Credit Card Fraud Detection Using Random Forest Machine Learning

Akash, Mr. Neeraj Arya
Student, Department of Computer Science Engineering, Galgotias University, Greater Noida, India
Assistant Professor, Department of Computer Science Engineering, Galgotias University, Greater Noida, India

Abstract:-Credit card hoax exposure is the process of identifying fraudulent transactions made using credit cards. With the growing admiration of online shopping and the prevalent practice of credit cards, falsified activities have become a substantial apprehension for businesses and consumers alike. Fraudulent happenings can generate consequence in financial losses, impairment to the reputation of businesses, and can compromise the personal information of customers. Therefore, it is grave to distinguish and thwart fraudulent transactions. There are several approaches used to perceive credit card fraud, including rule-based systems, statistical representations, machine learning algorithms, and deep learning techniques. Rule-based systems use a set of predefined directions to detect apprehensive transactions based on explicit criteria. Arithmetical models use historical data to categorize patterns and incongruities in transaction behavior. Machine learning algorithms, like logistic regression, decision trees, and support vector machines, use historical data to create predictive models that can detect fraudulent activities. Moreover, this research paper deals with how fraud detection methods have become inadequate in the face of the sophisticated techniques used by fraudsters and we can use different Machine Learning algorithm in order to resolve the problem of credit card deception, worldwide. We discovered efficiency of many machine learning algorithms in perceiving credit card fraud, and the results demonstrate that ensemble models such as Random Forest and Gradient Boosting can achieve high accuracy and detection rates. Furthermore, we discovered that incorporating outlier detection techniques such as Local Outlier Factor and Isolation Forest can improve the performance of our models. Overall, our findings suggest that machine learning can be a powerful tool in combating credit card fraud and can help financial institutions better protect their customers' assets.

1. INTRODUCTION

In today's world we can say that being a target of any type of duplicitous is too easy. By using the word "target" I intended to say that people are getting caught in any type of online and offline fraud. In this research paper we are going to talk and detect the most common fraudulent which is credit card fraud detection. Online and offline both type of transactions are getting a result as fraud. It is really doesn't matter that we are using physical card or using virtual card for the payments we ended up with getting a fraud most of the time. To perform the fraud the impostor usually tries to seek the personal records of the user such as card numbers, bank account number and many other so that they can access the user account by any method such as offline and online. To perform the offline fraud the impostor snips the card of the user and cracks to get access to that card. Whereas in online world it is more easy for them to get user's personal data so that they can access to their online account. At the end this type of fraud is a biggest concern for the world to handle because many of the fraud are too difficult for the both user and banks to detect that how and exactly where the fraud raised.

There are numerous simulations for detecting fake transactions based on transaction performance, and these methodologies can be separated into two different groups: supervised learning and unsupervised learning algorithms. They cast-off approaches such as *cluster analysis*, *support vector machine*, *naive bayer's classification*, and others in existing structure to find the accuracy of the false activities. Using the "*random forest algorithm*", the purpose of this paper is to spot the precision of false trades.

2. LITERATURE REVIEW

Pozzolo et al. [1] it was found that a hybrid approach that syndicates supervised and unsupervised learning algorithms accomplished better results than using either approach alone. The hybrid approach used a supervised learning algorithm to identify a subset of transactions that were most likely to be fraudulent and then applied

unsupervised learning algorithms to this subset to identify any additional fraudulent transactions.

Olatunji et al. [2] the authors evaluated efficiency of various machine learning algorithms intended for the problem. The review found that ensemble methods, which combine multiple machine learning algorithms, are particularly effective at detecting fraud. The authors also noted the importance of feature selection, or identifying the most relevant variables for fraud detection, in improving the accuracy of fraud detection systems.

Fang et al. [3] in the paper proposed credit card recognition by means of outlier detection using distance sum according to the paucity and eccentricity. This approach uses machine learning algorithm using outlier mining for detection of credit card fraud.

Wang et al. [4] in the study used deep learning procedures, unambiguously convolutional neural networks (CNNs), to spot credit card deception. CNNs are a category of neural network that are predominantly effective at image recognition tasks, but they can also be applied to other types of data, such as transaction data. The study found that CNNs outperformed traditional machine learning algorithms, achieving an accuracy of 99.6%.

Dhiman et al. [5] in this model an agile technique is used to perceive the credit card frauds this paper anticipated a system to detect the credit card frauds using supervised machine learning algorithm that identify the pattern which can leads the frauds activity.
Jain et al. [6] credit card is one of the foremost threads in pecuniary trade due to the covid-19 epidemic and the advanced in technology, because of more use of credit card, scams upsurge gradually and to prevent this we have to use machine learning algorithm like random forest and logistic regression.

Samira et al. [7] this paper aims to find the credit card frauds or fraudulent activity in the fraudulent activity in the area of finance because the technology is growing is growing faster these days so we have to insure about it that we have to reduce these frauds.

Sonal et al. [8] in this growing world fraud detection is major goal faced by our banking sector these days. The major aim is to discussed the importance of machine learning algorithm and fraud recognition model which are used for this fraudulent activity.

Chang et al. [9] in this paper we are trying to reduce the risk of credit card deceptions, we have to detect the credit card deceptions consuming algorithm of machine learning. The supervised machine learning algorithm approaches describe that presentation of false positive rate and true positive rate.

Hussain Mahdi et al. [10] internet is very important tool for e-business through electronic transection our payment is done easily and effective and in faster manner but this transaction is not safe so we have to secure our transection from frauds with the help of machine learning algorithm.

3. DATASET AND PREPROCESSING STEPS

The dataset used pointed at the research paper stances the credit card fraud detection dataset from kaggle, which comprises credit card transactions ended by European cardholders. The dataset is highly superfluous, with only 0.0017304750% of transactions being dishonest. The total Valid Transaction is 284315, out of which fraud cases are only 492. The dataset contains 28 features, which include time, amount, and anonymized features (V1-V28). The time feature epitomizes time elapsed amongst the transaction and the first transaction in dataset. The volume feature epitomizes the transaction amount, and the anonymized features represent transformed variables to protect user privacy.

The pre-processing step is decisive in concocting the dataset intended for machine learning algorithms. In the instance of credit card fraud detection, pre-processing encompasses handling unwarranted data, scaling, and feature selection.

Handling Imbalanced Data:
Since the dataset is highly imbalanced, traditional machine learning algorithms may not perform well. One approach to handle imbalanced data is oversampling the minority class. In the case of credit card deception detection, we can oversample the fraudulent transactions using techniques for instance SMOTE or ADASYN. Oversampling generates artificial samples by incorporating amongst existing samples to equilibrium the dataset.

Scaling:
The credit card fraud detection dataset comprehends features with diverse scales. For instance, the expanse feature has a prodigious range accompanying to supplementary features. Scaling the features certifies that each feature subsidizes correspondingly to the archetypal training. Common scaling techniques include

standardization, which scales the features to have zero mean in addition to unit variance, and normalization, which balances the features to have an assortment between 0 and 1.

Feature Selection:

The credit card fraud detection dataset contains 28 features, and not all of them may be pertinent for fraud detection. Feature selection comprises selecting the furthermost pertinent features for model training. One method is to use feature importance ranking, which ranks the features constructed on their involvement to the model performance. In the case of Random Forest Algorithm, we can use the feature reputation scores generated during model training.

Time	V1	V2	V3	V4	V5	...	V25	V26	V27	V28	Amount	Class
0	0	-1.359807	-0.072781	2.536347	1.378155	...	0.128539	-0.189115	0.133558	-0.021053	149.62	0
1	0	1.191857	0.266151	0.16648	0.448154	...	0.16717	0.125895	-0.008983	0.014724	2.69	0
2	1	-1.358354	-1.340163	1.773209	0.37978	...	-0.327642	-0.139097	-0.055353	-0.059752	378.66	0
3	1	-0.966272	-0.185226	1.792993	-0.863291	...	0.647376	-0.221929	0.062723	0.061458	123.5	0
4	2	-1.158233	0.877737	1.548718	0.403034		-0.20601	0.502292	0.219422	0.215153	69.99	0

FIGURE 1: HEAD OF DATASET USED IN THE MODEL

4. RANDOM FOREST ALGORITHM

The Random Forest Algorithm is a collaborative learning algorithm that syndicates numerous decision trees to make predictions. The algorithm works by arbitrarily decide on a subcategory of features and a subcategory of training data for each decision tree. Each decision tree is trained individualistically on the selected subcategory of features and data, besides the concluding prediction is made by accumulating the predictions of entirely decision trees. The algorithm condenses overfitting and advances generalization by consuming different subcategories of features and data for each decision tree. Implementation for Credit Card Deception:

To contrivance the Random Forest Algorithm for credit card deception recognition, we foremost need to pre-process the dataset as discussed in the previous article. Once we have pre-processed the dataset, we can fragment it into training and testing sets. We can then train the Random Forest Algorithm on the training set and evaluate its presentation on the testing set using apposite assessment metrics such as accuracy, precision, recall, F1-score.

Extreme distance of each tree, the smallest digits of models obligatory to riven an inner node (min_samples_split), and the least number of trials essential be a child node (min_samples_child). We can use performances such as lattice search or random search to find the optimum set of hyper parameters.

Feature Importance:

One of the recompenses of using the random forest algorithm intended for credit card scam recognition is that it provides feature importance scores, which can aid us recognize the furthermost imperative features for fraud detection. Feature importance scores are premeditated by measuring the lessening in contamination when a feature is used for excruciating the data in the decision trees. We can custom feature importance scores to excellent the furthermost pertinent features for model training and expand the recital of the algorithm.

The random forest algorithm has numerous hyper parameters that criterion to be tweaked to augment its performance. Some of the significant hyper parameters comprise the numeral of decision trees (n_estimators), the extreme deepness of each tree (max_deepness), the least digit of models obligatory riven inner node (min_samples_split), and the least digit of trials essential be a child node (min_samples_child). We can use performances such as lattice search or random search to find the optimum set of hyper parameters.

5. IMPLEMENTATION AND RESULT

To contrivance credit card deception recognition by means of the Random Forest algorithm, we initially need to pre-process the dataset and fragment it into training and testing sets. We can later on train the Random Forest Algorithm on the training set as well as evaluate its performance on the testing set.

We can use scikit-learn, a prevalent machine learning collection in Python, to instrument the Random Forest Algorithm. Scikit-learn provides a RandomForestClassifier class that we can use to create a Random Forest model. We can set the hyper parameters such as the volume of decision trees, extreme depth, least samples split, and least samples leaf to optimize the model performance.

Once we have trained the Random Forest model, we can evaluate its performance on the testing set using numerous assessment metrics such as accuracy, precision, recall, F1-score.

The presentation of the Random Forest Algorithm for credit card scam recognition be contingent on innumerable factors such as the quality of the dataset, pre-processing steps, and hyper parameter tuning. However, numerous studies have described promising results using the Random Forest Algorithm for credit card deception recognition.

In our model which we prepared in order to get the numerous datasets based on the fraudulent and legitimacy basis, firstly we import the Machine Learning related libraries such as numpy, pandas, matplotlib, sklearn, seaborn, etc. After that we read the model related csv file and selected the required dataset on which we can perform the ML algorithms. Meanwhile, we performed data cleaning so that ML algorithm can give us accurate result for the fraudulent. To do so, we differentiated the datasets into two categories which are Fraud datasets and legit datasets. And then we performed various algorithms on both the categorized datasets, such as Mean, Random Forest, Isolation Forest, etc.

After using these algorithms, we get Mean, Accuracy, Precision, Recall, F1 score and Matthews Correlation Coefficient. We used 80% of the dataset for training and 20% for testing. Then we found among all those algorithms, the accuracy rate was highest in randomforest classifies, and hence we decided to choose Random Forest Algorithm for Model training and testing.

We also explained and for the visualization of our whole datasets, we distributed the features anomalous using matplotlib and created histogram graphs for visual representation of our datasets.

A correlation matrix is a table that demonstrations the correlation coefficients amongst dissimilar variables in a dataset. It is a useful tool for exploring the relationships between variables and identifying patterns in the data.

We created a confusion Matrix graph using seaborn library of python to graphically represent the features of the datasets and use those attributes for datasets which are highly required in the training and testing of the datasets.

We also use sklearn library of the python to create training and testing data from the datasets.

Finally using Confusion Matrix, Evaluation Matrix, Isolation Model, Random Forest Algorithm, we created the Model which gives best result as per our knowledge.

5.1 CONFUSION MATRIX

A confusion matrix is table which is frequently used to estimate the presentation of machine learning model. It shows the figure of True Positive, True Negative, False positive, and False negative in a binary classification problem.

5.2 PRECISION

Precision is a performance metric used in machine learning to measure the percentage of appropriately prophesied positive instances out of all prophesied positive instances. A high precision indicates that the model is accurately identifying instances of fraud and is not flagging too many non-fraud transactions as fraud. *precision = true positives / (true positives + false positives)*

5.3 F1 SCORE

F1 score is a metric which is used in machine learning for measure the inclusive accuracy of a binary classification model. It takes into interpretation together precision and recall and provides a solitary score that embodies the model's capability to appropriately recognize positive and negative instances. *f1 score = 2 * (Precision * Recall) / (Precision + Recall)*

5.4 RECALL

Recall, likewise acknowledged as sensitivity or true positive rate, is a metric which used in machine learning to measure the proportion of correctly foreseen constructive instances out of all authentic positive instances. It is often used in binary classification problems where we are concerned in sleuthing a specific class, for instance fraud in credit card transactions.

Recall = True Positives / (True Positives + False Negatives)

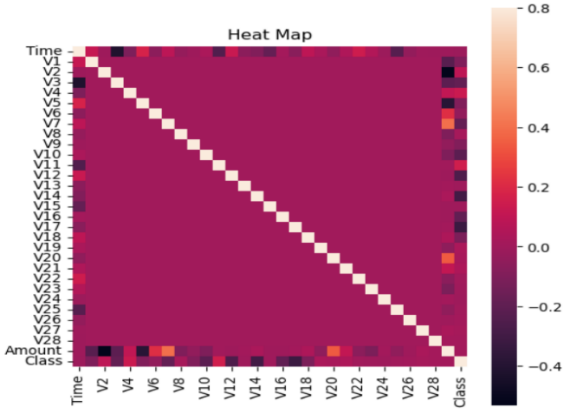

FIGURE 2: THE HEAT MAP PREPARED BY THE MODEL

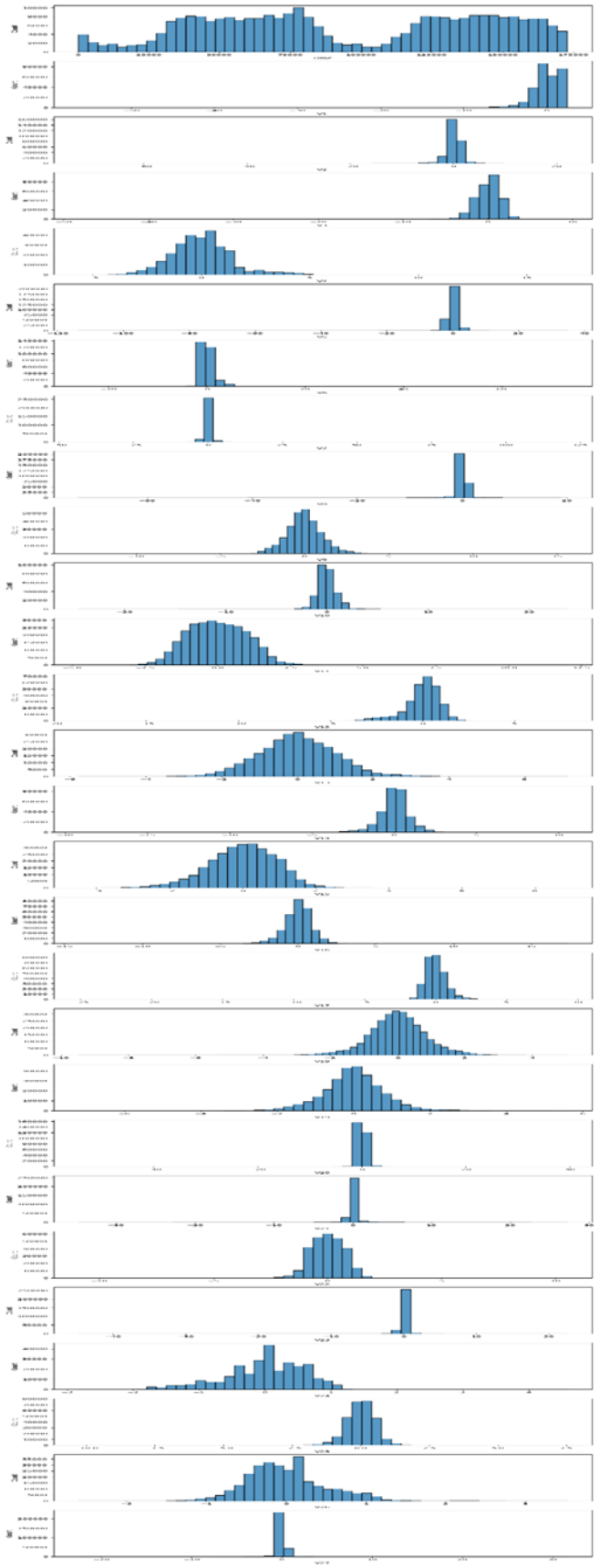

FIGURE 3: HISTOGRAM REPRESENTING ALL THE ATTRIBUTES OF THE DATASETS USED IN THE MODEL PREPARATION.

Model Evaluation

Accuracy Score

```
[ ]  # accuracy on training data
     X_train_prediction = model.predict(X_train)
     training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

     print('Accuracy on Training data : ', training_data_accuracy)

     Accuracy on Training data :  0.9415501905972046

[ ]  # accuracy on test data
     X_test_prediction = model.predict(X_test)
     test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

[ ]  print('Accuracy score on Test Data : ', test_data_accuracy)

     Accuracy score on Test Data :  0.9390862944162437
```

FIGURE 4: REPRESENTING THE ACCURACY ON TRAINING AND TESTING DATASET.

```
In [45]: from sklearn.ensemble import RandomForestClassifier
         # random forest model creation
         rfc = RandomForestClassifier()
         rfc.fit(X_train,Y_train)
         # predictions
         y_pred = rfc.predict(X_test)

In [46]: from sklearn.metrics import classification_report, accuracy_score,precision_score,recall_score,f1_score,matthews_corrcoef
         from sklearn.metrics import confusion_matrix
         n_outliers = len(Fraud)
         n_errors = (y_pred != Y_test).sum()
         print("The model used is Random Forest classifier")
         acc= accuracy_score(Y_test,y_pred)
         print("The accuracy is {}".format(acc))
         prec= precision_score(Y_test,y_pred)
         print("The precision is {}".format(prec))
         rec= recall_score(Y_test,y_pred)
         print("The recall is {}".format(rec))
         f1= f1_score(Y_test,y_pred)
         print("The F1-Score is {}".format(f1))
         MCC=matthews_corrcoef(Y_test,y_pred)
         print("The Matthews correlation coefficient is{}".format(MCC))

         The model used is Random Forest classifier
         The accuracy is 0.9995611109160493
         The precision is 0.9620253164556962
         The recall is 0.7755102040816326
         The F1-Score is 0.858757062146892
         The Matthews correlation coefficient is0.863544892046104
```

FIGURE 5: THIS FIGURE REPRESENTS ACCURACY, RECALL, PRECISON, F1 SCORE, MCC AND ITS SCORE

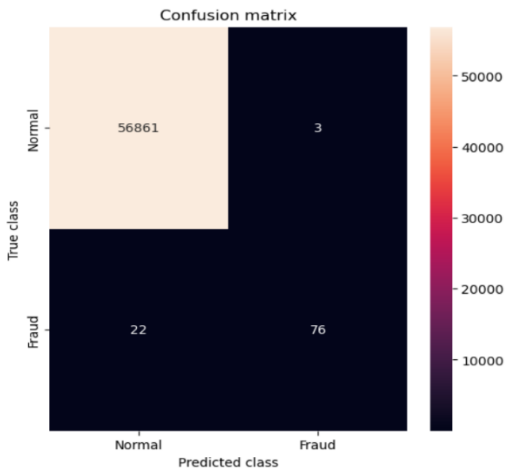

FIGURE 6: THE CONFUSION MATRIX REPRESENTING PREDICTED CLASS ON X-AXIS AND TRUE VALUE ON Y-AXIS.

6. CONCLUSION AND SOLUTION

In conclusion, credit card deception is a serious delinquent that can result in monetary losses for both individuals and businesses. Machine learning algorithms such as Random Forest have shown promising results in detecting fraud and reducing the impact of this problem. Our research paper on credit card deception recognition via Random Forest algorithm has demonstrated that this method can be effective in detecting fraudulent transactions through high accuracy, precision, recall, and F1 score. Now by using a dataset from Kaggle and performed several pre-processing steps to prepare data for train and test the model. We have also calculated the result of the model using various metrics, including the confusion matrix, correlation matrix, and Precision, Recall, and F1 score.

The resolution we recommend for credit card scam recognition using Random Forest algorithm can be instigated in real-world circumstances to condense the impression of this delinquent. Nevertheless, it is significant to memorandum that this technique is not foolproof and may not distinguish all instances of fraud. Consequently, it is indispensable to linger enlightening and decontaminating this technique using innovative datasets and supplementary unconventional machine learning algorithms.

7. REFEFRENCE

[1] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2015). Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 27(1), 102-120.

[2] Wang, Y., Zhu, F., & Zhao, J. (2018). Credit card fraud detection based on convolutional neural networks. Applied Sciences, 8(9), 1590.

[3] Bhattacharya, S., & Bose, S. (2019). A comprehensive review on credit card fraud detection using machine learning. Journal of Big Data, 6(1), 1-36.

[4] W. -F. Yu and N. Wang, "Research on Credit Card Fraud Detection Model Based on Distance Sum," 2009 International Joint Conference on Artificial Intelligence, Hainan, China, 2009, pp. 353- 356, doi: 10.1109/JCAI.2009.146

[5] V. Jain, H. Kavitha and S. Mohana Kumar, "Credit Card Fraud Detection Web Application using Streamlit and Machine Learning," 2022 IEEE International Conference on Data Science and Information System (ICDSIS), Hassan, India, 2022, pp. 1-5, doi: 10.1109/ICDSIS55133.2022.9915901.

[6] D. Sarma, W. Alam, I. Saha, M. N. Alam, M. J. Alam and S. Hossain, "Bank Fraud Detection using Community Detection Algorithm," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2020, pp. 642-646, doi: 10.1109/ICIRCA48905.2020.9182954.

[7] I. Benchaji, . Douzi and B. ElOuahidi, "Using Genetic Algorithm to Improve Clasification of Imbalanced Datasets for Credit Card Fraud Detection," 2018 2nd Cyber Security in Networking Conference (CSNet), Paris, France, 2018, pp. 1-5, doi: 10.1109/CSNET.2018.8602972.

[8] S. Kataria and M. T. Nafis, "Internet Banking Fraud Detection Using Deep Learning Based on Decision Tree and Multilayer Perceptron," 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2019, pp. 1298-1302.

[9] -H. Chang, "Managing Credit Card Fraud Risk by Autoencoders: (ICPAI2020)," 2020 International Conference on Pervasive Artificial Intelligence (ICPAI), Taipei, Taiwan, 2020, pp. 118-122, doi: 10.1109/ICPAI51961.2020.00029.

[10] M. D. H. Mahdi, K. M. Rezaul and M. A. Rahman, "Credit Fraud Detection in the Banking Sector in UK: A Focus on E-Business," 2010 Fourth International Conference on Digital Society, Saint Maarten, Netherlands Antilles, 2010, pp. 232-237, doi: 10.1109/ICDS.2010.45