Institut Riset dan Publikasi Indonesia (IRPI)
SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat
Homepage: https://journal.irpi.or.id/index.php/sentimas
Prosiding Seminar Nasional/ ISSN(E): 2830-3083
25 Agustus 2022/ Pages: 215-221

Classification Determining Employee Data Work Schedules Using C4.5 and K-Nearest Neighbor Algorithm

Klasifikasi Menentukan Jadwal Kerja Data Karyawan Menggunakan Algoritma C4.5 dan K-nearest Neighbor

Syarifah Jasmine Putri[1], Qevin Attaqwa[2], Adetia Pratama[3], Rahmaddeni[4*]

[1,2,3,4]Program Studi Teknik Informatika, STMIK AMIK Riau, Indonesia

E-Mail: [1]2110031802155@sar.ac.id, [2]2110031802124@sar.ac.id,
[3]2110031802143@sar.ac.id, [4]rahmaddeni@sar.ac.id

Corresponding Author: Rahmaddeni

Abstract

Data mining is the process of finding useful information from a large number of databases. The technique that is often used in data mining methods is classification, with the classification of employee data to prepare a work schedule, it is expected to produce certain patterns. Making a work schedule is indeed not easy. Companies want their employees to be able to work as much as possible according to their expertise. It takes accuracy in managing time management so that the schedule of all employees is regular to maximize company performance. By using the C4.5 algorithm, the decision tree technique has the advantage of being able to process numerical data (continuous) and variables, can handle missing attribute values, and produce rules that are easy to interpret. Meanwhile, using K-nearest Neighbor to classify data and employee work schedules. To overcome these obstacles, the application of the C4.5 (Decision Tree) and K-nearest Neighbor algorithms was made to determine the schedule with 501 data used. The software used to manage the data is Rapid miner and the accuracy results from the C.45 algorithm are 54% while KNN 16%.

Keyword: Accuracy, C4.5, Data Mining, KNN, Rapid Miner

Abstrak

Data mining merupakan proses dalam menemukan informasi yang bermanfaat dari banyaknya database dengan jumlah besar. Teknik yang sering digunakan pada metode data mining adalah klasifikasi, dengan adanya klasifikasi data karyawan untuk menyusun sebuah jadwal kerja diharapkan menghasilkan pola-pola tertentu. Membuat jadwal kerja memanglah tidak mudah Perusahaan menginginkan karyawannya dapat bekerja semaksimal mungkin sesuai dengan keahliannya. Dibutuhkan ketelitian dalam mengatur waktu agar jadwal seluruh karyawan teratur untuk memaksimalkan performa perusahaan. Algoritma C4.5 berupa teknik pohon keputusan yang memiliki kelebihan dapat mengolah data numerik (kontinyu) dan variabel, dapat menangani nilai atribut yang hilang, dan menghasilkan aturan-aturan yang mudah diintrepetasikan. Sedangkan dengan menggunakan *K-nearest Neighbor* untuk melakukan klasifikasi pada data dan jadwal kerja karyawan. Algoritma C4.5 *(Decision Tree)* dan *K-nearest Neighbor* digunakan untuk menentukan jadwal menggunakan data sebanyak 501 data. *Software* yang digunakan untuk mengelolah data tersebut adalah Rapid Miner dengan hasil akurasi dari algoritma C.45 yaitu 54% sedangkan KNN 16%.

Kata Kunci: Akurasi, C4.5, Data mining, KNN, Rapid Miner

1. PENDAHULUAN

Berbagai perusahaan menginginkan karyawannya dapat bekerja semaksimal mungkin sesuai dengan keahliannya. Banyak faktor yang dapat mempengaruhi kinerja karyawan diataranya umur, gender, pendidikan terakhir, agama, keterangan sehat, dan lain-lain. Berlandaskan faktor tersebut perusahaan harus menyusun sebuah jadwal kerja yang sesuai dengan kriteria. Penyusunan jadwal kerja secara manual menghadapi beberapa kendala seperti hilangnya beberapa informasi, kewalahan dalam menghitung banyaknya data dan sulit mempertimbahkan kriteria karyawan.

Membuat jadwal kerja membutuhkan ketelitian dalam *time management* agar jadwal seluruh karyawan teratur untuk memaksimalkan performa perusahaan dengan sumber daya manusia yang terbatas. Pembagian

waktu kerja menghadapi kendala apabila terjadinya ketidakteraturan jam kerja terutama jika perusahaan memiliki sistem shift, jika jadwal tidak diatur dengan benar berakibat pada waktu yang berbenturan maupun menimbulkan efek lelah dan tidak konsentrasi dalam bekerja.

Algoritma C4.5 merupakan teknik pohon keputusan yang terkenal karena memiliki kelebihan diantaranya dapat mengolah data numerik dan diskret, dapat menangani nilai atribut yang hilang, menghasilkan aturan-aturan yang mudah diintrepetasikan dan tercepat diantara algoritma-algoritma yang lain[1]. Kelebihan lain yang dimiliki yaitu keakuratan prediksi berupa kemampuan model untuk dapat memprediksi dengan baik label kelas terhadap data baru atau yang belum diketahui sebelumnya. Selain itu algoritma C4.5 dapat memprediksi dengan benar walaupun pada data terdapat nilai dari atribut yang hilang serta kemampuan skalabilitas yaitu kemampuan untuk membangun model secara efisien dalam data yang cukup besar dengan menggunakan interpretabilitas yang bermakna model yang dihasilkan mudah dipahami[2].

Metode *K-nearest Neighbor* (KNN) merupakan algoritma non parametrik yaitu algoritma yang tidak membuat asumsi apapun terhadap data. Algoritma KNN dapat digunakan untuk kasus klasifikasi (*classification*) maupun regresi (*regression*). Hasil dari *query instance* yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada KNN[3]. Algoritma KNN memiliki kelebihan yaitu dapat menghasilkan data yang kuat atau jelas serta efektif jika digunakan pada data yang cukup besar.

Penelitian terdahulu yang dilakukan oleh Novianti dan kawan-kawan berjudul *Penentuan Jadwal Kerja Berdasarkan Klasifikasi Data Karyawan Menggunakan Metode Decision Tree C4.5 (Studi Kasus Universitas Muhammadiyah Surabaya)* menghasilkan akurasi yang cukup baik yaitu 70%. Teknik memperoleh data didapat dari observasi data karyawan UMSurabaya. Beberapa tahapan pada penelitian ini antara lain eksplorasi data karyawan UM Surabaya tahun 2015 yakni aktivasi pembersihan data hingga transformasi data[4].

Penelitian lainnya dengan topik *Analisa Judul Skripsi untuk Menentukan Peminatan Mahasiswa Menggunakan Vector Space Model dan Metode K-nearest Neighbor* menyatakan bahwa penggunaan metode KNN sangat efektif dalam melakukan klasifikasi dan prediksi hanya pada data yang terdiri dari dua kelas namun tidak efektif apabila diterapkan pada data yang memiliki banyak kelas. Adapun atribut yang sangat berpengaruh untuk penentuan klasifikasi judul pada penelitian tersebut yaitu atribut ‘topik’ dengan jumlah akurasi sebanyak 96,85 %[5].

Berdasarkan latar belakang yang dijabarkan, penelitian ini melakukan evaluasi pada masing masing algoritma C4.5 dan KNN untuk mendapatkan hasil perbandingan akurasi pada setiap metode. Data yang digunakan pada penelitian ini adalah data *private* sebanyak 501 data. Atribut yang digunakan adalah binominal, label, polynominal untuk jadwal kerja karyawan menggunakan algoritma C4.5 dan K-nearest Neighbor.

2. METODE PENELITIAN

Metode adalah suatu kegiatan yang berkaitan dengan cara kerja (sistematis) yang dapat dipahami pada suatu objek ataupun subjek untuk menemukan suatu jawaban yang dapat dipertanggung jawabkan secara ilmiah. Penelitian adalah suatu kegiatan penulis dalam menganalisa serta kontruksi yang dilakukan secara sistematis, metodologis, dan konsisten yang bertujuan untuk mengetahui hasil yang sedang dikerjakan.

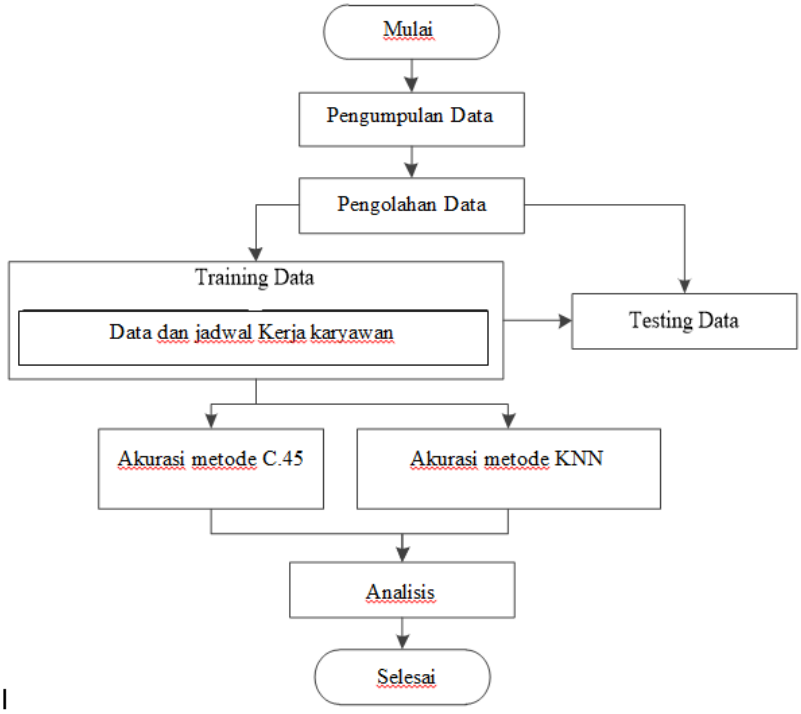
2.1 Metode pada Algoritma C45 dan *Decision Tree* (Pohon Keputusan)

1) Pembersihan data (*data cleaning*)
   Pembersihan data merupakan proses menghilangkan noise dan data yang tidak konsisten atau data tidak relevan. Dari data awal yang digunakan yaitu Nama, Nip, Umur, Jenis kelamin, Agama, Pendidikan terakhir, Posisi, Kesehatan, dan Jadwal Kerja (Pagi atau Sore). Dalam penulisan penelitian ini hanya menggunakan data yaitu Umur, Jenis Kelamin, Pendidikan terakhir, Posisi, dan Jadwal Kerja (Pagi atau Sore)

2) Pemilihan Data
   Data yang ada pada database sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari *database* karena tidak semua data karyawam digunakan maka perlu dilakukan pembersihan data agar data yang akan diolah benar-benar relevan dengan yang dibutuhkan.

3) Transformasi data (*data transformation*).
   Data digabung ke dalam format yang sesuai untuk diproses dalam data mining. Transformasi data merupakan proses pengubahan atau penggabungan data ke dalam format yang sesuai untuk diproses dalam data mining.

4) Proses Mining
   Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data. Pohon keputusan (Decision Tree) merupakan teknik yang akan digunakan pada proses mining. Secara keseluruhan, proses penelitian ditunjukkan pada gambar 1.



Gambar 1. Proses Penelitian

Konsep algoritma yang digunakan untuk membentuk pohon keputusan (Decision Tree) dengan algoritma C4.5 terdiri dari 4 langkah.

1) Pilih atribut sebagai akar
2) Buat cabang untuk tiap tiap nilai
3) Bagi kasus didalam cabang
4) Ulangi proses untuk setiap cabang samapai semua khasus pada cabang memiliki kelas yang sama

2.2 Metode pada K-nearest Neighbor

1) Menentukan jumlah *(k)* untuk mempertimbangkan penentuan data
2) Hitung data baru ke masing masing data point
3) Ambil jumlah *(k)* dengan jarak terdekat
4) Tentukan jumlah dari data baru tersebut

2.3. Persamaan

Untuk memilih atribut akar di dasarkan pada nilai *gain* tertinggi dari atribut yang ada. untuk mendapatkan nilai *gain* tertinggi maka nilai *entropy* harus ditentukan terlebih dahulu. *Entropy* adalah suatu parameter untuk mengukur tingkat keberagaman (heterogenitas) dari kumpulan data. Jika nilai dari *entropy* semakin besar maka tingkat keberagaman suatu kumpulan data semakin besar. Rumus dalam menghitung *entropy* ditunjukkan pada rumus (1).

$$Entropy\ (S)\ =\ -\sum plog2p \quad (1)$$

Keterangan:

S = himpunan khasus
n = jumlah partisi s
pi = proposisi dari si terhadap s

Gain adalah ukuran efektifitas suatu variabel dalam mengklasifikasikan data. *Gain* dari suatu variabel merupakan selisih antara nilai *entropy* total dengan *entropy* dari variabel tersebut. Formula dalam menghitung *Gain* ditunjukkan pada rumus (2).

$$Gain\ (S,)\ =\ Entropy(S)\ -\ \sum Entropy(S\ ) \quad (2)$$

Keterangan:

S = Kasus

A = Atribut
n = Jumlah partisi atribut A
Ai = Jumlah kasus pada partisi ke-i.
S = Jumlah kasus

Pada metode K-nearest Neighbor (KNN), untuk mencari dekat atau jauhnya jarak antar titik pada kelas *k=3* pada umumnya dihitung menggunakan jarak *Euclidean*. Jarak *Euclidean* adalah formula untuk mencari jarak antara 2 titik dalam ruang dua dimensi. Formula untuk menghitung jarak *Euclidean* ditunjukkan pada rumus (3).

$$d_{Euclidian\ (x,)} = \sqrt{\sum(x_i - {}_i)^2} \quad (3)$$

2.4. Rapid Miner

Rapid Miner merupakan perangakat lunak dengan kode sumber yang bersifat terbuka (*open source*) yang dapat menjadi sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. *Tool* ini menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik.

Rapid Miner memiliki kurang lebih 500 operator data mining, termasuk operator untuk *input, output, data preprocessing* serta visualisasi. Rapid Miner merupakan *software* yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. Rapid Miner ditulis dengan munggunakan bahasa java sehingga dapat bekerja di semua sistem operasi.

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini metode yang digunakan untuk menemukan pengetahuan pada studi kasus ini menggunakan data yang berjumlah 501 data karyawan tetapi hanya digunakan dengan jumlah tabel sebanyak 5 yaitu umur, jk(jenis kelamin), pt (pendidikan terakhir), posisi, dan jadwal (pagi atau malam). Dilakukan pengujian validitas model yaitu metode Decision Tree beserta total akurasi. Penggunaan Decision Tree pada penelitian dimaksudkan agar dapat memudahkan pengambilam keputusan yang kompleks sehingga menjadi solusi dalam sebuah permasalahan jadwal kerja karyawan.

3.1 Penjelasan Data

Pada tabel yang digunakan adalah umur yang terdiri dari 20> dan 30, JK (Jenis kelamin) yang terdiri dari L (lelaki) dan P (Perempuan), Pendidikan yang terdiri dari S1, SLTA, dan D3, Posisi yang terdiri dari Senior dan Junior, dan label penentu adalah jadwal yang terdiri dari pagi dan malam. Penjelasan lengkap mengenai data serta jumlah total data yang digunakan dapat ditunjukkan pada tabel 1 dan 2.

Tabel 1. Penjelasan Data

Data	Keterangan	Type	Kategori
Umur	20>	Binominal	
	30		
Jk	L	Binominal	
	P		
	S1		
Pendidikan	D3	Polynominal	
	SLTA		
Posisi	Junior	Binominal	
	Senior		
	Pagi		
Jadwal	Sore	Polynominal	Label
	Malam		

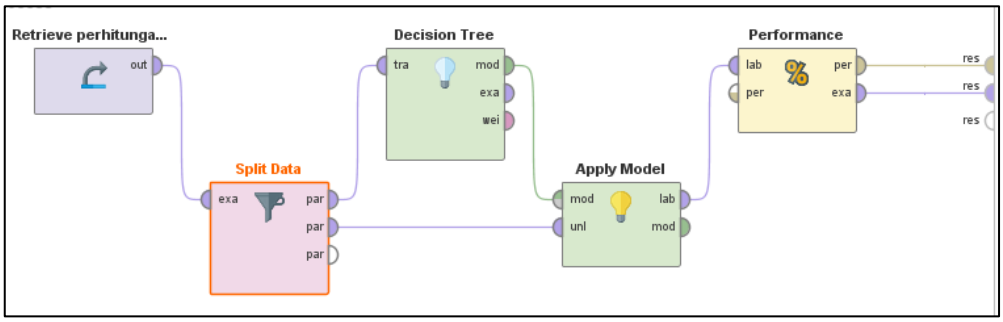
Table 2. Jumlah Total Tiap Data

		Jumlah	Pagi	Malam	Sore	Entropy	Gain
Total		501	213	229	59	1,40430201	
Umur							0,004906785
	20>	251	99	125	27	1,376279396	
	30	250	114	104	32	1,422603515	
Jk							0,065819343
	L	311	97	175	39	1,366673739	
	P	190	116	54	20	1,292338332	
Pendidikan							0,06556642

		Jumlah	Pagi	Malam	Sore	Entropy	Gain
Posisi	S1	312	151	142	19	1,269460458	0,004906785
	D3	130	50	63	17	1,420457713	
	SLTA	59	12	24	23	1,525006185	
	JUNIOR	251	99	125	27	1,376279396	
	SENIOR	250	114	104	32	1,422603515	

3.2 Hasil Akurasi Algoritma C.45 dan *Decision Tree* (pohon keputusan)

Pada perhitungan akurasi menggunakan metode perhitungan algoritma C4.5, diperoleh akurasi sebesar 54% sedangkan hasil akurasi dari *decision tree* (pohon keputusan) adalah jenis kelamin laki-laki akan mendapatkan jadwal malam dan jenis kelamin perempuan dengan pendidikan S1 dan D3 akan mendapatkan jadwal pagi sementara itu pendidikan SLTA akan mendapatkan jadwal sore. Visualisasi hasil perhitungan dengan Rapid Miner beserta pohon keputusan yang dihasilkan dapat ditunjukkan pada gambar 3, 4, dan 5.



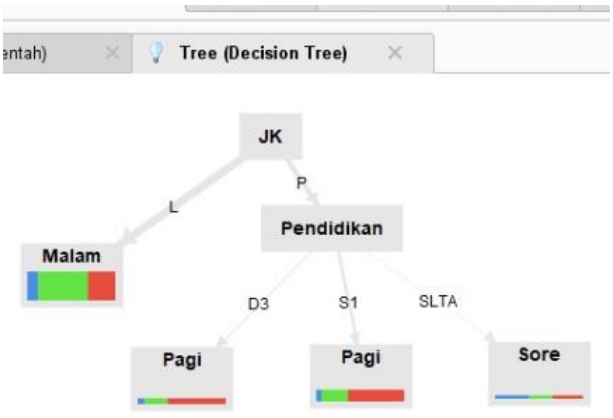
Gambar 3. Perhitungan C4.5 dengan Rapid Miner

Table View Plot View

accuracy: 54.00%

	true Sore	true Malam	true Pagi	class precision
pred. Sore	2	7	4	15.38%
pred. Malam	11	46	25	56.10%
pred. Pagi	2	20	33	60.00%
class recall	13.33%	63.01%	53.23%	

Gambar 4. Hasil *Accuracy* C4.5



Gambar 5. Hasil *Decision Tree*

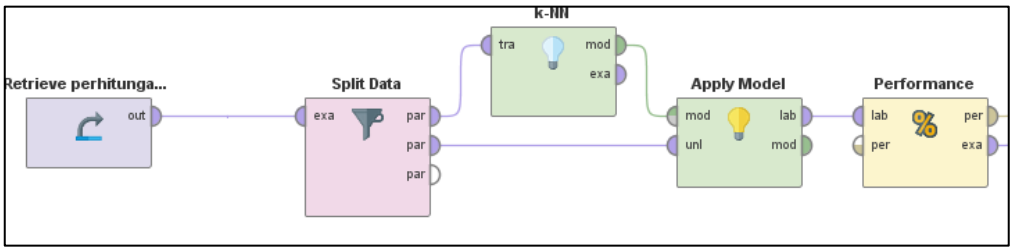
3.3 Hasil Akurasi K-nearest Neighbor (KNN)

Pada perhitungan *accuracy* menggunakan metode perhitungan K-nearset Neighbor diperoleh hasil akurasi sebesar 16%. Adapun atribut yang digunakan yaitu Umur, Jenis Kelamin, Pendidikan, Posisi, dan Jadwal. Data yang dilakukan pemrosesan KNN dapat dilihat pada tabel 3.

Tabel 3. Penjelasan Data KNN

Data	Keterangan	Type	Kategori
Umur	20> 30	Binominal	
Jk	1 2	Binominal	
Pendidikan	3 4 5	Polynominal	
Posisi	6 7	Binominal	
Jadwal	Pagi Sore Malam	Polynominal	Label

Perhitungan yang dilakukan pada algoritma KNN dimulai dengan pembagian data dan implementasi algoritma KNN, selanjutnya diperoleh hasil performa dari model yang diterapkan yang kemudian dibandingkan dengan algoritma C4.5 yang menghasilkan perbandingan kinerja antara kedua model. Jadwal pada jenis kelamin lelaki dengan jadwal malam berjumlah 190 sedangkan pada jenis kelamin perempuan dengan jadwal pagi berjumlah 140 menghasilkan total akurasi KNN adalah sekitar 16% dengan *error* 84% sedangkan algoritma C4.5 adalah 54% dengan *error* 46%. Visualisasi hasil KNN dapat dilihat pada gambar 5 dan 6.



Gambar 5. Perhitungan KNN dengan Rapid Miner

accuracy: 16.00%

	true Sore	true Malam	true Pagi	class precision
pred. Sore	13	62	54	10.08%
pred. Malam	2	11	8	52.38%
pred. Pagi	0	0	0	0.00%
class recall	86.67%	15.07%	0.00%	

Gambar 6. Hasil *Accuracy* KNN

4. KESIMPULAN

Kesimpulan yang diperoleh dari penelitian menggunakan algoritma C.45 dan Pohon Keputusan beserta KNN yaitu metode tersebut dapat digunakan untuk menampilkan sebuah informasi yang berguna tentang Penentuan Jadwal Kerja Karyawan dengan teknik data mining dengan memberikan informasi berupa hubungan antara Jadwal Kerja Karyawan dengan Umur, Jenis Kelamin, Pendidikan Terakhir, Posisi, dan Jadwal. Hal ini terlihat pada pohon keputusan dan KNN yang dilakukan perhitungan dengan aplikasi Rapid Miner menggunakan 501 data yang sebelumnya dilakukan pembersihan.

REFERENSI

[1] J. Han, M. Kamber, and M. Kaufmann, “Data Mining: Concepts and Techniques (2nd edition) Classification and Prediction,” 2006.

[2] H. Dhika and F. Destiawati, “Penerapan Algoritma C45 Untuk Penilaian Karyawan Pada Restoran

Cepat Saji," no. September, pp. 55–59, 2018.
[3] M. Fansyuri, "Analisa algoritma klasifikasi k-nearest neighbor dalam menentukan nilai akurasi terhadap kepuasan pelanggan (study kasus pt. Trigatra komunikatama)," *Humanika J. Ilmu Sos. Pendidikan, dan Hum.*, vol. 3, no. 1, pp. 29–33, 2020.
[4] T. Novianti and I. Santosa, "PENENTUAN JADWAL KERJA BERDASARKAN KLASIFIKASI DATA KARYAWAN MENGGUNAKAN METODE DECISION TREE C4.5 (Studi Kasus Universitas Muhammadiyah Surabaya)," *J. Komunika J. Komunikasi, Media dan Inform.*, vol. 5, no. 1, p. 1, 2016, doi: 10.31504/komunika.v5i1.633.
[5] D. M. U. Atmaja and R. Mandala, "Analisa Judul Skripsi untuk Menentukan Peminatan Mahasiswa Menggunakan Vector Space Model dan Metode K-Nearest Neighbor," *IT Soc.*, vol. 4, no. 2, pp. 1–6, 2020, doi: 10.33021/itfs.v4i2.1182.