# Perbandingan Metode Klasifikasi *Naive* Bayes dan *K-Nearest Neighbor* (Studi Kasus : Status Kerja Penduduk Di Kabupaten Kutai Kartanegara Tahun 2018)

## *The Comparison Method Of Classification Naive Bayes and K-Nearest Neighbor (Case Study: Employment Status Of Citizen In Kutai Kartanegara Regency 2018)*

**Viona Novalia, Rito Goejantoro, dan Sifriyani**
Laboratorium Statistika Komputasi, FMIPA Universitas Mulawarman
E-mail: vionanovalia@gmail.com

***Abstract***

*Classification is a technique to build a model and assess an object to put in a particular class. Naive Bayes is one of algorithm in the classification based on the Bayesian theorem, which assumes the independencies of one class with another class. K-nearest neighbor is an algorithm in the classification method for classifiying based on data that has a closest distance between one object and another object. Naive Bayes and k-nearest neighbor methods are used in classification of the employment status of citizen in Kutai Kartanegara regency because has a good accuracy and produce a small error rate when using large data sets. This research aim to compared optimal performance accuracy of both methods on the classifiying of the employment status of citizen. The data used are employment status of citizen in Kutai Kartanegara Regency based on SAKERNAS of East Kalimantan Province in 2018 and used 5 factors namely age, sex, status in the household, marital status, and education to predict employment status of citizen. Based on the analysis, classification the employment status of citizen with naive Bayes method has accuracy of 90,08% and in the k-nearest neighbor has accuracy of 94,66%. To evaluate the accuracy of classification used calculation of Press's Q. Based on Press's Q value showed that both of classification methods are accurate. From that analysis, can be concluded that the k-nearest neighbor method works better compared with the naive Bayes method for the case of the employment status of citizen in Kutai Kartanegara Regency.*

***Keywords:*** *classification, naive Bayes, k-nearest neighbor, employment status of citizen.*

**Pendahuluan**

*Data mining* adalah suatu istilah yang digunakan untuk menemukan pengetahuan yang tersembunyi di dalam *database*. *Data mining* merupakan proses semi otomatis yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi pengetahuan potensial dan berguna yang bermanfaat, dan tersimpan di dalam *database* besar (Turban, 2005).

Terdapat banyak jenis teknik analisis yang digolongkan dalam *data mining*. Salah satu teknik analisis yang akan dibahas lebih lanjut dalam penelitian ini adalah klasifikasi. Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi terdapat proses yang dilakukan yaitu membangun model untuk melakukan klasifikasi pada suatu data lain, agar diketahui di kelas mana objek data tersebut dimasukkan berdasarkan model yang telah disimpan dalam memori (Prasetyo, 2012).

*Naive* Bayes merupakan pengklasifikasian dengan metode probabilitas yang ditemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan sebelumnya sehingga dikenal sebagai teorema Bayes. Teorema tersebut dikombinasikan dengan *naive* di mana diasumsikan kondisi antar petunjuk (atribut) saling bebas. Klasifikasi *naive* Bayes diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya (Bustami, 2013). Sedangkan algoritma *k-nearest neighbor* merupakan sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut.

Keunggulan dari pendekatan *naive* Bayes dan *k-nearest neighbor* adalah terkenal memiliki tingkat keakuratan yang baik. Pada pengklasifikasian *Naive* Bayes nilai *error* yang dihasilkan akan lebih kecil ketika data set berjumlah besar (Berry, 2006). Sedangkan pada pengklasifikasian dengan algoritma *k-nearest neighbor* akan lebih efektif ketika data *training* berjumlah besar, sehingga menghasilkan ketepatan klasifikasi yang lebih akurat. Selain itu, menurut Putri (2014), berbeda dengan metode pengklasifikasian regresi logistik, pada metode klasifikasi *naive* Bayes dan *k-neare neighbor* tidak diperlukan adanya pemodelan. Metode klasifikasi *naive* Bayes dan *k-nearest neighbor* juga merupakan metode yang cocok digunakan pada data kelas *C* bertipe kategorik dimana untuk penelitian ini digunakan data status kerja penduduk yaitu kelas pengangguran dan bukan pengangguran.

Indonesia merupakan negara kepulauan yang memiliki ribuan pulau di dalamnya. Hal ini

menyebabkan Indonesia menjadi salah satu negara yang memiliki jumlah penduduk terbesar di dunia. Dengan jumlah penduduk yang besar, maka Indonesia sangat erat kaitannya dengan status kerja penduduknya yaitu sebagai pengangguran ataupun bukan pengangguran. Menurut Putri (2014), jumlah penduduk yang besar apabila tidak diimbangi dengan lapangan kerja yang memadai akan menyebabkan semakin tingginya tingkat pengangguran.

Penelitian ini terkait dengan penelitian sebelumnya yaitu, Claudy (2018) yang telah melakukan penelitian tentang klasifikasi dokumen *twitter* untuk mengetahui karakter calon karyawan menggunakan algoritma *k-nearest neighbor* didapatkan hasil bahwa algoritma *k-nearest neighbor* dapat diimplementasikan pada sistem klasifikasi kepribadian atau karakter calon karyawan, dengan tingkat akurasi sebesar 66%. Sementara itu Sholihah (2016) telah melakukan penelitian tentang klasifikasi perubahan harga obligasi korporasi di Indonesia menggunakan metode *naive* Bayes *classification* yang menyatakan bahwa metode *naive* Bayes bekerja cukup baik untuk mengklasifikasi perubahan harga obligasi korporasi di Indonesia.

Berdasarkan latar belakang tersebut, maka penulis tertarik untuk mencoba mengaplikasikan metode klasifikasi *naive* Bayes dan *k-nearest neighbor* dengan mengangkat permasalahan kependudukan serta membandingkan keoptimalan kedua metode tersebut dalam mengklasifikasikan data status kerja penduduk di Kabupaten Kutai Kartanegara.

### *Data mining*

*Data mining* adalah suatu istilah yang digunakan untuk menemukan pengetahuan yang tersembunyi di dalam *database*. Data *mining* merupakan proses semi otomatis yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi pengetahuan potensial dan berguna yang bermanfaat yang tersimpan di dalam *database* besar (Turban, 2005).

### Klasifikasi

Menurut Prasetyo (2012), klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi terdapat suatu proses yang dilakukan, yaitu dengan membangun model untuk melakukan pengenalan atau klasifikasi atau prediksi pada suatu data lain supaya diketahui di kelas mana objek data tersebut dimasukkan berdasarkan model yang telah disimpan dalam memori.

### Metode Klasifikasi *Naive* Bayes

Menurut Bustami (2013), algoritma *naive* Bayes merupakan salah satu algoritma yang terdapat pada teknik klasifikasi. *naive* Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai teorema Bayes. Teorema tersebut dikombinasikan dengan *naive* dimana diasumsikan kondisi antar atribut saling bebas. Klasifikasi *naive* Bayes diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya.

Persamaan dari teorema Bayes adalah sebagai berikut:

$$P(C \mid F) = \frac{P(C)P(F \mid C)}{P(F)} \tag{1}$$

dimana :

| | | |
|---|---|---|
| $P(C\|F)$ | : | Probabilitas akhir bersyarat (*posterior*) suatu kelas *C* terjadi jika diberikan petunjuk. |
| $P(C)$ | : | Probabilitas awal (*prior*) kelas *C* terjadi tanpa memandang petunjuk (atribut) apapun. |
| $P(F\|C)$ | : | Probabilitas sebuah petunjuk (atribut) *F* terjadi akan mempengaruhi kelas *C*. |
| $P(F)$ | : | Probabilitas awal (*prior*) petunjuk (atribut) *F* terjadi tanpa memandang kelas apapun . |

Untuk menjelaskan teorema *naive* Bayes, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk (atribut) untuk menentukan kelas yang tepat bagi objek yang dianalisis tersebut. Oleh karena itu, teorema Bayes di atas disesuaikan sebagai berikut:

$$P(C \mid F_1 ..., F_n) = \frac{P(C)P(F_1 ..., F_n \mid C)}{P(F_1 ..., F_n)} \tag{2}$$

Dimana variabel *C* mempresentasikan kelas, sementara variabel $F_1 ..., F_n$ mempresentasikan sejumlah petunjuk (atribut) yang dibutuhkan untuk melakukan klasifikasi. Maka rumus tersebut menjelaskan bahwa peluang masuknya objek dengan petunjuk tertentu (atribut tertentu) dalam kelas *C* (*posterior*) adalah peluang munculnya kelas *C* (sebelum masuknya objek tersebut, seringkali disebut *prior*), dikali dengan peluang kemunculan petunjuk-petunjuk (atribut) objek pada kelas *C* (disebut juga *likelihood*), dibagi dengan peluang kemunculan petunjuk (atribut) objek secara global (disebut juga *evidence*). Oleh karena itu, rumus diatas dapat pula ditulis secara sederhana sebagai berikut:

$$Posterior = \frac{prior \times likelihood}{evidence} \quad (3)$$

Nilai dari *posterior* tersebut nantinya akan dibandingkan dengan nilai-nilai *posterior* kelas lainnya untuk menentukan kelas suatu objek dapat ditentukan dengan memilih kelas yang memiliki *posterior* terbesar karena nilai *evidence* selalu tetap dan merupakan pembagi pada setiap kelasnya sehingga dalam perhitungan *posterior* cukup mengalikan nilai *prior* dengan *likelihood*. Penjabaran lebih lanjut rumus Bayes tersebut dilakukan dengan menjabarkan $P(C \mid F_1 ..., F_n)$ menggunakan aturan perkalian berikut:

$$\begin{aligned} P(C \mid F_1 ..., F_n) &= P(C)P(F_1 ..., F_n \mid C) \\ &= P(C)P(F_1 \mid C)P(C)P(F_2 ..., F_n \mid C, F_1) \\ &= P(C)P(F_1 \mid C)P(F_2 \mid C, F_1)P(F_3 \mid C, F_1, F_2) \\ &\quad ... P(F_n \mid C, F_1, F_2, F_3, ..., F_{n-1}) \end{aligned} \quad (4)$$

Dapat dilihat bahwa hasil penjabaran tersebut menyebabkan semakin banyak dan semakin kompleksnya faktor-faktor syarat yang mempengaruhi nilai probabilitas, yang hampir mustahil untuk dianalisis satu per satu. Akibatnya mustahil perhitungan tersebut menjadi sulit untuk dilakukan. Disinilah digunakan asumsi independensi yang sangat tinggi (*naive*), bahwa masing-masing petunjuk $F_1, F_2, ..., F_n$ saling bebas (*independent*) satu sama lain. Dengan asumsi tersebut, berlaku suatu kesamaan sebagai berikut:

$$P(F_1 \mid F_j) = \frac{P(F_i \cap F_j)}{F_j} = \frac{P(F_i) \cap P(F_j)}{P(F_j)} = P(F_i) \quad (5)$$

Untuk $i \neq j$, sehingga

$$P(F_1 \mid C, F_j) = P(F_i \mid C) \quad (6)$$

Persamaan diatas dapat disimpulkan bahwa asumsi independensi *naive* tersebut membuat syarat peluang menjadi sederhana, sehingga perhitungan menjadi mungkin untuk dilakukan. Selanjutnya penjabara $P(C \mid F_1, ..., F_n)$ dapat disederhanakan menjadi,

$$\begin{aligned} P(C \mid F_1, ..., F_n) &= P(C)P(F_1 \mid C)P(F_2 \mid C)...P(F_n \mid C) \\ &= P(C)\prod_{i=1}^{n} P(F_i \mid C) \end{aligned} \quad (7)$$

Persamaan diatas merupakan model dari teorema *naive* Bayes yang selanjutnya akan digunakan dalam proses klasifikasi.

Menurut Saleh (2015), alur dari metode *naive* Bayes adalah sebagai berikut :

1. Membaca data *training*
2. Menghitung nilai probabilitas setiap atribut pada setiap kelasnya
3. Membuat tabel probabilitas
4. Menentukan probabilitas akhir untuk setiap kelas. Persamaan yang digunakan adalah sebagai berikut :
   a. Menghitung probabilitas akhir untuk setiap kelas. Persamaan yang digunakan sebagai berikut :

$$\prod_{i=1}^{n} P(F_i \mid C) \quad (8)$$

   b. Nilai dari persamaan diatas dimasukkan untuk mendapatkan probabilitas akhir. Adapun persamaan yang digunakan berdasarkan persamaan (7).

**Metode Klasifikasi *K-Nearest Neighbor***

Menurut Islam, Wu, Ahmadi, dan Sid-Ahmed (2007) algoritma *k-nearest neighbor* adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. *k-nearest neighbor* adalah sebuah metode untuk mencari kasus dengan menghitung kedekatan antara kasus baru dan kasus lama yaitu berdasarkan pada kecocokan bobot dari sejumlah fitur yang ada. Untuk mendefinisikan jarak antara dua titik pada data *training* ($x$) dan titik pada data *testing* ($y$) maka digunakan rumus *euclidean*, seperti yang ditunjukkan pada persamaan (8).

$$d(x, y) = \sqrt{\sum_{i=1}^{r} (x_{ik} - y_{ik})^2} \quad (9)$$

Dengan *d* adalah jarak *euclidean* antara titik pada data *training* (*x*) dan data *testing* (*y*) yang akan diklasifikasi, dimana $x_{ik}$ adalah nilai ke-*i* variabel ke-*k* dari *x*, $y_{ik}$ adalah nilai ke-*i* variabel ke-*k* dari *y*, dan *r* adalah jumlah variabel.

Menurut Prasetyo (2012) pada metode *k-nearest neighbor* (K-NN), nilai *k* menyatakan jumlah tetangga terdekat yang dilibatkan dalam penentuan prediksi label kelas pada data *testing*. Dari *k* tetangga terdekat yang terpilih kemudian dilakukan *voting* kelas dari *k* tetangga terdekat tersebut. Kelas dengan jumlah suara terbanyak yang diberikan sebagai label kelas hasil prediksi pada data *training* tersebut. Algoritma K-NN dapat dilihat pada tabel 1.

**Tabel** 1. Algoritma KNN

| No | Algoritma K-NN |
|---|---|
| 1 | Menentukan jumlah K tetangga terdekat |
| 2 | Melakukan perhitungan jarak antar data baru dan semua data yang di *training* dengan menggunakan rumus jarak *euclidean* (*d*). |
| 3 | Mengurutkan jarak (*ranking*) |
| 4 | Gunakan *voting* kelas sebagai prediksi dari data baru tersebut |

**Teknik Validasi**

Ada beberapa metode dalam validasi silang diantaranya yang pertama metode *k-fold*. *K-fold cross validation* adalah salah satu teknik untuk memperkirakan kinerja klasifikasi. Pendekatan ini memecah set data menjadi *k* bagian set data dengan ukuran yang sama. Setiap kali berjalan, satu pecahan berperan sebagai set data uji sedangkan pecahan lainnya menjadi set data latih. Prosedur tersebut dilakukan sebanyak *k* kali sehingga setiap data berkesempatan menjadi data uji tepat satu kali dan menjadi data latih sebanyak *k*-1 kali (Prasetyo, 2012).

Metode validasi silang yang kedua yaitu metode *holdout*. Dalam metode *holdout*, data awal yang diberi label dipartisi ke dalam dua himpunan secara random yang dinamakan data *training* dan data *testing*. Proporsi data yang dicadangkan untuk data *training* dan data *testing* tergantung pada analisis misalnya 50%-50% atau 2/3 untuk *training* dan 1/3 untuk *testing*, namun menurut Written (2005) serta Han and Kamber (2006) pada umumnya perbandingan yang digunakan yaitu 2:1 untuk data *training* berbanding data *testing*.

**Pengukuran Akurasi Metode**

Menurut Prasetyo (2012) sebuah sistem yang melakukan klasifikasi diharapkan dapat melakukan klasifikasi semua set data dengan benar, tetapi tidak dapat dipungkiri bahwa kinerja suatu sistem tidak bisa 100% benar sehingga sebuah sistem klasifikasi juga harus diukur kinerjanya. Umumnya, pengukuran kinerja klasifikasi dilakukan dengan matriks konfusi (*confusion matrix*).

Matriks konfusi merupakan tabel pencatat hasil kerja klasifikasi. Tabel 2 merupakan contoh matriks konfusi yang melakukan klasifikasi masalah biner (dua kelas), hanya ada dua kelas, yaitu kelas 0 dan kelas 1. Setiap sel $f_{ij}$ dalam matriks menyatakan jumlah *record* atau data dari kelas *i* yang hasil prediksinya masuk ke kelas *j*. Misalnya, sel $f_{11}$ adalah jumlah data dalam kelas 1 yang secara benar dipetakan ke kelas 1, dan $f_{10}$ adalah data dalam kelas 1 yang dipetakan secara salah ke kelas 0.

**Tabel** 2. Matriks Konfusi Untuk Klasifikasi Dua Kelas

| $f_{ij}$ | | Kelas hasil prediksi (*j*) | |
|---|---|---|---|
| | | Kelas = 1 | Kelas = 0 |
| Kelas asli (*i*) | Kelas = 1 | $f_{11}$ | $f_{10}$ |
| | Kelas = 0 | $f_{01}$ | $f_{00}$ |

Berdasarkan isi matriks konfusi, kita dapat mengetahui jumlah data dari masing-masing kelas yang diprediksi secara benar, yaitu $(f_{11} + f_{00})$, dan data yang diklasifikasi secara salah, yaitu $(f_{10} + f_{01})$. Kuantitas matriks konfusi dapat diringkas menjadi dua nilai, yaitu akurasi dan laju eror. Dengan mengetahui jumlah data yang diklasifikasi secara benar, kita dapat mengetahui akurasi hasil prediksi, dan mengetahui jumlah data yang diklasifikasi secara salah, kita dapat mengetahui laju eror dari prediksi yang dilakukan. Dua kuantitas ini digunakan sebagai matriks kinerja klasifikasi. Untuk menghitung akurasi dapat digunakan rumus sebagai berikut (Prasetyo, 2012):

$$\text{Akurasi} = \frac{\text{Jumlah data yang diprediksi secara benar}}{\text{Jumlah prediksi yang dilakukan}} \times 100\% = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}} \times 100\% \quad (10)$$

APER (*Appearent Error Rate*) atau yang disebut laju error merupakan ukuran yang digunakan untuk melihat peluang kesalahan klasifikasi yang dihasilkan oleh suatu fungsi klasifikasi. Semakin kecil nilai APER maka hasil pengklasifikasian semakin baik (Prasetyo, 2012).

Rumus untuk menghitung APER (Prasetyo, 2012) yaitu :

$$\text{Laju } \textit{Error} = \frac{\text{Jumlah data yang diprediksi secara salah}}{\text{Jumlah prediksi yang dilakukan}} \times 100\% = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}} \times 100\% \quad (11)$$

Semua algoritma klasifikasi berusaha membentuk model yang mempunyai akurasi tinggi (laju *error* yang rendah). Umumnya, model yang dibangun dapat memprediksi dengan benar pada semua data yang menjadi data latihnya, tetapi ketika model berhadapan dengan data uji, barulah kinerja model dari sebuah algoritma klasifikasi ditentukan.

**Evaluasi Ketepatan Klasifikasi**

Untuk mengetahui apakah hasil klasifikasi yang didapatkan sudah akurat atau belum maka perlu dilakukan uji ketepatan hasil klasifikasi. Untuk mengetahui ketepatan tersebut dapat dilakukan dengan menghitung nilai *Press's Q*. *Press's Q* merupakan pengujian untuk mengukur apakah klasifikasi yang dilakukan sudah akurat atau belum. Rumus untuk menghitung *Press's Q* menurut Hair (2006) adalah:

$$\textit{Press's } Q = \frac{[N - (nK)]^2}{N(K-1)}, \quad (12)$$

dimana:
$N$ = Ukuran total sampel
$n$ = Banyak kasus yang diklasifikasi secara tepat
$K$ = Banyak grup

Pengklasifikasian dikatakan akurat apabila nilai *Press's Q* lebih besar daripada nilai kritis yang diambil dari tabel *chisquare* ($\chi^2_{\alpha,1}$) dengan derajat bebas bernilai satu dan tingkat keyakinan sesuai yang diinginkan.

**Status Kerja**

Menurut Putri (2014) status kerja dibedakan menjadi dua yaitu pengangguran dan bukan pengangguran yang merupakan salah satu faktor indikasi suatu negara tersebut dikatakan sudah sejahtera atau belum. Menurut Raharja dan Manurung (2004) bekerja penuh (*employment*) yaitu mereka yang bekerja penuh atau jam kerjanya lebih dari 35 jam/minggu. Sedangkan menganggur (*unemployment*) yaitu mereka yang sama sekali tidak bekerja atau sedang mencari pekerjaan.

Data pengangguran dikumpulkan BPS melalui survei rumah tangga, seperti Survei Angkatan Kerja Nasional (SAKERNAS), Sensus Penduduk (SP), Survei Penduduk Antar Sensus (SUPAS), dan Survei Sosial Ekonomi Nasional (SUSENAS). Diantara sensus/survei tersebut SAKERNAS merupakan survei yang dirancang untuk mengumpulkan data ketenagakerjaan secara periodik. Saat ini SAKERNAS diselenggarakan dua kali setahun yaitu pada bulan Februari dan Agustus (Badan Pusat Statistik Kalimantan Timur, 2017).

**Hasil Penelitian dan Pembahasan**

**1. Data Penelitian**

Data yang digunakan dalam penelitian ini merupakan data Survei Angkatan Kerja Nasional Kabupaten Kutai Kartanegara pada bulan Agustus Tahun 2018 . Dari data tersebut dilakukan analisis dengan menggunakan metode *naive* Bayes dan *k-nearest neighbor*. Adapun variabel terikat adalah status kerja dan faktor yang mempengaruhi terdiri dari umur, jenis kelamin, status dalam rumah tangga, status perkawinan dan pendidikan.

**2. Metode Klasifikasi *Naive* Bayes**

Dalam proses menghitung klasifikasi *naive* Bayes, terdapat tiga alur yaitu membaca data *training*, menghitung nilai probabilitas setiap atribut pada setiap kelasnya dan menentukan probabilitas akhir. Hasil perhitungan probabilitas akhir digunakan untuk menentukan termasuk dalam kelas pengangguran atau bukan pengangguran data yang akan diuji. Pengolahan data menggunakan teknik validasi *holdout*. Adapun data yang digunakan pada proses perhitungan klasifikasi adalah dengan data *training* 90% sebanyak 1178 sampel, sedangkan untuk data *testing* 10% sebanyak 131 sampel. Perhitungan diulang sebanyak 9 kali yang menghasilkan akurasi terbaik pada metode *naïve* Bayes dengan matriks konfusi dapat dilihat pada Tabel 3.

Berdasarkan Tabel 3 pada metode *naive* Bayes dapat diketahui bahwa jumlah data dalam kelas 1 atau kelas bukan pengangguran yang di petakan secara benar ke kelas 1 atau kelas bukan pengangguran adalah sebanyak 115 penduduk, dan yang dipetakan secara salah ke kelas 0 atau kelas pengangguran adalah sebanyak 9 penduduk. Sedangkan jumlah data dalam kelas 0 atau kelas pengangguran yang di petakan secara benar ke kelas 0 atau kelas pengangguran adalah sebanyak 3 penduduk, dan yang dipetakan secara salah ke kelas 1 atau kelas bukan pengangguran adalah sebanyak 4 penduduk. Sehingga didapatkan keakurasian dengan menggunakan metode *naive* Bayes adalah sebesar 90,08%.

**Tabel** 3. Matriks Konfusi *Naive* Bayes

| $f_{ij}$ | | **Kelas Hasil Prediksi (*j*)** | |
|---|---|---|---|
| | | Kelas = 1 | Kelas = 0 |
| Kelas asli (*i*) | Kelas = 1 | 115 | 9 |
| | Kelas = 0 | 4 | 3 |

**3. Metode Klasifikasi *K-Nearest Neighbor***

Dalam proses perhitungan *k-nearest neighbor*, langkah pertama yang dilakukan adalah menentukan nilai *k* optimal. Penentuan nilai *k* optimal dilakukan agar hasil prediksi menjadi lebih akurat. Nilai *k* optimal dapat diketahui dari laju *error* masing-masing *k*. Langkah berikutnya adalah mencari nilai *euclidean* dari data, jarak *Euclidean* yang telah terbentuk kemudian diurutkan (*ranking*) dari jarak yang memiliki nilai terkecil. Kelas yang memiliki anggota terbanyak pada data uji merupakan hasil klasifikasi akhir yang digunakan untuk menentukan termasuk dalam kelas pengangguran atau bukan pengangguran data yang akan diuj. Pengolahan data menggunakan teknik validasi *holdout*. Adapun data yang digunakan pada proses perhitungan klasifikasi adalah dengan data *training* 90% sebanyak 1178 sampel, sedangkan untuk data *testing* 10% sebanyak 131 sampel, serta digunaakan jumlah tetangga terdekat sebanyak 3,5,7, dan 9 yang diulang sebanyak 9 kali menghasilkan laju *error* seperti pada Tabel 4.

**Tabel** 4. Laju *Error* untuk Masing-masing Nilai *K*

| **Jumlah *K*** | **Laju *Error*** |
|---|---|
| 3 | 5,34% |
| 5 | 5,34% |
| 7 | 5,34% |
| 9 | 5,34% |

Dari Tabel 4 dapat diketahui bahwa laju error bernilai konstan pada nilai $k = 3$, sehingga dapat diketahui klasifikasi *k-nearest neighbor* akan menghasilkan akurasi yang optimal apabila

menggunakan nilai $k = 3$. Pada kasus ini diperoleh hasil yang akurasi terbaik dengan metode *k-nearest neighbor* untuk jumlah tetangga terdekat yaitu 3 dengan matriks konfusi seperti pada Tabel 5.

**Tabel** 5. Matriks Konfusi *K-Nearest Neigbor*

| $f_{ij}$ | | Kelas Hasil Prediksi (*j*) | |
|---|---|---|---|
| | | Kelas = 1 | Kelas = 0 |
| Kelas asli (*i*) | Kelas = 1 | 124 | 0 |
| | Kelas = 0 | 7 | 0 |

Dari Tabel 5 dapat diketahui bahwa pada metode *k-nearest neighbor* jumlah data dalam kelas 1 atau kelas bukan pengangguran yang di petakan secara benar ke kelas 1 atau kelas bukan pengangguran adalah sebanyak 124 penduduk, dan tidak ada data yang dipetakan secara salah ke kelas 0 atau kelas pengangguran. Sedangkan jumlah data dalam kelas 0 atau kelas pengangguran yang di petakan secara benar ke kelas 0 atau kelas pengangguran adalah sebanyak 7 penduduk, dan tidak ada data yang dipetakan secara salah ke kelas 1 atau kelas bukan pengangguran. Sehingga didapatkan keakurasian dengan menggunakan metode *k-nearest neighbor* adalah sebesar 94,66%.

**4. Evaluasi Ketepatan Klasifikasi**

Untuk mengevaluasi ketepatan hasil prediksi (klasifikasi) digunakan nilai *Press's Q*. Adapun nilai *Press's Q* pada kedua metode dapat dilihat pada Tabel 6.

**Tabel** 6. Nilai *Press's Q*

| Metode | Nilai *Press's Q* |
|---|---|
| *Naive Bayes* | 84,16 |
| *K-Nearest Neighbor* | 104,49 |

Pada Tabel 6 dapat diketahui bahwa pengklasifikasian menggunakan metode *naive* Bayes dan *k-nearest neighbor* tersebut dikatakan akurat, karena memiliki nilai *Press's Q* yang lebih besar daripada nilai *chisquare* dengan derajat bebas bernilai satu dan tingkat kepercayaan 5% yaitu 3,841.

**5. Perbandingan Ketepatan Kinerja Metode**

Setelah diperoleh akurasi dan nilai *Press's Q* dengan menggunakan metode *naive* Bayes dan *k-nearest neighbor*, hasil perbandingan ketepatan kinerja kedua metode dalam melakukan prediksi (klasifikasi) dapat dilihat pada Tabel 7.

**Tabel** 7. Perbandingan Ketepatan Kinerja Metode

| Metode | Akurasi | *Press's Q* |
|---|---|---|
| *Naive Bayes* | 90,08% | 84,16 |
| *K-Nearest Neighbor* | 94,66% | 104,49 |

Berdasarkan Tabel 7 dapat disimpulkan bahwa metode *k-nearest neighbor* bekerja lebih baik dibandingkan metode *naive* Bayes dalam mengklasifikasikan status kerja penduduk di Kabupaten Kutai Kartanegara dilihat dari akurasi dan nilai *Press's Q* yang lebih tinggi.

**Kesimpulan**

Berdasarkan analisis dan pembahasan yang telah dilakukan, maka dapat diambil kesimpulan sebagai berikut:

1. Metode klasifikasi *naive* Bayes dan *k-nearest neighbor* dapat dipergunakan dalam mengklasifikasikan data status kerja penduduk.
2. Pengklasifikasian status kerja penduduk dengan metode *naive* Bayes menghasilkan akurasi sebesar 90,08% dan pada metode *k-nearest neighbor* menghasilkan akurasi sebesar 94,66%. Berdasarkan perhitungan nilai *Press's Q*, menunjukkan bahwa pengklasifikasian status kerja penduduk di Kabupaten Kutai Kartanegara dengan metode *naive* Bayes dan *k-nearest neighbor* sudah akurat. Berdasarkan hasil analisis tersebut, dapat disimpulkan bahwa pengklasifikasian menggunakan metode *k-nearest neighbor* memiliki ketepatan kinerja yang lebih baik dibandingkan dengan metode *naive* Bayes dalam mengklasifikasikan status kerja penduduk.

**Daftar Pustaka**

Berry, M. W and Browne, M. (2006). *Lecture* Publishing. *Notes in Data Mining*. New Jersey: World Scientific

BPS Provinsi Kalimantan Timur, (2017). *Keadaan Angkatan Kerja Provinsi Kalimantan Timur 2017*. Samarinda: BPS Provinsi Kalimantan Timur.

Bustami. (2013). Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi. *TECHSI : Jurnal Penelitian Teknik Informatika*, 3(2), 129-132.

Claudy, Yessivha., Pradana, Rizal Setya dan Fauzi, M. Ali. (2018). Klasifikasi Dokumen Twitter Untuk Mengetahui Karakter Calon Karyawan Menggunakan Algoritma K-*Nearest Neigbor*. *Jurnal Pengembangan Teknologi dan Ilmu Komputer,* 2(8), 2761-2765.

Hair, J. F., Black. W. C., Babin. B. J. and Anderson R. E. 2006. *Multivariate Data Analysis. Seventh Edition.* Pearson Education Prentice Hall. Inc

Han, J and Kamber, M. (2006). *Data Mining Concepts and Techniques, second edition.* California: Morgan Kaufmann Publishers

Islam, M. J., Wu, Q. M. J., Ahmadi and Sid-Ahmed, M. A. (2007). Investigating the Performance of *Naive Bayes* Classifiers and *K-Nearest Neighbor Classifiers*.

*International Conference on Convergence*

Prasetyo, Eko. 2012. *Data Mining Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta: ANDI Yogyakarta

Putri, Riyan Eko. (2014). Perbandingan Metode Klasifikasi *Naive Bayes* dan *K-Nearest Neighbor* Pada Analisis Data Status Kerja di Kabupaten Demak Tahun 2012. *Jurnal Gaussian*, 3(4), 831-838.

Saleh, A. (2015). Implementasi Metode Klasifikasi *Naive* Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga. *Citec Journal*, 2(3), 207.

Sholihah, Khotimatus (2016). Klasifikasi Perubahan Harga Obligasi Korporasi Di Indonesia Menggunakan Metode Naïve Bayes Classificationn. *Jurnal Gaussian*. 5(2), 269-278.

*Information Technoloy*.

Turban, E., Aronson, J.E., and Liang, T. P. (2005). *Decision Support System and Intelegent System*. New Jersey: Pearson Education, Inc..

Witten, Ian H., Frank, Eibe., and Hal, M.A. (2011). *Data Mining: Pratical Machine Learning Tools and Techniques, Third Edition*. Burlington: Morgan Kaufmann Publishers.

Manurung, Mandala dan Rahardja, Pratama. (2004). *Uang, Perbankan, dan Ekonomi Moneter (Kajian Kontekstual Indonesia)*. Jakarta: Lembaga Penerbit FEUI.