JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DETECTION AND EVALUATION OF MACHINE LEARNING BIAS

[1]S. Durga Devi,[2]R.Rama Krishna Reddy,[3]S. Revanth Reddy,[4] K. Spandana

[1]Assistant Professor,[2]Student,[3]Student,[4]Assistant Professor

[1]Computer Science and Engineering,

[1]Chaitanya Bharathi Institute of Technology, Hyderabad,India

Abstract : Machine learning models are constructed using training data, which is biased and derived from human experience. Humans exhibit cognitive bias in their thoughts and conduct, which is reflected in the data acquired. The best machine learning models are said to mimic humans' cognitive ability, and thus such models are inclined towards bias. However, detecting and evaluating bias is important for better explainable models. In this paper, we aim to identify bias in learning models in relation to humans' cognitive bias and propose a novel technique for detection and evaluation of machine learning bias. In this paper, we try to detect the bias in the dataset. In the deployed dataset, the potentially biased attributes are observed. To detect bias, first we select some common biased attributes and then we use the concept of alternation function to swap the values of PBA's and evaluate the impact on prediction using KL Divergence. In this paper, we compare the KL divergence value obtained from different models used for training the dataset and predicting the output (average income).

Index Terms **– Bias in ML, cognitive Bias, bias detection and evaluation**

1. INTRODUCTION

Machine learning models are constructed using training data, which is biased and derived from human experience. Humans exhibit cognitive bias in their thoughts and conduct, which is reflected in the data acquired. When compared to the training data, bias is the amount by which a model's prediction departs from the target value. Bias in technology is a well-known problem. One potential avenue for continued bias is the code review process, it tends to rely more on past participation than anything else, and it can be a significant roadblock to people beginning their careers or joininga new organization. This paper provides an extensive insight into the technical and realtime 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 themachine learning model bias, it is still considered as a challenging problem.The objective of this paper is to detect whether an attribute impacts the prediction if we alternate the values of that attribute. If a predictor is dependent onone 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.

2. METHODOLOGY

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 analternation function to detect bias in the data set. We do this by using various machinelearning 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 Biased Attributes.

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 attributes based on the outcomes of these two methodologies.

3. LITERATURE SURVEY

[2] ***Age and gender bias in pedestrian detection algorithms*** *Authors - Martim Brandao*

Martim[2] analyzes different Pedestrian detection algorithms and states that the missed detections are more likely to happen on specific kinds of pedestrians due to algorithmic bias. Martim particularly focuses on age and gender bias evaluation and concludes that there is a clear worse performance of algorithms on children. Author evaluated bias qualitatively by comparing male/female and child/adult miss rates, as well as quantitatively by computing the average performance differences, and Wilcoxon rank-sum test p-values. They use the rank-sum test as a measure of whether the distribution of performance is the same for female/male and child/adult consistently over multiple algorithms or not.

The labeling system does not allow simultaneous male-and-female or child- and-adult labels. The Wilcoxon test may produce misleading results if many measurements are of the same value. When multiple values are similar, their relative ranks are also similar, diluting the test. The rank sum test is not used for large data sets.

[3] ***LOGAN: Local Group Bias Detection by Clustering, Authors - Jieyu Zhao, Kai-Wei Chang***

The authors argue that evaluating bias at the corpus level is not enough for understanding how biases are embedded in a model. A model that reports similar performance across two groups in a corpus may behave differently between these two groups in a local region. To detect local group bias, authors proposed LOGAN, a LOcal Group biAs detectioN algorithm to identify biases in local regions. LOGAN adapts a clustering algorithm to group instances based on their features. The purpose of LOGAN is to cluster instances in test corpus such that each cluster demonstrates local group bias contained in the trained model. In this, LOGAN is used to detect local group bias in texts. One limitation of LOGAN is that it considers only the binary attributes. LOGAN can help detect model biases that previously were hidden from the global bias metrics.

[4] ***Analyzing and Reducing the Damage of Dataset Bias to Face Recognition With Synthetic Data ,*** *Authors - Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig,Andreas Morel-Forster, Thomas Vetter.*

Adam Kortylewski, Bernhard Egger and Others[4] used synthetic data to investigate the negative effects of data set bias on deep face recognition systems, demonstrating the power of synthetic data. The face recognition rate is used as a function to investigate the effects of various types of bias on neural network architectures' generalization abilities. The neural network model is pre-trained with synthetic data, and after fine tuning with real-world data, the data set bias is reduced. The difficulty in studying the effects of data set bias on generalization performance is a fundamental issue with face recognition systems. The proposed system solves this issue by utilizing synthetic face images created with a parametric 3D Morphable Face Model.

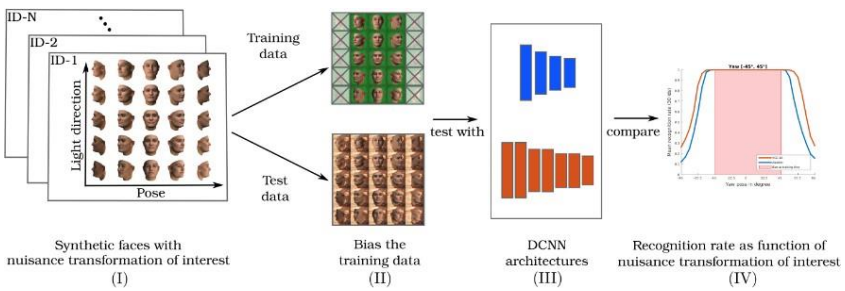

Fig. 2.1 Experimental Setup for analysis of biased training dataset

The proposed system works as shown in Fig. 2.1. Synthetic images of various facial identities are created and then transformed along the nuisance transformation axes. The training data is biased, for example, by deleting specific face postures after splitting the synthetic data into a training and test set. Following that, multiple DCNNarchitectures are trained on the biased training data, and the DCNNs' generalization tothe unbiased test data is evaluated. Because the synthetic data is fully parametric, it is possible to evaluate the recognition rate as a function of the biased nuisance transformation. In the proposed system, to reduce the impact of dataset bias on the generalization ability of neural networks, a large sample of synthetic images are generated.

4. DESIGN OF THE PROPOSED SYSTEM

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

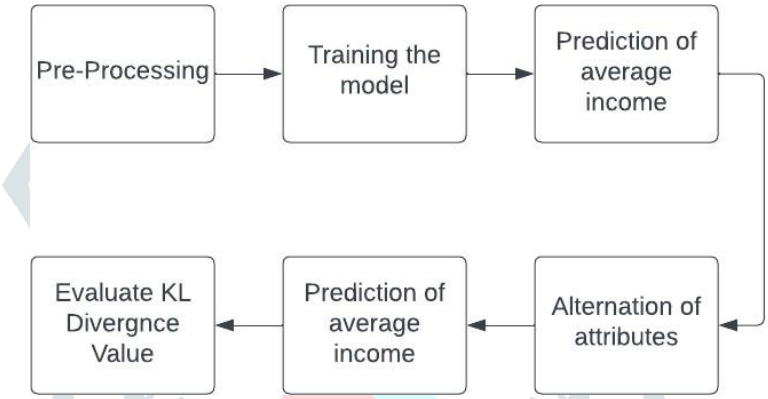

Fig. 4.1 Block Diagram of proposed system

The flow of the proposed system is depicted in below figure Fig. 4.2.

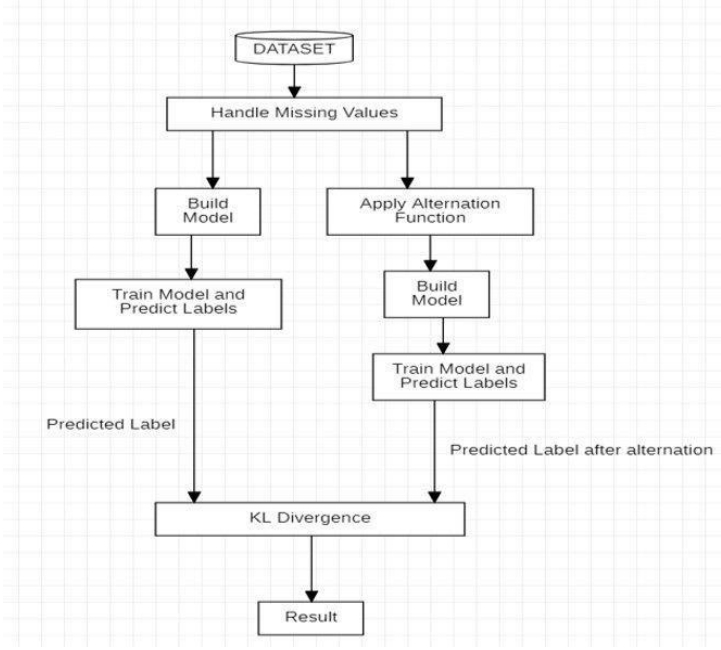

Fig. 4.2 Basic flowchart of Proposed System

4.1. 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.

4.2 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 therows with null values, replacing the invalid characters ' ?' with the mode of that column and balancing the data frame using SMOTE. After this, the data frame is divided into training and testing sets with the help of K-Fold cross validation. Then a model is trained on the training set and the class label is predicted using the testingset.

4.3 .Alternation Module

This module takes the dataframe 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. The working of the module is depicted below Fig. 4.3.

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

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

Fig 4.3 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.

4.4. Algorithms

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

- Layered Method
- 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 dataframe from these predictions and this dataframe is given as input to the sequential model. The basic design of this method is shown in fig. 4.4.

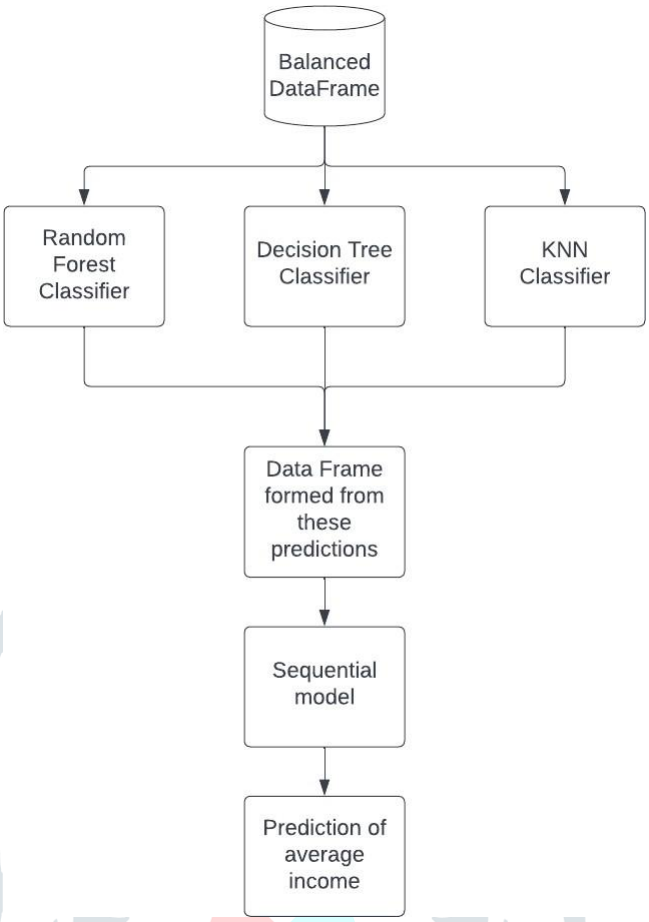

Fig. 4.4 Design of Layered Method

Sequential Model

For a simple stack of layers, a Sequential model is more suitable. Sequential API's fundamental concept is to arrange Keras layers in a sequential order, hence the name Sequential model. Most ANNs have layers that are arranged in a sequential order, with data flowing from one layer to the next in the specified 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 are 12, 8, 4 and 1 nodes respectively.The loss function is Binary Cross entropy and the optimizer is Adam Optimiser.

Decision Tree Classifier

Decision Tree is a supervised learning technique that can be used to address classification and regression issues, however it is most typically used for classification. It's a tree-structured classifier with core nodes that represent dataset attributes, branches that represent decision rules, and leaf nodes that reflect the result. Decision nodes and leaf nodes are the two types of nodes. Decision nodes can contain several branches and are used to make any decision. The outputs of those decisions are called leaf nodes, and they don't have any further branches. The below figure Fig. 4.5 depicts the working of decision tree classifier.

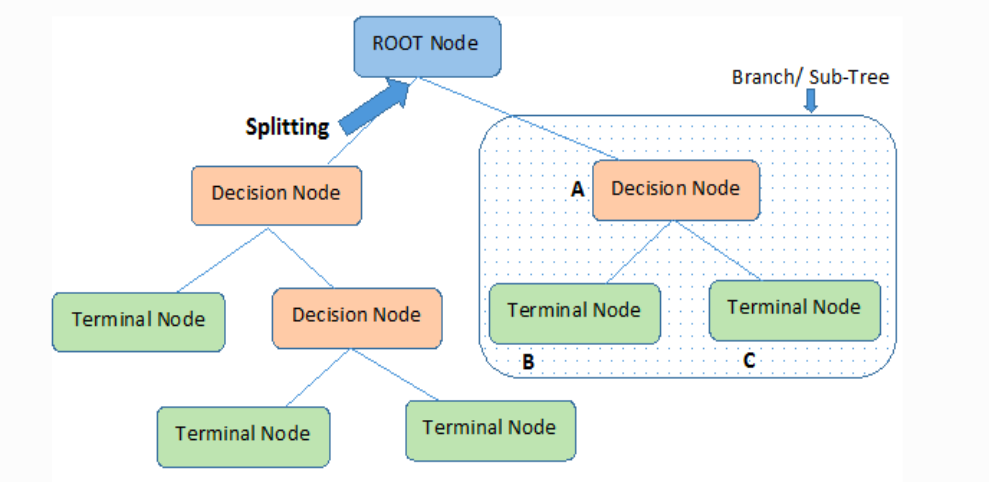

Fig. 4.5 Decision Tree Classifier

Random Forest Classifier

Random Forest is a technique for supervised learning. It is based on the notion of ensemble learning, which is a strategy for solving a complicated problem and improving the model's performance by merging multiple classifiers. Random Forest is a classifier that averages the results of several Decision Trees on distinct subsets of a dataset to improve the dataset's predictive accuracy.

As a result, rather than depending on a single Decision Tree, Random Forest considers the forecasts from each tree and predicts the final output based on the majority votes of predictions. As a result, accuracy improves and the problem of overfitting is avoided. The figure Fig.4.6 depicts the random forest classifier.

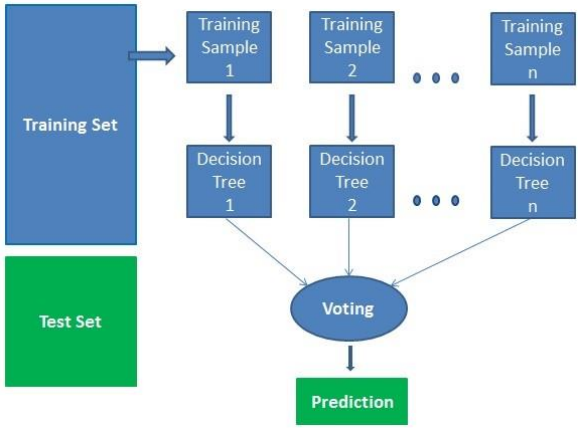

Fig. 4.6 Random Forest Classifier

KNN Classifier

K Nearest Neighbor is a technique for supervised learning. It assumes that the new case/data and old cases are comparable, and it assigns the new case to the category that is closest to the current ones. It keeps all of the available data and classifies new data points depending on how closely they resemble the existing data. During the training phase, the KNN algorithm saves the dataset, and when it receives new data, it classifies it into a category that is quite similar to the new data.

The Euclidean function can be used as a distance function as our prediction is a continuous variable.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

where : x , y are coordinates of points and d is the distance between points
i i

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 is shown in fig.4.7.

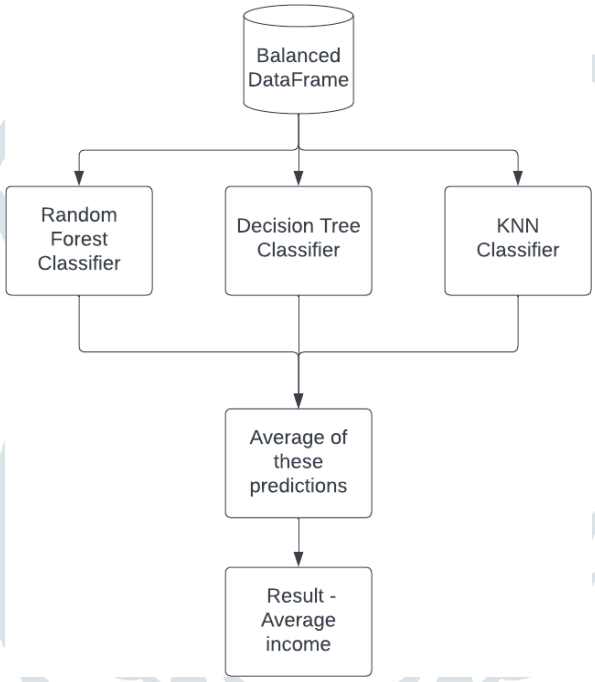

Fig. 4.7 Design of Averaging Method

KL Divergence

It is used to quantify the difference between probability distributions. The KL Divergence score measures how much one probability distribution differs from another. The formula for KL Divergence score is shown below.

$$D_{KL}(p||q) = log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

where μ1, μ2 are the mean of distributions p, q and σ1, σ2 are the standard deviations of p, q distributions respectively. If the value of KL divergence score is 0 for the ideal distributions. That means the bias is proportional to the value of KL Divergence score. The algorithm for the calculation of KL Divergence score is:Two lists of the same length are provided as input.

Output : KL Divergence scorefunction KL(p,q)

1. i=0, sum=0
2. n = length(p)
3. if (i < n), goto 5.
4. goto 7.

5. sum = sum + (p[i] * log(p[i]/q[i]))

6. goto 3.

7. return sum

5. RESULTS AND DISCUSSIONS

This enlists and displays all the outputs and results obtained from the training and execution of the model.

Results - Layered Method

The below figure Fig 5.1, shows the average predicted wage for male beforeapplying the alternation function.

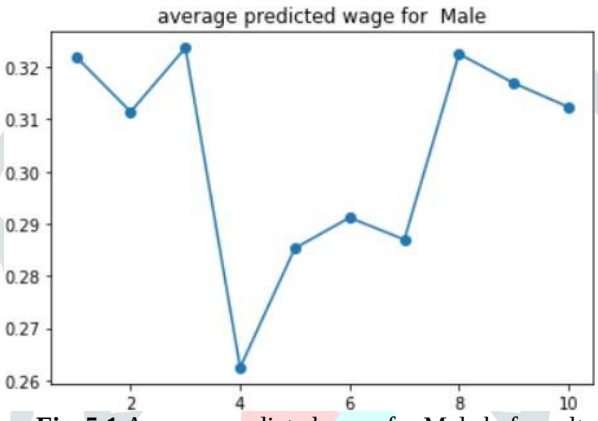

Fig. 5.1 Average predicted wage for Male before alternation

Figure Fig 5.2, depicts the average predicted wage for male after applying thealternation function (Male/Female alternation).

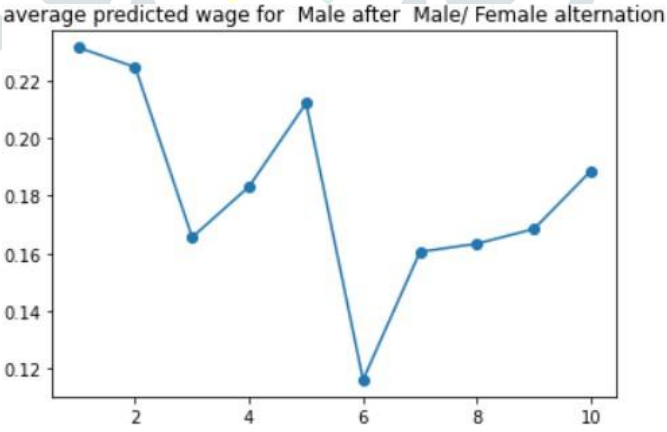

Fig. 5.2 Average predicted wage for Male after Male/Female alternation

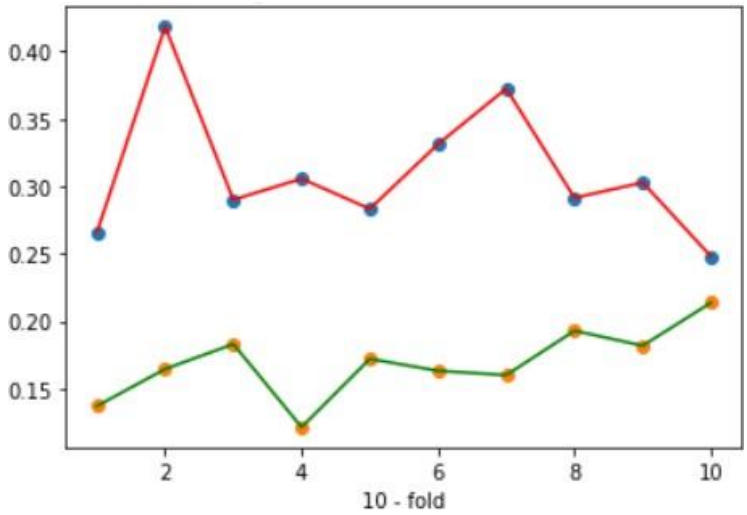

Fig. 5.3 Average predicted wage for male.

The above graph Fig. 5.3 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 Fig 5.4, shows the average predicted wage for females before applying the alternation function.

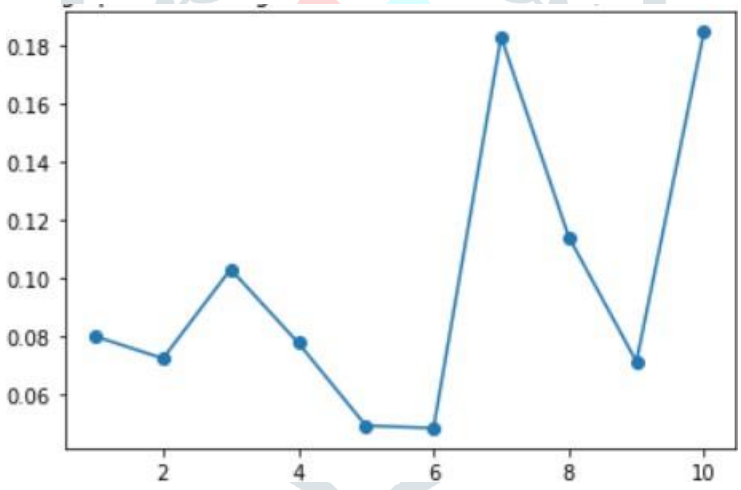

Fig. 5.4 Average predicted wage for Female before alternation

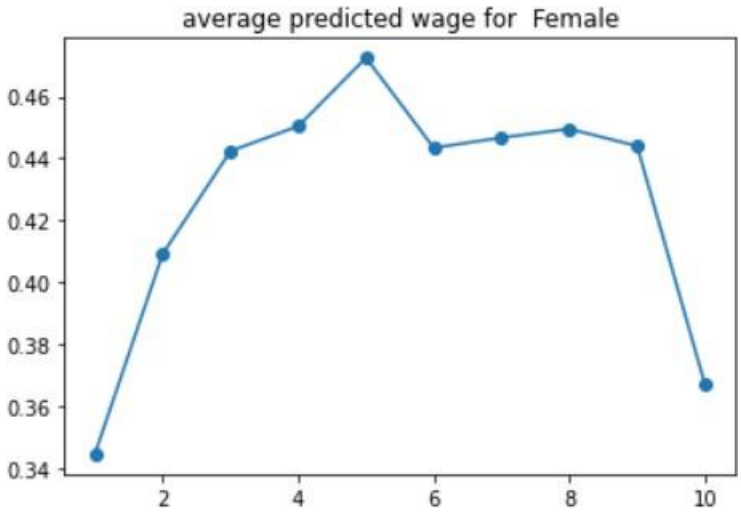

Figure Fig 5.5, depicts the average predicted wage for females after applying the alternation function (Female/Male alternation).

Fig. 5.5 Average predicted wage for Female after Male/Female alternation

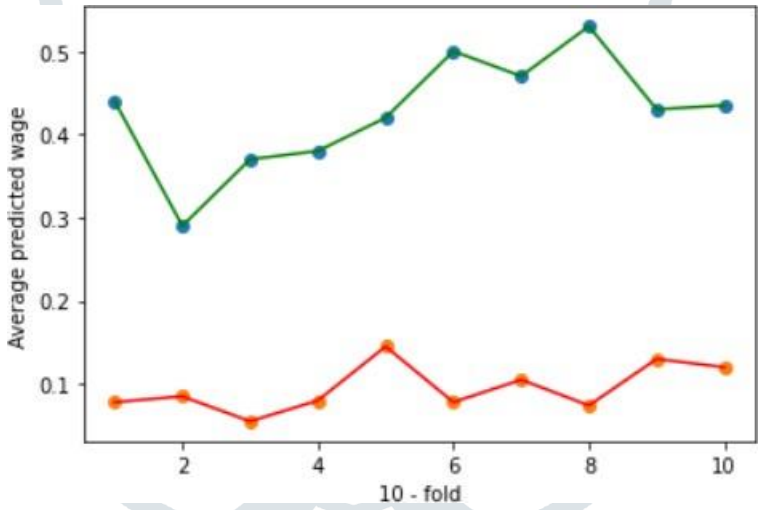

Fig. 5.6 Average predicted wage for Female.

The above graph Fig. 5.6 shows the average predicted wage for females beforeand 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.

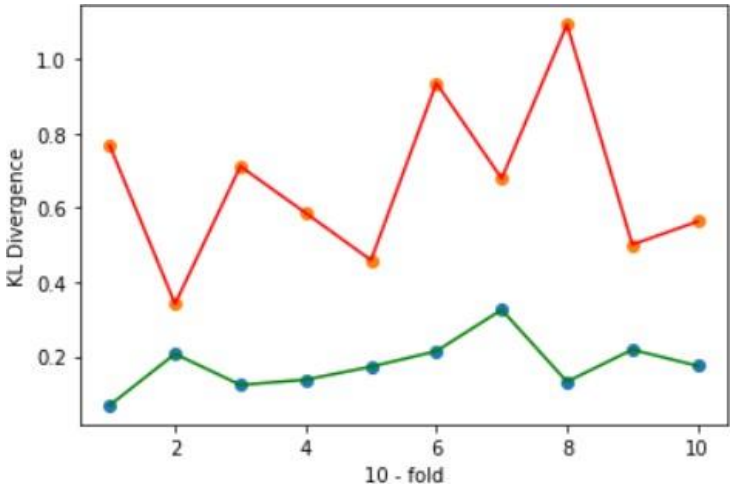

Fig. 5.7 KL divergence of male and female.

In the above graph Fig. 5.7, the green line represents the KL divergence between the predicted wage of the male in the original dataset and the predicted wage when changing male into female. The red line represents the KL divergence between the predicted wage of the female in the original dataset and the predicted wage when changing female into male. From the above graph, we can conclude that the bias is more against females than males.

The average KL-divergence over all folds was found to be 0.486.The accuracy of the model is observed as 82%.

Results - Averaging Method

The below figure Fig 5.8, shows the average predicted wage for male beforeapplying the alternation function.

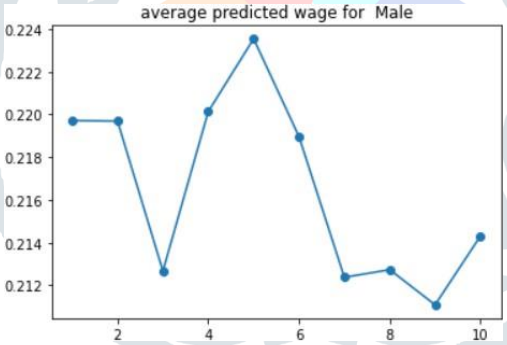

Fig. 5.8 Average predicted wage for Male before alternation

Figure Fig 5.9, depicts the average predicted wage for male after applying thealternation function (Male/Female alternation).

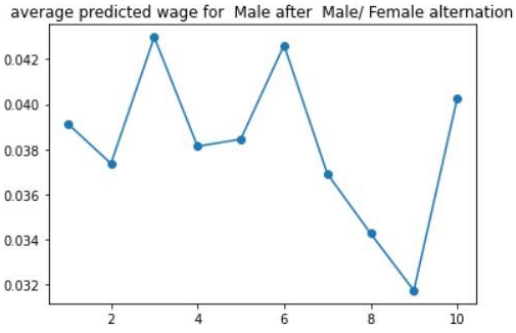

Fig. 5.9 Average predicted wage for Male after Male/Female alternation

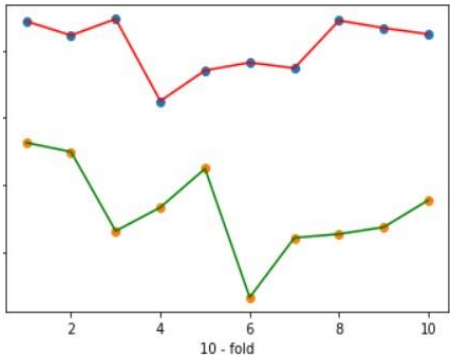

Fig. 5.10 Average Predicted wage for male.

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.

The below figure Fig 5.11, shows the average predicted wage for females before applying the alternation function.

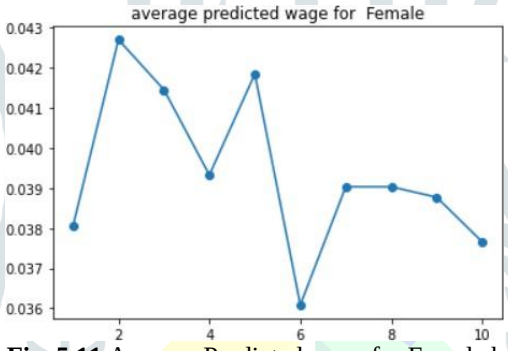

Fig. 5.11 Average Predicted wage for Female before alternation

Figure Fig 5.12, depicts the average predicted wage for females after applyingthe alternation function (Female/Male alternation).

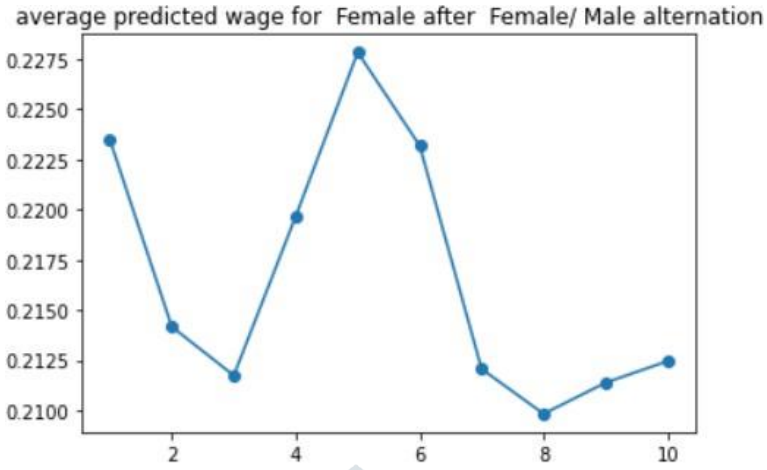

Fig. 5.12 Average predicted wage for Male after Male/Female alternation

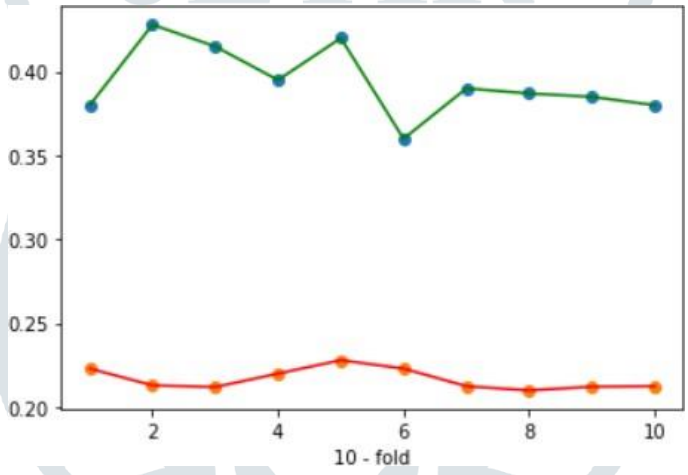

Fig. 5.13 Average Predicted wage for Female.

The above graph Fig. 5.13 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.

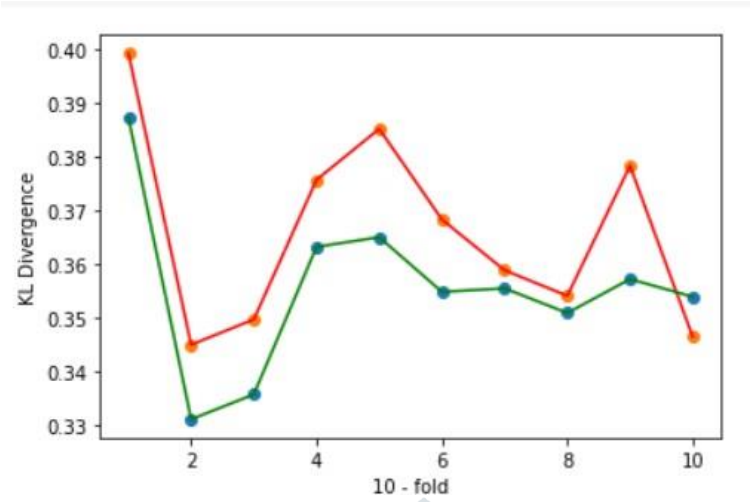

Fig. 5.14 KL divergence of male and female

In the above graph Fig. 5.14, the green line represents the KL divergence between the predicted wage of the male in the original dataset and the predicted wage when changing male into female. The red line represents the KL divergence between the predicted wage of the female in the original dataset and the predicted wage when changing female into male. From the above graph, we can conclude that the bias is more against females than males.

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 values are listed in the below table 5.1.

Table 5.1 KL Divergence values of two approaches

S. No.	Layered Method	Averaging Method
1	0.486	0.0106
2	0.452	0.0362
3	0.519	0.0221

Discussions

We proposed two different approaches for detection and evaluation of databias in the given data sample. We observed from above results that both the approaches clearly detect that data sample is biased towards which attribute value of the protected attribute. But, the KL Divergence values differ from each other.

We found that, for females, the average predicted wage increases if the samples of females are changed to male and vice-versa. Also, the average predicted wage for males decreases after applying the alternation function. The figure Fig. 5.7 and fig. 5.14 clearly indicates that the dataset is biased towards females. We tested thedataset to know whether the race attribute is potentially biased and we found that the dataset is more biased towards black than white.

The KL Divergence value in Layered approach is greater than the value in Averaging approach. We observed that the accuracy of the Averaging Method (94%) is more than Layered Method (82%). Because, in the Layered Method, the predictionsof the models (Random Forest Classifier, Decision Tree Classifier and KNN Classifier) are given as input to the Sequential model. Before giving it as input, we are balancing the data sample using SMOTE. So, the accuracy of the model decreasesand the KL Divergence value increases.

6. CONCLUSIONS AND FUTURE SCOPE

Conclusions

Data bias detection in machine learning models is particularly useful foraccurate 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. Datasets contain bias because they reflect human behavior, practice, experience, and actions. Due to bias in the training datasets, machine learning models are prone to bias, however it is critical to detect bias.

we try to detect bias by alternating values for PBAs. Then, we will evaluate the amount of bias by calculating the divergence between the original and the alternated mean of predicted class values with respect to each attribute's value. We contrasted the results using two different ways that we specified andassessed the outcomes of those approaches in our work.

Limitations

The limitations of our proposed system is listed below:

We are trying to evaluate bias using KL Divergence value but this value also includes the model prediction bias.The proposed system may not give better results if the size of the data sample is too small. Though we can detect the bias, we cannot explain the ways to mitigate it.

Future Scope

In future, an effort could be made to create a better approach, which would theoretically lead to more accurate results. A tool can be developed to find the different types of bias and their impact on the results. A better evaluation function canbe developed to evaluate the amount of bias by selecting some parameters andcriteria. A tool can be developed to mitigate the databias in the dataset if the bias exists in it.

References

[1] Alelyani, S. "Detection and Evaluation of Machine Learning Bias" *Appl. Sci.* 2021, *11*, 6271. https://doi.org/10.3390/app11146271.

[2] Martim Brandao, "Age and gender bias in pedestrian detection algorithms" , 2019, https://doi.org/10.48550/arXiv.1906.10490.

[3] Jieyu Zhao, Kai-Wei Chang, "LOGAN: Local Group Bias Detection by Clustering", 2020, https://doi.org/10.48550/arXiv.2010.02867.

[4] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster and T. Vetter,"Analyzing and Reducing the Damage of Dataset Bias to Face Recognition With Synthetic Data," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, pp. 2261-2268, doi: 10.1109/CVPRW.2019.00279.

[5]. Gianfrancesco, Milena A et al. "Potential Biases in Machine Learning Algorithms using Electronic Health Record Data." JAMA Internal Medicine 178 (2018): 1544- 1547.

[6]. Christopher G. Harris. 2020. "Methods to Evaluate Temporal Cognitive Biases in Machine Learning Prediction Models". In Companion Proceedings of the Web Conference 2020. DOI: https://doi.org/10.1145/3366424.3383418

[7]. Sun W, Nasraoui O, Shafto P (2020) "Evolution and impact of bias in human and machine learning algorithm interaction".https://doi.org/10.1371/journal.pone.0235502.

[8] I. Serna, A. Pena, A. Morales and J. Fierrez, "InsideBias: Measuring Bias in Deep Networks and Application to Face Gender Biometrics," 2020 25th International Conference on Pattern Recognition(ICPR), 2021, pp. 3720-3727, doi:10.1109/ICPR48806.2021.9412443.

[9] M. Atay, H. Gipson, T. Gwyn and K. Roy, "Evaluation of Gender Bias in Facial Recognition with Traditional Machine Learning Algorithms," 2021 IEEE Symposium Series on Computational Intelligence (SSCI), 2021, pp. 1-7,doi:10.1109/SSCI50451.2021.9660186.

[10] A. Puc, V. Struc and K. Grm, "Analysis of Race and Gender Bias in Deep Age Estimation Models," 2020 28th European Signal Processing Conference (EUSIPCO), 2021, pp. 830-834, doi: 10.23919/Eusipco47968.2020.9287219.

[11]. S. Leavy, "Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning," 2018 IEEE/ACM 1st International Workshop on Gender Equality in Software Engineering (GE), 2018, pp. 14-16.