Integrasi *Synthetic Minority Over-Sampling Technique (SMOTE)* dengan *Correlated Naïve Bayes Classifier (C-NBC)* pada Klasifikasi Siswa Berkesulitan Belajar

Zia Ulhaq, Teguh Bharata Adji
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Grafika No.2 Campus UGM, Yogyakarta 55281, Indonesia
zia.mti13@mail.ugm.ac.id, adji@ugm.ac.id

***Abstract*—Kesulitan belajar *(Learning Disabilities)* adalah cacat mental seumur hidup yang mengakitbatkan kesulitan belajar. Siswa berkesulitan belajar perlu diidentifikasi sedini mungkin agar mendapatkan perlakuan khusus yang tidak diperoleh di sekolah formal untuk mencegah kegagalan dalam pendidikan. Objek penelitian ini adalah data siswa berkesulitan belajar yang memiliki permasalahan ketidakseimbangan kelas dengan perbandingan jumlah kelas sebesar 1:5. Penelitian ini bertujuan mengkombinasikan *Synthetic Minority Over-Sampling Technique* (SMOTE) untuk mengatasi ketidakseimbangan kelas pada klasifikasi menggunakan metode *Correlated Naïve Bayes Classifier* (C-NBC). Pengukuran kinerja klasifikasi dilakukan menggukan *confussion matrix* berdasarkan nilai akurasi, sensitivitas, *spesificity*. Dari hasil pengujian yang dilakukan, kombinasi metode C-NBC dan SMOTE menghasilkan akurasi, sensitivitas dan *spesificity* lebih baik dari C-NBC yaitu masing – masing sebesar 96.53%, 99.44%, 95.09 %.**

Keywords—siswa berkesulitan belajar; C-NBC; SMOTE; ketidaksimbangan kelas; confusion matrix

I. PENDAHULUAN

Kesulitan belajar *(Learning Disabilities)* adalah cacat seumur hidup yang mempengaruhi kehidupan seseorang, dan ditandai perbedaan prestasi yang signifikan yang memiliki satu atau berapa karakteristik diantaranya mengalami kesulitan dalam membaca, menulis, ekspresi tertulis, ejaan, perhitungan matematika, atau penalaran matematika, atau memiliki masalah memori dan persepsi, gangguan bicara dan bahasa, dan menjadi mudah terganggu dan sulit untuk diam [1]. Jumlah mahasiswa di Amerika Serikat yang mengidentifikasi diri sebagai memiliki ketidakmampuan belajar telah meningkat dari 0.6 % menjadi 12,2% [2]. Di Indonesia diperkirakan terdapat 6 % siswa berkebutuhan khusus bersekolah disekolah formal yang sebagian besar dari mereka termasuk ke dalam siswa berkesulitan belajar [3]. Ketika siswa memiliki kesulitan belajar mengakibatkan gagal untuk memperoleh keterampilan keaksaraan yang memadai, hal itu dapat menyebabkan mereka untuk mengalami hasil akademik yang tidak memuaskan [4]. siswa berkesulitan belajar memerlukan pendidikan khusus yang tidak mereka peroleh di sekolah biasa. Jika tidak demikian, tidak sedikit yang terpaksa tinggal kelas pada setiap tingkat [3]. Sehingga diharapkan siswa berkesulitan belajar dapat diprediksi untuk mendapat penangan yang tepat.

Naive Bayes Classifier (NBC) merupakan salah satu metode *machine learning* yang digunakan untuk membuat model – model dengan kemampuan memprediksi [5]. Algoritme *Naive Bayes* termasuk dalam *supervised learning* dan salah satu algoritme pembelajaran tercepat yang dapat menangani sejumlah fitur atau kelas [6]. *Correlated-Naïve Bayes Classifier* (C-NBC) merupakan pengembangan dari metode *Naïve Bayes Classifier* yang menambahkan korelasi antar atribut - atribut dengan kelas [7].

Penelitian ini menggunakan *dataset* siswa berkesulitan belajar yang memiliki jumlah *instance* 208 yang terdiri dari dua kelas. *dataset* ini termasuk ke dalam *dataset* tidak seimbang karena memiliki selisih perbandingan antara kelas *yes* dengan kelas *no* 1:5 dimana kelas *Yes* terdiri dari 31 *instance* sedangkan kelas *no* terdiri dari 157 *instance* [8].

Permasalahan ketidakseimbangan kelas adalah kondisi dimana jumlah kelas dengan perbedaan tinggi. Dengan kata lain pengamatan pada suatu kelas kurang dari kelas yang lainnya. Masalah ini mempengaruhi kinerja prediksi model karena model cenderung memprediksi kelas dengan jumlah yang lebih besar dari *data sample* [9].

Secara umum terdapat dua pendekatan untuk menangani ketidakseimbangan kelas pada suatu *dataset* yaitu pendekatan level algoritme dan level data [10]. Pendekatan pada level data mencakup berbagai teknik resampling dan sintesis data untuk memperbaiki kecendrungan distribusi kelas data latih. Pada pendekatan level algoritme, metode utamanya adalah menyesuaikan operasi algoritme yang ada untuk membuat pengklasifikasi (*classifier*) agar lebih konduktif terhadap klasifikasi kelas minoritas[11]. Kelemahan level algoritme jika diaplikasikan dalam algoritme klasifikasi kuat adalah waktu yang lebih lama karena adanya penyesuaian bobot dan iterasi sampai mendapat nilai yang sesuai[12].

Salah satu metode untuk mengatasi masalah ketidakseimbangan kelas adalah SMOTE *(Synthetic Minority*

Over-sampling Technique). SMOTE bekerja dengan membuat data sintesis baru untuk menambah jumlah data pada kelas minoritas. Metode SMOTE dikenal mampu untuk mengindari *overfitting* ketika mensintesis data kelas minoritas [18].

Penelitian ini bertujuan membandingkan kinerja antara metode C-NBC dan metode C-NBC+SMOTE untuk menangani ketidakseimbangan kelas pada klasifikasi untuk memprediksi data siswa berkesulitan belajar yang memiliki data kelas yang tidak seimbang. Selain itu akan dibandingkan kinerja klasifikasi metode C-NBC terhadap dataset siswa berkesulitan belajar dengan kinerja klasifikasi NBC pada penelitian sebelumnya yang telah dilakukan oleh Muangnak [8].

II. PENELITIAN TERKAIT

Terdapat beberapa penelitian yang berkaitan dengan penelitian ini, baik itu dari penerapan metode klasifikasi yang digunakan, teknik penyeimbangan kelas dan dataset yang menjadi objek penelitian. Muangnak et al [8] melakukan perbandingan akurasi metode *NBC* dengan *Decision Tree* pada klasifikasi *dataset* siswa berkesulitan belajar. Hasilnya metode *Decision Tree* memberikan hasil yang lebih baik daripada *NBC* dengan perbandingan akurasi masing – masing 96.15 % and 94.23 %. Pada penelitian ini tidak menggunakan *cross fold validation* dan hanya mengunakan akurasi sebagai variabel perbandingan.

Muktamar et al [7] melakukan penelitian dengan membandingkan akurasi metode *Correlated-Naive Bayes Classifier (C-NBC)* dengan *NBC*. Penelitian ini dilakakukan menggunakan beberapa *dataset* diantaranya *dataset servo, dataset* balance-scale, *dataset haberman* dan *dataset* iris. dan penelitian ini menggunakan 30 *fold validation*. Hasilya, *C-NBC* menghasilkan akurasi yang lebih baik pada semua dataset yang diuji. Pada penelitian ini jumlah atribut *dataset* relative sedikit, terdiri antara empat sampai lima atribut.

Riquelme et al [13] membandingkan metode C4.5 dengan metode *Naïve Bayes Classifer* yang diintegerasikan dengan SMOTE untuk menangani ketidakseimbangan kelas. Hasil dari penelitian tersebut menyatakan bahwa *SMOTE* dapat meningkatkan nilai AUC. Selain itu NBC menghasilkan nilai *AUC* lebih baik dibandingkan dengan C4.5. Luis Roberto Mercado-Diaz et al melakukan penelitian strategi penyeimbangan kelas dengan membandingkan *SMOTE* dengan *Subsampling* pada metode *SVM* (*Support Vector Machine*) pada fungsi prediksi protein. Dari penelitian tersebut menyatakan *SMOTE* sangat efisien dan menghasilkan *spesificity* terbaik.

Chawla et al [14] Penggunaan teknik *SMOTE (Synthetic Minority OverSampling Technique)* menghasilkan hasil yang baik dan cara yang efektif untuk menangani ketidakseimbangan kelas yang mengalami *overfitting* pada teknik *oversampling* untuk memproses kelas minoritas.

III. METODE YANG DIUSULKAN

Penelitian ini dilakukan terhadap dataset siswa berkesulitan belajar yang terdiri 208 *instance* dan 12 atribut termasuk 1 atribut kelas seperti yang ditunjukkan oleh Tabel I.

TABEL I. *DATASET ATTRIBUTES*

Attributes	***Description***	***Possible values***
stdGender	*Gender*	*{MALE, FEMALE}*
stdAge	*Age*	{6, 7, 8, 9, 10}
stdClass	*Class of Student*	{1, 2, 3}
lookSmart	*Look smart*	*{TRUE, FALSE}*
noAnyDis	*No detact any Disabilites*	*{TRUE, FALSE}*
numReadYes	*Number of true in reading*	{0 – 9}
numReadNo	*Numbre of false in reading*	{0 – 9}
numWriteYes	*Number of true in writing*	{0 – 6}
numWriteNo	*Number of true in writing*	{0 – 6}
numCalYes	*Number of true in calculation*	{0 – 6}
numCalNo	*Number of false in calculation*	{0 – 6}
class	*Result of classification*	*{TRUE, FALSE}*

Metode yang diusulkan untuk klasifikasi *dataset* siswa berkesulitan belajar adalah *Correlated Naïve Bayes Classifer*. Metode *SMOTE* digunakan untuk menangani permasalahan ketidakseimbangan kelas pada *dataset*, sedangkan untuk mengukur kinerja klasifikasi menggunakan menggunakan *Confussion Matrix*.

Correlated Naïve Bayes Classifier(C-NBC) merupkan metode modifikasi dan pengembangan dari metode *Naïve Bayes Classifier* dengan menambahkan faktor korelasi masing – masing atribut terhadap kelas [7] seperti yang terlihat pada Persamaan 1 yaitu:

$$P(Y|X) = \frac{P(Y)\sum_{i=1}^{q} P(X_i|Y)^{l} R(X_i|Y)}{P(X)} \quad (1)$$

dengan:

- $P(Y|X)$ adalah *Posterior Probability,*
- $\sum_{i=1}^{q} P(X_i|Y)^{l} R(X_i|Y)$ adalah jumlah total dari perkalian antara probabilitas independen kelas Y dari fitur dalam *vector* X dengan nilai korelasi fitur dalam *vector* X terhadap kelas Y,
- l adalah *laplacian,*
- $P(Y)$ adalah *prior probability,*
- $P(X)$ adalah *probability* dari *vector* X.

Sedangkan korelasi antara atribut dengan kelas dapat dihitung berdasarkan nilai r *square* seperti yang ditunjukkan pada Persamaan 2 yaitu:

$$r = \frac{nx(\sum XY) - (\sum X)x(\sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)\,(n\sum Y^2 - (\sum Y)^2)}} \quad (2)$$

Synthetic Minority Oversampling Technique (SMOTE) merupakan teknik yang diajukan oleh Chawla et al [4] untuk mengatasi ketidakseimbangan kelas pada suatu *dataset*. SMOTE adalah pendekatan baru dengan cara kerja menggunakan pendekatan *oversampling* pada kelas minoritas dengan membuat sampel sintesis dihasilkan beroperasi dalam "*feature space"* bukan "*data space*". Sampel sintesis dibuat dengan cara menghitung nilai perbedaan (pengurangan) antara vektor atribut

yang dipilih dengan vektor tetangga yang terletak berdekatan. Kemudian nilai pengurangan dikalikan dengan dengan nomor acak antara 0 sampai dengan 1, dan kemudian ditambahkan pada nilai atribut vektor yang sebelumnya telah dipilih. Proses ini menggambarkan seleksi poin secara acak pada garis segmen antara dua atribut spesifik. Algoritme SMOTE ditulis dalam bentuk *pseudocode* sebagai berikut.

Algoritme SMOTE(T,N,k)
Input : T(jumlah sampel minoritas); N persentase SMOTE; k tetangga terdekat
Output : (N/100*T) sampel sintesis kelas minoritas

```
(* Apabila N kurang dari 100%, randomisasi sampel kelas minoritas akan diSMOTE)
if N<100
        Then Randomisasi T sampel minortias
        T = (N/100) * T
        N = 100
Endif

N = (int)(N/100)(*jumlah SMOTE diasumsikan integerdari perhitungan 100)
k = jumlah dari tetangga terdekat
numattr = jumlah attribut
Sample[][] = array dari sampel kelas minoritas awal
newIndex = jumlah dari sampel sintesis yang dibangkitkan, diawali dengan 0
Synthetic[][] = array dari sampel sintesis (* menghitung k tetangga yang berdekatan dari tiap sampel kelas minoritas)
for i ← to T
        Hitung k tetangga yang berdekatan pada i dan simpan indexnya ke nnarray
        Populate(N, i, nnarray)
Endfor

Populate(N, i, narray) (*fungsi untuk membangkitkan sampel sintesis)
While N <> 0
        Pilih nomor random antar i dan k, sebut sebagai nn, langkah ini memilih satu dari k tetangga yang berdekatan dengan i
        for attr ← 1 to numattrs
                Hitung: diff = Sampel(nnarray[nn][attr]-Sampel[i][attr]
                Hitung: gap = nomor acak antara 0 sampai dengan 1
                Sintesis [newindex][attr] = Sampel[i][attr]+gap * diff
        endfor
        newindex++
        N=N-1
endwhile
return (*End of populate*)
```

Kinerja klasifikasi diukur menggunakan *confussion matrix* dan *Area Under the ROC (Receiver Operating Characteristic) Curve (AUC)*. *Confusion matrix* digunakan untuk mengukur kinerja dari metode klasifikasi dalam mengenali tupel dari kelas yang berbeda [15]. *Confusion matrix* memberikan penilaian kinerja klasifikasi objek dengan benar atau salah [16]. Tabel confusion matrix dibuat berdasarkan nilai perbandingan fakta data dan hasil prediksi seperti yang terlihat pada Tabel II. AUC adalah ukuran numerik untuk membedakan kinerja model, dan menunjukkan seberapa sukses dan benar peringkat model dengan memisahkan pengamatan positif dan negatif [17]. AUC dikenal telah terbukti menjadi ukuran kinerja yang andal untuk permasalahan ketidakseimbangan kelas dan sensitivitas [19]. Persamaan untuk menentukan akurasi, sensitvitas, spesifisitas dan AUC masing – masing ditunjukkan oleh Persamaan 3, 4, 5 dan 6.

TABLE I. *CONFUSION MATRIX*

		Prediction	
		Positive	*Negative*
Actual	*Positive*	TP	FN
	Negative	FP	TN

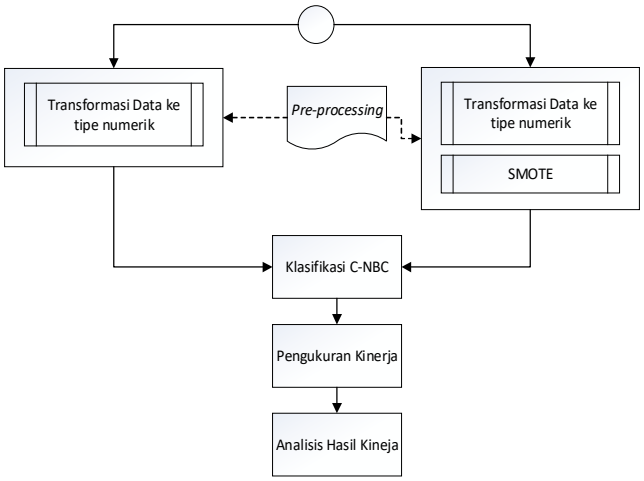
$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$\text{Sensitivitas } /Recall/TPrate = \frac{TP}{TP+FN} \quad (4)$$

$$\text{Spesifisitas}/TNrate = \frac{TN}{TN+FP} \quad (5)$$

$$AUC = \frac{TPrate+TNrate}{2} \quad (6)$$

Alur penelitian diawali dengan tahap *pre-processing* yang terdiri dari dua bagian. Bagian pertama adalah proses tranformasi dengan mengubah tipe data atribut - atribut *non numeric* menjadi *numeric* dengan metode konversi. Kemudian pada bagian kedua dari *pre-processing* adalah menyeimbangkan jumlah *instance* kelas minoritas terhadap *instance* kelas mayoritas dengan menggunakan metode SMOTE. Tahap selanjutnya adalah proses klasifikasi dengan menggunakan metode C-NBC. Klasifikasi dilakukan pada dua jenis *dataset*, yaitu *dataset* yang mengalami proses penyeimbangan kelas dengan teknik SMOTE dan *dataset* yang tidak mengalami proses penyeimbangan kelas. Proses selanjutnya adalah pengukuran kinerja. Hasil pengukuran kinerja diukur menggunakan *confusion matrix* dari hasil klasifikasi menggunakan metode C-NBC dengan teknik *30 fold cross validation*. Alur percobaan yang dilakukan ditunjukkan oleh Gambar 1.



Gambar 1 Alur Percobaan Penelitian

IV. HASIL DAN PEMBAHASAN

Penelitian ini dilakukan terhadap *dataset* siswa berkesulitan belajar. *Dataset* ini diperoleh berdasarkan penelitan yang dilakukan oleh muangnak [8] terhadap siswa sekolah dasar *DongMaFaiJaremSin* dan *BanPhonBokSaharatwittaya* di Thailand pada tahun 2010. *Dataset* yang dihasilkan terdiri dari 12 atribut 2 kelas dan terdiri dari 12 *instance*, yang tediri dari kelas *yes* dan *no* dengan perbandingan 1:5 seperti yang ditunjukan oleh Tabel I.

Penelitian ini dilakukan menggunakan laptop dengan prosesor intel *core i3*, RAM 8 GB, dan sistem operasi *Windows 10*. Sedangkan *tools* yang digunakan adalah *Software Weka data mining tools* dan *Microsoft Excel* 2013.

Pada tahap *preprocessing*, *dataset* siswa berkesulitan belajar yang bernilai nominal terlebih dahulu ditransformasikan menjadi *numeric* menggunakan *software Excel*. Hal ini dilakukan untuk memudahkan dalam menghitung nilai korelasi atribut - atribut pada tahap klasifikasi menggunakan metode C-NBC. Hasil tranformasi tipe data *non numeric* ke tipe *numeric* dengan metode konversi ditunjukkan oleh Tabel II.

TABEL II. *DATASET ATTRIBUTES AFTER TRANFORMATION*

Attributes	***Description***	***Possible values***
stdGender	*Gender*	*{1, 2}*
stdAge	*Age*	*{6, 7, 8, 9, 10}*
stdClass	*Class of Student*	*{1, 2, 3}*
lookSmart	*Look smart*	*{1, 2}*
noAnyDis	*No detact any Disabilites*	*{1, 2}*
numReadYes	*Number of true in reading*	*{1 – 10}*
numReadNo	*Numbre of false in reading*	*{1 – 10}*
numWriteYes	*Number of true in writing*	*{1 – 7}*
numWriteNo	*Number of true in writing*	*{1 – 7}*
numCalYes	*Number of true in calculation*	*{1 – 7}*
numCalNo	*Number of false in calculation*	*{1 – 7}*
class	*Result of classification*	*{1,2}*

Kemudian dilakukan tahap *pre-processing* berikutnya yaitu proses menyeimbangan jumlah kelas dari *dataset* menggunakan Metode SMOTE. Tahap selanjutnya adalah proses klasifikasi menggunakan metode C-NBC. Tahap berikutnya dilakukan proses pengujian menggunakan teknik *k-fold validation* yaitu dengan membagi data menjadi beberapa bagian sesuai jumlah *k-fold* yang ditentukan. Dalam penelitian ini menggunakan *30-fold cross validation*. Tahap selanjutnya adalah pengukuran kinerja klasifikasi dengan menggunakan *Confusion Matrix* berdasarkan nilai akurasi, sensitivitas, spesifisitas dan AUC yang ditunjukkan oleh Tabel III.

TABEL III. HASIL PENGUKURAN KINERJA KLASIFIKASI

Parameter Uji	**C-NBC**	**C-NBC+SMOTE**
Akurasi	94.33 %	96.53 %
Sensitivitas (*Recall*)	90 %	99.44%
spesificity	93 54 %	95.09 %
AUC	96.83` %	97.13 %

Berdasarkan Tabel III diatas, kombinasi metode C-NBC dengan SMOTE menunjukkan nilai akurasi yang lebih baik sebesar 96.53 % dari metode NBC (94.33 %). Persentase sensitivitas C-NBC menghasilkan nilai lebih baik dari C-NBC+SMOTE masing – masing 90 % dan 99.44 %. Hasil pengukuran spesifisitas C-NBC dan SMOTE menghasilkan persentase tertinggi yaitu 95.09 % dari C-NBC dengan persentase 93.54%. Sedangakan diukur dari nilai AUC maka CNBC dengan SMOTE menghasilkan nilai lebih baik yaitu 97.13% dibandingkan C-NBC.

V. KESIMPULAN DAN SARAN

Dari hasil penelitian yang telah dilakukan kombinasi C-NBC dengan SMOTE menghasilkan nilai akurasi, spesifisitas, dan AUC yang lebih baik. Dengan demikian dapat disimpulkan bahwa penerapan metode SMOTE dapat meningkatkan kinerja metode C-NBC dalam mengklasifikasi *dataset* siswa berkesulitan belajar.

Dibandingkan dengan penelitian sebelumnya yang dilakukan oleh muangnak [8] kinerja klasifikasi siswa berkesulitan belajar dengan metode C-NBC sedikit lebih baik dari NBC dengan persentase masing - masing 94.33 % dan 94,23% , namun metode C-NBC memilik waktu komputasi yang lebih lama karena dalam proses klasifikasi dilakukan proses tambahan untuk menghitung korelasi antara atribut dengan kelas.

Pada penelitian selanjutnya sebaiknya menggunakan model yang berbeda yang cocok dengan data kuantitatif dan menganalisa korelasi tiap atribut untuk meningkatkan kinerja klasifikasi. Selain itu juga diharapkan untuk dapat menerapkan metode klasifikas yang berbeda dan mengkombinasikan dengan teknik penyeimbangan kelas lainnya.

REFERENCES

[1] Ayala Gonen, Keren Grinberg. "Academic Students' Attitudes toward Students with Learning Disabilities". Journal of Education and Training Studies, 2016, vol 4 no 9

[2] White, W. J., Schumaker, J. B., Warner, M. M., Alley, G. R., & Deshler, D. D. "An epidemiological study of learning disabled adolescents in secondary schools".1980,. Achievement and ability, socioeconomic status, and school experiences, 1-56.

[3] Sambira Mambela. " Mainstreaming as an Alternative Treatment Education Children with Special Needs in Indonesia ". Sosiohumanika. 2010.

[4] Hitchcock, C. H., Prater, M. A., & Dowrick, P. W."Reading comprehension and fluency: examining the effects of tutoring and video self-modeling on first grade students with reading difficulties". Learning Disabilities Quarterly 27, 2004, pp 27, 89-103. http://dx.doi.org/10.2307/1593644

[5] Eibe Frank, Mark Hall, Bernhard Pfahringer. "Locally Weighted Naive Bayes". 2003. UAI'03 Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence, 2003, pp 249-256

[6] Lee, C.-H. "A Gradient Approach for Value Weighted Classification Learning in Naive Bayes". Knowledge-Based Systems, 2015, pp 1-9.

[7] Muktamar, B. A., Setiawan, N. A., Adji, T. B. "Analisis Perbandingan Tingkat Akurasi Algoritme Naïve Bayes Classifier dengan Correlated Naïve Bayes Classifier". Seminar Nasional Teknologi Informasi Dan Multimedia, 2015, pp 6–8.

[8] Muangnak Nittaya, Pukdee Wannapa, Hengsanunkun Thapani. "Classification Students with Learning Disabilities Using Naive Bayes Classifier and Decision Tree". The 6th International Conference on Networked Computing and Advanced Information Management, 2010, pp 189-192.

[9] Sanguanmak Yotsathon, Hanskunatai Anantaporn. "Auto-Tuning of Parameters in Hybrid Sampling Method for Class Imbalance Problem".. International Computer Science and Engineering Conference (ICSEC), 2016, pp 1-5

[10] Zhang, Z.-Z., Chen, Q., Ke, S.-F., Wu, Y.-J., Qi, F., & Zhang, Y.-P. "Ranking Potential Customers Based on GroupEnsemble", International Journal of Data Warehousing and Mining, 2008, pp 79–89.

[11] Zhang, D., Liu, W., Gong, X., & Jin, H.. "A Novel Improved SMOTE Resampling Algorithm Based on Fractal". Computational Information Systems, 2011, pp 2204-2211.

[12] Korada, N, K., Kumar, N, P., & Deekshitulu, Y. "Implementatioan Of Naive Bayesian Classifier and Ada-Boost Algoritm Using Maiz Expert System". International Journal Of Information Sciences And Techniques (IJIST), 2012, Vol.2, No.3, 63- 75, 2(3), 63–75.

[13] Riquelme, J. C., Ruiz, R., Rodriguez, D., & Moreno, J. "Finding Defective Modules From Highly Unbalanced Datasets". Actas de los Talleres de las Jornadas de Ingeniería del Software y Bases de Datos, Gijon, España: Sistedas, 2008, pp. 67-74.

[14] Eibe Frank, Mark Hall, Bernhard Pfahringer. "Locally Weighted Naive Bayes". UAI'03 Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence, 2003, pp 249-256

[15] Han, J., Kamber, M., & Pei, J. "Data Mining: Concepts and Techniques (3rd ed.)". San Francisco: Morgan Kaufmann Publishers Inc. 2011

[16] Gorunescu, F. Data Mining: Concepts, Models and Techniques. Berlin: Springer-Verlag. 2011

[17] Attenberg, J., & Ertekin, S. "Class Imbalance and Active Learning". In H. He, & Y. Ma, Imbalanced Learning: Foundations, Algorithms, and Application. New Jersey: John Wiley & Sons, 2013, pp. 101-149

[18] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas. "Handling imbalanced datasets: A review". GESTS International Transactions on Computer Science and Engineering, 2006, Vol.30

[19] T. Fawcett "ROC graphs: Notes and practical considerations for researchers". HP Labs, Palo Alto, CA, Tech. Rep. HPL-2003-4. 2003