International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 03, 2021

Differential Privacy Preservation Mechanism Using Bernstein Polynomial Function For Heart Disease Dataset

Kousika N[1], Premalatha K[2]

[1]Sri Krishna College of Engineering and Technology, Coimbatore, Tamil Nadu
[2]Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu

Email: [1]kousika.cs@gmail.com, [2]kpl_barath@yahoo.co.in

Abstract: Information is put away in various frameworks as emerging technologies allows influential data gathering and processing. Protection and security have turn into a longstanding challenging issue with advances in data and communication innovation. Privacy preserving data mining makes the customer information more secure by means of data perturbation and also it makes it harder to identify a person in an occurrence of data is spilled. Machine learning has got attention recently due to an energetic advancement of differential privacy (DP). DP is golden response scheme to address the privacy protection in analysis of data but it is quite hard to implement on real world data. The proposed system uses Bernstein polynomial function under differential privacy for perturbation. Heart disease dataset is used in this work to analyze the performance between the original and the modified dataset using DP with the classifier models decision tree, linear model, random forest, SVM, linear model and neural network. The experiment results show the minor variations in the accuracy, sensitivity and specificity measures.

Keywords: Privacy Preserving Data Mining, Machine Learning, Differential Privacy, Bernstein Model

1. INTRODUCTION

Data collectors and decision makers frequently need to examine data containing confidential information of individuals. The customer behavior identified from collected data can be shared among different tech companies for business purpose[15]. The necessity increases for powerful and conceptually meticulous algorithms that are suitable for preserving privacy. In recent days, differential privacy has established as a great data protection paradigm [1]. DP allows companies to access data for investigation by ensuring the privacy of user groups. DP is a technology which ensures strong privacy; prevent data leakage and re-identifies people within a dataset [2]. Differential privacy enhancing technologies are used by Facebook for efficient data accessing and to provide data to the public without apparent data loss. DP principles are applied for the fulfillment of data protection and overcome the limitations on data generalization [3].

Privacy can be measured by various means. Most efficient privacy protection method scan be implemented that are reliable for handling data source attackers. Differentially private systems are designed in a way that nobody can predict the information. Two basic categories of perturbation techniques are used for arbitrarily adjusting the source data. DP is the widely available solution for the applications with sensitive data, statistics release and accessing of sensitive data by opponents. It computationally requires multiple stages to enhance privacy. As a first step, the random statistical noise id added to the original data for protecting sensitive information. The noisy and unclear data is retrieved by the opponents so that privacy is not revealed. Then, simple masking method is applied on individual's data and thereby protect privacy and also to improve the variance of additional noise. Based on the sensitivity of the data, the quantity of the noise to be added will be decided. The quantity of noise in the database depends on the number of people [4].

In data preprocessing, one key aspect is feature selection. Data correlation is the tool to explain the connectivity among several different parameters in the statistical database. Statistical correlation helps to recognize the closeness of the attributes and identify the missing values in the dataset. Sensitivity analysis is the business strategy that measures the extent to which the input parameters are influenced because of randomly approximated model. The data analyst can predict how the adjustment in the parameter changes the results. Performing sensitivity analysis on the statistical database gives thorough knowledge of attributes and the area in which enhancements can be done [1]. Data Sanitization is the process of attempting to safeguard the sensitive data in the statistical data source. There are two ways in applying sanitization for the statistical dataset. First, the collected data can be supplied to the analysts as a waterproof version by applying some perturbation or data manipulation strategies. This process is called
anonymization. Second, the data collector provides a framework for accessing the data that provides sanitized data from dataset [1]. In the proposed work, the quazi identifiers are perturbed using Bernstein polynomial function which is used in DP mechanism[13]. The original data set and modified data set are analyzed with the classifiers decision tree, linear model, random forest, SVM, linear model and neural network. The remaining paper is organized in the following way. Section 2 presents the related work [14]. The proposed methodology details are provided in section
3. The results of our experiments are shown in section 4. Section 5 provides the conclusion.

2. LITERATURE SURVEY

Thaler et.al [7] proved that polynomial approach is helpful for introducing external function in differential privacy. They applied polynomial function directly to the dataset and generated perturbed records. The learning algorithms have not used that restricts the database access. The query results are improved by producing minimal error rate and the time taken for execution. Ivan Damgard et al., [8] proposed a protocol with three category players: client input, client output, and n servers. There are three roles for each player. The block of l different values is simultaneously divided between l distinct points by a polynomial. In packed secret sharing method is started with an arithmetic circuit C with a minimum of l gates[5]. Every layer in the circuit is having a unique type of gate. The values stored in the blocks lined up correctly by performing permutation function between blocks. The computational complexity depends on the number of players participating in the computation and the communication overhead is calculated as a large polynomial in k and n. Hall, Rinaldo et.al [9] introduces cryptographic based differential data protection mechanism to release querying data. Gaussian noise is added to disclose the summary results[6]. The result shows that using Gaussian process achieves weaker form of confidentiality and data utility rate. Christos Dimitrakakis et. al
[10] proposed posterior sampling to offer divergence query responses. Le Cam's method is

applied for obtaining lower levels in distinctiveness boundaries. It is shown that Bayesian inference and posterior distribution are reliable but low utility rates is received.

3. PROPOSED WORK:

The Bernstein polynomial mechanism is used for privatizing the values in the data source. In contrast to the existing perturbation methodologies, a Bernstein functional approximation mechanism is used for sanitization. Bernstein polynoms uses Stone-Weiestrass approximation theorem and it achieves high data utility and rapid rate of divergence [1]. Let B_n(f)be the continuous function of n^{th} polynomial between the thresholds [0 to 1] for the real life feature f.

$$B_n(f) = \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} f\left(\frac{k}{n}\right) \qquad (1)$$

Bernstein polynomial supports several realistic properties like positivity, normalization and exclusive restricted upper limit[12]. q-Bernstein polynomials take over some features from traditional Bernstein polynomial and gives good convergence rate [11]. The parametric curve which is used to represent the coefficients of Bernstein polynomial function is Bezier curve. Figure 1 shows the proposed work using Bernstein polynomial function.

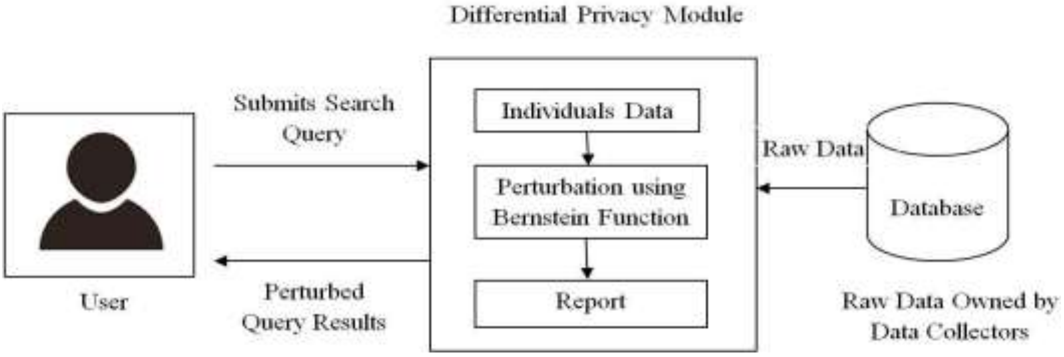

Differential Privacy Module
Submits Search Query
Individuals Data
Perturbation using Bernstein Function
Report
Raw Data
Database
User
Perturbed Query Results
Raw Data Owned by Data Collectors

Fig.1. Perturbation of Quazi identifiers using Bernstein polynomial function

4. RESULTS AND DISCUSSION

The experiments are conducted using a real life statistical Heart disease dataset from UCI machine learning repository. In the dataset, there are 14 attributes 270 records. The Age is modified using Bernstein polynomial function. An input to the Bernstein function is between 0 and 1. The Age is normalized between 0 and 1 and passed as argument to the Bernstein function. Fig.2. shows the quantile plot for Age and modified age using Bernstein function.

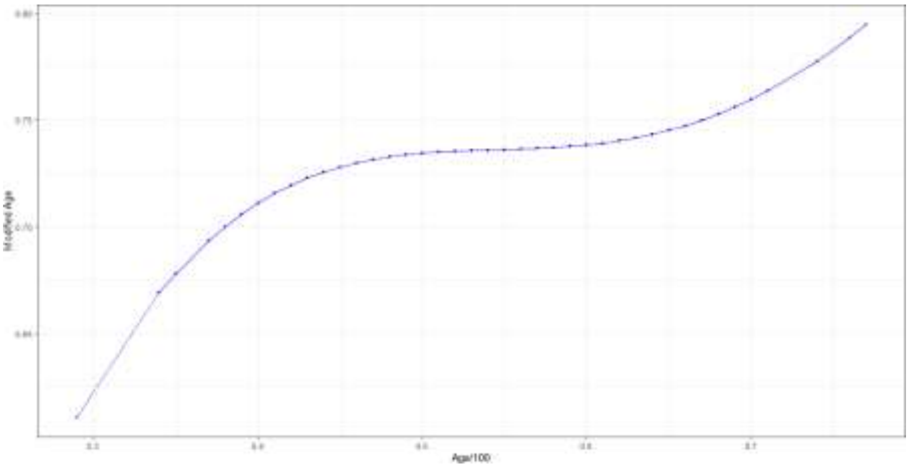


Fig.2. Qplot for age and modified age using Bernstein function

Fig.3. shows the line chart of Age and the modified age. The red color line shows the line chart drawn for Age and the Bernstein value of the age. The blue color line shows the age and the difference between the age and Bernstein value of the age is added with Bernstein age. In this work, the blue color value of age is considered as modified age.

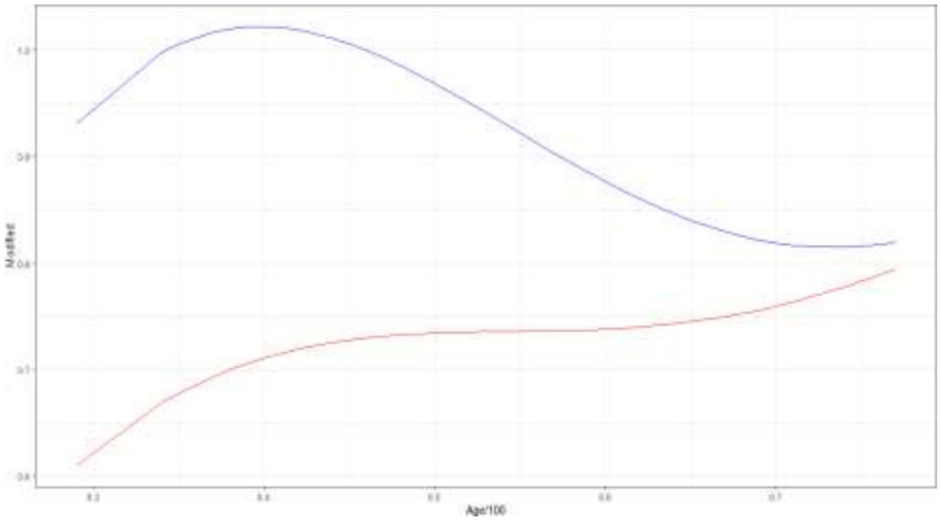


Fig.3. Line chart for age and modified age of the heart disease dataset

Before data mining it is important to review the distributions of value of original and perturbed datasets. It helps us the visual explorations and reveals the distributions of the data. Fig.4 and 5 show the Benford’s law applied for the attribute age in original data and perturbed data. The figures show that how the original data varies from perturbed data.

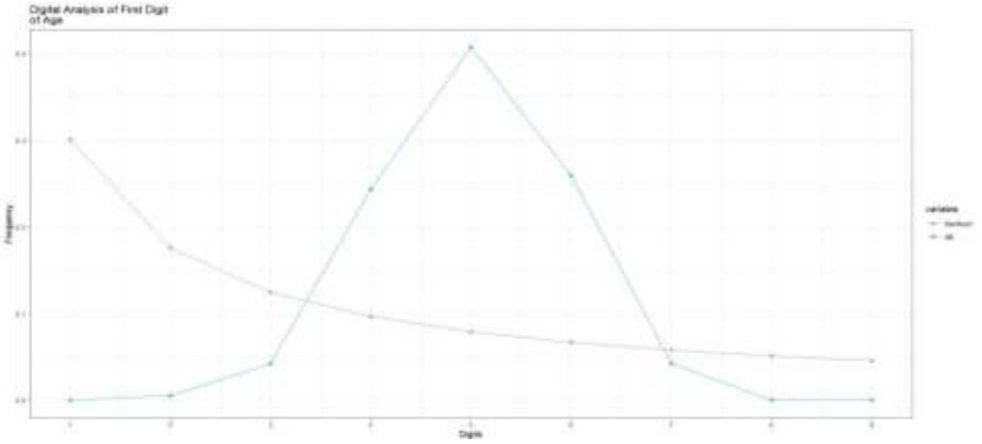


Fig.4. Benford's law for age of original data

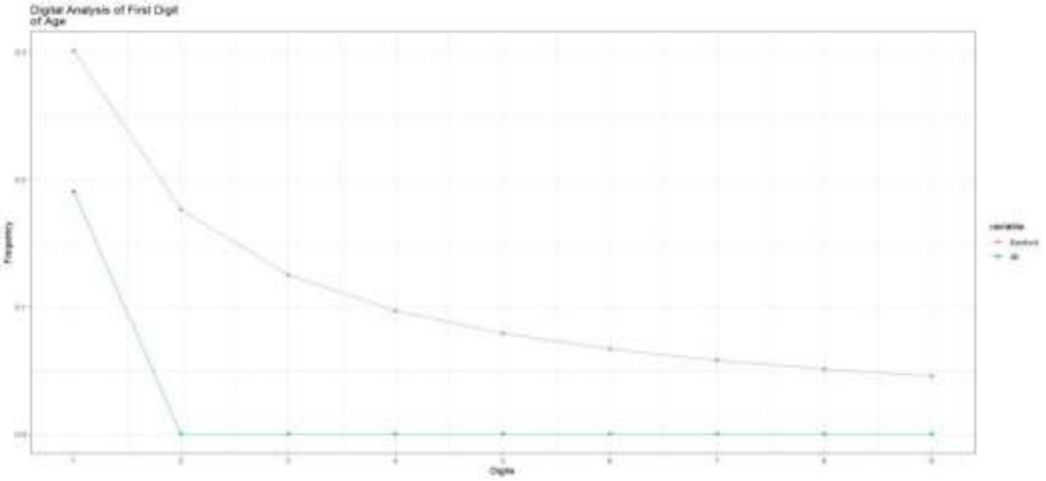
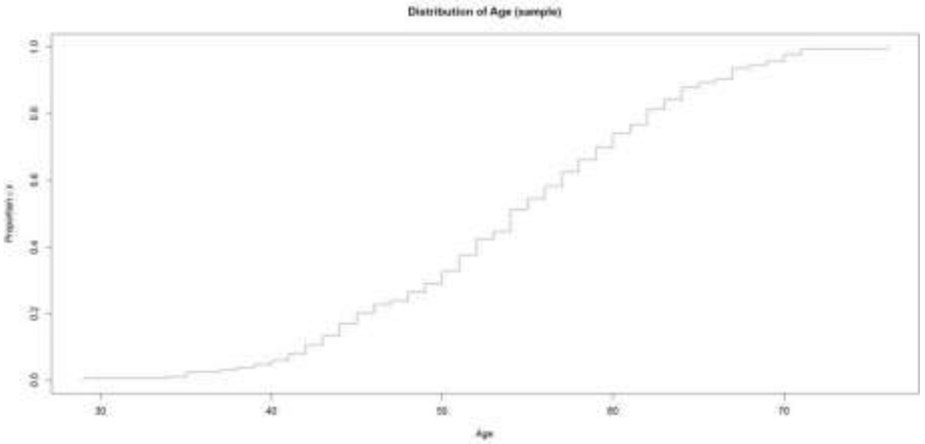


Fig.5. Benford's law for age of perturbed heart disease dataset

Fig.6. and 7 shows the cumulative distribution of age and the perturbed value of age respectively. In both the model it increases linearly from the lower value to upper value but



the range is different.

Fig.6. Cumulative chart for age of heart disease dataset

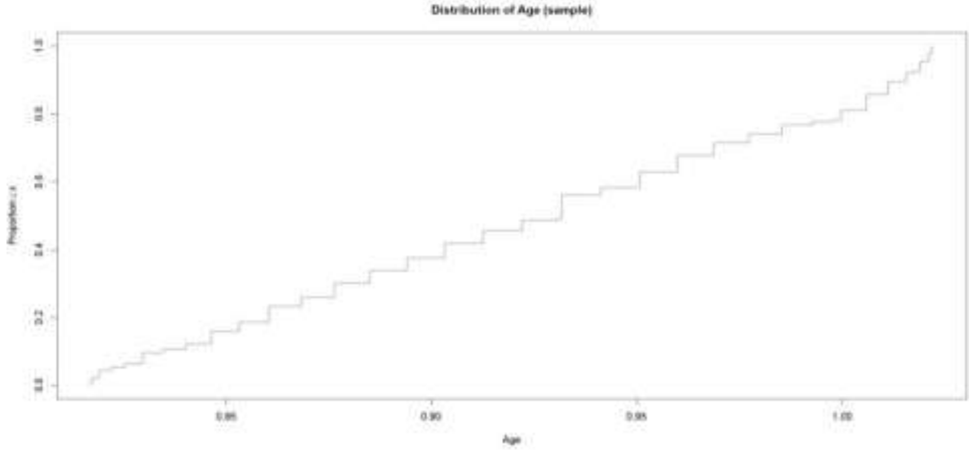


Fig.7. Cumulative chart for age of perturbed heart disease dataset

Fig.8. and 9 shows the density distribution of the age and perturbed value of age respectively.

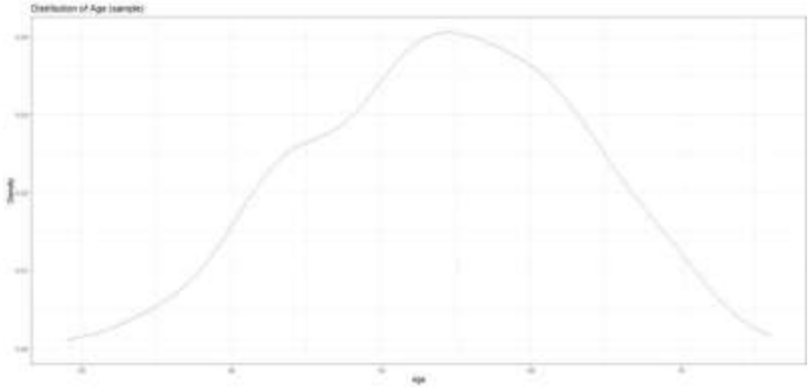


Fig.8. Density model for age of heart disease dataset

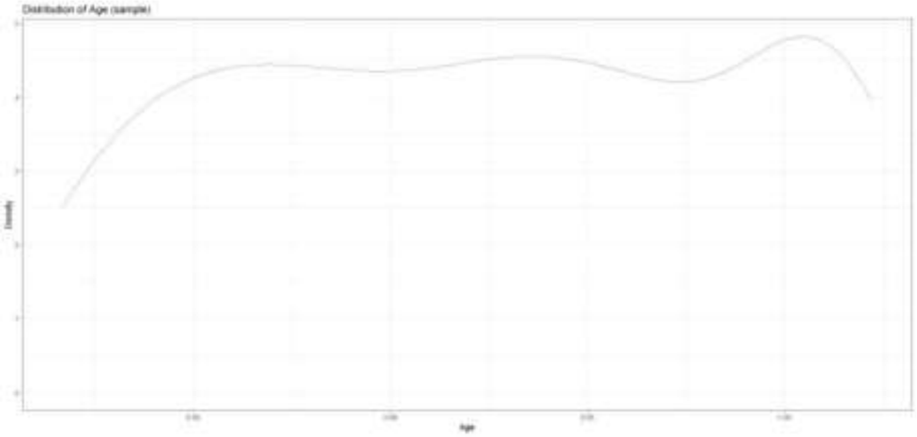


Fig.9. Density model for age of perturbed heart disease dataset

Fig.10. and 11 shows the correlation between the attributes in the original dataset and the perturbed dataset respectively. It is observed that the relationship between the age with the other attributes in the original dataset and perturbed dataset is different. From the above analysis, it is shown that how the original age is different from the modified age.

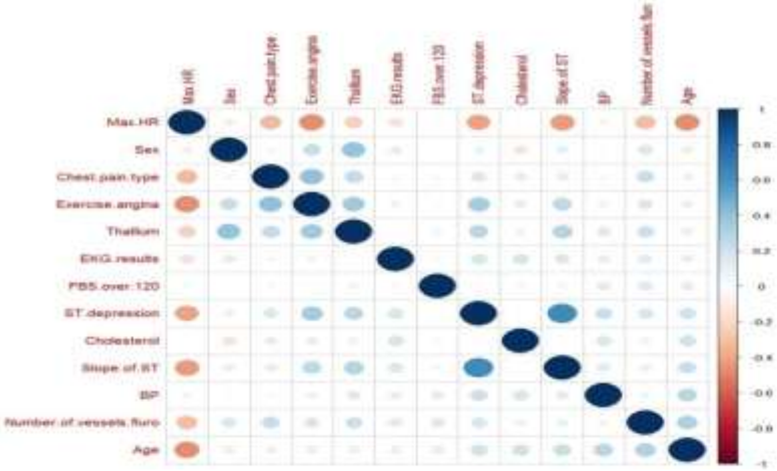


Fig.10. Correlation of attributes in heart disease dataset

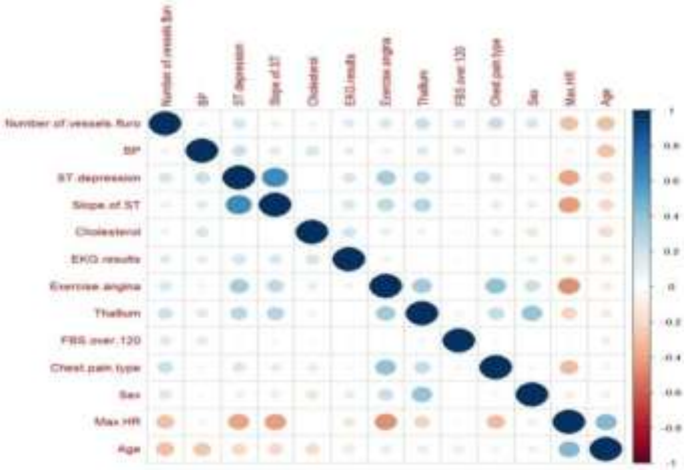


Fig.11. Correlation of attributes in perturbed heart disease dataset

The classifier models decision tree, linear model, random forest, SVM, linear model and neural network are applied in the original heart disease dataset and perturbed heart disease dataset. Fig.12. shows the decision tree obtained from the original dataset. The same decision tree is obtained for perturbed dataset also.

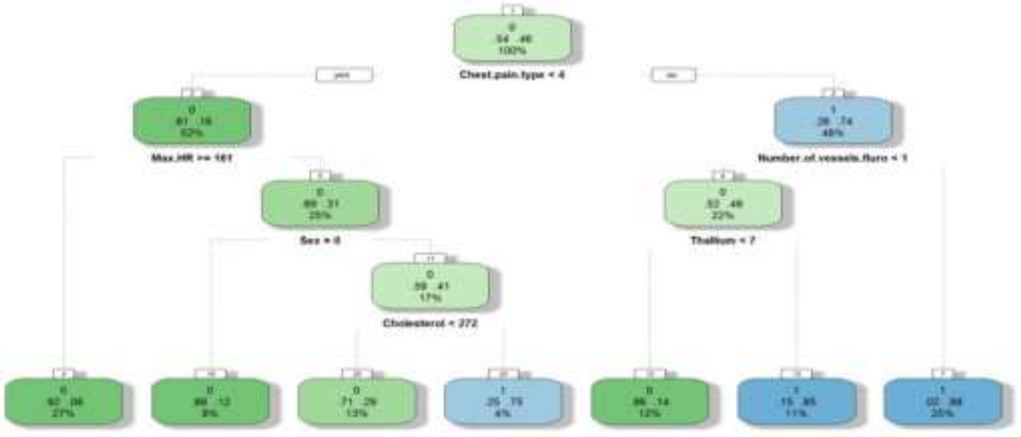


Fig.12. Decision tree model for heart disease dataset

Fig.13, 14 and 15 show the accuracy, sensitivity and specificity obtained from the original and perturbed heart disease dataset.

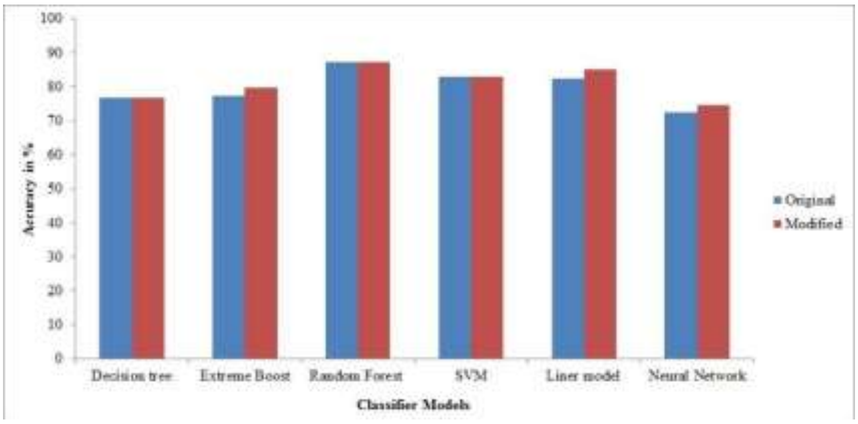


Fig.13. Accuracy Analysis for original and perturbed dataset

The sensitivity gives the proportion of true positive rate (TPR) and the specificity gives the proportion of true negative rate (TNR) of medical diagnosis from synthetic dataset. These two are the essential measures for accurate test results but it cannot be use to identify the probability of records having particular disease. The outcome shows that utility of the information is maintained after data perturbation.

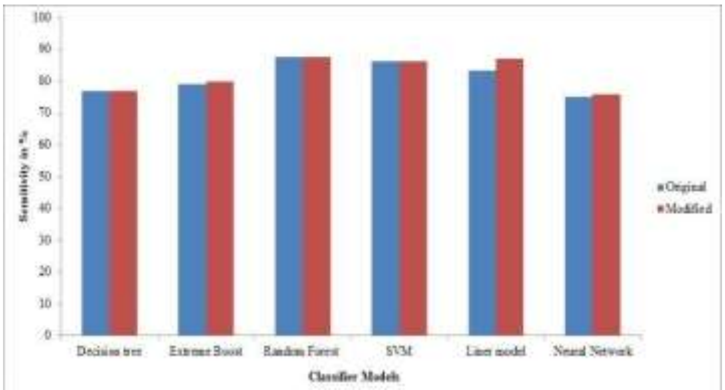


Fig.14. Sensitivity analysis for original and perturbed dataset

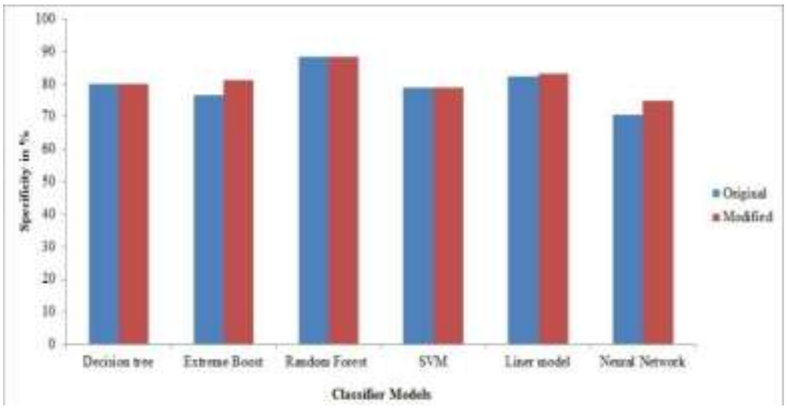


Fig.15. Specificity analysisfor original and perturbed dataset

5. CONCLUSION

In this work, heart disease dataset is used to measure the privacy preserving data mining with the classifier models decision tree, linear model, random forest, SVM, linear model and neural network. Bernstein function is used to perturb the data. Through this report, it is no harm to release the responsive training samples by protecting confidential information. The concept of quasi-identifier is used here to show the preservation of privacy of published data. An easy and efficient method that utilizes Bernstein polynomial is used to perturb the quazi attributes of the dataset. The appropriateness of the quasi-identifier is examined and it is observed that privacy is preserved. The proposed algorithm minimizes the probability of re-identification of record retrieval by perturbing the features that expose the record.

6. REFERENCES

[1] Dwork C., McSherry F., Nissim K., Smith A, "Calibrating Noise to Sensitivity in Private Data Analysis. In: Halevi S., Rabin

[2] T. (eds) Theory of Cryptography", Lecture Notes in Computer Science, vol 3876. Springer, Berlin, Heidelberg, https://doi.org/10.1007/11681878_14 (2006).

[3] A platform for differential privacy, Sarah Bird, Joshua Allen and Kathleen Walker, Microsoft (2020).

[4] Chin, Andrew and Klinefelter, Anne, Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study, North Carolina Law Review, Vol. 90, No. 5, UNC Legal Studies Research Paper No. 2062447, Available at SSRN: https://ssrn.com/abstract=2062447 (2012).

[5] Dwork, Cynthia, and Aaron Roth, "The algorithmic foundations of differential privacy", Foundations and Trends in

[6] Theoretical Computer Science 9.3-4, 211-407 (2014).

[7] C. Dwork, "The Promise of Differential Privacy: A Tutorial on Algorithmic Techniques," 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science, Palm Springs, CA, pp. 1-2, doi: 10.1109/FOCS.2011.88 (2011).

[8] Aldà, Francesco, and Benjamin IP Rubinstein, "The bernstein mechanism: Function release under differential privacy", Thirty-First AAAI Conference on Artificial Intelligence (2017).

[9] Thaler J, Ullman J, and Vadhan S., "Faster algorithms for privately releasing arginals. In Automata, Languages, and Programming", Springer. 810–821 (2012).

[10] Damgård, Ivan, Yuval Ishai, and Mikkel Krøigaard, "Perfectly secure multiparty computation and the computational overhead of cryptography", International conference on the theory and applications of cryptographic techniques, Springer (2010).

[11] Hall, Rob, Alessandro Rinaldo, and Larry Wasserman, "Differential privacy for functions and functional data", Journal of

[12] Machine Learning Research 703-727 (2013).

[13] Christos Dimitrakakis, Blaine Nelson, Zuhe Zhang, Aikaterini Mitrokotsa, and Benjamin IP Rubinstein, "Differential privacy for Bayesian inference through posterior sampling", The Journal of Machine Learning Research 18, No. 1, 343-381 (2017).

[14] Wu, Xuezhi, "Approximation by-Bernstein Polynomials in the Case", Abstract and Applied Analysis, Vol. 2014, Hindawi,

[15] (2014).

[16] https://mathworld.wolfram.com/BernsteinPolynomial.html

[17] http://www.2dcurves.com/polynomial/polynomialb.html

[18] Mao, J., Sun, Q., Wang, X., Muthu, B., & Sujatha Krishnamoorthy, S. (2020). The importance of public support in the implementation of green transportation in the smart cities. Journal of Computational Intelligence. Wiley publications .https://doi.org/10.1111/coin.12326. 26th April 2020

[19] V.R. Balaji, Maheswaran S, M. Rajesh Babu, M. Kowsigan, Prabhu E., Venkatachalam K,Combining statistical models using modified spectral subtraction method for embedded system,Microprocessors and Microsystems, Volume 73,2020.