JISKa, Vol. 6, No. 3, SEPTEMBER, 2021, Pp. 178 – 188
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Analisis Perbandingan Algoritma Decision Tree, kNN, dan Naive Bayes untuk Prediksi Kesuksesan *Start-up*

Adhitya Prayoga Permana [(1)], Kurniyatul Ainiyah [(2)*], Khadijah Fahmi Hayati Holle [(3)]
Teknik Informatika, Fakultas Sains dan Teknologi, UIN Maulana Malik Ibrahim, Malang
e-mail : {18650086,18650088}@student.uin-malang.ac.id, khadijah.holle@uin-malang.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 17 Juni 2021, direvisi 17 Agustus 2021, diterima 24 Agustus 2021, dan dipublikasikan 22 September 2021.

Abstract

Start-ups have a very important role in economic growth, the existence of a start-up can open up many new jobs. However, not all start-ups that are developing can become successful start-ups. This is because start-ups have a high failure rate, data shows that 75% of start-ups fail in their development. Therefore, it is important to classify the successful and failed start-ups, so that later it can be used to see the factors that most influence start-up success, and can also predict the success of a start-up. Among the many classifications in data mining, the Decision Tree, kNN, and Naïve Bayes algorithms are the algorithms that the authors chose to classify the 923 start-up data records that were previously obtained. The test results using cross-validation and T-test show that the Decision Tree Algorithm is the most appropriate algorithm for classifying in this case study. This is evidenced by the accuracy value obtained from the Decision Tree algorithm, which is greater than other algorithms, which is 79.29%, while the kNN algorithm has an accuracy value of 66.69%, and Naive Bayes is 64.21%.

Keywords: Classification, Decision Tree, kNN, Naive Bayes, Start-up

Abstrak

Start-up memiliki peran yang sangat penting dalam pertumbuhan ekonomi, adanya *start-up* dapat membuka banyak lapangan kerja baru. Namun, tidak semua *start-up* yang sedang berkembang dapat menjadi sebuah *start-up* yang sukses. Hal ini dikarenakan *start-up* memiliki tingkat kegagalan yang tinggi, data menunjukkan sebanyak 75% *start-up* mengalami kegagalan dalam perkembangannya. Oleh karena itu, penting untuk melakukan pengklasifikasian *start-up* sukses dan gagal, sehingga nantinya dapat digunakan untuk melihat faktor-faktor yang paling mempengaruhi keberhasilan *start-up*, dan juga dapat memprediksi keberhasilan suatu *start-up*. Di antara banyaknya metode klasifikasi dalam *data mining*, algoritma Decision Tree, kNN, dan Naïve Bayes merupakan algoritma yang penulis pilih untuk melakukan klasifikasi terhadap 923 *record* data *start-up* yang telah didapatkan sebelumnya. Hasil pengujian menggunakan *cross validation* dan *T-test* menunjukkan algoritma Decision Tree merupakan algoritma paling tepat untuk melakukan klasifikasi dalam studi kasus ini. Hal ini dibuktikan dengan nilai akurasi yang diperoleh oleh algoritma Decision Tree lebih besar diantara algoritma lainnya, yaitu sebesar 79,29%, sedangkan algoritma kNN memiliki nilai akurasi 66,69%, dan Naive Bayes sebesar 64,21%.

Kata Kunci: Decision Tree, Klasifikasi, kNN, Naive Bayes, *Start-up*

1. PENDAHULUAN

Perkembangan teknologi yang terjadi tentunya memberikan dampak dalam berbagai bidang, salah satunya ekonomi. Banyak perusahaan-perusahaan baru dalam bidang teknologi yang bermunculan dan menarik perhatian masyarakat, salah satunya *start-up*. *Start-up* merupakan organisasi yang dibangun untuk mendapatkan keuntungan yang maksimal dengan rancangan model bisnis yang tepat (Blank, 2013 dalam Afdi & Purwanggono, 2017). *Start-up* hadir dengan ide-ide dan inovasi baru yang dapat menciptakan banyak lapangan pekerjaan bagi masyarakat. Oleh karena itu, dapat dikatakan, *start-up* memiliki peran yang sangat penting dalam pertumbuhan ekonomi karena dapat menggerakkan roda perekonomian.

Di balik peran pentingnya, *start-up* memiliki tingkat kegagalan yang cukup tinggi. Menurut penelitian yang dilakukan Sikhar Ghosh, dosen senior di Harvard Business School, terhadap 2000 perusahaan dalam periode 2004 – 2010, penelitian dilakukan terhadap *start-up* - *start-up* yang telah menerima bantuan pendanaan, sebanyak 75% *start-up* mengalami kegagalan dalam perkembangannya (Ghosh, 2013). Kegagalan sebuah *start-up* tentunya akan menimbulkan kerugian bagi semua pihak yang berkontribusi, salah satunya yaitu investor. Oleh karena itu, para investor harus cermat dalam memilih dan melihat faktor-faktor yang mempengaruhi keberhasilan maupun kegagalan suatu *start-up*.

Tujuan utama investor adalah menemukan *start-up* yang memiliki potensi untuk berkembang pesat dan dapat memberikan keuntungan. Dengan ini, maka penulis memutuskan untuk melakukan proses klasifikasi terhadap faktor-faktor yang mempengaruhi keberhasilan maupun kegagalan sebuah *start-up*. Sehingga nantinya para investor dapat memprediksi *start-up* mana yang memiliki potensi besar untuk berhasil, dan juga melihat faktor apa saja yang sangat mempengaruhi keberhasilan *start-up*. Algoritma dalam *data mining* yang umum digunakan untuk melakukan klasifikasi yaitu Naïve Bayes, K-Nearest Neighbor, Decision Tree, Random Forest, dan Support Vector Machines (Wibisono & Fahrurozi, 2019). Pada penelitian ini, penulis menggunakan beberapa metode klasifikasi yang kemudian akan dilakukan perbandingan terhadap nilai performa (akurasi, presisi, dan *recall*) yang dihasilkan masing-masing algoritma.

Sebelumnya, Rahman et al. (2018) telah melakukan penelitian yaitu melakukan komparasi antara algoritma kNN dengan Naive Bayes untuk mengklasifikasi kualitas air bersih. Hasilnya k-NN memiliki nilai rata-rata akurasi lebih tinggi, yaitu sebesar 82,42%, sedangkan Naive Bayes hanya mendapatkan nilai rata-rata akurasi sebesar 70,32%. Penelitian sejenis lainnya dilakukan oleh Marutho (2019) yaitu melakukan perbandingan antara metode Naive Bayes, k-NN, dan Decision Tree dalam studi kasus laporan level ketinggian air di Jakarta. Hasilnya menunjukkan dengan menggunakan evaluasi model *cross validation* dengan *k-fold* 10 menghasilkan nilai akurasi sebesar 96,56% untuk algoritma Decision Tree, kemudian Naïve Bayes sebesar 94,32%, dan terakhir K-NN sebesar 95,98%. Sehingga kesimpulannya adalah Decision Tree merupakan metode terbaik untuk diimplementasikan ke dalam dataset ketinggian air di jakarta. Setiyorini & Asmono (2018) juga melakukan komparasi algoritma antara Decision Tree, Naïve Bayes, dan kNN untuk mengklasifikasi kinerja siswa. Hasil penelitian yang dilakukan pada *dataset student performance* ini menunjukkan algoritma kNN memiliki nilai akurasi tertinggi yaitu sebesar 79,31%, selanjutnya Decision Tree dengan nilai akurasi 78,85%, sedangkan algoritma Naïve Bayes nilai akurasinya sebesar 77,69%.

Penelitian lainnya dilakukan oleh Dellermann et al. (2018) yang memprediksi kesuksesan *start-up* di tahap awal melalui metode kecerdasan hibrida. Dalam penelitian ini penulis menggunakan *Hybrid Intelligence Method* di mana memperoleh kesimpulah bahwa manusia dapat melengkapi kelemahan ketika mesin gagal. Terutama mengenai informasi tersembunyi dan resiko yang tidak diketahui. Penelitian sejenis juga dilakukan oleh Glupker et al. (2019) melakukan prediksi kesuksesan investor menggunakan teori grafik dan pembelajaran mesin. Data yang digunakan bersumber dari basis *Crunch*, termasuk di dalamnya adalah karakteristik perusahaan rintisan, investor, dan individu di seluruh dunia di seluruh abad ke-20 dan hingga tahun 2013. Hasilnya akurasi keseluruhan yang diperoleh berkisar antara 55% dan 75%. Selanjutnya yaitu penelitian untuk identifikasi investor yang sukses di ekosistem *start-up* yang dilakukan oleh Gupta et al. (2015). Dalam penelitian ini, peneliti membuat *platform Investor Rank* dengan menggunakan pendekatan metode heuristik. Di mana hasilnya investor-investor potensial dapat diidentifikasi lebih dulu dan diverifikasi pengaruh jumlah pendanaan dan waktu pendanaan selama *start-up* berdiri.

Setiap metode yang klasifikasi memiliki kelebihan dan kekurangan masing-masing. Kelebihan yang dimiliki Decision Tree adalah sifatnya yang fleksibel sehingga mampu meningkatkan kualitas keputusan yang dihasilkan, sedangkan kekurangan dari algoritma ini adalah akan terjadi *overlap* jika menggunakan data yang memiliki kelas dan kriteria dengan jumlah yang sangat banyak. Sedangkan kelebihan dari metode Naïve Bayes adalah proses perhitungannya sederhana sehingga prosesnya lebih cepat dan efisian, namun dengan fakta bahwa masing-

masing variabel bersifat independen, hal ini dapat mengurangi besar akurasi yang dihasilkan. Terakhir adalah kNN, kelebihan metode ini yaitu dapat diterapkan pada data yang besar secara efektif dengan hasil yang akurat, namun kekurangannya yaitu membutuhkan biaya komputasi yang cukup tinggi karena harus melakukan perhitungan jarak pada setiap *query instance* secara bersama-sama.

Melihat penelitian-penelitian yang telah dilakukan sebelumnya, dan mempertimbangkan kelebihan dan kekurangan masing-masing metode, maka penulis memutuskan untuk melakukan perbandingan antara metode klasifikasi Decision Tree, Naive Bayes, dan kNN (K-Nearest Neighbor). Hal inilah yang juga menjadi keunikan dari penelitian ini, ketiga metode tersebut akan diimplementasikan kedalam dataset *start-up* sebanyak 923 *record* data dengan 19 atribut yang telah dipilih. Perbandingan ini dilakukan untuk menemukan algoritma terbaik yang dapat digunakan untuk melakukan klasifikasi *start-up*. *Dataset* tersebut akan dianalisis dengan menggunakan tahapan-tahapan yang ada di dalam proses CRISP-DM (*Cross Industry Standard Process for Data Mining*).

2. METODE PENELITIAN

2.1. *Dataset*

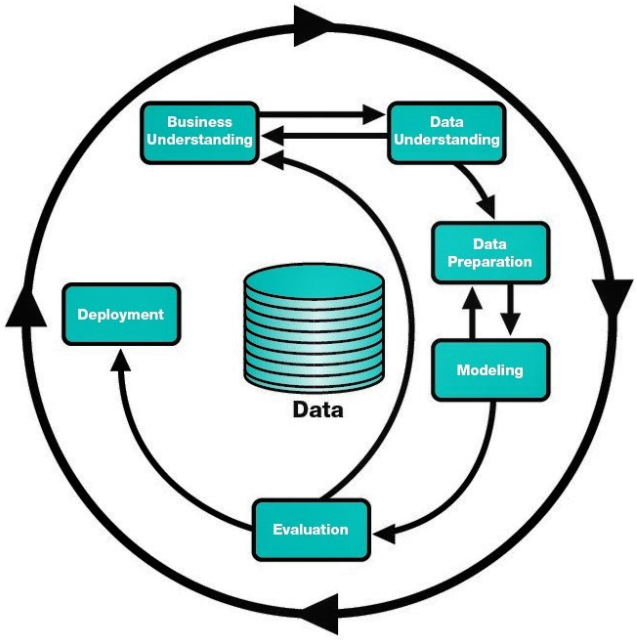
Dataset yang digunakan dalam penelitian ini adalah data sekunder. Data sekunder merupakan data yang didapatkan secara tidak langsung, data-data tersebut dapat bersumber dari buku, jurnal, dokumentasi, literatur, dan sumber informasi lainnya yang berhubungan dengan topik penelitian (Sabna & Muhardi, 2016). Dataset *start-up* yang digunakan dalam penelitian ini diperoleh dari *website* https://www.kaggle.com/manishkc06/startup-success-prediction.

2.2. Variabel (Atribut) Penelitian

Variabel yang digunakan terdiri dari 19 atribut yang dipilih dari 49 atribut yang ada pada *dataset*. 19 atribut ini dipilih berdasar atribut yang paling signifikan karena ada beberapa atribut seperti *category_code* yang dapat mewakili beberapa atribut lain seperti *is_software*, *is_web*, dll. 19 atribut yang dipilih yaitu Id sebagai nilai primer, 17 atribut biasa dan 1 atribut sebagai label. 17 atribut biasa ini terdiri dari *Age_first_funding_year* yaitu jumlah pendanaan pertama *start-up* tersebut, *Age_last_funding_year* adalah jumlah pendanaan terakhir *start-up*, *Age_first_milestone_year* merupakan rekor pendanaan pada waktu tersebut, *Age_last_milestone_year* adalah rekor pendanaan pada waktu terakhir kali, *Relationships* menggambarkan jumlah relasi kerjasama yang dilakukan *start-up* tersebut, *Funding_total_usd* merupakan total pendanaan *start-up*, *Milestones* menandakan kategori *event* penting ke sekian pada saat *start-up* berdiri, *Category_code* merupakan jenis kategori *start-up* tersebut. Sedangkan *Funding rounds, Has VC, Has angel, Has roundA, Has roundB, Has roundC, Has roundD*, merupakan tahapan perputaran pendanaan uang yang harus dilalui sebuah *start-up* untuk mencapai tahapan tertentu. *Avg_participants* merupakan rata-rata peserta ataupun pendukung pelaku *start-up*, Is top 500 menandakan apakah *start-up* tersebut termasuk 500 besar *start-up*, dan terakhir atribut yang berperan sebagai label adalah Status (*closed* dan *acquired*).

2.3. Proses Analisis Data

Metode yang digunakan dalam analisis data mengacu pada tahapan proses CRISP-DM dengan menggunakan *tools* RapidMiner. Pada Gambar 1 berikut ini merupakan tahapan dalam proses CRISP-DM.



Gambar 1. Alur Proses CRISP-DM.

2.3.1. Pemahaman Bisnis (*Business Understanding*)

Tahapan ini merupakan pemahaman terhadap permasalahan yang akan diteliti. Dalam penelitian ini permasalahan yang dibahas yaitu besarnya persentase kegagalan yang dimiliki sebuah *start-up* yang sedang berkembang, membuat investor atau tim pengembang *start-up* harus dapat mengetahui faktor apa saja yang paling berpengaruh dalam penentuan keberhasilan sebuah *start-up*. Selain itu, dengan dataset yang ada, investor juga dapat memprediksi *start-up* yang sedang berkembang dapat menjadi *start-up* sukses atau gagal.

2.3.2. Pemahaman Data (Data Understanding)

Data yang diperoleh dari *website* kaggle.com memiliki *record* sebanyak 923 data *start-up*. Atribut yang digunakan sebanyak 19, yaitu *Status*, *Id*, *Age first funding year*, *Age last funding year*, *Age first milestone year*, *Age last milestone year*, *Relationships*, *Funding rounds*, *Funding total usd*, *Milestones*, *Category code*, *Has VC*, *Has angel*, *Has roundA*, *Has roundB*, *Has roundC*, *Has roundD*, *Avg participants*, *Is top 500*. Data ini nantinya akan digunakan untuk melakukan prediksi keberhasilan *start-up*, sekaligus melihat faktor-faktor yang paling mempengaruhi keberhasilan tersebut.

2.3.3. Persiapan Data (*Data Preparation*)

Dari 923 data *start-up* memiliki 322 *missing value*, sehingga diperlukannya proses *data preparation* agar dapat memperbaiki kualitas data. *Missing value* dipecahkan dengan menambahkan data dengan nilai rata-rata pada atribut-atribut yang datanya hilang ataupun kosong (Praningki & Budi, 2018). Kemudian pada tahap ini juga menggunakan operator *filter example*, filter yang digunakan adalah *label is not missing* yang diterapkan pada atribut label.

2.3.4. Pemodelan (*Modeling*)

Tahapan pemodelan ini dilakukan untuk membangun data *training* dengan menggunakan pelatihan algoritma sehingga menghasilkan sejumlah aturan (Sabna & Muhardi, 2016). Pada penelitian ini, menggunakan algoritma Decision Tree, selain itu, algoritma lain yang digunakan untuk melakukan perbandingan, yaitu algoritma kNN dan Naive Bayes.

1) Decision Tree
Algoritma Decision Tree bersifat sangat kuat, populer, berbasis logika, dan mudah dipahami (Lakshmi et al., 2016). Hal yang menarik dari Decision Tree adalah penggunaan struktur pohon (*tree*) yang berfungsi untuk merepresentasikan aturan yang terbentuk dari hasil klasifikasi. Dalam *tree*, atribut direpresentasikan oleh sebuah *node*, dan kelas direpresentasikan oleh daun (*leaf*). Setiap pohon memiliki akar (*root*) yaitu *node* yang berada di paling atas (Anam & Santoso, 2018).

2) K-Nearest Neighbor
K-Nearest Neighbor merupakan salah satu metode klasifikasi dalam data mining yang termasuk ke dalam *supervised learning*. Pengklasifikasian yang dilakukan berdasarkan atribut dan data *training*, sehingga proses pengklasifikasian data baru dilakukan berdasarkan perbandingan kemiripan mayoritas pada data *training*. Dalam kNN nilai jarak ditentukan dengan pengujian data *testing* terhadap data *training* kemudian menggunakan nilai terkecil dari nilai ketetanggan terdekat (Krisandi et al., 2015). Penghitungan jarak umumnya menggunakan jarak *Euclidean Distance* sebagai berikut.

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n}(a_r(x_i) - a_r((x_j))^2} \qquad (1)$$

Di mana $d(x_i, x_j)$ merupakan jarak Euclidean, x_i dan x_j adalah dua titik yang akan dihitung jaraknya, dengan x_i merupakan *record* data ke i dan x_j adalah *record* data ke j, serta a_r adalah data ke-r dengan i,j bernilai 1,2,3,…n.

3) Naïve Bayes
Naïve Bayes merupakan algoritma yang berdasarkan teorema Bayes, di mana antar atributnya tidak memiliki hubungan atau ketergantungan, sehingga setiap atribut bersifat saling bebas. Klasifikasi Naïve Bayes merupakan metode klasifikasi yang menghitung probabilitas suatu kejadian berdasarkan kondisi tertentu, berikut persamaan dalam Naïve Bayes (Lukito & Chrismanto, 2015).

$$P(A|B) = \frac{P(A)\times P(B|A)}{P(B)} \qquad (2)$$

Dimana B adalah data yang akan dicari *class*-nya, sedangkan A merupakan hipotesis *class*. P(A|B) merupakan peluang terjadinya A terhadap kondisi B, P(A) merupakan peluang terjadinya A, dan P(B) adalah peluang terjadinya B.

2.3.5. Evaluasi (*Evaluation*)

Pada tahapan evaluasi, dilakukan pengujian menggunakan *cross validation* dari algoritma yang digunakan untuk klasifikasi. Hasil pengujian tersebut akan dibandingkan nilai akurasi, presisi, dan recall dari algoritma Decision Tree, kNN, dan Naive Bayes. Setelah itu, pengujian juga dilanjutkan dengan metode T-test untuk melihat algoritma yang terbaik.

1) *Cross Validation*
Cross validation adalah sebuah pengujian standar yang berfungsi untuk memprediksi *error rate*. Jumlah data training dan data testing yang mewakili setiap kelas harus memiliki jumlah yang sama, pembagian data dilakukan secara acak dengan perbandingan yang sama pada setiap kelasnya. Tingkat kesalahan pada setiap tingkat iterasi akan dihitung rata-ratanya sehingga dapat menghasilkan *error rate* keseluruhan (Hastuti, 2012).

2) T-Test
T-test merupakan metode pengujian hipotesis yang memperlakukan satu individu (objek penelitian) dengan dua perlakuan berbeda. Sampel yang berasal dari objek yang sama akan dibagi menjadi dua data, data pertama menggunakan perlakuan pertama, dan data kedua mendapat perlakuan kedua. Nilai *performance* didapatkan dengan membandingkan

kondisi objek pertama dan kondisi objek kedua (Hastuti, 2012). Rumus perhitungan pada T-test adalah sebagai berikut (Kadafi, 2018).

$$T_{hitung} = \frac{X-\mu_0}{\frac{\sigma}{sn}} \qquad (3)$$

Di mana T_{hitung} adalah nilai T yang dicari dan menunjukan standar deviasi pada distribusi normal (tabel t), kemudian X merupakan nilai rata-rata dari data yang diolah, μ_0 adalah rata-rata nilai yang menjadi sampel, dan s adalah standar devisiasi dari populasi yang telah diketahui, dengan n adalah jumlah pupolasi yang digunakan dalam penelitian. Jika nilai uji T adalah lebih kecil dari a maka Ho ditolak (Huda, 2013).

2.3.6. Penyebaran (*Deployment*)

Selanjutnya, hasil pemodelan atau hasil klasifikasi pada tahap sebelumnya dapat digunakan sebagai acuan untuk melakukan prediksi terhadap *start-up* yang sedang berkembang dapat menjadi *start-up* sukses atau gagal.

3. HASIL DAN PEMBAHASAN

3.1. *Data Understanding*

Pembahasan mengenai hasil dari penelitian ini dimulai dari tahapan pertama dalam proses CRISP-DM yaitu *data understanding*. Gambar 2 berikut merupakan tampilan dari *dataset* yang digunakan dalam penelitian ini.

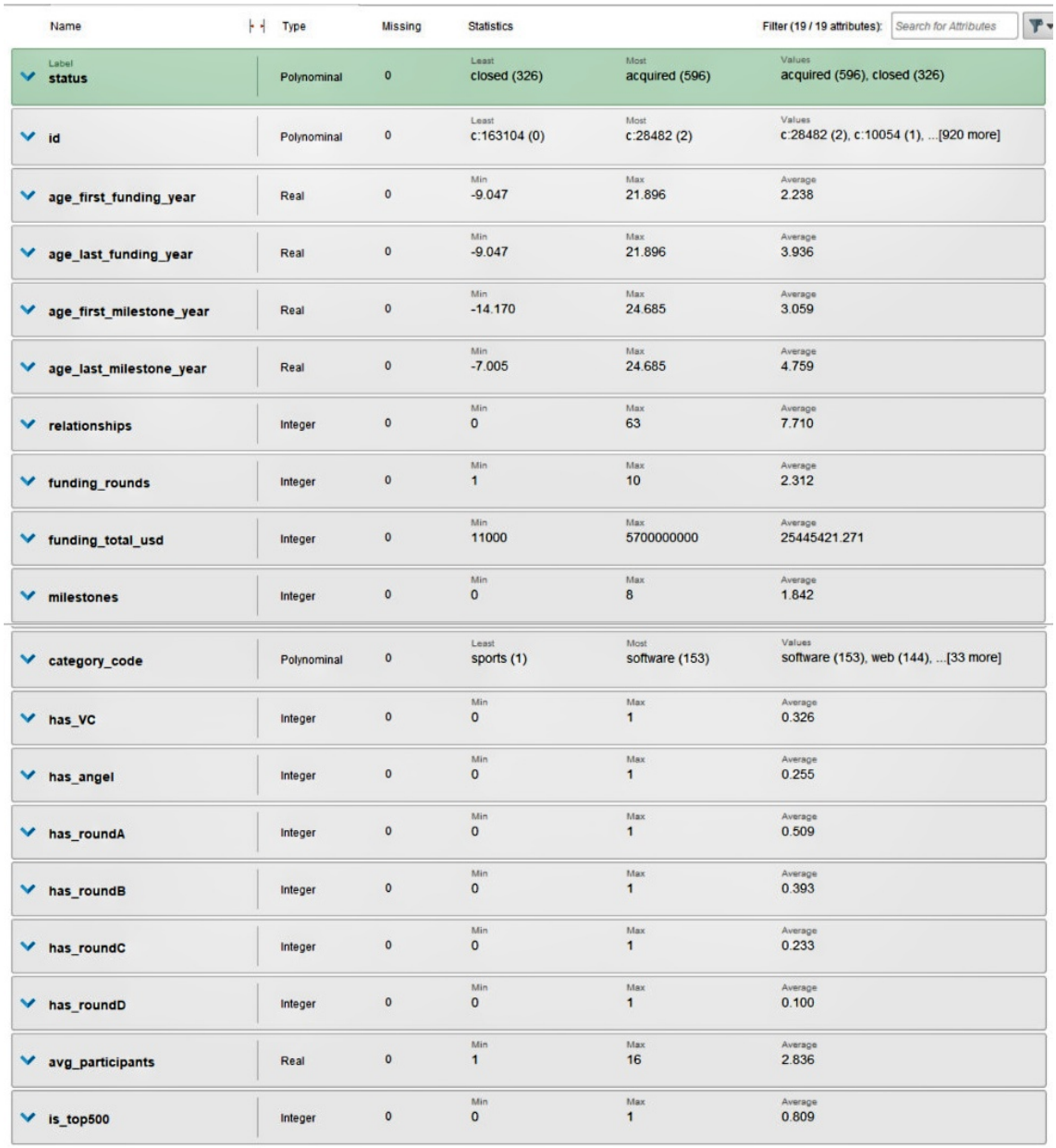
Row No.	status	id	age_first_fu...	age_last_fu...	age_first_m...	age_last_mi...	relationships	funding_rou...	funding_tota...	milestones
1	acquired	c:6669	2.249	3.003	4.668	6.704	3	3	375000	3
2	acquired	c:16283	5.126	9.997	7.005	7.005	9	4	40100000	1
3	acquired	c:65620	1.033	1.033	1.458	2.205	5	1	2600000	2
4	acquired	c:42668	3.131	5.315	6.003	6.003	5	3	40000000	1
5	closed	c:65806	0	1.669	0.038	0.038	2	2	1300000	1
6	closed	c:22898	4.545	4.545	5.003	5.003	3	1	7500000	1
7	acquired	c:16191	1.720	5.211	3	6.608	6	3	26000000	2
8	acquired	c:5192	1.647	6.762	5.606	7.362	25	3	34100000	3
9	acquired	c:1043	3.586	11.112	8.005	9.995	13	3	9650000	4
10	acquired	c:498	1.671	4.685	2.918	6.115	14	3	5750000	4
11	acquired	c:3949	4.627	9.449	10.134	10.649	22	3	27500000	3
12	closed	c:4829	1.085	5.337	-0.616	4.608	8	5	10400000	2
13	closed	c:30290	4.904	4.904	?	?	0	1	350000	0
14	acquired	c:1491	0.019	2.436	0.794	4.378	15	3	9950000	3
15	acquired	c:15645	4.666	8.997	8.838	8.838	12	5	10700000	1
16	closed	c:54177	6.608	6.608	?	?	0	1	200000	0
17	closed	c:16770	2.586	6.764	5.501	5.501	8	5	49000000	1
18	acquired	c:107	4.592	7.173	-0.499	12.680	7	4	25000000	3
19	acquired	c:50727	0.743	1.581	1.285	3.003	10	3	4575000	3

Gambar 2. *Dataset Start-up*.

3.2. *Data Preparation*

Tahapan selanjutnya yaitu *data preparation*, pada tahap ini dilakukan pembersihan data menggunakan operator *Replace Missing Value* yang ada di dalam RapidMiner, sehingga akan menghasilkan data yang bersih dari *missing value*. Dengan operator ini *missing value* dipecahkan dengan menambahkan data dengan nilai rata-rata. Tetapi operator *Replace Missing Value* tidak dapat merubah data dari label yang kosong, sehingga diperlukan lagi satu operator yaitu *Filter Example*, dengan filter label *is not missing*. Hasil dari tahapan data preparation ini dapat dilihat pada Gambar 3 berikut tepatnya pada kolom *Missing*, nilai yang ada di dalam kolom tersebut

adalah 0 untuk setiap data atributnya yang menandakan sudah tidak ada data yang terdapat *missing value* di dalamnya.

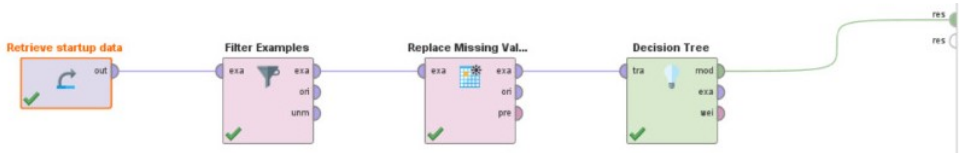


Name	Type	Missing	Statistics		
Label **status**	Polynominal	0	Least: closed (326)	Most: acquired (596)	Values: acquired (596), closed (326)
id	Polynominal	0	Least: c:163104 (0)	Most: c:28482 (2)	Values: c:28482 (2), c:10054 (1), ...[920 more]
age_first_funding_year	Real	0	Min: -9.047	Max: 21.896	Average: 2.238
age_last_funding_year	Real	0	Min: -9.047	Max: 21.896	Average: 3.936
age_first_milestone_year	Real	0	Min: -14.170	Max: 24.685	Average: 3.059
age_last_milestone_year	Real	0	Min: -7.005	Max: 24.685	Average: 4.759
relationships	Integer	0	Min: 0	Max: 63	Average: 7.710
funding_rounds	Integer	0	Min: 1	Max: 10	Average: 2.312
funding_total_usd	Integer	0	Min: 11000	Max: 5700000000	Average: 25445421.271
milestones	Integer	0	Min: 0	Max: 8	Average: 1.842
category_code	Polynominal	0	Least: sports (1)	Most: software (153)	Values: software (153), web (144), ...[33 more]
has_VC	Integer	0	Min: 0	Max: 1	Average: 0.326
has_angel	Integer	0	Min: 0	Max: 1	Average: 0.255
has_roundA	Integer	0	Min: 0	Max: 1	Average: 0.509
has_roundB	Integer	0	Min: 0	Max: 1	Average: 0.393
has_roundC	Integer	0	Min: 0	Max: 1	Average: 0.233
has_roundD	Integer	0	Min: 0	Max: 1	Average: 0.100
avg_participants	Real	0	Min: 1	Max: 16	Average: 2.836
is_top500	Integer	0	Min: 0	Max: 1	Average: 0.809

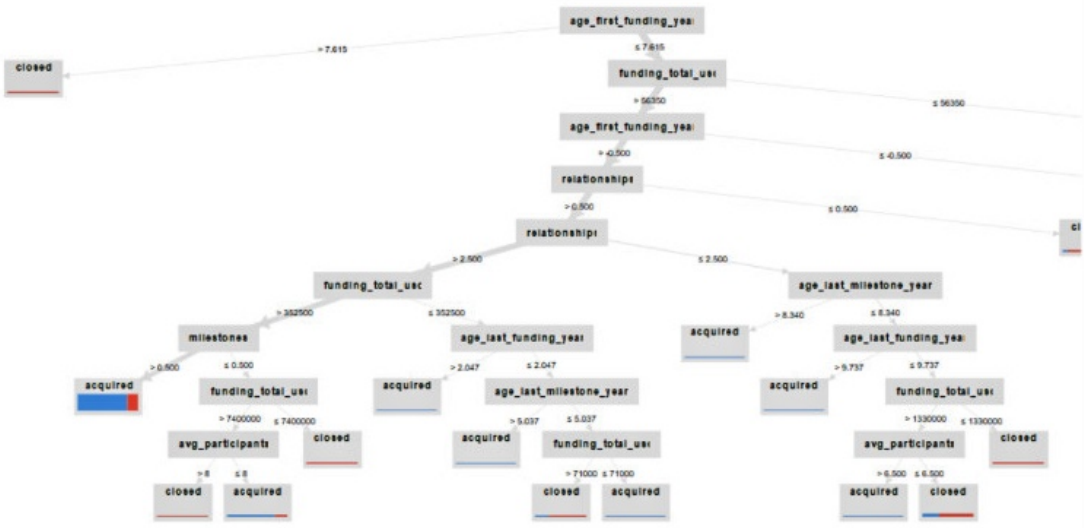
Gambar 3. Hasil *Data Preparation*.

3.3. *Modeling*

Kemudian memasuki tahapan selanjutnya yaitu *modeling*, pada tahap ini data training diklasifikasikan menggunakan algoritma Decision Tree. Hasil pembentukan model ini dapat diketahui faktor apa saja yang sangat mempengaruhi kesuksesan sebuah *start-up*. Pemodelan yang dilakukan pada penelitian ini dapat dilihat pada Gambar 4 berikut, kemudian hasil klasifikasinya yang berupa *tree* dapat dilihat pada Gambar 5.



Gambar 4. *Modeling*.

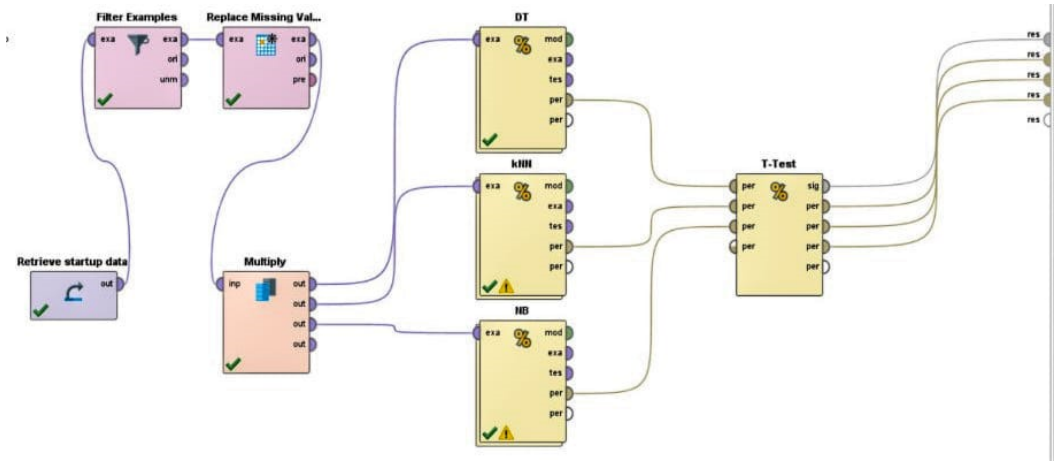


Gambar 5. *Tree* Hasil *Modeling*.

Dari hasil *tree* diatas, dapat ditarik kesimpulan bahwa faktor-faktor penting yang mempengaruhi kesuksesan sebuah *start-up* adalah *age_first_funding_year, total_funding*, serta *relationship*. Variabel *age_first_funding_year, total_funding*, serta *relationship* yang semakin besar maka semakin besar pula kesempatan sebuah *start-up* tersebut akan semakin sukses.

3.4. Evaluation

Tahap berikutnya yaitu evaluasi, pada tahap ini melakukan perbandingan hasil pengujian algoritma Decision Tree, kNN, dan Naive Bayes menggunakan *cross validation* dengan 10-*fold cross validation* dan dilanjutkan dengan pengujian T-test dengan $\alpha = 0{,}05$. Desain dari model pengukuran *performance* dapat dilihat pada Gambar 6 dibawah ini. Sedangkan hasil perbandingan nilai *performance* menggunakan *cross validation* dan T-test ada pada Tabel 1 dan Tabel 2 berikut.



Gambar 6. Desain Model Pengujian.

Tabel 1. Hasil Pengujian dengan *Cross Validation*.

No.	Algoritma	Akurasi	Presisi	Recall
1	Decision Tree	79,29%	78,99%	56,27%
2	kNN	66,69%	55,13%	40,14%
3	Naive Bayes	64,21%	51,32%	79,16%

Dari hasil perbandingan performa akurasi, terlihat algoritma Decision Tree memiliki nilai persentase paling tinggi yaitu 79,29%, sedangkan algoritma kNN dengan 66,69%, dan Naive Bayes dengan 64,21%. Pada performa presisi, algoritma Decision Tree masih menunjukkan nilai terbaik dengan 78,99%, diikuti algoritma kNN dengan 55,13%, dan Naive Bayes 51,32%. Dari hasil performa *recall*, ternyata algoritma Naive Bayes menunjukkan hasil paling baik dengan 79,16%, sedangkan Decision Tree 56,27% dan kNN dengan 40,14%. Dari keseluruhan hasil belum ada yang melebihi 80% sehingga belum bisa dikatakan baik, akan tetapi algoritma Decision Tree memiliki 2 nilai tertinggi pada performa akurasi dan presisi. Sehingga dari ketiga algoritma tersebut pada pengujian dengan *cross validation* ini, algoritma Decision Tree adalah pilihan terbaik.

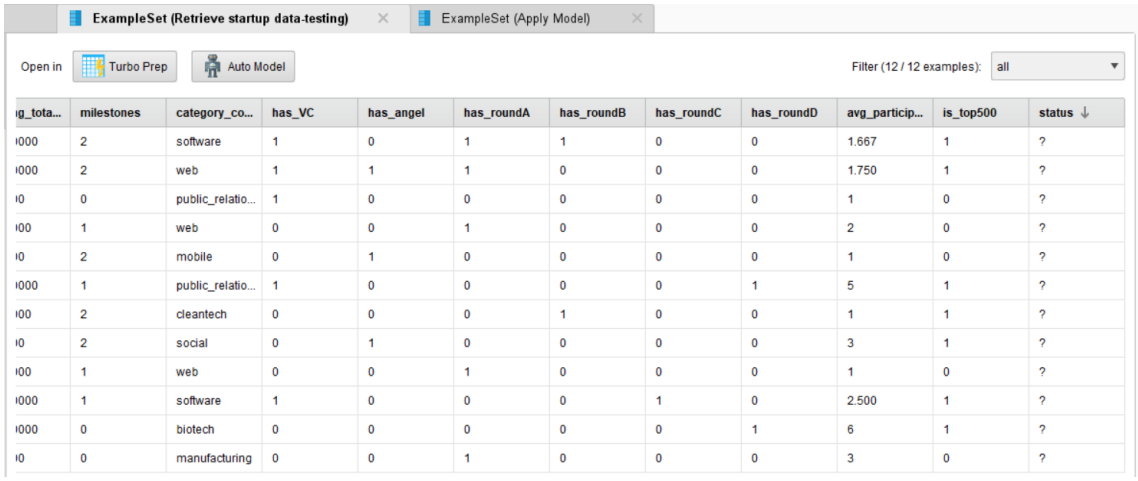
Tabel 2. Hasil Pengujian dengan T-test.

	DT	kNN	NB
DT		**0,000**	**0,000**
kNN	**0,000**		0,484
NB	**0,000**	0,484	

Berdasarkan tabel hasil T-test di atas, menunjukkan hasil bahwa algoritma Decision Tree merupakan algoritma yang paling dominan terhadap algoritma yang lain. Suatu algoritma bisa dikatakan dominan apabila nilai hasil perbandingan performa T-test dengan algoritma lain lebih kecil daripada nilai *alpha*, sedang nilai *alpha* disini adalah 0,050. Sehingga berdasarkan perbandingan nilai performa dan hasil T-test, dapat disimpulkan bahwa algoritma Decision Tree adalah algoritma terbaik yang dapat digunakan pada studi kasus pengklasifikasian *start-up*.

3.5. *Deployment*

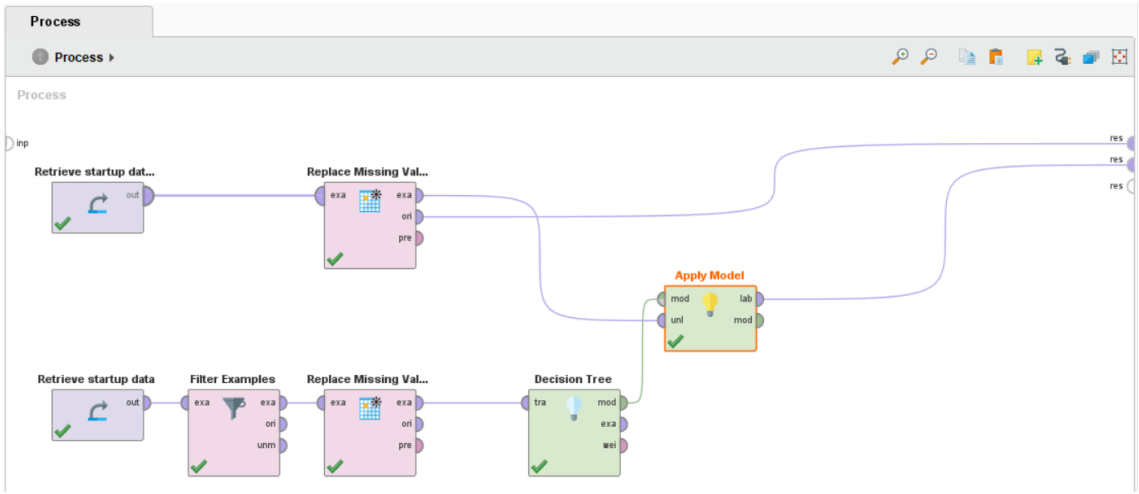
Tahapan terakhir yaitu *deployment*, pada tahap ini model yang telah dibentuk pada tahapan *modeling* digunakan untuk melakukan prediksi keberhasilan *start-up*. Untuk data *testing* yang akan digunakan dapat dilihat pada Gambar 7. Desain model untuk melakukan prediksi dapat dilihat pada Gambar 8 berikut, kemudian perhatikan Gambar 9 di bawah ini untuk hasil prediksi yang telah dilakukan.


ExampleSet (Retrieve startup data-testing) | ExampleSet (Apply Model)

Open in | Turbo Prep | Auto Model | Filter (12 / 12 examples): all

g_tota...	milestones	category_co...	has_VC	has_angel	has_roundA	has_roundB	has_roundC	has_roundD	avg_particip...	is_top500	status
000	2	software	1	0	1	1	0	0	1.667	1	?
000	2	web	1	1	1	0	0	0	1.750	1	?
0	0	public_relatio...	1	0	0	0	0	0	1	0	?
00	1	web	0	0	1	0	0	0	2	0	?
0	2	mobile	0	1	0	0	0	0	1	0	?
000	1	public_relatio...	1	0	0	0	0	1	5	1	?
00	2	cleantech	0	0	0	1	0	0	1	1	?
0	2	social	0	1	0	0	0	0	3	1	?
00	1	web	0	0	1	0	0	0	1	0	?
000	1	software	1	0	0	0	1	0	2.500	1	?
000	0	biotech	0	0	0	0	0	1	6	1	?
0	0	manufacturing	0	0	1	0	0	0	3	0	?

Gambar 7. *Data Testing*.



Gambar 8. Desain Model Prediksi.

ExampleSet (Retrieve startup data-testing) | ExampleSet (Apply Model)

Open in Turbo Prep | Auto Model | Filter (12 / 12 examples): all

Row No.	prediction(s...	confidence(...	confidence(...	id	age_first_fu...	age_last_fu...	age_first_m...	age_last_mi...	relationships	funding_rou...	funding_
1	acquired	0.825	0.175	c:16191	1.720	5.211	3	6.608	6	3	2600000
2	acquired	0.825	0.175	c:4829	1.085	5.337	-0.616	4.608	8	5	1040000
3	closed	0.097	0.903	c:54177	6.608	6.608	2.081	4.229	0	1	200000
4	closed	0.326	0.674	c:15888	5.490	5.490	0	0	1	1	3170000
5	closed	0.273	0.727	c:149809	0	0.584	0.584	0.595	9	2	125000
6	acquired	0.825	0.175	c:22027	5.186	6.814	7.334	7.334	9	2	2900000
7	acquired	0.825	0.175	c:25325	2.644	2.644	0	7.392	4	1	5000000
8	acquired	0.825	0.175	c:70586	4.285	4.285	2.416	5.515	7	1	400000
9	acquired	0.825	0.175	c:34028	1.956	1.956	1.003	1.003	5	1	5000000
10	acquired	0.825	0.175	c:45611	5.685	6.315	5.005	5.005	4	2	1505000
11	closed	0.326	0.674	c:31754	4.238	4.238	2.081	4.229	1	1	5000000
12	closed	0.097	0.903	c:158294	4.370	4.370	2.081	4.229	0	1	662000

Gambar 9. Hasil Prediksi.

4. KESIMPULAN

Berdasarkan hasil perbandingan antara algoritma Decision Tree, kNN, dan Naive Bayes, untuk melakukan klasifikasi terhadap 923 data *start-up*, menunjukkan algoritma Decision Tree merupakan algoritma yang paling cocok untuk digunakan di antara algoritma kNN dan Naive Bayes. Hasil akurasi Decision Tree adalah sebesar 79,29%, sedangkan algoritma kNN dengan 66,69%, dan Naive Bayes dengan 64,21%. Selanjutnya untuk nilai presisinya, Decision Tree masil lebih unggul dengan nilai 78,99%, diikuti algoritma kNN dengan 55,13%, dan Naive Bayes 51,32%. Dari hasil performa *recall*, ternyata algoritma Naive Bayes menunjukkan hasil paling baik dengan 79,16%, sedangkan Decision Tree 56,27% dan kNN dengan 40,14%. Hasil pengujian T-test juga menunjukkan algoritma Decision Tree adalah algoritma paling dominan di antara algoritma yang lain. Selain itu, faktor-faktor yang sangat mempengaruhi kesuksesan sebuah *start-up* adalah *age_first_funding_year*, *total_funding*, serta *relationship*. Variabel *age_first_funding_year, total_funding*, serta *relationship* yang semakin besar maka semakin besar pula kesempatan sebuah *start-up* tersebut akan sukses.

Dikarenakan pada penelitian ini penulis melakukan seleksi fitur/atribut secara manual, maka penulis menyarankan bagi penelitian selanjutnya untuk menggunakan metode *feature selection* untuk mengoptimalkan jumlah *feature*/*attributes* sehingga dapat meningkatkan nilai akurasi yang diperoleh. Selain itu penulis juga menyarankan penelitian selanjutnya untuk melakukan

percobaan dengan metode klasifikasi lain seperti Support Vector Machine, Neural Network, dan sebagainya.

DAFTAR PUSTAKA

Afdi, Z., & Purwanggono, B. (2017). Perancangan Strategi berbasis Metodologi Lean Startup untuk Mendorong Pertumbuhan Perusahaan Rintisan berbasis Teknologi di Indonesia. *Industrial Engineering Online*, *6*(4), 1–13.

Anam, C., & Santoso, H. B. (2018). Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa. *Jurnal Ilmiah Ilmu-Ilmu Teknik*, *8*(1), 13–19.

Blank, S. (2013, May). *Why the Lean Start-Up Changes Everything*. Harvard Business Review. https://hbr.org/2013/05/why-the-lean-start-up-changes-everything

Dellermann, D., Ebel, P., Lipusch, N., Popp, K. M., & Leimeister, J. M. (2017). Finding the Unicorn: Predicting Early Stage Startup Success Through a Hybrid Intelligence Method. *International Conference on Information Systems (ICIS)*, 1–12. https://doi.org/https://dx.doi.org/10.2139/ssrn.3159123

Glupker, J., Nair, V., Richman, B., Riener, K., & Sharma, A. (2019). Predicting investor success using graph theory and machine learning. *Journal of Investment Management*, *17*(1), 92–103.

Gupta, S., Pienta, R., Tamersoy, A., Chau, D. H., & Basole, R. C. (2015). Identifying Successful Investors in the Startup Ecosystem. *Proceedings of the 24th International Conference on World Wide Web*, 39–40. https://doi.org/10.1145/2740908.2742743

Hastuti, K. (2012). Analisis Komparasi Algoritma Klasifikasi Data Mining untuk Prediksi Mahasiswa Non Aktif. *Seminar Nasional Teknologi Informasi & Komunikasi Terapan 2012*, *14*(1), 241–249.

Huda, F. A. (2013). *t-Test*.

Kadafi, A. R. (2018). Perbandingan Algoritma Klasifikasi Untuk Penjurusan Siswa SMA. *Jurnal ELTIKOM*, *2*(2), 67–77. https://doi.org/10.31961/eltikom.v2i2.86

Krisandi, N., Helmi, & Prihando, B. (2015). Acute toxicity of zinc oxide nanoparticles and bulk ZnCl2 to rats. In *Information Technology* (Vol. 2, Issue 1, pp. 123–126). CRC Press. https://doi.org/10.1201/b18776-23

Lakshmi, B. N., Indumathi, T. S., & Ravi, N. (2016). A Study on C.5 Decision Tree Classification Algorithm for Risk Predictions During Pregnancy. *Procedia Technology*, *24*, 1542–1549. https://doi.org/10.1016/j.protcy.2016.05.128

Lukito, Y., & Chrismanto, A. R. (2015). Perbandingan Metode-Metode Klasifikasi untuk Indoor Positioning System. *Jurnal Teknik Informatika Dan Sistem Informasi*, *1*(2), 123–131. https://doi.org/10.28932/jutisi.v1i2.373

Marutho, D. (2019). Perbandingan Metode Naive Bayes , KNN , Decision Tree Pada Laporan Water Level Jakarta. *Manajemen Informatika AMIK JTC Semarang*, *15*(2), 90–97. https://doi.org/https://doi.org/10.53845/infokam.v15i2.175

Praningki, T., & Budi, I. (2018). Sistem Prediksi Penyakit Kanker Serviks Menggunakan CART, Naive Bayes, dan k-NN. *Creative Information Technology Journal*, *4*(2), 83. https://doi.org/10.24076/citec.2017v4i2.100

Rahman, M. A., Hidayat, N., & Afif Supianto, A. (2018). Komparasi Metode Data Mining K-Nearest Neighbor Dengan Naïve Bayes Untuk Klasifikasi Kualitas Air Bersih (Studi Kasus PDAM Tirta Kencana Kabupaten Jombang). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(12), 6346–6353.

Sabna, E., & Muhardi, M. (2016). Penerapan Data Mining Untuk Memprediksi Prestasi Akademik Mahasiswa Berdasarkan Dosen, Motivasi, Kedisiplinan, Ekonomi, dan Hasil Belajar. *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer Dan Teknologi Informasi*, *2*(2), 41. https://doi.org/10.24014/coreit.v2i2.2392

Setiyorini, T., & Asmono, R. T. (2018). Komparasi Metode Decision Tree, Naive Bayes Dan K-Nearest Neighbor Pada Klasifikasi Kinerja Siswa. *Jurnal Techno Nusa Mandiri*, *15*(2), 85. https://doi.org/10.33480/techno.v15i2.889

Wibisono, A. B., & Fahrurozi, A. (2019). PERBANDINGAN ALGORITMA KLASIFIKASI DALAM PENGKLASIFIKASIAN DATA PENYAKIT JANTUNG KORONER. *Jurnal Ilmiah Teknologi Dan Rekayasa*, *24*(3), 161–170. https://doi.org/10.35760/tr.2019.v24i3.2393