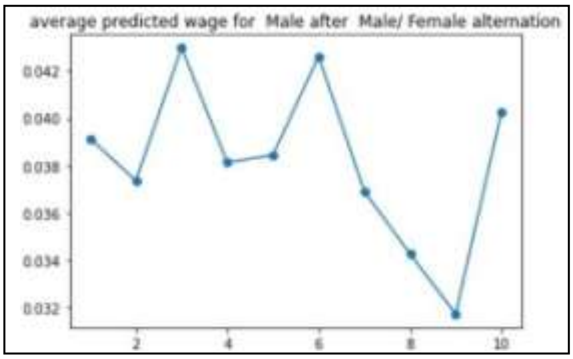


Figure 13: Average predicted wage for Male after Male/Female alternation

The above graph Fig. 5.10 shows the average predicted wage for males before and after alternation. The Red Line indicates the average predicted wage of male before alternation. Green Line indicates average predicted wage of male after alternation. From that, we can say that average predicted wage for males decreases after applying alternation.

Figure 14, depicts the average predicted wage for females after applying the alternation function (Female/Male alternation).

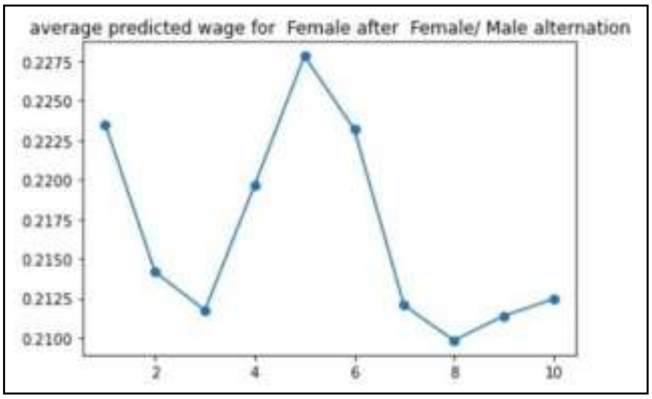


Figure 14: Average predicted wage for Male after Male/Female alternation

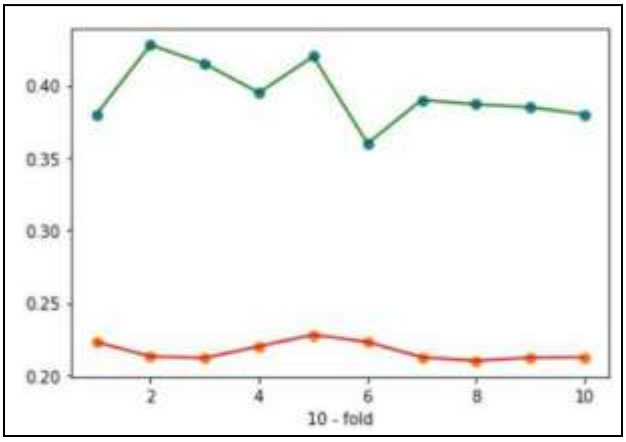


Figure 15: Average Predicted wage for Female.

The above graph shows the average predicted wage for females before and after alternation. The Red Line indicates the average predicted wage of female before alternation. Green Line indicates average predicted wage of female after alternation. From that, we can say that average predicted wage for female's increases after applying alternation.

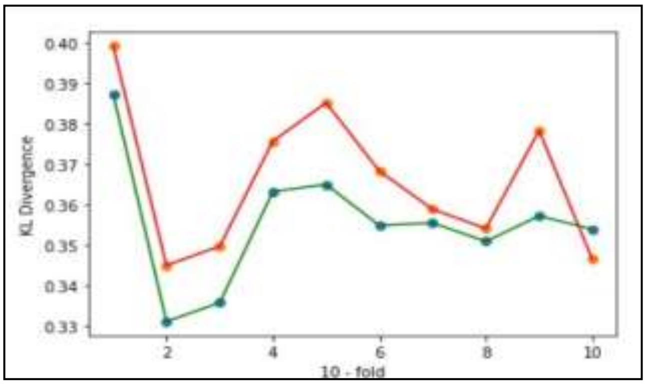


Figure 16: KL divergence of male and female

The green line in the graph above, represents the KL divergence between the original dataset's male and female pay predictions. The red line shows the KL divergence between the original dataset's anticipated wage for women and men and the expected wage for women. We can infer from the graph above that there is a greater prejudice towards women than men.

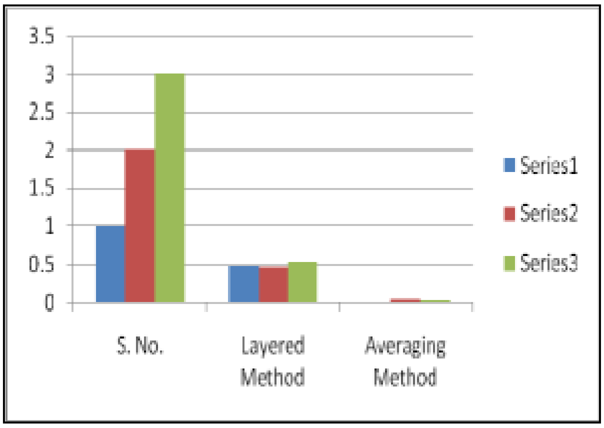


Figure 17: KL Divergence values of two approaches

The average KL-divergence over all folds was found to be 0.0106.The accuracy of the model is observed as 94%. We tested our proposed model several times to know its behavior and the KL-Divergence shown the above figure.

5. CONCLUSION

Data bias detection in machine learning models is particularly useful for accurate predictions because if we know that bias exists in the data set, we may employ approaches to minimize it if it exceeds the threshold value for the intended outcomes. Because they are based on human behavior, practice, experience, and acts, datasets contain bias. Machine learning models are susceptible to bias because of prejudice in the training datasets, yet it is crucial to identify bias. By changing the PBAs' values, try to identify bias. The degree of bias will then be determined by computing the difference between the original and alternate means of predicted class values for each attribute's value. In our work, we characterized and evaluated those approaches in two separate ways, and we then compared the results in those two ways.

REFERENCES

[1] Yadala Sucharitha, Yellasiri Vijayalata and Valurouthu K. Prasad, “Predicting Election Results from Twitter Using Machine Learning Algorithms”, Recent Advances in Computer Science and Communications (2021) 14 (1): pp: 246-256.

[2] Reddy, P.C.S., Yadala, S. and Goddumarri, S.N., 2022. Development of rainfall forecasting model using machine learning with singular spectrum analysis. IIUM Engineering Journal, 23(1), pp.172-186.

[3] Reddy PC, Sureshbabu A. An applied time series forecasting model for yield prediction of agricultural crop. InInternational Conference on Soft Computing and Signal Processing 2019 Jun 21 (pp. 177-187). Springer, Singapore.

[4] Liu, L., Shafiq, M., Sonawane, V.R., Murthy, M.Y.B., Reddy, P.C.S. and kumar Reddy, K.C., 2022. Spectrum trading and sharing in unmanned aerial vehicles based on distributed blockchain consortium system. Computers and Electrical Engineering, 103, p.108255.

[5] Singhal, A., Varshney, S., Mohanaprakash, T.A., Jayavadivel, R., Deepti, K., Reddy, P.C.S. and Mulat, M.B., 2022. Minimization of latency using multitask scheduling in industrial autonomous systems. Wireless Communications and Mobile Computing, 2022.

[6] Dhanalakshmi, R., Bhavani, N.P.G., Raju, S.S., Shaker Reddy, P.C., Marvaluru, D., Singh, D.P. and Batu, A., 2022. Onboard Pointing Error Detection and Estimation of Observation Satellite Data Using Extended Kalman Filter. Computational Intelligence and Neuroscience, 2022.

[7] Shaker Reddy PC, Sureshbabu A. An Enhanced Multiple Linear Regression Model for Seasonal Rainfall Prediction. International Journal of Sensors Wireless Communications and Control. 2020 Aug 1;10(4):473-83.

[8] Suharitha Y, Vijayalata Y, Prasad VK. Analysis of Early Detection of Emerging Patterns from Social Media Networks: A Data Mining Techniques Perspective. InSoft Computing and Signal Processing 2019 (pp. 15-25). Springer, Singapore.

[9] Sujihelen, L., Boddu, R., Murugaveni, S., Arnika, M., Haldorai, A., Reddy, P.C.S., Feng, S. and Qin, J., 2022. Node Replication Attack Detection in Distributed Wireless Sensor Networks. Wireless Communications and Mobile Computing, 2022.

[10] Reddy PC, Sureshbabu A. An adaptive model for forecasting seasonal rainfall using predictive analytics. International Journal of Intelligent Engineering and Systems. 2019:22-32.

[11] Wei B, Ren X, Zhang Y, Cai X, Su Q, Sun X. Regularizing output distribution of abstractive chinese social media text summarization for improved semantic consistency. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP). 2019 Apr 30;18(3):1-5.

[12] Balamurugan, D., Aravinth, S.S., Reddy, P., Rupani, A. and Manikandan, A., 2022. Multiview Objects Recognition Using Deep Learning-Based Wrap-CNN with Voting Scheme. Neural Processing Letters, pp.1-27.