A red and white striped flag

Description automatically generated with medium confidence

TUGAS BESAR MATA KULIAH DATA MINING & DATA WAREHOUSE

BAB 7 & k-Nearest Neighbors (kNN)

KELOMPOK 5

17110275 krisna gilang p

17110264 Azhara f silmi

17110271 wahyu sanjaya

17110170 yani oktapiani

20110150 kiki agustin

k-Nearest Neighbors

(kNN)  
 Dalam bab ini, kami menjelaskan algoritma k-nearest-neighbors yang dapat digunakan untuk klasifikasi (dari hasil kategorikal) atau prediksi (dari hasil numeric datang). Untuk mengklasifikasikan atau memprediksi rekor baru, metode ini mengandalkan pencarian yang "serupa"

catatan dalam data pelatihan. "neighbors" ini kemudian digunakan untuk mendapatkan klasifikasi kation atau prediksi untuk rekor baru dengan voting (untuk klasifikasi) atau rata-rata (untuk prediksi). Kami menjelaskan bagaimana kesamaan ditentukan, berapa jumlah tetangga dipilih, dan bagaimana klasifikasi atau prediksi dihitung. k-NN adalah metode berbasis data yang sangat otomatis. Kami membahas keuntungan dan kelemahan.

kebutuhan metode k-NN dalam hal kinerja dan pertimbangan praktis seperti waktu komputasi.

## 7.1 The k-NN Classifier (Categorical

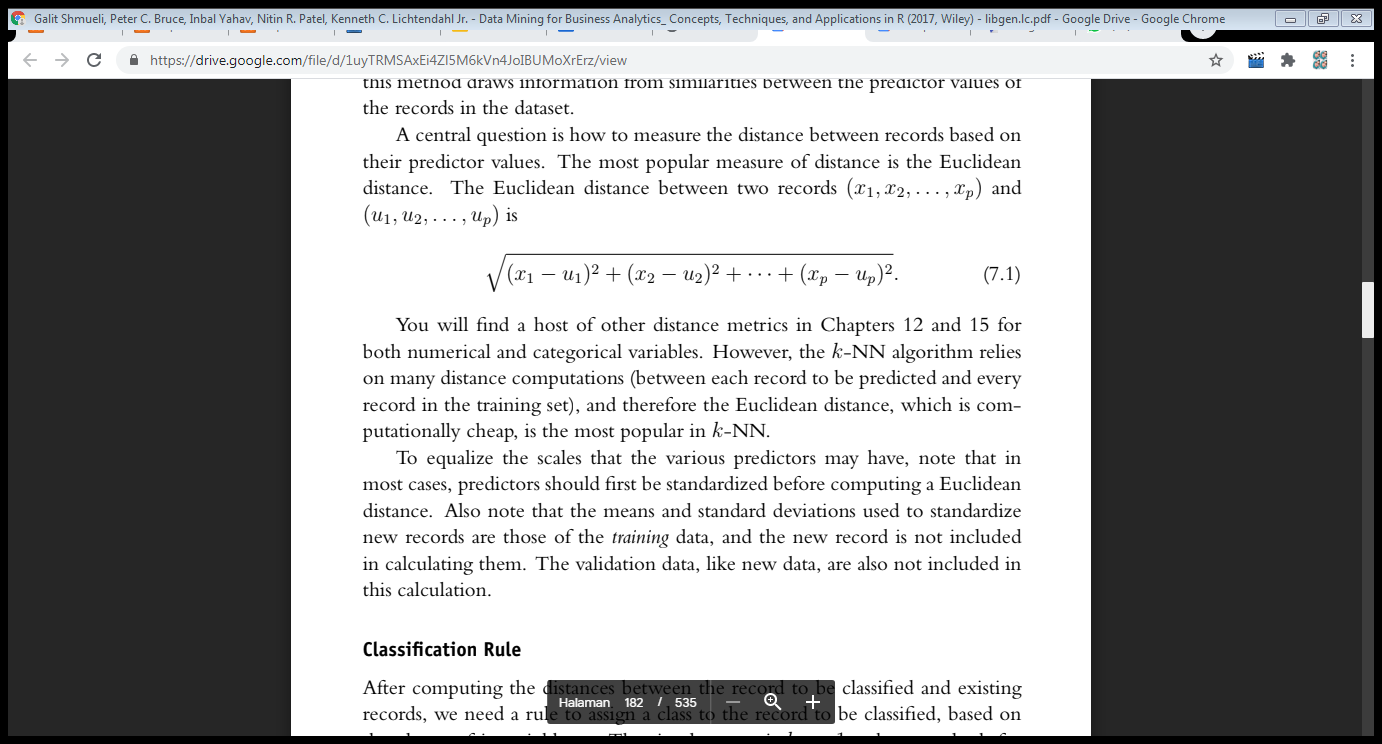
## Outcome)

Ide dalam metode k-nearest-neighbors adalah untuk mengidentifikasi record kedalam pelatihan dataset yang mirip dengan record baru yang ingin kita klasifikasikan. Kami kemudian menggunakan record serupa (bersebelahan) ini untuk mengklasifikasikan record baru ke dalam kelas, menugaskan record baru untuk kelas dominan di antara tetangga ini. Sebutkan nilainya dari prediktor untuk record baru ini dengan x1, x2,. . . , xp. Kami mencari record dalam data pelatihan kami yang serupa atau " near" dengan catatan untuk diklasifikasikan diruang prediktor (i.e.., record yang memiliki nilai mendekati x1, x2,..., xp). Kemudian berdasarkan kelas yang memiliki catatan terdekat tersebut, kami menetapkan kelas ke rekaman yang ingin kami klasifikasikan.

**Determining Neighbors**

Algoritma k-NEAREST NEIGHBORS merupakan salah satu metode klasifikasi yang tidak membuat asumsi tentang bentuk hubungan antar kelas keanggotaan k-NEAREST NEIGHBORS (k-NN) . (Y) dan prediktor X1, X2,. . . , Xp. Ini adalah metode nonparametrik karena tidak melibatkan estimasi parameter dalam bentuk fungsi yang diasumsikan, seperti bentuk linier yang diasumsikan dalam regresi linier (Bab 6). Sebaliknya, metode ini menarik informasi dari kesamaan antara nilai prediktor record dalam kumpulan data.

Pertanyaan utama adalah bagaimana mengukur jarak antar record berdasarkan nilai prediktornya. Ukuran jarak yang paling populer adalah jarak Euclidean. Jarak Euclidean antara dua record (x1, x2,.., Xp) dan (u1, u2,..., Up) adalah



Anda akan menemukan sejumlah metrik jarak lainnya di Bab 12 dan 15 untuk variabel numerik dan kategori. Bagaimanapun, algoritma k-NN bergantung pada banyak perhitungan jarak (antara setiap record yang akan diprediksi dan setiap record dalam set pelatihan), dan oleh karena itu jarak Euclidean, yang secara komputasi murah, adalah yang paling populer di k-NN.

Untuk menyamakan skala yang mungkin dimiliki oleh berbagai prediktor, perhatikan bahwa dalam banyak kasus, prediktor harus distandarisasi terlebih dahulu sebelum menghitung jarak Euclidean. Perhatikan juga bahwa sarana dan deviasi standar yang digunakan untuk membakukan record baru adalah yang ada pada data pelatihan, dan record baru tidak disertakan dalam penghitungannya. Data validasi, seperti data baru, juga tidak disertakan dalam perhitungan ini.

**Classification Rule**

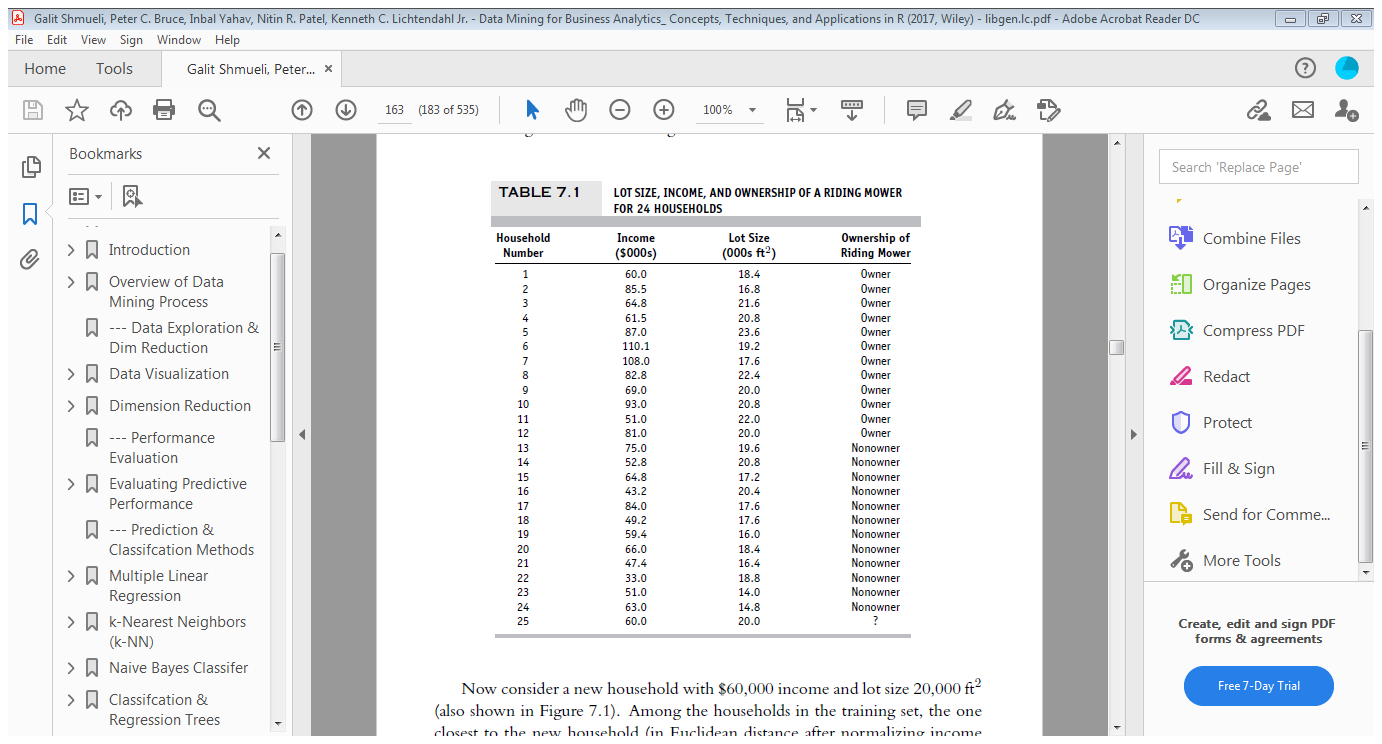
Setelah menghitung jarak antara record yang akan diklasifikasikan dan record yang sudah ada, kita memerlukan aturan untuk menetapkan kelas pada record yang akan diklasifikasikan, berdasarkan kelas tetangganya. Kasus paling sederhana adalah k = 1, di mana kita mencari record yang terdekat (tetangga terdekat) dan klasifikasikan record baru tersebut sebagai milik kelas yang sama dengan tetangga terdekatnya. Merupakan fakta yang luar biasa bahwa ide sederhana dan intuitif dalam menggunakan satu tetangga terdekat untuk mengklasifikasikan record bisa menjadi sangat berguna ketika kita memiliki banyak record dalam set pelatihan kita. Ternyata kesalahan klasifikasi kesalahan skema 1-terdekat tetangga memiliki tingkat kesalahan klasifikasi yang tidak lebih dari dua kali kesalahan ketika kita tahu persis fungsi kepadatan probabilitas untuk setiap kelas.

Gagasan tentang 1-nearest neighbor dapat diperluas ke k> 1 tetangga sebagai berikut:

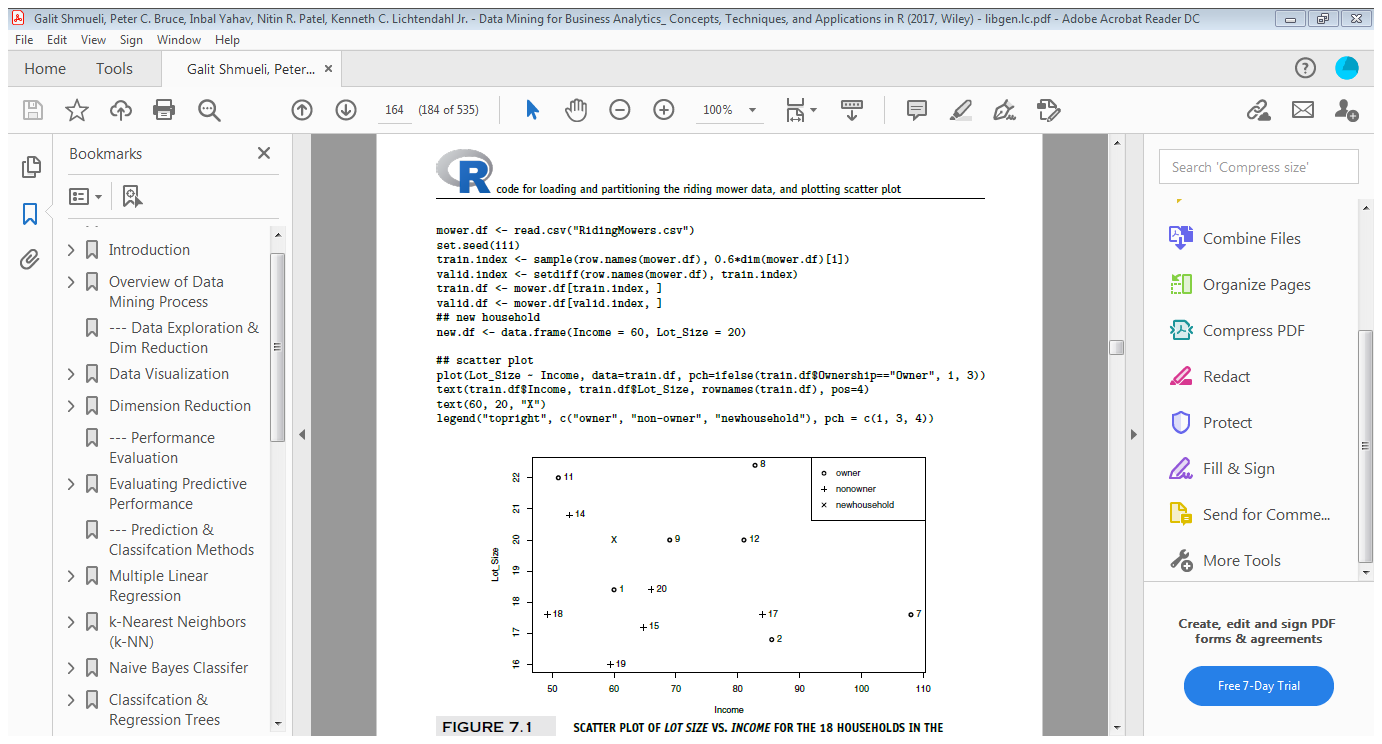
1. Temukan k- nearest neighbors ke catatan untuk diklasifikasikan.

2. Gunakan aturan keputusan mayoritas untuk mengklasifikasikan rekaman, di mana rekaman tersebut diklasifikasikan sebagai anggota kelas mayoritas k tetangga.

Contoh: Mengendarai Mesin Pemotong Produsen mesin pemotong tunggangan ingin menemukan cara untuk mengklasifikasikan keluarga di kota menjadi mereka yang kemungkinan besar akan membeli mesin pemotong tunggangan dan mereka yang kemungkinan tidak akan membelinya. Sampel acak percontohan dilakukan dari 12 pemilik dan 12 bukan pemilik di kota. Data tersebut ditunjukkan pada Tabel 7.1. Kami pertama kali mempartisi data menjadi data pelatihan (14 rumah tangga) dan data validasi (10 rumah tangga). Jelas, kumpulan data ini terlalu kecil untuk dipartisi, yang dapat mengakibatkan hasil yang tidak stabil, tetapi kami akan melanjutkan pemartisian ini untuk tujuan ilustrasi. Plot sebar dari data pelatihan ditunjukkan pada Gambar 7.1.



Sekarang pertimbangkan rumah tangga baru dengan pendapatan $ 60.000 dan ukuran lot 20.000 ft2 (juga ditunjukkan pada Gambar 7.1). Di antara rumah tangga di set pelatihan, yang paling dekat dengan rumah baru (dalam jarak Euclidean setelah menormalkan pendapatan dan ukuran lot) adalah rumah tangga 9, dengan pendapatan $ 69.000 dan ukuran lot 20.000 ft2. Jika kita menggunakan pengklasifikasi 1-NN, kita akan mengklasifikasikan rumah tangga baru sebagai pemilik, seperti rumah tangga 9. Jika kita menggunakan k = 3, maka tiga rumah tangga terdekat adalah 9, 14, dan 1, seperti terlihat secara visual di sebar plot, dan seperti yang dihitung oleh perangkat lunak (lihat keluaran pada Tabel 7.2). Dua dari tetangga ini adalah pemilik mesin pemotong rumput,

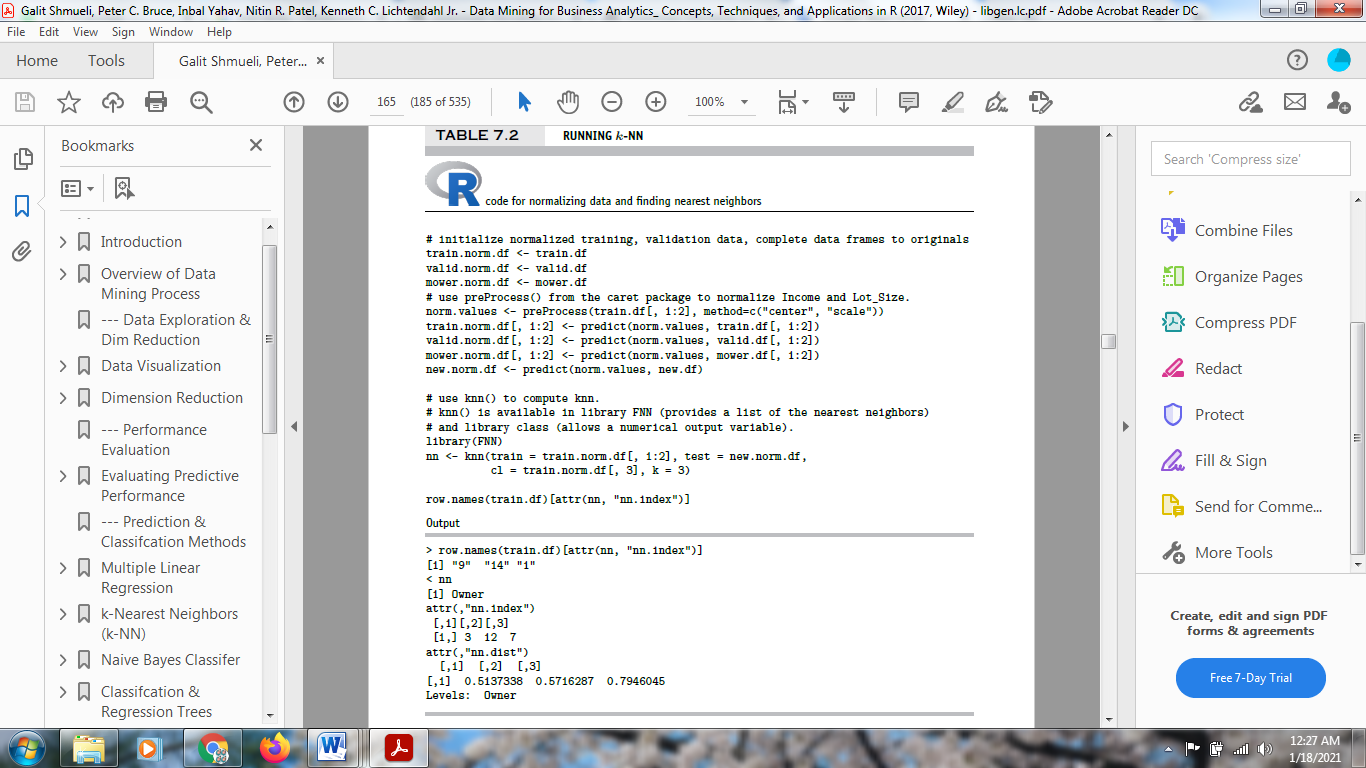


**GAMBAR 7.1 POTONGAN BANYAK UKURAN VS. PENGHASILAN UNTUK 18 HOUSEHOLDS DI SET PELATIHAN DAN RUMAH TANGGA BARU YANG AKAN DIKLASIFIKASI**

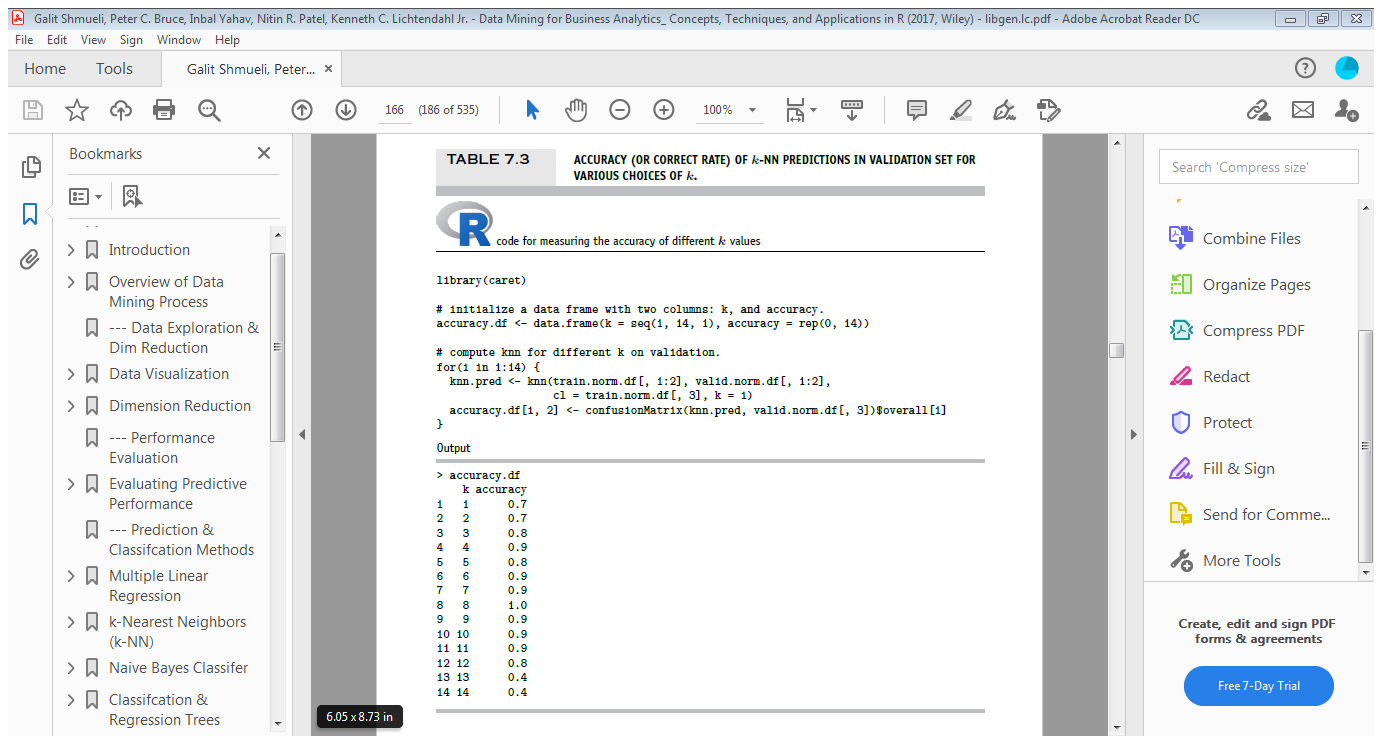
dan yang terakhir adalah bukan pemilik. Oleh karena itu, suara mayoritas adalah pemilik, dan household baru akan diklasifikasikan sebagai pemilik (lihat bagian bawah output pada Tabel 7.2)

**Choosing k**

Keuntungan memilih k> 1 adalah bahwa nilai k yang lebih tinggi memberikan penghalusan yang mengurangi risiko overfitting karena noise dalam data pelatihan. Secara umum, jika k terlalu rendah, kita mungkin cocok dengan derau dalam data. Namun, jika k terlalu tinggi, kita akan kehilangan kemampuan metode untuk menangkap struktur lokal dalam data, salah satu keunggulan utamanya. Secara ekstrim, k = n = jumlah record dalam dataset pelatihan. Dalam hal ini, kami hanya menetapkan



semua rekaman ke kelas mayoritas dalam data pelatihan, terlepas dari nilainya dari (x1; x2;:::; xp), yang bertepatan dengan aturan naif! Ini jelas sebuah kasus kelonggaran dengan tidak adanya informasi yang berguna dalam prediktor tentang keanggotaan kelas. Dengan kata lain, kami ingin menyeimbangkan antara overfitting ke informasi prediktor dan mengabaikan informasi ini sepenuhnya. Pilihan yang seimbang sangat bergantung pada sifat datanya. Semakin kompleks dan tidak beraturan struktur datanya, semakin rendah nilai k yang optimal. Biasanya, nilai k berada dalam rentang 1 sampai 20. Kita akan menggunakan angka ganjil untuk menghindari hubungan. Jadi bagaimana k dipilih? Jawaban: Kami memilih k dengan kinerja klasifikasi terbaik. Kami menggunakan data pelatihan untuk mengklasifikasikan catatan dalam data validasi, kemudian menghitung tingkat kesalahan untuk berbagai pilihan k. Sebagai contoh, jika kita memilih k = 1, kita akan mengklasifikasikan dengan cara yang sangat sensitif terhadap karakteristik lokal dari data pelatihan. Di sisi lain, jika kita memilih yang besar



nilai k, seperti k = 18, kami hanya akan memprediksi kelas yang paling sering dalam dataset dalam semua kasus. Ini adalah prediksi yang sangat stabil tetapi sepenuhnya mengabaikan informasi dalam prediktor. Untuk menemukan keseimbangan, kami memeriksa keakuratan (prediksi dalam set validasi) yang dihasilkan dari pilihan k yang berbeda antara 1 dan 14. Untuk bilangan genap k, jika ada seri dalam pengelompokan rumah tangga, dasi putus secara acak .1 Ini ditunjukkan pada Tabel 7.3. Kami akan memilih k = 4, yang memaksimalkan akurasi kami dalam set validasi.2 Namun, perlu diperhatikan bahwa sekarang set validasi digunakan sebagai bagian dari proses pelatihan (untuk menetapkan k) dan tidak mencerminkan set pisahan yang sebenarnya seperti sebelumnya. Idealnya, kami menginginkan set pengujian ketiga untuk mengevaluasi kinerja metode pada data yang tidak dilihatnya.

**Mengatur Nilai Cut-off**

K-NN memayoritaskan record yang baru untuk membuat keputusan, dimana record diklasifikasikan sebagai mayoritas dari kelas turunan, definisi dari mayoritas yaitu secara langsung ditautkan kedalam nilai Cutoff dan diterapkan ke kelas turunannya. Mari kita pertimbangkan sebuah binary dari hasil studi kasus. Untuk sebuah record baru, proporsikan sebuah kelas diantara turunannya untuk perkiraan dari probability untuk memiliki 1 buah kelas, untuk contohnya dengan *K* = 4 kita temukan kelas terdekat ke kelas 1 maka (didapatkan = $60.000 dan ukurannya = 20.000 ft) dan hasil akhir 9, 14, 1 dan 20. Karena dua diantaranya adalah owner dan dua lainnya adalah no owner. Menggunakan mayoritas yang sederhana sama dengan mengatur nilai Cutoff dari 0,5. Dari tabel 7.4 subuah aplikasi bisa melihat secara acak dari sebuah kelas.

Sebagaiman disebutkan dalam bab 5, mengubah nilai cutoff mempengaruhi sebuah matrix (I.e.., tingkat kesalahan) karena dibeberapa kasus memungkinkan memilih sebuah cutoff lain dengan default 0.5 dengan tujuan untuk memaksimalkan akurasi atau untuk menggabungka biaya kesalahan klasifikasi.

***K-*NN Dengan Lebih Dari 2 Kelas**

Sebuah *K*-NN tergolong mudah diaplikasikan untuk suatu hasil dengan *M* Classes, dimana *M* > 2. Sebuah aturan akan di klasifikasikan sebagai record baru sebagai anggota dari kelas turunannya. Sebuah Alternatif, ketika ada sebuah kelas tertentu dan untuk mengidentifikasinya ( bersedia mengidentifikasi seabagai catatan kelas ini), untuk mengkalkulasi sebuah proporsi dari *K* kelas turunannya yang termasuk kedalam kelas ini, gunakan sebagai estimasi dari sebuah probability dan akan memiliki sebuah kelas baru, dan kemudian untuk spesifikasi dari nilai Cutoff dan menentukan nilai baru dari sebuah kelas. Untuk lebih jelas tentang mengklasifikasikan sebuah nila Cutoff dan penggunaan single class, lihat di bab 5

**Mengkonversi Categorical Variable ke Binary Dummies**

Tidak biasanya untuk menghitung sebuah jarak lebih dari 2 dan bukan kategori numeric (e.g., cookbooks and maps, in a bookstore). Karena itu sebelum bisa menerapkan *K*-NN, sebuah variabel dan lalu dikonversi ke binari dummies. Berbeda dengan model regrasi statistic, semua *M* binary sebaiknya dibuat dengan menggunakan *K*-NN, sementara secara matematis ini adalah redundant, Sejak *M*-1 dummies berisi informasi yang sama sebagia *M* dummies, informasi redundant tidak dibuat sebuah masalah multicollinearity untuk model linear. Bahkan di *K*-NN digunakan untuk *M* dummies bisa menghasilkan klasifikasi dari penggunaan *M* dummies, dan menyebabkan ketidakseimbangan untuk kategori berbeda to sebuah model.

## 7.2 k-NN Untuk Hasil Numerik

Ide dari *K-*NN dapat dengan mudah diperluas untuk memprediksi nilai continuous (dengan tujuan sebagai mutlitple linear model regresi) Langkah peratama untuk menentukan kelas turunan denan komputer yang tidak berubah, Langkah kedua dengan memayoritaskan pilihan dari turunan untuk mementukan kelas, lalu modifikasi untuk mengambil nilai rata-rata dari *K*-nearest turunan untuk menentukan prediksi. Dengan sebuah penurunan dari bertambahnya jarak dari titik dimana prediksis diperlukan, kita bisa membuat sebuah function knn() didalam class ke komputer *K*-NN prediksi numerical untuk mengatur validasi.

Modifikasi lain adalah kesalahan metrik pengguna untuk menenukan nilat terbaik *R* , daripada kesalahan digunakan dalam klasifikasi RMS error or another prediction error metric should be used in prediction (lihat di bab 5)

**Pandora**

Pandora adalah layanan radio music yang memungkinkan pengguna mengubah “statition” lalu memutar musik yang mirip dengan artis yang mereka tentukan, pandora menggunakn sebuah *K*-NN style clustering/classification, proses membatalkan sebuah music dan mencari lagu baru atau menutup lagi seorang artis.

Pandora adalah bagian tim Westergren, yang bekerja sebagai pengasuh ketika dia lulus pada tahun 1980, berasama denan Nolan Gasser dia mempelajari music abad pertengahan dia mengembangkan mesin, dari memasukan data tentang karakteristik dari sebuah lagu kedalam spreadsheet, hasil pertama mengejutkan lagu beatle cocok dengan lagu Bee Gees, lalu ia membuat sebuah perusahaan, hari-hari awal sangat sulit, westergen mempunyai hutang pribadi lebih dari $300.000, dan berakhir dirumah sakit karena penyakit jantung, seorang investor akhirnya mengivestasikan dananya pada tahun 2004 untuk menyelamatkan perusahaan, dan pada 2013 terdaftar dibursa efek.

* Pendora telah menetapkan ratusan variabel untuk mengukur sebuah lagu pada sekala 0-5 dan berikut adalah daftarnya
* Acid Rock Qualities
* Accordion Playing
* Acousti-Lectric Sonority
* Acousti-Synthetic Sonority
* Pendora membayar musisi dan menganalisa puluhan ribu lagu, dan memberi peringkat pada masing-masing lagu, setiap lagu kemudian dikasih nilai dari 0 - 5, misalnya untuk Led Zeppelin's Kashmir

Kashmir 4 0 3 3 … (high on acid rock attributes, no accordion, etc.)

Langkah ini merupakan investasi yang mahal yang ada di pandora karena teknik tersebut telah diuji dan terpilih karena datanya akurat dari sebuah lagu. dan memberikan dasar untuk menentukan preferensi yang sangat individual

* Pengguna yang menggunakan dapat meentukan lagu yan dia suka(lagu yang berada didalam padora)
* Pandora kemudian menghitung statistik lagu yan sedang didengarkan yang ada didalam penyimpanan, lagu yang didengarkan dipilih oleh pengguna
* Pengguna kemudian memiliki pilihan lagu "saya menyukai lagu ini", "saya tidak menyukai lagu ini" atau diam saja
* jika lagu yang dipilih suka, lagu aslinya, ditambah lagu baru yang baru yang kemungkinan disukai
* Jika lagu yang dipilih tidak suka, lagu yang tidak disukai disimpan untuk referensi(jika pengguna tidak mengungkapkan pendapat tentang lagu yang dipilih, contoh sederhananya lagu baru tidak digunakan lagi untuk perbandingan)
* Pandora mencari lagu baru di penyimpanannya, dari statistik lagu yang disukai dan yang tidak disukai, tergantung yang dipilih pengguna, lagu baru akan ditambahkan baik yang disukai atau yang tidak disukai

Pandora mengembangkan kemampuan untuk membawakan lagu yang sesuai dengan selera pengguna, seorang pengguna akan membangun banyak stasiun dari kelompok lagu yang berbeda, ini adalah cara untuk membatasi pengguna memilih jenre lagu.

Proses yang dijelaskan diatas lebih sedikit kompleks dari proses biasanya, proses yang dijelaskan dalam bab ini, adalah pundamental proses untuk menklasifikasikan sebuah record, proses dalam pembelajaran ini yaitu memilih seorang pemimpin, dan pemilhannya dibuat oleh manusia.

Baca lebih lanjut, di [www.pandora.com](http://www.pandora.com/), Wikipedia artikel sebuah musik, genoa project, and Joyce John’s article “Pandora and the Music Genome Project,” Scientific Computing, vol. 23, no. 10: 14, p. 40–41, Sep. 2006.

## 7.3 Keuntungan dan Kekurangan Algoritma k-NN

Keuntungan utama metode k-NN adalah penggunaanya yang sederhana dan kurangnya parameter asumsi yang kurang baik. Dihadapkan pada pelatihan yang cukup besar, metode ini bekerja dengan sangat baik, terutama ketika setiap kelas ditandai oleh banyak kombinasi nilai prediktor. Misalnya, dalam database real estat, ada kecenderungan menjadikan beberapa kombinasi dari {tipe rumah, jumlah kamar,tetangga,harga yang diminta, dll.} yang ditandai rumah yang terjual dengan cepat dibandingkan dengan yang bertahan untuk waktu yang lama di pasar.

Ada tiga kesulitan untuk mengeksploitasi hal-hal yang berkaitan dengan

pendekatan k-NN. Pertama, meskipun tidak diperlukan waktu untuk memperkirakan parameter dari data pelatihan (seperti kasus model parametrik regresi),

waktu untuk menemukan tetangga terdekat dalam set pelatihan yang besar bisa menjadi penghalang.

Keunggulan dan Keuntungan Algoritma k-NN sejumlah ide telah diterapkan untuk mengatasi kesulitan ini. Contohnya adalah :

• Kurangi waktu yang dibutuhkan untuk menghitung jarak dengan bekerja dalam pengurangan dimensi menggunakan teknik reduksi dimensi seperti analisis komponen

• Gunakan struktur data yang canggih seperti pohon pencarian untuk mempercepat identifikasi lokasi tetangga terdekat. Pendekatan ini sering kali diselesaikan dengan "Kecepatan terdekat ”untuk meningkatkan kecepatan. Contohnya adalah menggunakan bucketing, dimana record dikelompokkan ke dalam bucket sehingga record di dalam setiap bucket bisa dekat satu sama lain. Untuk catatan yang akan diprediksi, dimulai dengan jarak yang terdekat. Mulai dari ember terdekat, jarak record dengan bucket diukur. Algoritma berhenti jika jarak ke ember lebih besar dari jarak ke record.

Kedua, jumlah record yang dibutuhkan dalam pelatihan agar memenuhi syarat sebagai cara untuk meningkatkatkan secara eksponensial dengan jumlah prediktor p. Ini karena jarak yang diharapkan ke tetangga terdekat naik secara drastis dengan p kecuali ukuran pelatihan meningkat secara eksponensial dengan p. Fenomena ini dikenal sebagai kutukan dimensionalitas, masalah mendasar yang berkaitan dengan semua kelas teknik sification, prediksi, dan clustering. Inilah sebabnya mengapa kami sering mencarinya mengurangi jumlah prediktor melalui metode seperti memilih subset dari prediktor untuk model ini atau dengan menggabungkannya menggunakan metode seperti principal analis, dekomposisi nilai singular, dan analisisis faktor (lihat Bab 4).

Ketiga, k-NN adalah "Pelajaran yang membosankan": Pencatatan prediksi yang banyak menghabiskan waktu. Untuk setiap catatan yang akan diprediksi, kami menghitung jaraknya dari seluruh rangkaian catatan pelatihan hanya pada saat prediksi. Perilaku ini melarang penggunaan algoritma ini untuk prediksi real-time yang besar dengan jumlah record secara bersamaan.

**PROBLEMS**

**7.1 Menghitung Jarak Dengan Kategori Prediksi.**

Latihan ini dengan sangat mudahmenggambarkan perhitungan jarak Euclidean, dan pembuatan bilangan biner. Perusahaan pendidikan online Statistics.com menyegmentasikan pelanggannya dengan prospek menjadi tiga kategori utama: profesional TI (TI), ahli statistik (Stat), dan lainnya(Lain). Hal Ini juga digunakan untuk mengetahui, jumlah setiap pelanggan, dan jumlah tahun sejak kontrak pertama(tahun). Pertimbangkan pelanggan berikut; informasi tentang apakah mereka telah mengambil

kursus atau tidak (hasil yang akan diprediksi) termasuk:

Pelanggan 1: ahli statistik, 1 tahun, tidak ikut kursus

Pelanggan 2: Lainnya, 1,1 tahun, mengambil kursus

**a.** Simak sekarang prospek baru berikut ini:

Prospek 1: IT, 1 tahun

Dengan menggunakan informasi di atas tentang dua pelanggan dan satu prospek, buat satu data untuk ketiganya dengan menggunakan variabel prediktor kategorikal diubah menjadi 2 bilangan, dan kumpulan data serupa dengan variabel prediktor kategorikal diubah menjadi 3 bilangan.

**b**. Untuk setiap set data turunan, hitung jarak Euclidean antara prospek dan

masing-masing dari dua pelanggan lainnya. (Catatan: meskipun normalisasi data untuk k-NN, ini bukan aturan berlapis besi dan Anda dapat melanjutkan di sini tanpa normalisasi.)

**c.** Menggunakan k-NN dengan k = 1, klasifikasikan prospek sebagai mengambil atau tidak mengambil kursus masing-masing dari dua kumpulan data turunan. Apakah ada bedanya ketika Anda menggunakan 2 atau 3 bilangan?

**7.2 Penerimaan Pinjaman Pribadi**.

Bank Universal adalah bank yang relatif baru berkembang cepat dalam hal akuisisi pelanggan secara keseluruhan. Mayoritas pelanggan ini adalah (deposan) dengan berbagai macam hubungan dengan bank. Itu basis pelanggan aset pelanggan (peminjam) cukup kecil, dan bank tertarik

dalam memperluas basis ini dengan cepat untuk menghasilkan lebih banyak bisnis pinjaman. Secara khusus, ia menginginkan untuk mencari cara untuk mengubah nasabah kewajibannya menjadi nasabah pinjaman pribadi (sementara

mempertahankan mereka sebagai deposan).

Promosi yang dijalankan bank tahun lalu untuk nasabah kewajiban menunjukkan hasil yang baik tingkat konversi lebih dari 9% sukses. Ini telah mendorong departemen pemasaran ritel management untuk merancang Promosi yang lebih cerdas dengan target pemasaran yang lebih baik. Tujuannya adalah untuk menggunakan k-NN untuk memprediksi apakah pelanggan baru akan menerima tawaran pinjaman. Ini akan berfungsi sebagai dasar untuk desain Promosi baru.

Data UniversalBank.csv berisi data tentang 5000 pelanggan. Datanya termasuk informasi demografis pelanggan (usia, pendapatan, dll.), hubungan pelanggan dengan bank (hipotek, rekening sekuritas, dll.), dan tanggapan pelanggan terhadap promosi pinjaman pribadi terakhir (Personal Loan). Di antara 5.000 pelanggan ini, hanya 480 (= 9,6%) menerima pinjaman pribadi yang ditawarkan kepada mereka di promosi sebelumnya.

Partisi data pelatihan (60%) dan validasi (40%).

**a**.Pertimbangkan pelanggan berikut:

Usia = 40, Pengalaman = 10, Pendapatan = 84, Keluarga = 2, CCAvg = 2, Pendidikan\_1 = 0, Pendidikan\_2 = 1, Pendidikan\_3 = 0, Hipotek = 0, Rekening Efek = 0, Rekening CD = 0, Online = 1, dan Kartu Kredit = 1. Lakukan klasifikasi k-NN

dengan semua prediktor kecuali ID dan kode ZIP menggunakan k = 1. Ingatlah untuk mengubah prediktor kategori dengan lebih dari dua kategori menjadi bilangan variabel terlebih dahulu.

Tentukan kelas sukses sebagai 1 (penerimaan pinjaman), dan gunakan nilai batas default 0,5. Bagaimana pelanggan ini diklasifikasikan?

**b**. Apa pilihan yang menyeimbangkan antara overfitting dan mengabaikan prediktor

informasi?

**c**. Tunjukkan matriks yang berhubungan untuk data validasi yang dihasilkan dari penggunaan terbaik.

**d**. Pertimbangkan pelanggan berikut ini: Usia = 40, Pengalaman = 10, Pendapatan = 84, Keluarga = 2, CCAvg = 2, Pendidikan\_1 = 0, Pendidikan\_2 = 1, Pendidikan\_3 = 0, Hipotek = 0, Rekening Efek = 0, Rekening CD = 0, Online = 1 dan Kredit Kartu = 1. Klasifikasikan pelanggan terbaik.

**e**. Partisi ulang data, kali ini menjadi set pelatihan, validasi, dan pengujian (50% : 30% : 20%). Terapkan metode k-NN dengan yang dipilih di atas. Bandingkan kekurangannya matriks pengujian dengan pelatihan dan validasi. Berikan komentar

perbedaan dan alasannya

**7.3 Memprediksi Harga Rata-Rata Perumahan**.

File BostonHousing.csv berisi informasi perumahan lebih dari 500 bidang sensus di Boston, di mana untuk setiap saluran memiliki banyak variabel record. Kolom terakhir (CAT.MEDV) diturunkan dari MEDV, sedemikian rupa mendapatkan nilai 1 jika MEDV> 30 dan 0 sebaliknya. Pertimbangkan tujuan memprediksi nilai median (MEDV) dari sebuah bidang, berdasarkan informasi dari 12 kolom pertama.

Partisi data menjadi set pelatihan (60%) dan validasi (40%).

**a**. Lakukan prediksi k-NN dengan semua 12 prediktor (abaikan kolom CAT.MEDV

umn), mencoba nilai dari 1 sampai 5. Pastikan untuk menormalkan data, dan pilih

fungsi knn () dari kelas paket daripada paket FNN. Untuk memastikan R adalah

menggunakan paket kelas (ketika kedua paket dimuat), gunakan class :: knn ().

**b**. Prediksi MEDV untuk menjelaskan dengan informasi berikut, menggunakan K terbaik:

**c**. Jika kita menggunakan algoritma k-NN di atas untuk menilai data pelatihan, apa akan terjadi kesalahan dalam pelatihan?

**d**. Mengapa kesalahan data validasi terlalu optimis dibandingkan dengan tingkat kesalahan saat menerapkan prediktor k-NN ini ke data baru?

**e.** Jika tujuannya adalah untuk memprediksi MEDV untuk beberapa ribu traktat baru, apakah akan terjadi kerugian menggunakan prediksi k-NN? Buat daftar operasi algoritma berjalan untuk menghasilkan setiap prediksi.

|  |  |  |
| --- | --- | --- |
|  |  |  |