ANALISIS FAKTOR- FAKTOR YANG MEMPENGARUHI SESEORANG TERKENA PENYAKIT DIABETES MELITUS MENGGUNAKAN REGRESI RANDOM FOREST

(Studi Kasus : Data Diabetes di Virginia Amerika Serikat)

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana
Jurusan Statistika

Disusun Oleh:
Maulida Jannati Wulansari
(14611211)

PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS ISLAM INDONESIA
YOGYAKARTA
2018

HALAMAN PERSETUJUAN DOSEN PEMBIMBING

TUGAS AKHIR

Judul	:	Analisis Faktor- Faktor yang Mempengaruhi Seseorang Terkena Penyakit Diabetes Melitus Menggunakan Regresi Random Forest
Nama Mahasiswa	:	Maulida Jannati Wulansari
Nomor Mahasiswa	:	14 611 211

TUGAS AKHIR INI TELAH DIPERIKSA DAN DISETUJUI UNTUK DIUJIKAN

Yogyakarta, Mei 2018

Pembimbing

(Ayundyah Kesumawati, S.Si., M.Si.)

HALAMAN PENGESAHAN

TUGAS AKHIR

ANALISIS FAKTOR-FAKTOR YANG MEMPENGARUHI SESEORANG TERKENA PENYAKIT DIABETES MELITUS MENGGUNAKAN REGRESI RANDOM FOREST

Nama Mahasiswa : Maulida Jannati Wulansari

Nomor Mahasiswa : 146112111

TUGAS AKHIR INI TELAH DIUJIKAN

PADA TANGGAL, 15 Mei 2018

Nama Penguji	**Tanda Tangan**
1. Saepudin, M.Si., Ph.D., Apt	.....
2. Tuti Purwaningsih, S.Stat, M.Si	.....
3. Ayundyah Kesumawati, S.Si., M.Si	.....

Mengetahui,

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

(Drs. Allwar, M.Sc., Ph.D)

KATA PENGANTAR

Assalamualaikum Wr. Wb

Puji Syukur Kehadirat Allah SWT yang telah melimpahkan rahmat dan hidayahnya sehingga tugas akhir yang berjudul "Analisis Faktor- Faktor yang Mempengaruhi Seseorang Terkena Penyakit Diabetes Melitus Menggunakan Regresi Random Forest" dapat diselesaikan. Shalawat serta salam semoga selalu tercurah kepada junjungan Nabi Besar Muhammad SAW serta para sahabat dan pengikutnya sampai akhir jaman.

Tugas akhir ini disusun sebagai salah satu persyaratan yang harus dipenuhi dalam menyelesaikan jenjang Strata Satu atau S1 di Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia. Penyelesaian tugas akhir ini tidak terlepas dari bantuan, arahan, dan bimbingan dari berbagai pihak. Untuk itu Pada kesempatan ini penulis ingin mengucapkan terima kasih yang sebesar-besarnya kepada :

1. Allah SWT, atas kekuatan dan kemudahan yang telah diberikan, terima kasih ya Allah atas semua nikmat dan karunia yang telah Engkau Berikan.
2. Bapak Drs. Allwar, M.Sc., Ph.D selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Islam Indonesia.
3. Bapak Dr. Fajriya Hakim, S.Si., M.Si selaku Ketua Jurusan Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Islam Indonesia.
4. Ibu Ayundyah Kesumawati, S.Si., M.Si selaku Dosen Pembimbing penulis yang telah memberikan kritik dan saran bimbingan maupun arahan yang sangat berguna dalam penyusunan Tugas Akhir ini.
5. Kepada Kedua Orang Penulis, Bapak Sadali dan Ibu Supriyati yang selalu memberikan do'a, ridho, membimbing dan memberi semangat kepada penulis. Terima Kasih atas semua yang bapak dan Ibu berikan.

6. Kakak Anisa Nur Nia Rahmah sebagai saudara perempuan yang senantiasa memberikan semangat kepada penulis, serta adik-adik Muhammad Nur Abdurahman Wahid, Muhammad Aji Nurul Yaqin dan Muhammad Wahyu Wijaya yang selalu memberikan semangat dan do'a kepada penulis.
7. Teman- Teman satu bimbingan tugas akhir Siti Rahmi Kurniasari, Erdwika Putri, Suci Yulianawati, Herlina Catur Sulistya Ningrum, Riza Indriani Rakhmalia, Afifah Mukhtaroh, Reny Roswita Nazar, Zia Ayu Nuansa Gumilang, Molydah S., Gustiara Dayu Amrinda, Elsa Murni Nasution, Dian Purnama Sari, Purwina Kowala, Galih Rahmatdona Sugito, Yayan Dwi Septian, Achmad Kurniansyah Thalib, dan Syauqi Amri Yahya yang senantiasa selalu berbagi ilmu, cerita dan pengalaman.
8. Sahabat – sahabat yang sangat luar biasa hebat dan menjadi teman seperjuangan sekaligus keluarga sendiri, Rabiatul Adawiyah, Siti Rahmi Kurniasari, Erdwika Putri, Reny Roswita Nazar, Zia Ayu Nuansa Gumilang, Annisa Al Wasi'a, Khusnul Hajar Nuansari, Dhea Andryani Awitasari, Ditia Yosmita Praptiwi serta een juliani yang selalu memberikan dukungan, tempat berkeluh kesah, dan doa.
9. Sahabat Jalan- jalan , belajar dan juga bercerita tentang berbagai hal , Suri Islamiah dan juga Suci Yulianawati yang selalu memberikan semangat, doa dan juga kebahagiaan.
10. Amry Wicaksana dan Wahyu Hanna Rahmana yang senantiasa memberikan semangat dan juga Doa.
11. Mantan Penghuni Kos Hartono, Yuli Widyastuti, Avidah Amalia Zahro, Salsabila, Teja Arum, Eka S, dan Ana yang tidak pernah lupa memberikan dukungan dan Doa.
12. Anak- anak anti wacana magang telkomsel, Una, Adel, Dewa, Ofa, Ovi, Iqbal, Haris, Erra, Diaz, Rimud, Rizang, Hafiz, Regina, Tara, Deny, Febrianing, dan Vanisya Ramadhani yang selalu mengajak jalan untuk merefresh otak dan juga semangat dan doa.

13. Sahabat Statistika 2014 yang selama ini memberikan dukungan kepada penulis untuk menyelesaikan tugas akhir
14. Teman – Teman KKN , Riska Khairunnisa, Bella Sanyta Artina, Maghfira, Dana Aprilia, Afifah Nur Fauziah, Ryo Rizky Ananda, Rian Dwi Putraa, M. Khoiril Yusron, Azwar Hanik.
15. Sahabat dari SMA , Atika Damayanti, Nur Lailatul Mufidah, dan Nur Muslimah yang tidak henti- hentinya memberikan support dan juga doa kepada penulis untuk menyelesaikan tugas akhir.
16. Semua Pihak yang terlibat dalam pembuatan Tugas Akhir ini yang tidak bisa penulis sebutkan satu per satu.

Penulis Menyadari bahwa tugas akhir masih jauh dari kesempurnaan, oleh karenanya segala kritik dan saran yang bersifat membangun selalu penulis harapkan. Semoga Tugas Akhir ini dapat bermanfaat bagi penulis khususnya dan bagi yang membutuhkan umumnya. Sekian, semoga Allah SWT selalu melimpahkan karunia serta Hidayah-NYA kepada kita semua. Aamiin Ya Rabbal Alamin

Wassalamu'alaikum Warahmatullahi Wabarakaatuh.

Yogyakarta, April 2018

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Tugas Akhir ini tidak terdapat karya yang sebelumnya pernah diajukan untuk memperoleh gelar kesarjanaan di suatu perguruan tinggi dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang di acu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, April 2018

Penulis

Analisis Faktor-Faktor Yang Mempengaruhi Seseorang Terkena Penyakit Diabetes Melitus Menggunakan Regresi Random Forest

Maulida Jannati Wulansari
Program Studi Statistika Fakultas MIPA
Universitas Islam Indonesia

INTISARI

Diabetes Melitus Merupakan Penyakit yang disebabkan karena tubuh tidak mampu mengubah glukosa (gula) menjadi energi. Prevalensi secara global Penyakit Diabetes Melitus meningkat dari 4,7% menjadi 8,5% dua kali lipat sejak tahun 1980. Tingginya Penderita diabetes melitus disebabkan oleh beberapa faktor diantaranya keturunan, pola makan dan gaya hidup. Penelitian ini bertujuan untuk mengetahui faktor-faktor yang mempengaruhi seseorang terkena penyakit diabetes melitus dengan menggunakan regresi random Forest. Dalam penelitian ini Regresi Random Forest digunakan untuk melihat hubungan variabel dependen dengan variabel independen, variabel dependen adalah glikogen hemoglobin dengan variabel independen adalah kolesterol, Gula Darah, lemak baik, Usia, Berat Badan, Tinggi Badan, Tekanan Darah Sistol, Tekanan Darah Diastol, Lingkar Pinggang dan Lingkar Pinggul. Berdasarkan analisis regresi random forest didapatkan hasil bahwa banyaknya pohon yang terpilih adalah 500 dengan *m* yang dicobakan adalah 5 dan variabel yang paling berpengaruh seseorang terkena penyakit diabetes adalah gula darah 2 jam setelah makan dengan nilai kepentingan sebesar 57.56 paling tinggi diantara variabel independen yang lainnya dan juga nilai MSE dari hasil prediksi sebesar 0.456015.

Kata Kunci : Diabetes Melitus, Regresi Random Forest, Gula Darah

An Analysis of The Factors Which Affects Someone with Diabetes Mellitus Disease Using Random Forest Regression

Maulida Jannati Wulansari
Statistics Study Program at Faculty of Mathematics and Natural Sciences
Islamic University of Indonesia

ABSTRACT

Diabetes Mellitus is a disease caused by an inability of body to convert glucose (sugar) into energy. Prevalence Global of Diabetes Mellitus disease increased from 4.7% to 8.5% doubled since 1980. High Diabetes mellitus is caused by several factors including heredity, diet and lifestyle. This study aims to identify factors which affects somebody with diabetes disease by using Random Forest regression. Random Forest regression is applied to examine the association of dependent variables with independent variables. The dependent variables are hemoglobin glycogen while independent variables are cholesterol, Blood Sugar, good fats, Age, Weight, Height, Systolic Blood Pressure, Diastolic Blood Pressure, Waist Circumference and Hip Circle. The result analysis shows that the number of trees selected is 500, with m used for trial is 5. Additionally, the most influential variable on a person affected by diabetes is blood sugar two hours after eating with the importance value of 57.56, the highest among the other independent variables, and also the MSE value from the result of prediction is 0.456015.

Keywords: Diabetes Mellitus, Random Forest Regression, Blood Sugar

BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Diabetes Melitus (DM) merupakan penyakit yang disebabkan karena tubuh tidak mampu mengubah glukosa (gula) menjadi energi. Gula adalah sumber utama bahan bakar untuk tubuh ketika makanan dicerna, glukosa akan diubah menjadi lemak, protein dan karbohidrat. Badan Kesehatan Dunia (WHO) menyebutkan 422 juta orang dewasa di seluruh dunia hidup dengan diabetes. Jumlah itu meningkat empat kali lipat sejak 1980, dan sebagian besar penderita diabetes tersebut hidup di India, Cina, Amerika Serikat, Brasil dan Indonesia. (Avin, 2016)

Di Indonesia, WHO menyebutkan jumlah pasien DM pada tahun 1995 berjumlah 5 juta orang dan akan meningkat menjadi 25 juta orang pada tahun 2025. Selain itu, Perkeni (perkumpulan Endokrinologi Indonesia) memperkirakan pada tahun 2020 akan terdapat 178 juta orang terkena DM. DM berusia di atas 20 tahun berjumlah 7 juta orang dengan prevalensi DM sebesar 5 %. Peningkatan ini lebih disebabkan oleh pola makan yang tidak sehat, kurangnya aktivitas fisik serta meningkatnya harapan hidup.

Estimasi Terakhir Menurut International Diabetes Federation (IDF) terdapat 382 juta orang hidup dengan diabetes pada tahun 2013. Pada tahun 2013 Jumlah tersebut diperkirakan akan meningkat menjadi 592 juta orang. Diperkirakan dari 382 juta orang tersebut, 175 juta diantaranya belum terdiagnosis, sehingga terancam berkembang progresif menjadi komplikasi tanpa disadari dan tanpa pencegahan. (Infodatin, 2014)

Amerika Serikat sebagai negara maju menempati posisi pertama untuk penderita diabetes. Amerika Serikat memiliki Prevalansi tertinggi diantara semua negara maju diseluruh dunia, berdasarkan IDF hampir 11 persen warga Amerika Serikat yang berusia antara 20 sampai 79 tahun mengidap diabetes, yaitu sekitar 30 juta orang dewasa di seluruh AS. Jumlah Penderita diabetes tersebut menjadi

masalah besar di AS. Para ahli mengatakan tingginya jumlah pengidap diabetes di AS sebagian besar adalah pengidap diabetes tipe 2. (CNN Indonesia, 2015)

Penelitian terkait diabetes telah dilakukan oleh National Statistical Report pada tahun 2017, dalam penelitian tersebut disebutkan bahwa jumlah penderita diabetes di Amerika Serikat adalah 30,3 juta atau 9,4 % dari total populasi di Amerika Serikat dengan 23,1 juta orang diantaranya terdiagnosis dan 7,2 juta atau 23,8% tidak terdiagnosa diabetes. (Centers For Disease Control and Prevention, 2017)

Penelitian terkait diabetes telah dilakukan oleh WHO, dalam penelitian tersebut ditemukan bahwa secara global, diperkirakan 422 juta orang dewasa hidup dengan diabetes pada tahun 2014, dibandingkan dengan 108 juta pada tahun 1980. Prevalensi global (usia standar) diabetes telah hampir dua kali lipat sejak tahun 1980, meningkat dari 4,7% menjadi 8,5% dalam populasi orang dewasa. Ini mencerminkan peningkatan faktor risiko terkait seperti kelebihan berat badan. Selama dekade terakhir, prevalensi diabetes telah meningkat lebih cepat di negara-negara berpenghasilan rendah dan menengah daripada di negara-negara berpenghasilan tinggi. Diabetes menyebabkan 1,5 juta kematian pada tahun 2012. Glukosa darah yang lebih tinggi dari yang optimal menyebabkan tambahan 2,2 juta kematian, dengan meningkatkan risiko penyakit kardiovaskular dan lainnya. Sebesar 43% dari total 3,7 juta kematian ini terjadi sebelum usia 70 tahun. Persentase kematian disebabkan tinggi glukosa darah atau diabetes yang terjadi sebelum usia 70 lebih tinggi di negara-negara berpenghasilan rendah dan menengah daripada di negara-negara berpenghasilan tinggi.

Berdasarkan penjelasan diatas didapatkan bahwa penderita diabetes di Amerika Serikat memiliki prevelansi tertinggi sehingga perlu untuk diketahui faktor- faktor yang mempengaruhi seseorang terkena diabetes melitus. Dengan Menggunakan Regresi Random Forest yang digunakan untuk melihat hubungan variabel prediktor dengan variabel penjelas dan didalam regresi random forest juga terdapat nilai kepentingan masing- masing variabel penelitian sehingga bisa dilihat faktor – faktor yang paling berpengaruh seseorang terkena penyakit diabetes.

1.2 Rumusan Masalah

Berdasarkan uraian pada latar belakang permasalahan diatas didapatkan beberapa rumusan masalah yang akan dibahas dalam penelitian ini yaitu :

1. Bagaimana Statistika Deskriptif Profilisasi Penderita Diabetes Melitus dan Non Diabetes Melitus ?
2. Apakah Faktor – faktor yang mempengaruhi seseorang Terkena Penyakit Diabetes Melitus Berdasarkan Regresi Random Forest?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk menjawab rumusan masalah, yaitu :

1. Mendapatkan Statistika Deskriptif dari Profil Penderita Diabetes Melitus dan Non Diabetes Melitus.
2. Mengetahui Faktor- faktor yang mempengaruhi seseorang Terkena Penyakit Diabetes Melitus Berdasarkan Regresi Random Forest.

1.4 Manfaat Penelitian

Manfaat Penelitian ini adalah :

1. Mengimplementasikan Regresi Random Forest untuk mengetahui Faktor – faktor yang mempengaruhi Seseorang Terkena Penyakit Diabetes Melitus.
2. Memperluas Wawasan dan Ilmu terutama teknik penyelesaian regresi dengan cara yang berbeda dengan biasanya yaitu dengan Regresi Random Forest.
3. Mengetahui Faktor yang paling mempengaruhi seseorang terkena penyakit Diabetes Melitus.

1.5 Batasan Penelitian

Mengingat luasnya ruang lingkup penelitian dalam implementasi teknik - teknik *Algoritma*, maka penelitian ini dibatasi pada :

1. Sumber data untuk penelitian ini, diperoleh dari Website Kaggle.
2. Pendekatan dalam analisis data dalam penelitian ini akan menggunakan metode Regresi Random Forest Untuk Mengetahui Faktor- Faktor Terkena Seseorang Terkena Penyakit Diabetes Melitus.

3. Variabel yang digunakan dalam penelitian ini adalah Kolesterol, Gula Darah, Lemak Baik, Glikogen Hemoglobin, Umur, Berat Badan, Tinggi Badan, Tekanan Darah Sistol, Tekanan Darah Diastol, Lingkar Pinggang dan Lingkar Pinggul.
4. Jumlah pohon atau *ntree* yang digunakan adalah 500 Pohon.
5. Untuk mendukung analisis data dalam penelitian ini akan menggunakan bantuan Microsoft excel, dan *softwar*e R. 3.2.3

BAB II
TINJAUAN PUSTAKA

Penelitian tentang Random Forest sudah banyak dilakukan oleh Peneliti-peneliti sebelumnya, Seperti Penelitian dengan Judul "Prediksi Lama Studi Mahasiswa Dengan Metode Random Forest" (Adnyana, 2015) . Penelitian Ini Menunjukkan bahwa klasifikasi Random Forest Ini telah diaplikasikan di bidang pendidikan di Perguruan tinggi dalam hal ini mahasiswa berpotensi no-aktif. Dari hasil penelitian itu menunjukkan nilai *producer accuracy* untuk kelas diperoleh nilai sebesar 81.81% dan kelas " Lulus Lewat Waktu Studi" diperoleh nilai *procedur accuracy* sebesar 84,98%. Sedangkan *overall accuracy* yang didapatkan adalah 83.54% (kappa Coefficient 0.564).

Penelitian Lain tentang Random Forest dengan judul " Penerapan Metode Random Forest Dalam Driver Analysis" (Dewi,dkk). Penelitian ini menunjukkan bahwa Random Forest dapat diaplikasikan dalam bidang biostatistika. Dari hasil penelitian ini menunjukkan penyusunan *driver analysis* berdasarkan MDG menghasilkan *driver analysis* yang stabil jika ukuran Random Forest lebih dari 500. Untuk penyususnan *driver analysis* berdasarkan rataan MDG dari 1000 *driver analysis* cukup stabil meskipun menggunakan Random Forest dengan ukuran yang kecil. Hasil analisis juga stabil pada ukuran pada contoh peubah penjelas.

Penelitian lain tentang Regresi Random Forest dengan judul "Penyelesaian Regresi Semiparametrik Menggunakan Regresi Random Forest" (Firmani, 2016). Penelitian ini menunjukkan bahwa Random Forest dapat digunakan untuk menyelesaikan regresi. Dari Hasil Penelitian ini menunjukkan bahwa estimasi presentasse kemiskinan dan variabel yang paling berpengaruh terhadap presentase kemiskinan adalah angka melek huruf, MSE yang dihasilkan adalah 0.000308.

Penelitian lain tentang diabetes melitus dengan judul "Faktor- Faktor yang berhubungan dengan penyakit diabetes melitus Daerah Perkotaan di Indonesia Tahun 2007" (Wahyuni, 2010). Penelitian ini menunjukkan bahwa diketahui faktor yang paling dominan mempengaruhi penyakit diabetes melitus apda

penduduk perkotaan di Indonesia tahun 2007 secara berururutan adalah obesitas, pekerjaan, hipertensi, umur, konsumsi kafein dan konsumsi alkohol. Berikut ini adalah tabel hasil penelitian yang telah dilakukan, disajikan pada **tabel 2.1**

Tabel 2.1 Tabel Hasil Penelitian Sebelumnya

Nama Peneliti	Judul	Variabel Yang digunakan
Silvia Ikmalia Fernanda dkk	Identifikasi Penyakit Diabetes Melitus Menggunakan Metode *Modified K-Nearest Neighbor* (MKNN)	- Nafsu Makan Meningkat - Sering buang air kecil - Peningkatan kehausan - Turunnya berat badan - Usia (15-20) - Faktor Keturunan - Mulut Kering - Mudah Kelelahan/ Kurangnya aktivitas fisik - Sering mengantuk - Mual/Muntah- Muntah - Timbulnya luka yang tak kunjung sembuh - Gatal-gatal - Mengonsumsi makanan berkolesterol tinggi - Obesitas - Kadar Glukosa Darah Meningkat
Tahani Daghistani dan Riyad Alshammari	*Diagnosis Of Diabetes by Applying Data Mining Classification Techiques*	- Jenis kelamin, Usia dan Wilayah sebagai demografi - Pengukuran pasien seperti BMI dan tekanan darah dan 11 tes laboratorium
Aiswarya Iyer, S. Jelayalatha dan Ronak	*Diagnosis of Diabetes Using Classification Mining Techniques*	- Konsentrasi Glukosa Plasma - Indeks Massa tubuh (kg/m2) - Fungsi Pedigree Diabetes

Nama Peneliti	Judul	Variabel Yang digunakan
Sumbaly		- Umur (Tahun) - Variabel Kelas
Rizky Adhi Nugroho, Tarno dan Alan Prahutama	Klasifikasi Pasien Diabetes Melitus Menggunakan Metode *Smooth Support Vector Machine* (SSVM)	- Diabetes - Jenis Kelamin - Usia Pasien - Glukosa darah puasa (mg/dL) - Glukosa darah 2 jam (mg/dL) - HDL (mg/dL) - LDL (mg/dL) - Trigliserida (mg/dL) - HbA1c
I Made Budi Adyana	Prediksi Lama Studi Mahasiswa Dengan Metode Random Forest	- IPK (pada semester VI) - Jumlah total SKS (pada semester VI) - Jumlah tidak aktif dan cuti (sampai semester VI) - Jumlah matakuliah dengan nilai buruk (nilai D dan E) e) - Jumlah matakuliah dengan nilai bagus (nilai A dan B)
Laras Binarwati, Imam Mukhlas dan Soetrisno	Implementasi Algoritma Genetika untuk Optimalisasi Random Forest Dalam Proses Klasifikasi Penerimaan Tenaga Kerja Baru: Studi Kasus PT.XYZ	- Variabel yang dipakai sebanyak 4 macam yang berasal dari faktor internal karyawan dan sistem perusahaan itu sendiri.
Ulla B.	*Evaluating Random*	- Jenis Kelamin

Nama Peneliti	Judul	Variabel Yang digunakan
Mogensen, Hemant Ishwaran dan Thomas A. Gerds	*Forests for Survival Analysis Using Prediction Error Curves*	- Tekanan Darah - Penyakit jantung isekemik - Stroke sebelumnya - Penyakit Penonaktifan Lainnya - Diabetes - Status Merokok - Atrial brillation - Hemoragi - Skor stroke Skandinavia - Usia - Kolesterol

Pada Penelitian kali ini penulis mengambil judul "Analisis Faktor- faktor yang mempengaruhi Seseorang Terkena Diabetes Melitus menggunakan Regresi Random Forest". Penelitian terkait diabetes dengan menggunakan metode regresi Random Forest hampir jarang ditemukan. Pada Penelitian ini diharapkan dapat bermanfaat untuk khalayak umum guna mengetahui faktor-faktor apa saja yang mempengaruhi seseorang terkena penyakit Diabetes Melitus.

BAB III
LANDASAN TEORI

3.1 Diabetes Melitus

Dalam buku Hartini (2009) mendefinisikan Diabetes Melitus adalah Penyakit yang ditandai oleh tingginya kadar gula dalam darah. Pada Dasarnya, hal ini terjadi karena tubuh "kekurangan" hormon insulin. Hormon insulin adalah zat yang diproduksi oleh kelenjar pankreas. Kekurangan disini bisa berupa jumlah insulin yang memang kurang atau jumlahnya cukup tetapi kerjanya kurang baik. Pankreas merupakan satu organ di dalam tubuh dengan tugas menjaga kadar gula darah selalu dalam batas aman. Gula yang melebihi batas normal akan meracuni dan mengganggu mesin kehidupan pada umumnya. Oleh karena itu, didalam darah kadar gula selalu fluktuatif bergantung pada asupan makanan. Kadar paling tinggi tercapai pada 1 jam sesudah makan. Satu jam setelah makan, gula di dalam darah akan mencapai kadar paling tinggi, normalnya tidak akan melebihi 180mg per cc darah (=180mg/dl). Kadar 180mg/dl disebut nilai ambang ginjal. Ginjal, tempat pembuatan urine, hanya dapat menahan gula kalau kadarnya hanya sampai angka tersebut. Lebih tinggi dari itu, ginjal tidak akan menahan gula dan kelebihan gula akan keluar bersama urine. Insulin akan menurunkan gula darah dengan cara mendistribusikan gula masuk ke dalam sel-sel, yang akan diolah lebih lanjut menjadi energi. Selama pankreas sehat, semua akan berjalan dengan lancar tanpa kelainan.

3.2 Tipe- Tipe Diabetes Melitus

Menurut Hartini (2009) Penyakit Diabetes dibagi menjadi beberapa golongan atau tipe, berikut ini adalah tipe- tipenya :

1. Diabetes Tipe – 1

Diabetes tipe 1 adalah diabetes yang sakit pankreasnya menyeluruh. Karena Pankreas tidak dapat menghasilkan insulin sama sekali. Diabetes ini biasanya mengenai anak- anak dan remaja, namun tidak menutup kemungkinan diabetes ini terjadi pada usia dewasa.

Diabetes Tipe-1, untuk dapat bertahan hidup, bergantung pada pemberian insulin dari luar. Oleh Karena itu, pada waktu yang lalu, istilah yang dipakai adalah Insulin Dependent Diabetes Melitus (IDDM). Jumlah kejadiannya hanya 1-10% dari semua penderita diabetes di dunia.

Faktor penyebab terjadinya diabetes Tipe-1 adalah infeksi Virus atau reaksi auto-imun (rusaknya sistem kekebalan tubuh), yang merusak sel-sel penghasil insulin, yaitu sel-β pada pankreas secara menyeluruh. Oleh Karena itu, pada tipe ini pankreas sama sekali tidak menghasilkan insulin. Untuk bertahan hidup, insulin harus diberikan dari luar dengan cara disuntikkan. Sampai sekarang belum ada cara lain, karena jika diminum insulin akan merusak asam lambung. Gejala dan tanda-tandanya muncul mendadak, diantaranya cepat merasa haus, sering kencing, badan mengurus dan lemah.

2. Diabetes Tipe – 2

Diabetes Tipe- 2, bisa juga disebut diabetes *life style* karena penyebabnya selain faktor keturunan, yang terutama adalah gaya hidup yang tidak sehat. Biasanya tipe ini mengenai orang usia dewasa. Dahulu, diabetes ini pernah disebut *adult onset* atau *rityonset* diabetes. Namun karena diabetes ini ternyata juga dapat mengenai mereka yang lebih muda, maka istilah diabetes Tipe-2 dianggap lebih cocok.

Diabetes Tipe – 2 berkembang sangat lambat, bisa sampai bertahun-tahun. Oleh Karena itu, gejala dan tanda- tandanya sering kali tidak jelas. Diabetes Tipe -2 biasanya memiliki riwayat keturunan diabetes. Apabila tidak ada gejala klasik, yang biasa dikeluhkan adalah: cepat lelah, berat badan turun walaupun banyak makan, atau rasa kesemutan di tungkai. Kadang- kadang, bahkan ada diabetisi yang sama sekali tidak merasakan perubahan.

Diabetes Tipe- 2 tidak multak memerlukan suntikan insulin karena pankreasnya masih menghasilkan insulin walaupun jumlahnya tidak mencukupi dan yang terpenting kerja insulin tidak efektif karena adanya hambatan pada kerja insulin atau dalam istilah medisnya adalah resistensi insulin.

Sebenarnya resisten insulin ini mendahului terjadinya penurunan produksi insulin. Selama resistensi insulin belum diperbaiki pankreas harus bekerja keras

menghasilkan insulin sebanyak-banyaknya untuk dapat menggempur resistensi tersebut agar gula bisa juga masuk. Namun karena pankreas harus bekerja secara keras akhirnya timbul kelelahan dan akhirnya insulin yang dihasilkan berkurang. Oleh karena itu, obat yang diberiada diabetesi Tipe -2 tidak hanya obat untuk memperbaiki resistensi insulin, tetapi juga untuk membantu pankreas meningkatkan kembali produksi insulin.

3. Diabetes Tipe Lain

Yang dimaksut diabetes dengan nama "diabetes tipe lain" adalah diabetes yang tidak termasuk Tipe-1 atau Tipe-2 yang disebabkan oleh kelainan tertentu. Misalnya, diabetes yang timbul karena kenaikan hormon- hormon yang kerjanya berlawanan dengan insulin atau hormon kontra insulin. Misalnya, diabetes yang muncul pada penyakit kelebihan hormon tiroid.

1.1 Variabel Random

Suatu Variabel Random *X* yang memetakan ruang sampel *S*, yaitu *S*= $\{e_1, e_2, \ldots e_n\}$ sehingga menghasilkan *X* (e) = *x* , dengan e ∈ S dan x ∈ R.

Terdapat dua jenis variabel random yaitu variabel random diskret dan variabel random kontinu. Jika himpunan semua hasil yang mungkin dari variabel random *X* berhingga atau tak berhingga tetapi masih bisa dihitung maka *X* disebut sebagai variabel random diskrit. Sedangkan jika semua hasil yang mungkin dari variabel random *X* merupakan nilai dalam suatu interval *X* disebut Variabel random kontinu. (Bain L & Engelhardt, 1992)

1.1.1 Variabel Random Diskrit

Misal X merupakan variabel random. Variabel random *X* disebut variabel random diskrit jika himpunan nilai yang muncul dari *X* merupakan nilai yang dapat dihitung (countable). (Bain L & Engelhardt, 1992)

Jika *X* adalah variabel random diskret, diberikan fungsi sebagai berikut

$$f\ (x) = P(X = x)\ x = x\text{i}, x2, \ldots \qquad (3.1)$$

Yang memberikan probabilitas untuk setiap nilai *x* yang mungkin disebut fungsi densitas probabilitas diskrit (pdf diskrit).

Suatu fungsi dapat dikatakan sebagai distribusi probabilitas dari variabel random diskrit jika dan hanya jika memenuhi syarat berikut untuk semua nilai *x* :

$$f(xi) \geq 0 \text{ untuk semua } x_i \quad (3.2)$$

$$dan$$

$$\sum_{all} x_i f(xi) = 1 \quad (3.3)$$

1.1.2 Variabel Random Kontinu

Variabel random *X* merupakan nilai dalam suatu interval *X* disebut Variabel random kontinu.

Jika *X* adalah variabel random kontinu, maka fungsi distribusi kumulatifnya adalah sebagai berikut

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt \quad (3.4)$$

Untuk $f(t)$ merupakan pdf dari *x.* (Bain L & Engelhardt, 1992)

1.2 Analisis Regresi Random Forest

Analisis regresi adalah salah satu metode untuk menentukan hubungan sebab-akibat antara satu variabel dengan variable yang lain. Variabel "penyebab" biasa disebut *variabel independen*, atau secara bebas, *variabel X* . Variabel terkena akibat dikenal sebagai *variable yang dipengaruhi, variabel dependen, variabel terikat*, atau *variabel Y.* Kedua variabel ini dapat merupakan variabel acak (random), namun variabel yang dipengaruhi harus selalu variabel acak. Wikipedia, Analisis Regresi, (online)

Dalam analisis regresi terdapat tiga pendekatan yaitu dengan menggunakan regresi parametrik, regresi nonparametrik dan regresi semiparametrik**.** Pada regresi parametrik terdapat asumsi yang harus terpenuhi yaitu kurva regresi diketahui bentuknya misalnya linear, kubik, polynomial derajat p, kuadratik, eksponensial dan lain-lain. Namun pada kenyataannya tidak semua data memiliki pola tertentu. Terdapat data yang memiliki scatter plot namun tidak memiliki pola tertentu, sehingga data tidak diketahui polanya. Oleh karen itu jika dimodelkan dengan regresi parametrik terjadi *error* yang sangat tinggi dan menghasilkan kesimpulan yang bias. Pada regresi nonparametrik, digunakan untuk menduga kurva regresi yang tidak diketahui bentuk kurva regresinya dan tidak terikat dengan asumsi seperti pada regresi parametrik. Dalam

regresi nonparametrik data diharapkan mencari sendiri bentuk penduganya, sehingga memiliki fleksibilitas tinggi.

Seiring berkembangnya waktu dan juga berkembangnya ilmu pengetahuan, semakin kompleks pula pola hubungan yang ditemukan yang terjadi pada dua variabel. Karena dalam analisis regresi pola hubungan anatara dua variabel atau lebih tidak selalu berpola parametrik dan nonparametrik. Bahkan dalam beberapa kasus yang lain sering berpola semiparametrik. Regresi semeparametrik adalah gabungan dari parametrik dan juga non parametrik. Model regresi semiparametrik mulai berkembang pesat. Berbagai macam metode regresi semiparametrik yang sering digunakan untuk penelitian diantaranya model regresi semiparametrik *spline*, *kernel*, dan sebagainya. Metode – metode regresi tersebut juga masih belum memberikan solusi secara menyeluruh karena metode tersebut masih tergolong rumit dan memiliki banyak asumsi dan standar metode yang harus dipenuhi untuk menghasilkan hasil yang akurat pada metode tersebut.

Oleh karena semakin maju nya teknologi menyebabkan semakin banyak data yang muncul dari teknologi informasi dan kekuatan komputasi yang terus maju, maka munculah berbagai metode algoritma yang dapat mengatasi permasalahan pada big data. Salah satu nya untuk mengatasi metode regresi pada pola hubungan yang rumit dan data yang tidak memenuhi asumsi untuk dilakukan metode regresi. Dengan penyelesaian yang fleksibel, mudah dan sederhana namun menghasilkan akurasi yang tinggi, metode *Random Forest* dapat menyelesaikan masalah tersebut.

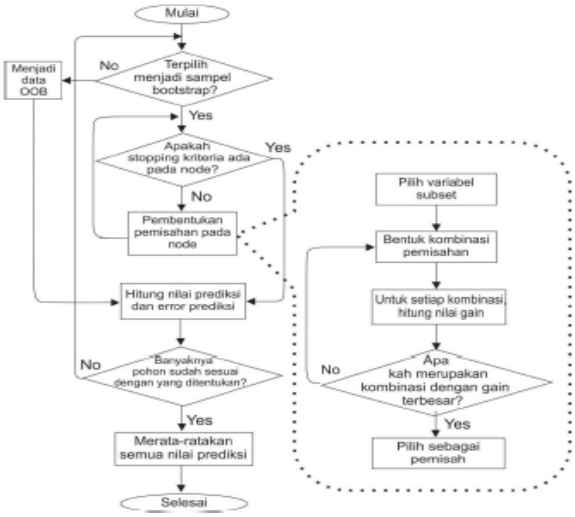
Regresi dengan *Random Forest* merupakan penyelesaian regresi dengan cara yang berbeda, tidak menggunakan *Smoothing* seperti metode regresi *kernel* ataupun *spline*, namun dengan melakukan penumbuhan pohon keputusan dengan jumlah banyak pada sampel *Bootstrap* (replikasi resampling). Metode *Bootstrap* mengambil sampel dari sampel yang sudah ada atau yang disebut replikasi, karena resample dilakukan berkali-kali. Oleh karena itu regresi *Random Forest* juga dapat menangani data dengan jumlah yang lebih sedikit.

1.2.1 Pengertian Random Forest

Algoritma *Random Forest* pertama kali diusulkan oleh Breiman (2001). Algoritma Random Forest dapat digunakan untuk menyelesaikan persoalan klasifikasi dan juga regresi selain itu, Random Forest dapat digunakan untuk berbagai jenis variabel respon seperti kontinu, diskrit, data survival maupun data kombinasi multivariat. Selain itu tidak ada asumsi yang harus dipenuhi pada *Random Forest*. Metode ini dapat mengestimasi berbagai bentuk fungsi yang terbentuk antara variabel respon dan variabel penjelas d mempermudah menentukan hubungan nonlinear yang kompleks yang mungkin akan sulit ditemukan tanpa adanya spesifikasi tertentu dan tanpa menggunakan standar metode tertentu. Intinya, *Random Forest* dapat dan mampu mendeteksi berbagai interaksi antara respon dan prediktor. Dengan fleksibilitas dari *Random Forest*, membuat metode ini sangat berguna sebagai metode eksplorasi data. *Random forest* biasa juga disebut sebagai metode *ensamble* atau metode gabungan. Disebut metode gabungan karena terbentuk dari model model kecil namun hasil prediksi nya ditentukan dengan mengkombinasikan semua *output* pada model kecil tersebut atau yang bisa disebut sub-model. Sub-model pada metode *Random Forest* adalah *classification and regression trees* (CART).

Metode Random Forest adalah pengembangan dari metode CART, yaitu dengan menerapkan metode *Bootstrap Aggregating* (bagging) dan *Random Feature Selection*. Dalam Random Forest, banyak pohon ditumbuhkan sehingga terbentuk hutan (forest), kemudian analisis dilakukan pada gugus data yang terdiri dari *n* amatan dan *p* peubah penjelasan. Sehingga dengan kata lain, *Random Forest* adalah metode yang merupakan gabungan dari CART. Oleh karena itu, sebelum membahas algoritma *Random Forest*, terlebih dahulu dibahas mengenai *Classification and Regression Trees* (CART). (Breiman L. , 2001).

Berikut adalah algoritma random forestnya



Gambar 3.1 *Flowchart* Algoritma *Random Forest*

1.3 *Classification and Regression Trees* (CART)

Classification and Regression Trees (CART) merupakan metode eksplorasi data yang didasarkan pada teknik pohon keputusan. Pohon Klasifikasi dihasilkan saat peubah respons berupa data kategorik, sedangkan pohon regresi dihasilkan saat peubah respons berupa data numerik (Breiman t al.1984).

Pengelompokkan pada CART ini bekerja dengan cara menentukan variabel prediktor dan nilai pemisahan nya yang merupakan nilai pada prediktor tersebut untuk dijadikan sebagai kandidat pemisahan. Pada setiap pemisahan yang dibentuk, dihitung berapa *error* yang dihasilkan jika variabel prediktor beserta nilai pemisahnya dijadikan sebagai kriteria pemisahan. Untuk memilih variabel mana yang dijadikan sebagai pemisahan, dipilih dengan menghitung penurunan *error* yang didapatkan jika variabel tersebut dijadikan sebagai variabel pemisah. (Breiman t al.1984).

1.4 *Algoritma Random Forest*

Random Forest adalah Pengembangan dari metode CART atau dengan kata lain *Random Forest* adalah metode *ensamble* (gabungan). Berikut ini adalah algoritma dari *Random Forest.*

1.4.1 Tahap Boostrap

Tahap Boostrap dilakukan untuk mendapatkan model kecil (submodel) pada *Random Forest* bukan dengan melakukan penaksiran pada semua data yang ada tetapi sampel *Boostrap* dari data asli. Data hasil *Boostrap* dinamakan data *out of bag* atau yang biasa disebut OOB. Data OOB berfungsi untuk menghitung nilai prediksi pada masing-masing CART. Hasil prediksi dari model CART dihasilkan dari memasukkan data OOB ke dalam pohon yang telah tebentuk dari sampel *Boostrap*. Sehingga, setiap observasi memiliki nilai prediksi yang dihasilkan dari setiap pohon CART data tersebut tidak menjadi sampel *Boostrap* pada pohon tersebut dan akan mengurangi resiko *Overfitting* yang ditimbulkan dari CART.

1.4.2 Penumbuhan CART pada setiap Sampel Boostrap

Setelah dilakukan resampling boostrap dari data asli, maka selanjutnya menumbuhkan pohon CART dari sampel *Boostrap*. Prosedur pemilihan variabel pemisahan dan nilai pemisahan pada CART ditentukan dengan besarnya penurunan *error* atau gain yang dihasilkan.

Pemilihan variabel pemisah dari semua variabel penjelas yang ada pada data asli, hanya variabel penelas yang terpilih pada random subset yang bisa dijadikan sebagai variabel pemisah. Banyaknya jumlah variabel penjelas yang terdapat pada pemilihan random subset dinotasikan (*m*). Nilai *m* yang dijadikan standar untuk regresi adalah p/3 ini tidak berlaku untuk semua kasus dan ukuran minimum node adalah 5. Menurut Breimen *default* yang digunakan untuk *m try* adalah 5.

Variabel penjelas yang menghasilkan nilai penurunan *error* terkecil setelah dilakukan *random subset* pada variabel penjelas memiliki kesempatan lebih besar untuk terpilih menjadi variabel pemisah.

1.4.3 Tahap Prediksi

Didalam regresi tahapan prediksi menjadi penting untuk mengetahui model terbaik. Semakin dekat dengan data asli maka hasil prediksi semakin baik. Untuk satu data observasi dapat menjadi data OOB berulang kali, sehingga dengan kata lain data tersebut masuk ke dalam pohon yang diestimasi. Pada regresi Random Forest, nilai prediksi yang dihasilkan dari kumpulan pohon adalah rata- rata nilai prediksi pada setiap pohonnya.

1.5 *Variable Importance*

CART merupakan pohon individu yang mudah dinterpretasikan sedangkan *random forest* merupakan kumpulan dari banyak pohon yang ditumbuhkan sehingga hasil interpretasinya tidak semudah CART. Namun, dengan beberapa metode yang telah dikembangkan untuk mendapatkan informasi yang lebih dari sekedar prediksi salah satunya datpat menggunakan *Variable Importance.*

Nilai *variable importance* diukur berdasarkan pemilihan variabel prediktor atau penjelasnya menjadi variabel pemisah terhadap *error*-nya. Pada data *X* misalkan x_j dikatakan tidak berhubungan dengan *y*, jika x_j pada *X* yang menjadi variabel pemisah tidak memiliki penurunan error yang signifikan dalam memprediksi variabel *y*. Sebaliknya x_j dikatakan kuat mempengaruhi variabel y jika menghasilkan penurunan error yang signifikan dalam memilih x_j dari hasil random subset untuk dijadikan pemisahan. Untuk menghitung nilai *variable importance* x_j pada pohon t € T adalah sebagai berikut :

$$VI^{(t)}(x_j) = L\,(y^{(t)}, \hat{y}^{(t)} - y^{(t)}, \hat{y}_{\pi}^{(t)}) \qquad (3.4)$$

Dimana t adalah indeks dari pohon x_j prediktornya $\hat{y}^{(t)}$ hasil prediksi dari pohon sebelum x_j dijadikan variabel pemisah dan $\hat{y}_{\pi}^{(t)}$ adalah hasil prediksi dari pohon setelah x_j dijadikan variabel pemisah. Untuk menghitung nilai *variable importance* pada pohon t. Nilai *variable importance* pada pohon individu hanya menghitung nilai selisih error yang dihasilkan sebelum pemisahan dengan x_j dan setelah pemisahan dengan x_j. Untuk menghitung nilai variable importance pada seluruh pohon yaitu dengan menghitung nilai rata-rata pada nilai *variable importance* yang dihitung pada setiap pohon, berikut adalah persamaannya :

$$VI(x_j) = \frac{1}{T}\sum_{t=1}^{T} VI^{(t)}(x_j) \qquad (3.5)$$

Dari persamaan diatas dapat dilihat bahwa untuk menghitung nilai kepentingan variabel prediktor tersebut dalam mempengaruhi responnya. Nilai *variable importance* hanya bisa menghitung seberapa penting variabel predktor dalam mempengaruhi variabel responnya, tetapi tidak tahu bagaimana hubungan yang terjadi antara kedua variabel tersebut.

1.6 Keakurasian Regresi Random Forest

Random forest untuk regresi dibentuk dari pohon yang bergantung pada vektor random θ dimana { θ_t} adalah vektor random i.i.d. Hasil prediksi respon dari beberapa pohon yang ditumbuhkan pada metode random forest dinotasikan dengan h_1 (x), $h_2\,(x), \ldots, h_T(x)$. Pada random forest, nilai $h_t\,(x) = h(X, \theta_t)$. Hasil prediksi random forest terbentuk dari rata- rata nilai {h(X, θ_t)} untuk semua T pohon. Keakurasian prediksi regersi random forest dilihat dari *nilai mean square error* pada hasil prediksi numerik h(x) sebagai berikut

$$E\text{x,y}\,(Y - h(X))^2 \tag{3.6}$$

Jika nilai *mean square error* semakin kecil maka semakin baik tingkat konfidensi hasil regresi dengan menggunakan metode Random Forest. Oleh karena itu, fungsi *mean square error* menentukkan nilai general *error* yang dihasilkan pada metode ini.

1.7 Regresi Logistik

Menurut Hosmer dan Lemeshow (2000) tujuan melakukan analisis data kategori menggunakan regresi logistik adalah mendapatkan model terbaik dan sederhana untuk menjelaskan hubungan antara keluaran dari variabel respons (X) dengan variabel-variabel prediktornya (Y). Variabel respons dalam regresi logistik dapat berupa kategori atau kualitatif, sedangkan variabel prediktornya dapat berupa kualitatif dan kuantitatif. Jika variabel merupakan variabel biner atau dikotomi dalam artian variabel respons terdiri dari dua kategori yaitu “sukses” ($Y = 1$) atau “gagal” ($X = 0$), maka variabel mengikuti sebaran Bernoulli yang memiliki fungsi densitas peluang:

$$f(y_i) = \pi(x_i)^{yi}(1 - \pi(x_i))^{1-yi}, \quad ; \text{y}_\text{i} = 0{,}1 \tag{3.7}$$

Sehingga diperoleh :

$$y_i = 0, \text{ maka } f(0) = \pi\,(x_i)^0\,(1 - \pi(x_i))^{1-0} = 1 - \pi(x_i\,) \tag{3.8}$$

Untuk

$$y_i = 1, \text{ maka } f(1) = \pi(x_i\,)^1(1 - \pi(x_i))^{1-1} = \pi(x_i) \tag{3.9}$$

Misalkan probabilitas dari variabel respons untuk nilai yang diberikan, dinotasikan sebagai $\pi(x)$. Model umum $\pi(x)$ dinotasikan sebagai berikut :

$$\pi(x) = \frac{\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}{1+\exp(\beta_0+\beta_1 x_1+\beta_2 x_2 +\cdots+\beta_p x_p)'} \quad (3.10)$$

Persamaan 3.10 disebut fungsi regresi logistik yang menunjukkan hubungan antara variabel prediktor dan probabilitas yang tidak linear, sehingga untuk mendapatkan hubungan yang linear dilakukan transformasi yang sering disebut dengan transformasi logit. Bentuk logit dari $\pi(x)$ dinyatakan sebagai $g(x)$ yaitu :

$$logit[\pi(x)] = g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (3.11)$$

Persamaan (3.11) merupakan bentuk fungsi hubungan model regresi logistik yang disebut model regresi logistik berganda (Hosmer dan Lemeshow, 2000)

1.8 Regresi Logistik Biner

Regresi logistik biner merupakan suatu metode analisis data yang digunakan untuk mencari hubungan antara variabel respon (y) yang bersifat biner atau dikotomus dengan variabel prediktor (x) yang bersifat polikotomus (Hosmer dan Lemeshow, 2000). Keluaran dari variabel respon y terdiri dari 2 kategori yang biasanya dinotasikan dengan y =1 (sukses) dan y=0 (gagal). Hosmer dan Lemeshow (2000) menjelaskan bahwa model regresi logistik dibentuk dengan menyatakan nilai $P(Y = 1 \mid x)$ sebagai $\pi(x)$ yang dinotasikan sebagai berikut :

$$g(x) = \frac{\exp(g(x))}{1 + \exp(g(x))}$$

Suatu fungsi dari π(x) dicari dengan menggunakan transformasi logit, yaitu g(x) yang dapat dinyatakan sebagai berikut :

$$g(x) = ln\frac{[\pi(x)]}{[1-\pi(x)]} = \beta_0 + \beta_1 x_1 + \cdots + \beta_2 x_2 \quad (3.12)$$

Pengujian terhadap parameter-parameter model dilakukan baik secara simultan maupun secara parsial. Menurut Hosmer dan Lemeshow (2000), pengujian parameter model secara simultan menggunakan uji nisbah kemungkinan (*Likelihood Ratio Tesis),* dengan hipotesis. Menurut Kleinbaum dan Klein (2002), regresi logistik adalah suatu model pendekatan matematika yang

digunakan untuk menggambarkan hubungan antara beberapa vaiabel penjelas dengan suatu variabel dikotomi. Variabel dikotomi mempunyai dua kemungkinan yang biasanya dinyatakan dengan 0 (gagal) dan 1 (sukses). Diberikan model sebagai berikut:

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_{pi} + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2) \tag{3.13}$$

Jika Y diberi 0 dan 1, maka

$$P(Y_i = 1|X = x_i) = \pi(x_i)\ dan\ P(Y_i = 0|X = x_i) = 1 - \pi(x_i) \tag{3.14}$$

Nilai harapan dari $(yi|xi)$ adalah $E(yi|xi) = \pi(xi)$

Persamaan umum regresi biner adalah

$$\pi(x_i) = \frac{e^{\beta 0 + \beta 1 x 1 i + \cdots \ldots + \beta p x p i}}{1 + e^{\beta 0 + \beta 1 x 1 i + \cdots \ldots + \beta p x p i}} \tag{3.15}$$

Persamaan (3.15) mempunyai bentuk yang tidak linier. Untuk membuatnya menjadi persamaan yang linier, maka digunakan transformasi log dari *odd rasio* atau disebut juga transformasi logit. Berikut ini adalah logit dari persamaan (3.15):

$$\ln \frac{\pi(x_i)}{1 - \pi(x_i)} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \tag{3.16}$$

Dengan hipotesis

H_0 : βi =…..=βp = 0

H_1 : minimal ada satu βi ≠ 0 ; i = 1,2,….p statistik uji G dirumuskan:

$$G = -2ln\frac{Lo}{Lp}$$

Diketahui Lo adalah fungsi kemungkinan tanpa peubah penjelas dan merupakan kemungkinan dengan peubah penjelas. Mengasumsikan H0 benar, statistik uji *G* akan mengikuti sebaran khi kuadrat dengan derajat bebas *p*. Keputusan tolak H0 jika $G > \alpha^2_{p(\alpha)}$. Interpretasi koefisien untuk model regresi logistik biner dapat dilakukan dengan menggunakan nilai rasio oddsnya. Odds sendiri dapat diartikan sebagai rasio peluang kejadian sukses dengan kejadian tidak sukses dari peubah respon. Rasio odds mengindikasikan seberapa lebih mungkin munculnya kejadian sukses pada suatu kelompok dibandingkan dengan kelompok lainnya. Rasio odds didefinisikan sebagai:

$$\varphi = \exp(\widehat{\beta_\iota}) = \exp[g(1) - g(0) \quad (3.17)$$

Interpretasi dari rasio odds ini adalah kecenderungan untuk Y= 1 pada X=1 sebesar Ψ kali dibandingkan pada X= 0

1.9 Penaksiran Parameter

Untuk penaksiran parameter regresi logistik dapat dilakukan dengan dua cara yaitu:

1.9.1 Maximum Likelihood Estimation (MLE)

Metode ini pada dasarnya memberikan nilai estimasi β untuk memaksimumkan fungsi likelihood (Hosmer dan Lemeshow 1989). Secara matematis fungsi likelihood (*xi.yi)* dapat dinyatakan:

$$f(x_i) = \pi(x_i)^{yi}[1 - \pi(x_i)]^{1-yi} \quad (3.18)$$

karena setiap pengamatan diasumsikan independen maka fungsi likelihoodnya merupakan perkalian antara masing-masing fungsi likelihood yaitu:

$$l = (\beta) = \prod_{i=1}^{n} f(xi) \quad (3.19)$$

Dan logaritma likelihoodnya dinyatakan sebagai :

$$L(\beta) = \ln[l(\beta)]$$

$$= \sum_{i=1}^{n}\{yi.\ln(\pi(xi) + (1 - yi)\ln\big(1 - \pi(xi)\big)\} \quad (3.20)$$

Untuk memperoleh nilai *β* maka dengan memaksimumkan nilai L*(β)* dan mendiferensialkan L*(β)* terhadap *β* dan menyamakannya dengan nol. Persamaan ini dapat ditulis dalam bentuk sebagai berikut:

$$\sum_{i=1}^{n}[yi - \pi(xi)] = 0$$

3.21

Dan persamaan likelihood :

$$\sum_{i=1}^{n}[yi - \pi(xi)] = 0 \quad 3.22$$

1.9.2 Metode Newton Rhapson

Metode ini merupakan metode untuk menyelesaikan persamaan nonlinier seperti menyelesaikan persamaan likelihood dalam model regresi logistik (Agresti, 1990). Metode newton rhapson memerlukan taksiran awal untuk nilai fungsi maksimumnya, yang mana fungsi tersebut merupakan taksiran yang menggunakan pendekatan polynomial berderajat dua. Dalam hal ini untuk menentukan nilai $\widehat{\beta}$ dari β yang merupakan fungsi maksimum dari $g = (\beta)$ Andaikan $\dot{q} = \left(\frac{\partial_g}{\partial_{\beta 1}} + \frac{\partial_g}{\partial_{\beta 2}}, \ldots\right)$ dan andaikan **H** dinotasikan sebagai matriks yang mempunyai anggota .

1.10 Pengujian Signifikansi

Setelah menaksir parameter maka langkah selanjutnya yang dilakukan adalah menguji signifikansi parameter tersebut. Untuk itu digunakan uji hipotesis statistik untuk menentukan apakah variabel terikat berpengaruh signifikansi parameter dilakukan sebagai berikut:

1.10.1 Uji Parsial

Digunakan untuk menguji apakah setiap βi secara individual. Hasil pengujian secara parsial/individual akan menunjukan apakah suatu variabel

terikat layak untuk masuk dalam model atau tidak (Agresti, 1990).

Hipotesis :

H0 : βi = 0

H0 : βi ≠ 0

Statistik Uji : Wald (W) = $= \frac{\beta i}{SE(\beta i)}$

Rasio yang dihasilkan dari statistik uji, dibawah hipotesis H0 akan mengikuti sebaran normal baku (Hosmer dan Lemeshow, 1989). Sehingga untuk memperoleh keputusan dilakukan perbandingan dengan distribusi normal baku (Z). kriteria penolakan (tolak H0) jika nilai $W > Z\alpha/2$

1.10.2 Uji Overall

Uji Overall disebut juga uji model chi-square, dilakukan sebagai upaya memeriksa peranan variabel terikat dalam model secara bersama-sama.

Hipotesis :

H0 : β1 = β2=……..= β k = 0

H1 : paling sedikit ada satu β1 ≠ 0 (i = 1,2,….,k)

Statistik uji yang digunakan adalah statistic uji G atau Likelihood Ratio Test :

$$G^2 = -2ln\frac{L_1}{L_0}$$

Statistic uji G2 mengikuti distribusi chi-square, sehingga untuk memperoleh keputusan dilakukan perbandingan dengan *X*2 tabel. Dimana derajat bebas = k (banyaknya variabel terikat). Kriteria penolakan (tolak H0) jika nilai G > *X*2($db,\ \alpha$).

1.11 Interpretasi Pengujian Model

Proses selanjutnya adalah mendapatkan interpretasi terhadap model pengujian signifikansi parameter tersebut. Interpretasi koefisien parameter diharapkan dapat menjelaskan tiga hal:

a. Menjelaskan hubungan fungsional antara variabel bebas dan variabe

terikat

b. Menentukan unit-unit perubahan setiap variabel independen

c. Mendapatkan nilai odds rasio yang menunjukk perbandingan tingkat

BAB IV
METODOLOGI PENELITIAN

4.1 Data Penelitian

Penelitian ini menggunakan data sekunder yang didapatkan dari website kaggle dataset diabetes melitus di Amerika Serikat pada tahun 2018. Populasi yang digunakan dalam penelitian ini adalah semua orang yang sudah terdiagnosis Diabetes Melitus dan Yang Tidak Terdiagnosis Diabetes Melitus. Sedangkan Sampel yang digunakan disini adalah Dua Daerah yang berada di kota Virginia Amerika Serikat yaitu Buckingham dan Louisa dengan jumlah Responden dalam dataset berjumlah 366 Orang baik laki- laki Maupun Perempuan. Berikut adalah linknya https://storage.googleapis.com/kaggle diakses pada tanggal 8 maret 2018.

4.2 Variabel Penelitian

Variabel yang digunakan dalam penelitian ini adalah :

a. Variabel Prediktor atau variabel bebas, yaitu kolesterol, gula darah, lemak baik (LDL), Usia, berat badan, tinggi badan, tekanan darah sistol, tekanan darah diastol, lingkar pinggang, dan lingkar pinggul.
b. Variabel Respon atau variabel terikat, yaitu glikogen hemoglobin

4.3 Definisi Operasional Peubah

Definisi Operasional variabel pada penelitian ini sebagai berikut :

a. Glikogen Hemoglobin (Y)

   Glikogen Hemogblobin dalam penelitian ini digunakan untuk mendiagnosis diabetes melitus terhitung sejak dilakukan pemeriksaan kadar glikogen hemoglobin dengan satuan mmol/L.

b. Kolesterol (X1)

   Kolesterol dalam penelitian ini merupakan uji untuk pengendalian apakah diabetes melitus atau bukan diabetes melitus terhitung sejak dilakukan pemeriksaan kadar kolesterol dengan satuan mg/dl.

c. Gula Darah (X2)

Gula darah dalam penelitian ini merupakan uji untuk pengendalian pasien penderita Diabetes Melitus maupun non diabetes melitus dihitung sejak dilakukannya pemeriksaan kadar gula darah setelah makan dengan satuan mm/dl.

d. Lemak Baik (X3)

   Lemak baik dalam penelitian ini digunakan untuk pengendalian diabetes melitus maupun non diabetes melitus dihitung sejak dilakukannya pemeriksaan lemak baik dengan satuan mg/dl.

e. Usia (X4)

   Usia dalam penelitian ini merupakan faktor yang tidak bisa dimodifikasi dan dikategorikan dari usia remaja, dewasa, lansia dan manula penderita diabetes melitus dan non diabetes melitus dengan satuan tahun.

f. Berat Badan (X5)

   Berat badan merupakan berat penderita Diabetes melitus dan non diabetes meitus pada saat dilakukannya pemeriksaan dengan satuan Kg.

g. Tinggi Badan (X6)

   Tinggi badan merupakan tinggi penderita Diabetes melitus dan non diabetes meitus pada saat dilakukannya pemeriksaan dengan satuan Kg.

h. Tekanan Darah Sistol (X7)

   Tekanan darah sistol atau sistolik merupakan tekanan darah atas yang tercatat pada saat pemeriksaan. Tekanan darah sistol tercipta karena adanya kontraksi jantung sehingga mendorong darah melalui arteri ke seluruh tubuh dan dengan satuan mmHg

i. Tekanan Darah Diastol (X8)

   Tekanan darah diastol atau diastolik merupakan tekanan darah bawah yang tercatat pada saat pemeriksaan. Tekanan darah diastol menunjukkan jumlah tekanan darah di dalam arteri ketika jantunng sedang beristirahat dengan satuan mmHg.

j. Lingkar Pinggang (X9)

Lingkar Pinggang merupakan lingkar pinggang penderita Diabetes melitus dan non diabetes meitus pada saat dilakukannya pemeriksaan dengan satuan cm.

k. Lingkar Pinggul (X10)

Lingkar Pinggul merupakan lingkar pinggul penderita Diabetes melitus dan non diabetes meitus pada saat dilakukannya pemeriksaan dengan satuan cm.

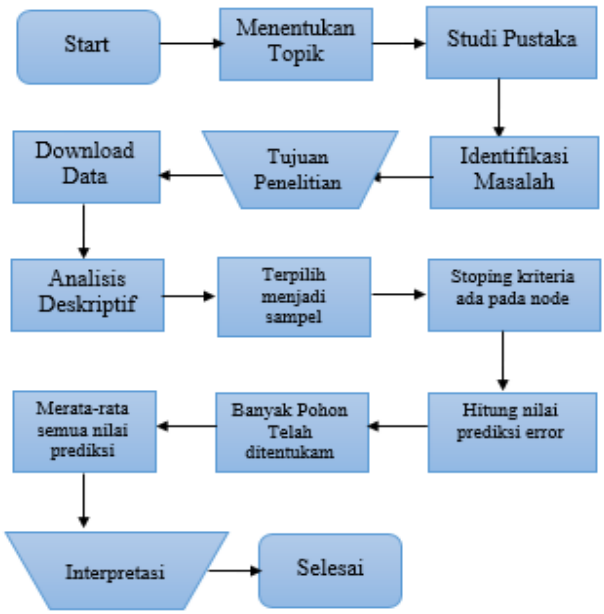
4.4 Metode Analisis Data

Proses analisis dalam penelitian ini menggunakan bantuan *software Excel* 2013, dan R 3.2.5. Ada Beberapa Metode yang digunakan dalam penelitian ini, diantaranya sebagai berikut :

1. Analisis Regresi, digunakan untuk mendapatkan model yang terbaik dan mengetahui faktor apa saja yang mempengaruhi seseorang terkena diabetes melitus.
2. Analisi Deskriptif, digunakan untuk memberikan gambaran umum tentang persebaran penyakit diabetes di Virginia Amerika Serikat.
3. Regresi Random Forest digunakan untuk menyelesaikan kasus dengan variabel penjelasnya bersifat numerik.
4. CART bekerja dengan cara menentukan variabel prediktor dan nilai pemisahan nya yang merupakan nilai pada prediktor tersebut untuk dijadikan sebagai kandidat pemisahan.
5. *Variable Importance* digunakan untuk mengeksplor informasi lebih dari sekedar prediksi.

4.5 Langkah Penelitian

Tahapan atau langkah dalam penelitian ini digambarkan dalam *flowchart* melalui **Gambar 4.1** Berikut ini:



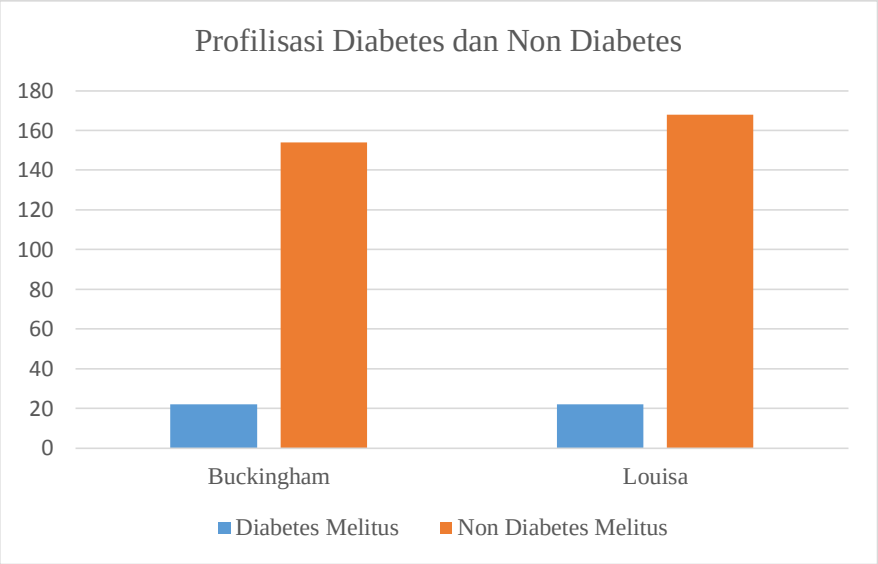
Gambar 4.1 Flowchart Penelitian

BAB V
HASIL DAN PEMBAHASAN

Pada bab ini akan dibahas hasil Analisis Statistika Deskriptif dan juga Regresi Random Forest.

5.1 Perbandingan Profil Penderita dan Bukan Penderita Diabetes Melitus

Dalam sub bab ini akan dibahas tentang hasil analisis deskriptif dari data yang digunakan yaitu data yang didapatkan dari website *kaggle* dataset penyakit diabetes. Variabel yang digunakan dalam penulisan ini menggunakan 11 Variabel diantaranya Kolesterol, Gula Darah, Lemak Baik, Glikogen Hemoglobin, Umur, Berat Badan, Tinggi Badan, Tekanan Darah Diastol, Tekanan Darah Sistol, Lingkar Pinggul, dan Lingkar Pinggang dan berikut ini adalah gambaran dataset yang ada di virginia Amerika Serikat dengan dua wilayah yaitu Louisa dan Buckingham :



Gambar 5.1 Profilisasi Diabetes melitus dan Non Diabetes Melitus

Berdasarkan Gambar 5.1 dapat dilihat bahwa jumlah penderita Diabetes Melitus di wilayah buckingham berjumlah 22 orang dan untuk wilayah louisa 22 orang dari kedua wilayah tersebut tidak jauh berbeda jumlah penderita Diabetes Melitus. Kemudian untuk jumlah penderita Non Diabetes Melitus wilayah louisa memiliki jumlah yang banyak yaitu 168 orang dibandingkan di wilayah Buckingham yaitu berjumlah 154 Orang.

Kemudian selanjutnya dilakukan uji proporsi untuk melihat proporsi masing- masing jumlah penderita diabetes melitus dan bukan penderita diabetes melitus , berikut adalah hasilnya :

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,073[a]	1	,787		
Continuity Correction[b]	,012	1	,913		
Likelihood Ratio	,073	1	,787		
Fisher's Exact Test				,873	,456
Linear-by-Linear Association	,073	1	,787		
N of Valid Cases	366				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 21,16.

b. Computed only for a 2x2 table

Gambar 5.2 Hasil Uji Proporsi

Berdasarkan gambar diatas dapat dilakukan uji proporsi menggunakan chi-square sebagai berikut :

a) Hipotesis

H_0 : $P_B = P_L$ (Tidak ada perbedaan jumlah penderita diabetes melitus dan bukan diabetes melitus diwilayah buckingham dan louisa)

H_1 : $P_B < P_L$ (Ada perbedaan jumlah penderita diabetes melitus dan bukan diabetes melitus diwilayah buckingham dan louisa)

b) Tingkat signifikansi

α=5% =0.05

c) Daerah Kritis

H_0 ditolak jika *p-value* < α

d) Statistik uji

(*p-value*) = 0.78

e) Keputusan

p-value > α (0,78< 0,05) → Gagal Tolak H_0

f) Kesimpulan

Dengan menggunakan tingkat kepercayaan 95%, maka tidak ada perbedaan jumlah penderita diabetes melitus dan bukan diabetes melitus diwilayah buckingham dan louisa.

1.5.1 Uji Statistik Perbedaan Penderita Diabetes dan Bukan Penderita Diabetes

Data yang diperoleh dari hasil penelitian dianalisis menggunakan uji statistik dengan menggunakan independen t-test . Hasil Perhitungannya dapat dilihat dari tabel berikut ini :

Tabel 4.1 Hasil Uji Independet T-test

Variabel	t-hitung	p-value	Keputusan
Glikogen Hemoglobin	18.97	0.000	Tolak H0
Kolesterol	3.09	0.003	Tolak H0
Gula Darah	9.062	0.01	Tolak H0
Lemak Baik	-2.64	0.000	Tolak H0
Usia	5.633	0.000	Tolak H0
Berat Badan	0.14	0.88	Gagal Tolak H0
Tinggi Badan	2.40	0.01	Tolak H0
Tekanan Darah Sistol	2.73	0.48	Tolak H0
Tekanan Darah Diastol	0.70	0.48	Gagal Tolak H0
Lingkar Pinggang	3.75	0.000	Tolak H0
Lingkar Pinggul	2.62	0.01	Tolak H0

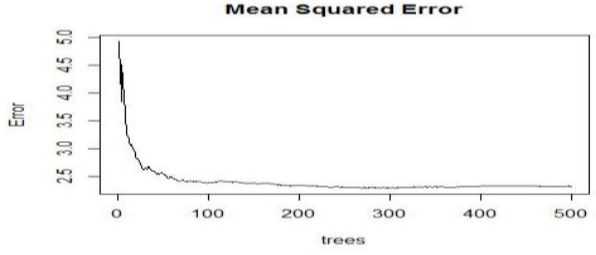
Berdasarkan tabel 4.1 didapatkan hasil bahwa variabel glikogen hemoglobin, Kolesterol, Gula Darah, Lemak Baik, Usia, Tinggi badan, tekanan darah sistol, lingkar pinggang dan lingkar pinggul bila dilihat dari nilai *p-value* uji beda penderita diabetes melitus dan bukan penderita diabetes melitus ke sembilan variabel itu nilai p-valuenya kurang dari 0.05 sehingga dapat disimpulkan bahwa ada perbedaan nilai antara penderita diabetes dan bukan penderita diabetes. Sementara itu, untuk variabel tekanan darah diastol dan berat badan dilihat dari nilai *p-value* uji beda penderita diabetes melitus dan bukan penderita diabetes melitus kedua variabel itu nilai *p-value*nya lebih besar dari 0.05 sehingga dapat disimpulkan bahwa tidak ada perbedaan nilai antara penderita diabetes dan bukan penderita.

5.2 Regresi Random Forest

Untuk Melakukan Regresi Random Forest langkah yang pertama adalah menumbuhkan jumlah pohon yang akan digunakan sebagai dasar dengan melihat nilai MSE yang paling kecil. Kedua, pemilihan nilai *m* atau random subset nilai *m* dicobakan 3 kali yang pertama dengan *default m*=5, setengah dari *default* m =2.5 dan dua kali *default*=10 (Breimen, 2001) setelah dicobakan nilai m atau random subsetnya pilih dari ketiga nilai *m* tersebut yang memiliki MSE yang paling kecil Ketiga, Setelah didapatkan jumlah pohon dan *m* atau random subset selanjutnya mencari nilai *variable importance*. Variable Importance ini menunjukkan kepentingan antara *variable* prediktor dan juga *variable* respon tetapi tidak memberikan nilai seberapa besar hubungan antara masing masing variabelnya. Keempat, Untuk melihat seberapa bagus hasil prediksinya dilihat melalui kurva semakin baik dengan data asli maka model akan semakin bagus.

5.2.1 Menumbuhkan Jumlah Pohon

Sesuai dengan namanya *Random Forest* maka langkah awal untuk melakukan analisis ini yaitu memilih pohon yang akan ditumbuhkan. Didalam *Random Forest* perlu dilakukan memilih jumlah pohon yang akan ditumbuhkan karena *Random Forest* merupakan metode *ensamble* (gabungan) dan juga karena metode ini merupakan penggabungan dari banyak pohon yang ditumbuhkan untuk menghasilkan prediksi yang akurat. Untuk melihat hasil keakuratan maka digunakan jumlah pohon yang tumbuh. Semakin banyak pohon yang tumbuh maka semakin akurat pula hasil prediksinya. Pemilihan banyaknya pohon yang ditumbuhkan akan mempengaruhi nilai estimasi *error* yang dihitung dari OOB *error* yang dihasilkan pada setiap pohon.



Gambar 5.3 *Grafik* Mean Square Error

Berdasarkan **Gambar 5.3** diatas dapat dilihat pada saat pohon ditumbuhkan dari 0-200 terlihat naik turun secara signifikan tetapi ketika jumlah pohon yang tumbuh dimulai dari pohon 200 hingga 500 terlihat nilainya sudah semakin stabil dan kecil nilai *error*nya. Sehingga dengan demikian untuk mengetahui pohon berapa banyak pohon yang akan ditumbuhkan, peneliti akan mencoba menumbuhkan masing- masing pohon dari pohon 200 hingga 500 dan melihat nilai MSE nya. Semakin kecil nilai MSE pada pohon yang ditumbuhkan maka hasil prediksinya semakin akurat.

Tabel 5.1 Nilai MSE pada masing-masing *n-tree* yang dicobakan

Jumlah Pohon Yang Ditumbuhkan	MSE OOB
200	2.317982
300	2.302475
400	2.302052
500	2.285527

Dari **tabel 5.1** diatas maka dapat disimpulkan bahwa jumlah pohon yang harus ditumbuhkan adalah 500 pohon ini dilihat dari nilai MSE OOB untuk pohon 500. Penentuan jumlah pohon bisa menggunakan *default* atau ketetapan yang ada menurut breimen *ntree*= 500 , tetapi terlebih dahulu dibandingkan agar hasilnya akurat.

5.2.2 Pemilihan Nila*i m* (Random Subset) berdasarkan Nilai MSE

Pada algoritma *random forest* terdapat pemilihan variabel penjelas secara random subset untuk menentukan variabel pemisahnya. Banyaknya sampel pada random subset dinotasikan dengan *m*. Pada Penelitian ini akan dicobakan nilai *m* yang berbeda untuk mendapatkan nilai prediksi yang terbaik. Nilai yang dicobakan yaitu *m*=5, *m*=2.5 dan *m*=10. Berikut adalah tabel perbandingan dengan menggunakan nilai MSE OOB yang dihasilkan pada pohon yang telah ditumbuhkan.

Tabel 5.2 Nilai MSE pada Masing-Masing *m* yang dicobakan dengan jumlah pohon yang telah ditentukan sebelumnya

Nilai *m*	MSE OOB
m = 5	2.30546
m = 2.5	2.409421
m = 10	2.416784

Berdasarkan **tabel 5.2** diatas, dari semua nilai *m* yang telah diujikan didapatkan hasil *m*=5 dengan nilai *Mean Square of Residual*nya paling kecil yaitu 2.305, ini menunjukkan bahwa nilai *m* = 5 keakurasiannya lebih tinggi diantara nilai m yang diuji cobakan.

Dari hasil perbandingan ini dapat disimpulkan bahwa dari 10 variabel prediksi jumlah variabel yang dipilih pada random subset dengan hasil yang paling akurat adalah memilih 5 kemungkinan variabel. Hal ini berarti, dengan jumlah prediktor dengan jumlah 10 variabel akan menghasilkan prediksi yang beragam jika hanya memilih 5 variabel secara random untuk dijadikan pemisahan dan dari kelimas variabel tersebut akan dipilih satu variabel untuk dijadikan pemisah. Sehingga, variabel penjelas dengan nilai penurunan *error* yang kecil dan lemah memiliki peluang lebih besar untuk terpilih menjadi variabel pemisah. Dengan demikian akan mengurangi terjadinya *overfitting* dan jika dirata- ratakan hasil setiap pohonnya akan mendekati nilai aslinya. Selanjutnya untuk memprediksi digunakan pohon regresi dengan nilai *m*=5 untuk memprediksi.

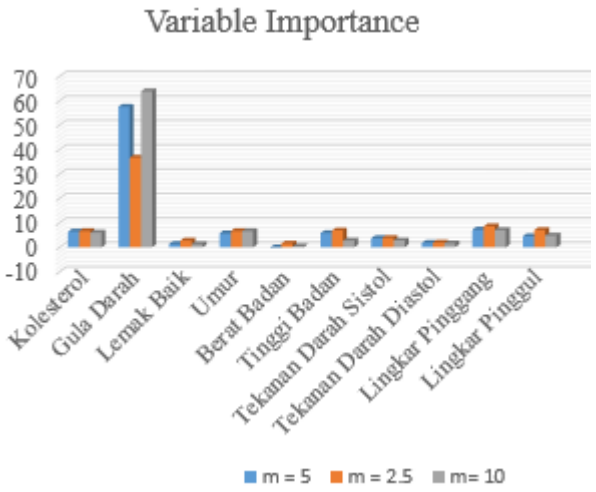
5.2.3 *Variable Importance*

Setelah selesai menentukan nilai *m* maka dalam analisis regresi *Random Forest* yaitu salah satu *output* yang dihasilkan dalam analisis Regresi *Random Forest* adalah nilai kepentingan masing-masing variabel penjelas dalam memprediksi nilai variabel bebasnya. *Variable Importance* mutlak adanya untuk menentukan faktor mana yang paling mempengaruhi seseorang beresiko terkena penyakit diabetes. Berikut adalah tabel *Variable Importance* yang dihasilkan *Random Forest* pada semua nilai *m*.

Tabel 5.3 *Variable Importance*

Variabel	*Variable Importance*		
	m = 5	*m* =2.5	*m*= 10
Kolesterol	6.46	6.62	6.01
Gula Darah	57.56	36.73	63.86
Lemak Baik	1.37	2.73	1.31
Umur	5.73	6.64	6.61
Berat Badan	-0.04	1.54	0.33
Tinggi Badan	5.85	6.88	2.71
Tekanan Darah Sistol	3.73	3.81	2.77
Tekanan Darah Diastol	1.83	1.94	1.60
Lingkar Pinggang	7.21	8.55	7.07
Lingkar Pinggul	4.52	7.03	4.78

Berikut ini adalah gambar diagram batang yang menggambarkan nilai *Variable Importance* masing – masing *m* yang dicobakan sebagai penjelas.



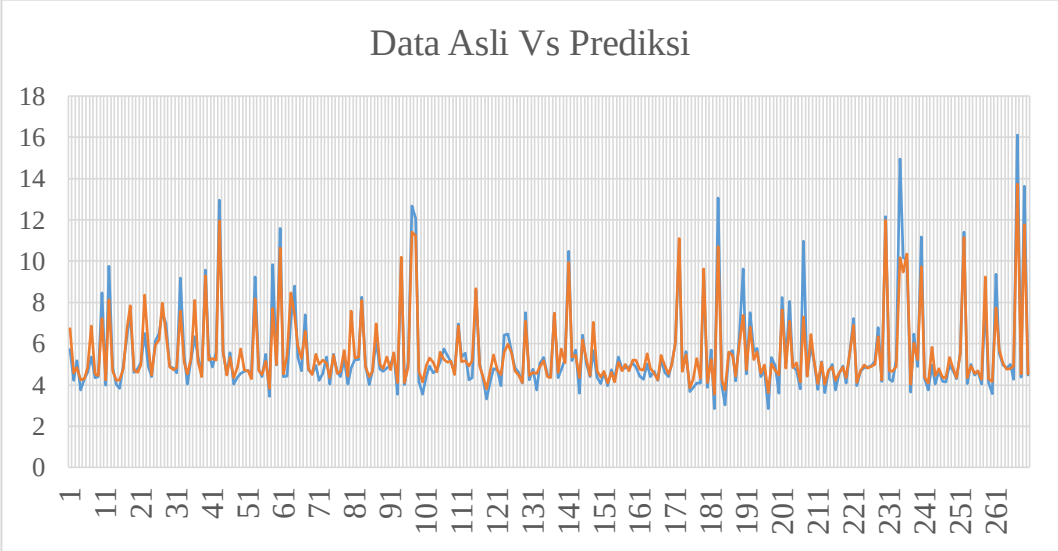
Gambar 5.4 Diagram Batang *Variable Importance*

Dari hasil diagram batang dan juga tabel variabel *Importance* diatas nilai *m* yang dicobakan tidak mempengaruhi urutan kepentingan suatu variabel namun hanya berpengaruh nilai kepentingan variabel saja. Dari semua nilai *m* yang dicobakan penulis menggunakan *m*=5 karena pada waktu pembentukkan pohon

nilai MSE dari *m*=5 memiliki nilai yang paling kecil diantara yang lainnya sehingga didapatkan variabel yang berpengaruh pada penyakit diabetes adalah kadar Gula Darah. Hal ini berarti variabel Gula Darah dapat menurunkan resiko seseorang terkena penyakit diabetes terbanyak dibandingkan 9 variabel yang lainnya. Kepentingan variabel lain setelah variabel Gula Darah adalah lingkar pinggang, Kolesterol, Tinggi Badan, Umur, lingkar pinggul, Tekanan darah sistol, Tekanan darah diastol, Lemak Baik dan Berat Badan.

5.2.4 Hasil Prediksi Regresi Random Forest

Tujuan utama regresi adalah untuk memprediksi respon dari model yang telah dihasilkan. Semakin dekat data prediksi dengan data sesungguhnya maka akan baik modelnya. Berikut ini akan ditampilkan grafik data asli dengan data prediksi



Gambar 5.5 Grafik Asli Vs Grafik Prediksi

Berdasarkan **gambar 5.5** diatas telihat bahwa nilai prediksi hampir mendekati nilai pada data asli dengan nilai MSEnya adalah 0.456015 Dari penjelasan ini dapat ditunjukkan bahwa dengan menggunakan algoritma *random forest* dapat menyelesaikan kasus regresi pada regresi dengan cara yang mudah. Dari kumpulan pohon keputusan dengan algoritma yang tepat, dapat digunakan untuk mengeksplorasi data. Sehingga dapat mengetahui hubungan antar variabel

respon dan prediktor nya dan selanjutnya dapat menghasilkan prediksi dengan keakuratan yang tinggi. Keakuratan yang tinggi dilihat dari nilai MSEnya semakin kecil nilai MSEnya maka semakin akurat hasil yang didapatkan, Berikut ini adalah tabel hasil prediksiya

Tabel 5.4 Hasil Prediksi

No	Glikogen Hemoglobin (Y)	$\hat{y}$
1.	4.31	4.41
2.	4.44	4.56
3.	4.64	5.26
4.	4.63	5.26
5.	7.72	6.75
6.	4.81	4.82
7.	4.84	4.78
8.	4.84	4.82
9.	5.78	5.52
10. 366	4.77 4.49	4.89 4.54

Berdasarkan **Tabel 5.4** hasil prediksi diatas dapat diketahui bahwa nilai Glikogen Hemoglobin dengan nilai prediksinya itu tidak jauh berbeda sebagai contoh data pertama nilai glikogen hemoglobinnya adalah 4.31 dan hasil prediksinya sebesar 4.41. Glikogen dihemoglobin atau biasa disebut HbA1c dikatakan dalam batas baik ketika nilainya < 6.5 mmol/L dan begitu seterusnya untuk data kedua hingga data yang ke 366.

5.3 Regresi Logistik Biner

Tidak seperti regresi linier biasa, regresi logistik tidak mengasumsikan hubungan antara variabel independen dan dependen secara linier. Didalam analisis regresi logistik biner variabel yang digunakan yaitu klasifikasi Diabetes Melitus sebagai variabel Dependen dengan 7 variabel independennya yaitu Kolesterol (X1), Gula darah (X2) Lemak Baik (X3), Lokasi (X4), Jenis Kelamin (X5), Tekanan Darah Diastol (X6), Tekanan Darah Sistol (X7). Berikut ini adalah hasil analisis regresi logistik biner

5.3.1 Uji Signifikansi

Pengujian signifikansi parameter dilakukan untuk mengetahui faktor-faktor yang memiliki pengaruh signifikan terhadap status penyakit diabetes melitus seseorang. Pengujian Signifikansi parameter dilakukan secara *Overall* dan Parsial.

1. Uji Overall

Pengujian secara Overall dilakukan dengan statistik likelihood rasio untuk mengetahui pengaruh keseluruhan faktor pada model

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	117,785	7	,000
	Block	117,785	7	,000
	Model	117,785	7	,000

Gambar 5.6 Hasil Uji Overall

Berdasarkan **Gambar 5.6,** maka dapat dapat dilakukan uji *overall* sebagai berikut :

a) Hipotesis

H_0 : $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$

H_1 : ada salah satu $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8 \neq 0$

b) Tingkat signifikansi

$\alpha = 5\% = 0.05$

c) Daerah Kritis

H_0 ditolak jika *p-value* < α

d) Statistik uji

(*p-value*) = 0.000

e) Keputusan

p-value < α (0,000< 0,05) → Tolak H_0

f) Kesimpulan

Dengan menggunakan tingkat kepercayaan 95%, maka berdasarkan uji yang telah dilakukan dapat disimpulkan bahwa ada minimal salah satu β_0, β_1, β_2, β_3, β_4, β_5, β_6, β_7, $\beta_8 \neq 0$ yang berpengaruh terhadap variabel terikat yaitu variabel glikogen hemoglobin sehingga model tersebut sesuai.

2. Uji Parsial

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Kolesterol	-,558	,287	3,770	1	,052	,573	,326	1,005
	Gula_Darah	-2,486	,311	63,982	1	,000	,083	,045	,153
	Lemak_Baik	-,206	,474	,189	1	,664	,814	,321	2,061
	Lokasi	-,617	,487	1,608	1	,205	,539	,208	1,400
	Jenis_Kelamin	,609	,502	1,475	1	,225	1,839	,688	4,918
	Tekanan_Darah_Sistol	,728	1,104	,434	1	,510	2,070	,238	18,032
	Tekanan_Darah_Diastol	-,001	,880	,000	1	,999	,999	,178	5,603
	Constant	3,153	1,287	6,000	1	,014	23,398		

a. Variable(s) entered on step 1: Kolesterol, Gula_Darah, Lemak_Baik, Lokasi, Jenis_Kelamin, Tekanan_Darah_Sistol, Tekanan_Darah_Diastol.

Gambar 5.7 Hasil Uji Parsial

Berdasarkan **Gambar 5.7** dapat dilakukan uji parsial sebagai berikut :

a) Hipotesis

H_0 : $\beta_0 = 0, \beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0, \beta_5 = 0, \beta_6 = 0, \beta_7 = 0,$

H_1 : $\beta_0 \neq 0, \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0, \beta_4 \neq 0, \beta_5 \neq 0, \beta_6 \neq 0, \beta_7 \neq 0,$

b) Tingkat signifikansi

α=5% =0.05

c) Daerah Kritis

H_0 ditolak jika *p-value* < α

d) Statistik uji

***P-value* untuk β_0**	**0.014**
***P-value* untuk β_1**	**0.052**
***P-value* untuk β_2**	**0.000**
***P-value* untuk β_3**	**0.664**
***P-value* untuk β_4**	**0.205**
***P-value* untuk β_5**	**0.225**
***P-value* untuk β_6**	0.510
***P-value* untuk β_7**	0.999

e) Keputusan

β_0	***p-value* < α (0.014 < 0.05)**	**Tolak H_0**
β_1	***p-value* > α (0.052 > 0.05)**	**Gagal tolak H_0**
β_2	***p-value* > α (0.000 < 0.05)**	**Tolak H_0**
β_3	***p-value* > α (0.664 > 0.05)**	**Gagal tolak H_0**
β_4	***p-value* > α (0.205 > 0.05)**	**Gagal tolak H_0**
β_5	***p-value* > α (0.225 > 0.05)**	**Gagal tolak H_0**
β_6	***p-value* > α (0.510 > 0.05)**	**Gagal tolak H_0**
β_7	***p-value* > α (0.999 > 0.05)**	**Gagal tolak H_0**

f) Kesimpulan

Dengan menggunakan tingkat kepercayaan 95%, maka dari uji di atas dapat disimpulkan bahwa $\beta_0 \neq 0, \beta_1 = 0, \beta_2 \neq 0, \beta_3 = 0, \beta_4 = 0, \beta_5 = 0, \beta_6 = 0, \beta_7 = 0, \beta_8 = 0$ sehingga model tersebut tidak sesuai

Selanjutnya melakukan uji *hosmer and lemeshow test* untuk melihat apakah data empiris cocok atau tidak dengan model, dengan kata lain diharapkan tidak ada perbedaan antara data empiris dengan model. Berikut uji hipotesisnya :

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7,338	8	,501

Gambar 5.8 Hasil *Uji Hosmer and Lemeshow Test*

a) Hipotesis

H_0 : Model sesuai dengan data.

H_1 : Model tidak sesuai dengan data.

b) Tingkat signifikansi

α=5% =0.05

c) Daerah Kritis

H_0 ditolak jika *p-value* < α

d) Statistik uji:

(*p-value*) = 0,363

e) Keputusan:

p-value > α (0,501 > 0,05) → Gagal Tolak H_0

f) Kesimpulan

Dengan menggunakan tingkat kepercayaan 95%, maka dari uji di atas dapat disimpulkan bahwa model yang didapatkan sesuai dengan data.

Selanjutnya untuk melihat tingkat kebenaran dari model yang telah dihasilkan maka dapat diinterpretasikan dari *model summary* berikut

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	151,122[a]	,275	,529

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

Gambar 5.9 *Output* Model *Summary*

Berdasarkan **Gambar 5.9** dapat dilihat bahwa nilai *Nagelkerke R Square* adalah 0,529 yang berarti bahwa variabel-variabel bebas mampu menjelaskan varians faktor-faktor yang mempengaruhi seseorang terkena diabetes Melitus sebesar 52,9 % dan sisanya 47,1% dijelaskan oleh faktor lain.

Selanjutnya dikarenakan model yang diperkirakan tidak sesuai maka dilakukan kembali analisis regresi seperti sebelumnya dengan mengeluarkan variabel *independent* yang paling tidak signifikan yaitu yang memiliki nilai *p-value* $> \alpha$. Apabila terdapat lebih dari satu variabel yang tidak siginifikan maka variabel yang paling tidak signifikan yang dikeluarkan.

Pengujian diulangi sebanyak 5 kali dengan mengeluarkan variabel yang tidak signifikan. Maka setelah dilakukan analisis regresi sebanyak 6 kali diperoleh perkiraan model regresi kembali yaitu :

$$\pi(X) = \frac{\exp(3.153 - 0.558X1 - 2.486X2)}{1 + \exp((3.153 - 0.558X1 - 2.486X2))}$$

Berdasarkan perkiraan model regresi yang telah didapatkan maka dapat dilakukan uji kesesuaian model sebagai berikut :

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	114,327	2	,000
	Block	114,327	2	,000
	Model	114,327	2	,000

Gambar 5.10 *Omnibus Tests of* Model *Coefficients*

1. Uji *Overall*

Berdasarkan **Gambar 5.10** maka dilakukan uji *overall* dengan hipotesis sebagai berikut :

a) Hipotesis

H_0 : $\beta_0 = \beta_1 = \beta_2 = 0$

H_1 : ada salah satu $\beta_0, \beta_1, \beta_2 \neq 0$

b) Tingkat signifikansi

α=5% =0.05

c) Daerah Kritis

H_0 ditolak jika *p-value* < α

d) Statistik uji:

(*p-value*) = 0,000

e) Keputusan:

p-value < α (0,000 < 0,05) → Tolak H_0

f) Kesimpulan

Dengan menggunakan tingkat kepercayaan 95%, maka dari uji di atas dapat disimpulkan bahwa ada salah satu $\beta_0, \beta_1, \beta_2 \neq 0$ sehingga model tersebut sesuai.

Selanjutnya, ntuk mengetahui parameter mana yang tidak sama dengan 0 maka dapat dilakukan uji parsial dari *output* berikut :

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Kolesterol	-,646	,281	5,299	1	,021	,524	,302	,908
	Gula_Darah	-2,343	,279	70,477	1	,000	,096	,056	,166
	Constant	3,661	,397	85,262	1	,000	38,913		

a. Variable(s) entered on step 1: Kolesterol, Gula_Darah.

Gambar 5.11 *Output Variables in the Equation*

2. Uji parsial

Berdasarkan **Gambar 5.11** dilakukan uji parsial dengan hipotesis sebagai berikut :

a) Hipotesis

H_0 : $\beta_0 = 0, \beta_1 = 0, \beta_2 = 0$

H_1 : $\beta_0 \neq 0, \beta_1 \neq 0, \beta_2 \neq 0$

b) Tingkat signifikansi

α=5% =0.05

c) Daerah Kritis:

H_0 ditolak jika *p-value* < α

d) Statistik uji

***P-value* untuk β_0**	**0.000**
***P-value* untuk β_1**	**0.021**
***P-value* untuk β_2**	**0.000**

e) Keputusan

β_0	***p-value* > α (0.000 < 0.05)**	**Tolak H_0**
β_3	***p-value* < α (0.012 < 0.05)**	**H_0 ditolak**
β_7	***p-value* < α (0.019 < 0.05)**	**H_0 ditolak**

f) Kesimpulan

Dengan menggunakan tingkat kepercayaan 95%, maka dari uji di atas dapat disimpulkan bahwa $\beta_1 \neq 0$, $\beta_2 \neq 0$ dan $\beta_0 \neq 0$, sehingga model tersebut telah sesuai.

Berdasarkan uji kesesuaian model yang telah dilakukan maka didapatkan dua variabel yang sesuai dengan model tersebut yaitu "Kolesterol dan Gula Darah". Sehingga, didapatkan model yang dapat digunakan untuk memprediksi adalah :

$$\pi(X) = \frac{\exp(3.661 - 0.646X1 - 2.343X2)}{1 + \exp((3.661 - 0.646X1 - 2.343X2))}$$

dimana :

X_1 = Kolesterol

X_2 = Gula Darah

Berdasarkan pengujian yang telah dilakukan, model tersebut merupakan model terbaik yang diperoleh dari kasus tersebut. Sehingga praktikan akan melakukan uji *hosmer and lemeshow test* untuk melihat apakah data empiris cocok atau tidak dengan model sehingga diharapkan tidak ada perbedaan antara data empiris dengan model. Berikut adalah uji hipotesisnya :

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	2,224	2	,329

Gambar 5.12 *Output Hosmer and Lemeshow Test*

a) Hipotesis

H_0 : Model sesuai dengan data

H_1 : Model tidak sesuai dengan data

b) Tingkat signifikansi

α=5% =0.05

c) Daerah Kritis

H_0 ditolak jika *p-value* < α

d) Statistik uji:

(*p-value*) = 0,329

e) Keputusan:

p-value > α (0,329 > 0,05) → Gagal Tolak H_0

f) Kesimpulan

Dengan menggunakan tingkat kepercayaan 95%, maka dari uji di atas dapat disimpulkan bahwa model yang didapatkan sesuai dengan data.

Berdasarkan uji *hosmer and lemeshow test* yang telah dilakukan, maka keputusan yang diperoleh praktikan semakin yakin bahwa model yang dihasilkan sesuai dengan data aslinya. Sehingga model yang telah dihasilkan dapat dipakai untuk memprediksi apakah faktor-faktor yang mempengaruhi seseorang terkena penyakit diabetes melitus. Selanjutnya untuk melihat tingkat kebenaran dari model yang telah didapatkan, maka dapat diinterpretasikan dari *model summary* berikut :

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	154,580[a]	,268	,516

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

Gambar 5.13 *Output* Model Summary

Berdasarkan **Gambar 5.13** dapat dilihat bahwa nilai *Nagelkerke R Square* adalah 0,516, artinya variabel-variabel bebas mampu faktor-faktor yang mempengaruhi seseorang terkena penyakit diabetes melitus sebesar 51,6% dan sisanya yaitu sebesar 48,4% dijelaskan oleh faktor lain. Jadi, dari model yang didapatkan hanya 51,6% tingkat kebenaran yang didapat jika melakukan prediksi dari model yang telah signifikan.

Odds ratio merupakan ukuran rasio perbandingan kemungkinan peristiwa terjadi dalam satu kelompok dengan kemungkinan hal yang sama terjadi di kelompok lain. Berdasarkan **Gambar 5.11** dapat dilihat nilai *odds ratio* untuk

parameter kolesterol adalah -0,646 (dilihat dari nilai *Exp(B)*), artinya apabila variabel lain bernilai konstan maka nilai resiko seseorang terkena penyakit diabetes melitus berubah sebesar -0,646 setiap satu satuan kolesterol dengan nilai estimasinya yaitu: $e^{-0.646}-1=-0.475$, artinya semakin tinggi nilai kolesterol diperkirakan nilai *odds* untuk peluang seseorang beresiko terkena penyakit diabetes melitus akan bertambah sebesar -047,5% pada variabel X_2 yang tetap.

Kemudian untuk nilai *odds ratio* parameter Gula Darah adalah -2,343, artinya apabila variabel lain bernilai konstan maka resiko seseorang terkena penyakit diabetes melitus berubah sebesar -2,343 setiap satu satuan gula darah. dengan nilai estimasinya adalah $e^{-2,343}-1=-0.9123$, artinya semakin tinggi gula darah diperkirakan nilai *odds* untuk peluang seseorang beresiko terkena penyakit diabetes melitus akan bertambah sebesar -091,23% pada variabel X_1 yang tetap.

Dengan demikian *odds ratio* antara kolesterol dan gula darah adalah $\frac{-0,646}{-2,343}=0,275$, artinya kadar kolesterol akan mempengaruhi kadar gula darah pada seseora ng yang beresiko terkena penyakit diabetes melitus sebesar 0,275 kali. Kolesterol lamanya bertempat tinggal akan mempengaruhi seseorang terkena penyakit diabetes melitus sebesar 0,275 kali jika dibandingkan dengan gula darah. Sedangkan untuk *odds ratio* antara gula darah dengan kolesterol adalah $\frac{-2,343}{-0,646}=3,626$, artinya gula darah dapat mempengaruhi seseorang terkena penyakit diabetes melitus sebesar 3,626 kali dibandingkan dengan kadar kolesterolnya.

BAB VI
KESIMPULAN DAN SARAN

6.1 Kesimpulan

Berdasarkan hasil analisis dan pembahasan pada bab sebelumnya, penulis dapat menarik kesimpulan sebagai berikut :

1. Berdasarkan hasil statistika deskriptif profilisasi penderita diabetes dan bukan penderita diabetes melitus didapatkan bahwa jumlah penderita Diabetes Melitus di wilayah buckingham berjumlah 22 orang dan untuk wilayah louisa 22 orang dari kedua wilayah tersebut tidak jauh berbeda jumlah penderita Diabetes Melitus. Kemudian untuk jumlah bukan penderita Diabetes Melitus wilayah louisa memiliki jumlah yang banyak yaitu 168 orang dibandingkan di wilayah Loiusa yaitu berjumlah 154 Orang. Berdasarkan uji proporsi didapatkan bahwa tidak ada perbedaan jumlah penderita diabetes melitus dan jumlah bukan penderita diabetes melitus di buckingham dan louisa. Berdasarkan uji independen t-test dari 11 variabel 2 diantaranya nilai berat badan dan tekanan darah diastol didapatkan hasil tidak ada perbedaan penderita diabetes melitus dan bukan penderita diabetes melitus.
2. Berdasarkan analisis Regresi Random Forest, banyaknya pohon yang terpilih adalah 500 dengan *m* yang dicobakan adalah 5. Faktor yang mempengaruhi seseorang terkena penyakit diabetes melitus berdasarkan hasil analisis regresi random forest adalah gula darah dengan nilai kepentingan sebesar 57.56 paling tinggi diantara ke sepuluh variabel yang lainnya dan juga nilai MSE dari hasil prediksi sebesar 0.456015.

6.2 Saran

Berdasarkan analisis dan kesimpulan , dapat diberikan saran sebagai berikut :

1. Untuk penelitian selanjutnya diharapkan bisa membandingkan regresi random forest dengan metode lainnya. Dan juga dapat menggunakan sub model algoritma yang lain selain CART.

2. Data yang digunakan untuk selanjutnya bisa menggunakan data penyakit diabetes tertinggi di asia tenggara. Misalnya Indonesia, china atau negara lain yang tingkat penderita diabetesnya tinggi.
3. Untuk Penelitian selanjutnya bisa menggunakan pohon lebih dari 500 pohon.

DAFTAR PUSTAKA

(2018, March 28). Diambil kembali dari https://www.cnnindonesia.com/gaya-hidup/20151202130942-255-95451/idf-2040-pengidap-diabetes-dunia-642-juta-orang

Aiswarya Iyer, S. J. (2015). Diagnosis Of Diabetes Using Classification Mining Techniques. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, Vol.5, No.1.

American. Diabetes Assosiation. (2014). Diagnosis and Classification of Diabetes Melitus. *Diabetes Care*, -.

Avin, R. (2016, April). Diambil kembali dari https://media.iyaa.com/article/2016/04/penderita-diabetes-tertinggi-ada-di-5-negara-3438864.html

Bain L, J., & Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistic Second Edition.* California: Duxbury Press.

Binarwati, L., Mukhlash, I., & Soetrisno. (2017). Implementasi Algoritma Genetika untuk Optimalisasi Random Forest Dalam Proses Klasifikasi Penerimaan Tenaga Kerja Baru Studi Kasus PT.XYZ. *JURNAL SAINS DAN SENI ITS Vol. 6, No. 2* , 2337-3520.

Breiman, L. (2001). Random Forest. *Machine Learning*, 5-32.

Breiman, L., & Friedman JH, O. R. (1984). *Classification and Regression Trees.* New York: Chapman & Hall.

Breimen, L., & Cutler, A. (2012). *Package "randomForest".*

Centers for Disease Control and Prevention (CDC). (2017). *Estimates of Diabetes and Its Burden in the United States.* United States: National Diabetes Statistics Report.

Centers For Disease Control and Prevention. (2017). *National Diabetes Statistic Report.* United States: Centers For Disease Control and Prevention (CDC).

CNN Indonesia. (2015, December 2). Diambil kembali dari https://www.cnnindonesia.com/gaya-hidup/20151202130942-255-95451/idf-2040-pengidap-diabetes-dunia-642-juta-orang

Dewi, K. (2011). Penerapan Metode Random Forest dalam Driver Analysis. *Forum Statistika dan Komputasi*, 35-43.

Eubank, R. (1988). *Spline Smoothing and Nonparametric Regression.* New York: Marcel Dekker.

Fernanda, S. I., Ratnawati, D. E., & Adikara, P. P. (2017). Identifikasi Penyakit Diabetes Mellitus Menggunakan Metode Modified K-Nearest Neighbor (MKNN). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 507-513.

Firmani, A. N. (2016). Penyelesaian Regresi Semiparametrik dengan Menggunakan Regresi Random Forest. *Skripsi*. Yogyakarta: Universitas Gadjah Mada.

Hartini, S. (2009). *Diabetes ? Siapa Takut!!* Bandung: Qanita.

Jones, Z., & Linder, F. (2015). *Exploratory Data Analysis Using Random Forest.* Chicago: MPSA Conference.

Mogensen, U. B., Ishwaran, H., & Gerds, T. A. (2012). Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. *Journal of Statistical Software Volume 50, Issue 11*, Issues 11.

Nugroho, R. A., Tarno, & Prahutama, A. (2017). Klasifikasi Pasien Diabetes Melitus Menggunakan Metode Smooth Support Vector Machine (SSVM). *JURNAL GAUSSIAN Volume 6, Nomor 3*, 439-448.

Nugroho, R., Tarno, & Prahutama, A. (2017). *Klasifikasi Pasien Diabetes Melitus Menggunakan Metode Smooth Support Vector Machine (SSVM)*. *JURNAL GAUSSIAN Volume 6 Nomor 3* , 439-488.

Riyad, A., & Tahani, D. (2016). Diagnosis of Diabetes by Applying Data Mining Classification Techniques. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 7.

Wahyuni, S. (2017). *Faktor-Faktor Yang Berhubungan Dengan Penyakit Diabetes Melitus (DM) Daerah Perkotaan di Indonesia Tahun 2017*. *Skripsi*. Jakarta: Universitas Islam Syarif Hidayatullah.

Wikipedia, Analisis Regresi, (online) https://id.wikipedia.org/wiki/Analisis_regresi Diakses Kamis, 26 April 2018 Pukul 12.23 WIB

www.kaggle.com , Diakses pada hari sabtu , 10 Maret 2018 pada Jam 13.00 WIB.

LAMPIRAN

Lampiran 1. Dataset Diabetes

Kolesterol	Gula Darah	Lemak Baik	Glikogen Hemoglobin	Lokasi	Umur	Jenis Kelamin	Berat Badan	Tinggi Badan	Tekanan Darah Sistol	Tekanan Darah Diastol	Lingkar Pinggang	Lingkar Pinggul
203	82	56	4.31	0	46	0	62	121	118	59	29	38
165	97	24	4.44	0	29	0	64	218	112	68	46	48
228	92	37	4.64	0	58	0	61	256	190	92	49	57
78	93	12	4.63	0	67	1	67	119	110	50	33	38
249	90	28	7.72	0	64	1	68	183	138	80	44	41
248	94	69	4.81	0	34	1	71	190	132	86	36	42
195	92	41	4.84	0	30	1	69	191	161	112	46	49
177	87	49	4.84	0	45	1	69	166	160	80	34	40
263	89	40	5.78	0	55	0	63	202	108	72	45	50
242	82	54	4.77	1	60	0	65	156	130	90	39	45
215	128	34	4.97	1	38	0	58	195	102	68	42	50
238	75	36	4.47	1	27	0	60	170	130	80	35	41
191	76	30	4.67	1	36	1	69	183	100	66	36	40
213	83	47	3.41	1	33	0	65	157	130	90	37	41
255	78	38	4.33	1	50	0	65	183	130	100	37	43

230	112	64	4.53	1	20	1	67	159	100	90	31	39
194	81	36	5.28	1	36	1	64	126	110	76	30	34
196	206	41	11.24	0	62	0	65	196	178	90	46	51
186	97	50	6.49	0	70	1	67	178	148	88	42	41
234	65	76	4.67	0	47	1	67	230	137	100	45	46
203	299	43	12.74	0	38	0	69	288	136	83	48	55
281	92	41	5.56	0	66	0	62	185	158	88	48	44
228	66	45	4.61	0	24	0	61	113	100	70	33	38
179	80	92	4.18	0	41	0	72	118	144	112	28	36
232	87	30	5.1	0	37	1	68	252	140	95	43	47
254	84	52	4.52	0	43	0	62	145	125	70	31	38
215	72	42	4.37	1	40	1	70	189	180	122	37	39
177	101	36	5.11	0	42	0	65	174	146	94	37	40
182	85	43	4.47	0	52	1	68	139	130	90	29	35
265	330	34	15.52	0	61	1	74	191	170	88	39	41
182	85	37	5.66	0	61	0	69	174	176	86	49	43
199	87	63	3.67	0	25	1	66	118	120	78	32	34
183	81	60	4.03	0	47	0	66	186	140	97	39	44
194	86	67	2.68	0	35	1	66	159	115	64	31	35
173	80	57	6.21	0	57	1	71	145	124	64	31	36
182	206	43	7.91	0	70	1	69	214	158	90	45	48
136	81	51	4.58	0	22	0	66	160	105	85	35	40

218	68	46	3.89	0	52	0	62	170	142	79	40	43
225	83	42	4.38	0	36	1	67	192	149	89	40	42
213	76	40	5.96	0	72	0	59	137	130	60	40	40
243	52	59	4.41	0	37	0	64	233	110	82	49	57
148	193	14	6.14	0	54	0	67	165	140	65	42	42
128	223	24	10.9	0	60	1	67	196	110	68	42	43
169	85	51	6.14	0	40	0	65	180	106	82	40	44
157	74	47	5.57	0	55	0	66	219	150	82	43	52
237	87	41	5.35	0	43	0	64	181	104	90	36	46
212	97	45	6.33	0	65	0	61	187	158	94	43	47
233	92	39	4.56	0	45	0	64	167	124	86	39	44
289	111	50	9.39	0	70	0	60	220	126	80	51	54
193	106	63	6.35	0	20	0	68	274	165	110	49	58
204	128	61	5.2	0	62	1	68	180	141	81	38	41
165	94	69	4.98	0	92	0	62	217	160	82	51	51
237	233	58	13.7	0	49	0	62	189	130	90	43	47
296	262	60	10.93	0	74	0	63	183	159	99	42	48
178	78	59	5.23	0	36	1	70	161	130	79	34	40
443	185	23	14.31	0	51	0	70	235	158	98	43	48
146	77	60	4.27	0	28	0	64	126	120	90	28	32
223	75	85	4.25	0	22	0	62	137	120	70	28	35
213	203	75	11.41	0	71	0	63	165	150	80	34	42

173	131	69	4.44	0	76	0	61	102	160	60	31	33
171	92	54	4.59	0	40	1	71	214	138	94	41	39
164	86	40	5.23	0	23	0	69	245	126	75	44	47
170	69	64	4.39	0	20	0	64	161	108	70	37	40
180	84	69	5.2	0	40	0	68	264	142	98	43	54
204	57	74	6.11	0	52	1	75	142	140	90	31	35
209	113	65	7.44	0	76	0	60	143	156	78	35	40
242	108	53	5.47	0	46	0	62	183	130	86	37	45
134	105	42	4.29	0	48	1	70	173	178	120	36	40
217	81	60	3.93	0	22	0	71	223	120	75	46	50
251	94	36	6.96	0	58	0	63	154	174	75	38	41
217	88	40	4.84	0	34	1	73	219	145	100	41	42
300	103	44	5.18	1	61	0	67	169	138	78	40	44
218	87	38	5.52	1	40	1	73	200	120	76	38	41
189	96	47	4.38	1	28	0	64	200	136	52	38	45
185	84	52	5.28	1	53	0	61	145	147	72	37	40
206	85	46	4.82	1	67	1	67	178	119	68	37	41
189	75	72	4.86	1	49	0	62	205	120	80	40	49
229	95	74	4.86	1	65	0	62	151	125	64	37	42
228	76	53	4.11	1	54	1	66	170	121	62	36	41
159	88	43	5.02	1	38	1	68	169	138	79	34	40
249	197	44	9.17	1	64	0	63	159	151	85	33	41

170	106	42	5.11	1	41	0	61	110	103	64	29	30
174	125	44	5.07	1	67	1	68	198	119	72	36	43
204	62	70	4.84	1	27	0	67	185	110	90	35	44
203	84	75	4.1	1	21	0	63	142	125	85	28	39
241	86	63	4.79	1	41	0	59	139	112	72	29	39
245	120	39	7.79	1	47	0	63	156	142	102	35	39
143	91	37	5.15	1	61	0	65	220	160	92	40	50
224	341	33	10.15	1	65	1	67	197	160	80	42	43
168	69	45	4.17	1	28	0	63	200	111	65	42	46
184	79	39	4.05	1	41	1	69	154	136	96	34	39
199	130	48	5.44	1	37	0	61	203	136	84	42	51
158	91	48	4.31	1	50	1	71	180	136	90	36	40
209	176	55	9.77	1	57	0	61	150	115	68	36	39
214	111	59	3.89	1	28	1	68	204	130	90	40	41
293	85	94	5.17	1	31	0	67	200	110	90	41	42
227	105	44	5.71	1	83	0	59	125	150	90	35	40
283	83	74	4.22	1	26	1	72	227	158	104	41	44
186	74	76	5.17	1	36	1	69	150	138	82	31	38
273	94	49	3.76	1	53	0	64	174	160	96	34	43
193	77	49	4.31	1	19	0	61	119	118	70	32	38
194	80	34	4.61	0	63	1	73	175	131	88	34	39
174	173	34	5.35	0	50	1	70	263	159	99	51	64

225	84	82	4.36	0	41	1	71	156	150	80	31	40
268	85	51	4.41	1	48	1	70	120	150	105	32	35
195	108	46	8.45	1	59	0	67	172	150	102	38	43
179	70	52	3.98	1	34	1	72	170	138	82	31	39
215	119	44	9.76	1	63	0	63	158	160	68	34	42
185	76	58	4.83	0	23	1	76	164	124	78	32	40
132	99	34	4.01	0	21	0	65	169	112	62	39	43
175	91	42	3.84	1	23	0	65	235	110	80	44	50
179	81	35	4.95	0	36	0	63	125	110	76	33	36
228	115	61	6.39	0	71	0	63	244	170	92	48	51
181	177	24	7.53	0	64	1	71	225	130	66	44	47
160	100	36	4.62	1	43	0	64	140	180	110	37	40
188	77	45	4.79	1	31	0	67	227	122	70	47	53
168	101	59	5.09	1	44	0	64	160	130	88	40	43
318	270	108	6.51	1	60	0	65	167	132	72	38	44
192	109	44	4.86	0	43	0	64	325	141	79	53	62
209	87	34	4.41	0	48	0	63	121	111	62	32	38
129	110	42	6.13	0	56	1	74	151	140	75	34	38
160	122	41	6.49	0	55	0	67	223	136	83	43	48
160	196	33	7.51	0	49	1	71	266	150	98	49	45
211	48	34	6.97	1	58	1	67	177	162	78	38	43
262	93	43	4.9	1	33	0	63	170	110	68	33	46

201	81	87	4.81	0	48	0	68	146	145	95	32	41
263	82	92	4.58	0	66	0	66	121	104	64	31	33
219	112	73	9.18	0	59	1	66	170	146	92	37	40
191	83	88	5.46	0	45	0	67	151	130	90	33	38
171	97	69	4.04	0	52	1	71	159	125	72	33	39
219	112	73	5.23	0	76	1	64	105	125	82	29	33
347	197	42	6.34	0	36	1	70	277	140	86	51	49
269	73	34	5.37	0	41	0	62	160	126	90	39	41
164	71	63	4.51	0	20	1	72	145	108	78	29	36
181	255	26	9.58	0	50	1	71	320	140	86	56	49
190	84	44	5.55	0	43	0	62	163	135	88	40	45
218	126	32	4.87	1	35	1	69	169	139	90	39	41
223	90	48	5.6	0	47	0	65	232	120	86	46	54
254	342	37	12.97	0	75	1	68	210	151	87	44	45
236	102	36	5.63	0	62	1	76	160	150	80	35	39
176	92	55	4.5	0	31	0	62	145	110	72	36	42
158	91	31	5.56	1	50	1	70	215	138	89	40	45
181	83	44	4.03	1	39	0	66	255	140	98	46	54
151	85	48	4.38	1	33	1	69	308	110	90	52	58
271	121	40	4.57	1	81	0	64	158	146	76	36	43
190	92	44	4.66	1	27	0	65	210	150	106	39	47
118	95	39	4.71	1	47	0	64	123	140	76	30	36

168	82	44	4.4	0	33	0	66	118	98	66	29	35
254	121	39	9.25	0	67	1	68	167	161	118	36	39
193	77	45	4.74	0	42	0	75	186	125	90	37	46
187	84	64	4.4	0	21	0	63	158	138		39	43
212	79	49	5.49	0	51	0	65	145	230	120	38	42
170	76	60	3.44	0	27	0	63	119	122	86	28	37
215	110	36	9.82	1	51	0	67	282	142	78	52	59
199	85	59	4.96	1	71	1	69	171	136	86	38	40
140	385	31	11.59	1	50	1	69	172	138	66	37	41
216	79	46	4.41	1	54	0	65	138	132	80	33	39
204	113	35	4.44	0	59	1	73	187	148	76	38	37
193	248	24	7.14	0	59	0	66	189	140	90	38	45
267	133	34	8.81	1	40	0	59	204	118	69	40	47
201	106	53	5.35	1	58	1	66	215	186	102	46	44
204	120	44	4.69	1	72	1	65	167	140	72	45	46
246	104	62	7.4	1	66	0	66	189	200	94	45	46
229	91	43	4.73	1	23	1	72	180	110	78	34	41
172	101	46	4.52	1	42	0	65	165	118	68	33	45
197	120	37	4.95	1	43	1	71	179	146	98	37	44
205	79	32	4.21	1	75	1	69	204	136	90	44	42
219	106	50	4.56	1	65	0	63	233	140	90	40	53
174	90	36	5.35	1	34	1	71	210	142	92	37	43

192	89	30	4.04	1	37	1	71	195	136	96	36	43
206	94	44	5.49	1	61	0	63	199	180	96	41	47
160	71	44	4.64	1	36	0	64	185	110	80	39	45
216	109	86	4.4	1	45	0	67	147	140	102	32	38
236	111	82	5.24	1	68	0	61	119	142	96	29	37
206	112	33	4.03	1	41	0	62	184	104	80	39	44
143	371	46	4.81	1	68	1	67	158	138	82	37	43
235	91	37	5.23	0	79	0	65	134	142	70	34	38
169	95	29	5.22	0	62	1	66	251	118	72	50	47
283	145	39	8.25	0	63	0	61	200	190	110	44	48
174	93	77	4.95	0	55	1	70	140	118	86	32	33
271	103	90	4.01	0	55	0	63	114	180	105	30	37
203	94	62	4.67	0	27	0	67	209	140	80	34	43
188	174	24	6.17	1	66	1	68	210	160	78	45	48
293	87	120	4.76	1	63	0	64	179	142	80	47	45
215	80	100	4.66	1	78	1	65	109	170	88	33	34
207	77	46	4.82	0	68	1	55	130	199	115	29	33
179	77	72	4.97	0	31	1	66	145	131	79	33	38
202	81	55	5.5	0	64	0	62	167	190	118	44	47
211	98	40	3.55	0	40	0	68	179	110	76	37	43
211	225	29	10.09	0	61	0	63	144	190	100	40	42
151	74	47	4.01	0	28	1	69	130	135	75	29	35

171	85	61	5.1	0	34	0	63	164	120	80	34	43
342	251	48	12.67	0	63	0	65	201	178	88	45	46
179	236	63	12.07	0	55	1	75	186	122	74	38	38
155	58	69	4.17	0	26	1	73	174	110	76	30	35
200	56	51	3.55	0	40	0	62	105	125	64	26	33
198	118	46	4.44	0	68	0	63	124	130	70	32	38
240	88	49	4.92	0	82	0	63	170	180	86	41	46
192	56	42	4.59	0	60	0	62	134	130	70	31	40
145	84	54	4.73	0	30	0	65	165	102	56	33	42
269	59	66	5.14	0	41	1	67	191	130	73	38	41
240	96	57	5.74	0	54	0	65	175	152	100	37	43
266	82	54	5.41	0	47	1	68	142	118	78	35	39
188	88	51	5.13	0	50	0	61	147	160	66	34	41
222	82	87	4.64	0	51	0	66	110	150	110	28	37
142	155	25	6.96	0	45	1	69	204	165	115	40	43
268	90	48	5.36	0	38	0	63	181	142	100	38	46
174	105	117	5.53	0	20	1	70	187	132	86	37	41
194	54	57	4.26	1	63	1	70	181	184	76	37	42
196	115	62	4.34	1	50	1	67	140	176	110	35	37
207	187	46	8.57	1	44	0	67	201	150	74	46	49
204	89	56	5.02	1	48	1	68	196	170	96	38	42
189	84	46	4.36	1	41	0	63	153	130	80	32	40

179	77	50	3.33	0	29	1	68	170	122	68	38	39
159	100	54	4.18	0	76	1	66	188	116	53	40	41
260	68	60	4.78	0	69	0	59	179	158	98	45	48
228	79	37	4.74	0	26	1	72	259	122	90	48	49
242	74	55	3.97	0	70	0	66	200	140	65	41	47
227	98	66	6.42	0	25	1	71	162	123	82	35	39
208	122	51	6.48	0	42	0	62	141	118	78	33	40
208	95	32	5.6	0	56	1	68	183	131	75	36	39
209	89	43	4.85	0	31	0	67	160	108	58	30	44
163	83	57	4.61	0	31	0	65	120	136	86	29	40
201	100	46	4.1	0	27	0	65	145	121	75	32	35
237	118	45	7.51	0	73	0	64	174	162	75	38	44
176	90	34	4.24	0	32	0	63	252	100	72	45	58
146	79	41	4.76	0	19	0	60	135	108	58	33	40
231	70	110	3.75	0	71	0	63	155	150	78	33	41
241	92	40	5.04	0	27	0	63	179	120	75	40	42
305	91	44	5.34	0	31	1	71	211	100	60	40	45
149	77	49	4.5	0	20	0	62	115	105	82	31	37
183	69	51	4.37	0	31	0	66	190	125	70	41	47
235	109	59	7.48	0	62	0	63	290	175	80	55	62
244	101	39	4.36	0	44	1	71	168	140	89	36	39
199	153	77	4.74	0	36	0	66	255	118	66	47	52

224	85	30	5.26	0	36	1	69	205	150	99	37	41
173	225	31	10.47	0	47	1	73	260	150	98	42	47
192	124	31	5.17	0	30	1	72	250	142	79	43	51
157	91	34	5.7	0	63	1	69	166	106	82	39	38
172	117	56	3.59	0	48	0	63	170	130	82	35	42
170	67	33	6.42	0	65	1	69	182	140	65	42	39
215	97	46	5.03	0	59	0	63	176	140	70	34	44
214	67	47	4.41	0	37	0	64	145	108	76	34	42
195	171	29	5.68	0	78	1	66	172	130	82	40	40
230	86	37	4.39	0	23	1	71	277	150	99	50	49
206	90	38	4.07	0	38	0	69	167	138	90	36	47
147	86	34	4.62	0	38	1	69	205	130	96	39	41
135	88	34	3.96	0	29	0	65	123	118	61	26	37
179	75	36	4.75	0	23	0	65	183	120	80	43	45
163	69	48	4.31	0	29	0	62	99	125	60	30	36
191	74	33	5.35	1	40	1	72	270	136	70	45	49
138	95	40	4.8	1	38	0	60	138	140	90	31	39
184	92	36	4.81	1	40	0	63	285	142	98	50	60
181	101	44	4.88	1	29	1	68	180	130	78	38	42
224	98	44	5.05	1	78	0	63	160	150	81	36	45
293	115	54	4.87	0	50	1	71	170	131	75	34	39
198	92	62	4.43	1	60	1	70	163	126	78	36	40

152	103	32	4.27	1	40	0	52	187	148	82	38	49
277	119	62	5.03	1	60	0	61	128	140	86	33	39
219	105	63	4.4	1	40	0	62	153	106	82	36	44
182	74	44	4.67	1	30	0	62	125	132	80	31	39
135	88	47	4.21	1	21	1	69	155	110	68	31	39
277	88	45	5.24	1	63	0	64	223	220	100	45	54
212	82	68	4.61	1	63	1	70	161	180	110	37	40
162	76	40	4.4	1	43	1	67	216	100	70	41	44
207	102	43	5.01	1	46	0	63	179	212	114	38	46
255	100	34	6.06	1	64	1	68	227	134	74	44	47
404	206	33	10.75	1	56	1	69	159	162	88	38	39
239	97	55	4.69	1	35	1	74	170	122	62	32	38
220	95	58	5.63	1	59	0	66	138	138	80	32	38
165	76	46	3.69	1	22	0	63	114	112	78	28	35
243	74	42	3.85	1	43	0	64	239	128	90	48	53
149	138	50	4.09	1	26	0	62	174	148	92	38	46
178	64	52	4.1	1	41	0	65	188	130	76	35	46
190	228	57	9.28	1	43	0	65	198	110	64	40	49
226	97	70	3.88	1	20	0	64	114	122	64	31	39
132	83	40	5.7	1	28	0	68	225	136	86	41	52
160	82	41	2.85	1	30	0	63	143	172	124	33	40
204	173	37	13.06	1	66	1	67	146	138	78	36	48

164	91	67	3.97	1	20	0	70	141	122	86	32	39
155	81	70	3.03	1	32	0	65	151	120	68	33	40
251	118	38	5.51	1	38	0	64	248	110	80	49	58
198	86	66	5.68	1	61	1	74	152	138	76	33	38
179	90	60	4.2	1	26	0	60	130	138	84	32	40
223	88	42	6.44	1	74	0	62	165	250	100	41	46
207	71	41	9.62	1	72	1	70	180	138	88	39	40
244	89	92	4.54	1	21	1	71	163	116	76	34	39
245	119	26	7.51	1	36	1	66	179	150	92	37	42
191	81	53	5.63	1	42	0	61	156	138	84	36	42
221	120	83	5.77	1	66	0	64	130	110	64	31	38
173	85	58	4.4	0	43	0	69	210	130	75	44	47
138	81	45	4.7	0	57	1	73	164	148	81	31	37
203	71	78	2.85	1	45	1	66	115	135	88	30	34
260	67	46	5.34	1	44	0	62	159	140	94	36	43
166	77	68	4.95	1	27	1	72	141	110	58	33	38
180	92	34	3.59	0	63	1	69	169	145	72	35	39
159	172	28	8.23	0	65	1	70	181	142	81	43	49
207	75	44	5.06	0	30	1	72	180	118	62	35	41
191	155	58	8.06	0	31	0	62	237	140	87	53	56
231	84	91	4.9	0	33	1	69	163	140	70	35	38
184	76	42	4.71	0	66	1	74	185	130	75	40	41

164	94	58	3.8	0	28	0	67	180	128	94	39	43
220	60	66	10.97	0	26	1	70	150	136	88	33	39
180	76	46	4.43	1	40	0	64	146	128	82	37	43
216	155	30	5.91	1	38	1	68	145	110	60	34	37
261	101	83	5.12	1	52	0	64	198	152	92	42	49
172	70	36	3.78	1	22	0	64	148	90	48	35	38
249	81	28	5.12	1	51	0	65	200	122	90	43	46
189	80	40	3.62	1	45	1	69	190	140	75	39	44
225	74	36	4.66	1	53	0	63	182	126	80	38	46
193	75	49	5.01	1	21	0	61	220	130	82	40	52
219	78	67	3.75	1	53	0	64	179	135	100	39	47
156	86	34	4.55	1	37	0	67	212	122	74	48	51
224	71	42	4.92	1	34	0	60	165	135	80	34	46
181	77	46	4.09	1	30	0	66	257	162	108	47	55
306	92	56	5.58	1	74	1	69	184	140	72	39	41
219	130	44	7.22	1	45	1	67	218	172	110	41	45
150	80	38	3.97	1	35	1	73	179	138	92	32	37
185	67	59	4.65	1	50	0	64	228	142	90	42	54
226	100	65	4.83	1	27	1	69	289	130	100	48	51
206	83	68	4.88	1	52	1	69	153	140	98	36	40
199	81	36	4.93	1	42	0	67	235	178	100	47	52
239	85	63	5.16	1	39	1	60	144	162	90	33	42

235	106	37	6.78	1	73	1	65	183	134	78	43	46
184	99	36	4.16	1	28	1	67	154	124	94	35	38
242	297	34	12.16	1	53	1	69	216	142	96	43	45
307	87	58	4.28	1	49	1	67	181	120	80	41	42
204	94	54	4.16	1	55	0	66	202	140	90	43	47
212	88	36	5.22	1	37	0	64	160	124	82	37	45
203	90	51	14.94	1	60	0	59	123	130	72	36	41
219	173	31	10.16	1	56	0	65	197	100	50	41	50
226	279	52	10.07	1	84	0	60	192	144	88	41	48
217	75	54	3.66	1	20	0	67	187	110	72	40	45
157	92	47	6.48	1	80	1	71	212	156	88	47	48
235	102	42	4.9	1	60	1	69	186	148	98	40	42
252	161	87	11.18	1	80	0	62	162	160	100	44	41
204	71	55	4.33	1	29	0	64	120	110	70	33	38
188	84	46	3.75	1	43	0	66	152	122	80	37	41
194	95	36	4.97	1	63	0	58	210	140	100	44	53
215	64	84	4.04	1	37	0	59	148	140	100	32	42
179	105	60	4.68	1	20	0	58	170	140	100	34	46
202	84	33	4.17	1	44	1	68	157	125	80	33	37
194	87	65	4.14	1	54	1	69	129	170	96	30	37
227	85	26	4.98	1	58	1	70	211	144	82	38	43
337	85	62	4.66	1	35	1	72	189	124	84	36	44

255	83	90	4.29	1	52	1	70	120	170	110	30	33
162	90	46	5.56	1	60	0	63	121	110	64	32	34
289	267	38	11.41	1	59	1	68	169	142	79	36	38
217	87	40	4.07	1	33	0	62	186	140	90	42	46
209	91	36	5.01	1	37	1	70	262	130	94	42	48
214	77	48	4.48	1	40	1	72	222	120	84	40	44
302	81	57	4.65	1	38	0	67	222	128	82	41	51
179	85	52	4.05	1	32	0	62	179	140	96	37	47
279	270	40	8.11	1	60	0	68	224	174	90	48	50
144	81	28	4.13	1	30	1	72	165	118	78	31	38
270	73	40	3.58	1	42	1	66	185	146	94	39	41
196	120	67	9.37	1	52	0	62	147	144	94	34	42
221	126	48	5.53	1	59	0	62	177	130	78	39	45
210	81	81	4.96	1	78	1	66	145	110	70	38	39
169	104	58	4.82	1	25	0	60	154	140	95	40	42
179	85	50	4.99	1	37	1	66	136	190	94	33	39
301	90	118	4.28	1	89	0	61	115	218	90	31	41
296	369	46	16.11	1	53	1	69	173	138	94	35	39
284	89	54	4.39	1	51	0	63	154	140	100	32	43
194	269	38	13.63	1	29	0	69	167	120	70	33	40
199	76	52	4.49	1	41	0	63	197	120	78	41	48

Keterangan :

Lokasi : 0 = Buckingham

1 = Louisa

Jenis Kelamin : 0 = Perempuan

1 = Laki- Laki

Lampiran 2. ***Syntax Regresi Random Forest***

```
library(randomForest)
data=read.delim("clipboard")
data
#untuk m=5 dan ntree=500
set.seed(234)
rf1=Glikogen.Hemoglobin.rf<-
randomForest(Glikogen.Hemoglobin~ ., data=data, mtry=5,
ntree=500, importance=TRUE, na.action=na.omit)
print(rf1)
plot(rf1)
round(importance(rf1),2)

#untuk m=2.5 dan ntree=500
set.seed(234)
rf2=Glikogen.Hemoglobin.rf<-
randomForest(Glikogen.Hemoglobin~ ., data=data,
mtry=2.5, ntree=500, importance=TRUE,
na.action=na.omit)
print(rf2)
round(importance(rf2),2)


#untuk m=10 dan ntree=500
set.seed(234)
rf3=Glikogen.Hemoglobin.rf<-
randomForest(Glikogen.Hemoglobin~ ., data=data,
mtry=10, ntree=500, importance=TRUE, na.action=na.omit)
print(rf3)
round(importance(rf3),2)

#plot MSE OOB
```

```
plot(randomForest(Glikogen.Hemoglobin~.,
data,keep.forest=FALSE),main="Mean Squared Error")

#prediksi
newdata=read.delim("clipboard")
newdata
p<-predict(rf1,newdata)
p
```

Lampiran 3. Hasil Prediksi

No	Glikogen Hemoglobin (Y)	$\hat{y}$
1.	4.31	4.41
2.	4.44	4.56
3.	4.64	5.26
4.	4.63	5.26
5.	7.72	6.75
6.	4.81	4.82
7.	4.84	4.78
8.	4.84	4.82
9.	5.78	5.52
10.	4.77	4.89
11.	4.97	5.65
12.	4.47	4.65
13.	4.67	4.67
14.	3.41	3.93
15.	4.33	4.55
16.	4.53	4.62
17.	5.28	4.81
18.	11.24	10.19
19.	6.49	6.06

20.	4.67	4.71
21.	12.74	11.33
22.	5.56	5.71
23.	4.61	4.68
24.	4.18	4.26
25.	5.1	4.98
26.	4.52	4.59
27.	4.37	4.47
28.	5.11	4.87
29.	4.47	4.52
30.	15.52	13.36
31.	5.66	5.58
32.	3.67	3.85
33.	4.03	4.21
34.	2.68	3.6
35.	6.21	5.61
36.	7.91	8.35
37.	4.58	4.59
38.	3.89	4.41
39.	4.38	4.48
40.	5.96	6.55
41.	4.41	5.14
42.	6.14	7.5
43.	10.9	9.83
44.	6.14	5.36
45.	5.57	5.28
46.	5.35	5.12
47.	6.33	5.97
48.	4.56	4.75
49.	9.39	8.11
50.	6.35	5.84

51.	5.2	5.51
52.	4.98	5.41
53.	13.7	12.09
54.	10.93	10.88
55.	5.23	4.95
56.	14.31	12.34
57.	4.27	4.12
58.	4.25	4.21
59.	11.41	10.61
60.	4.44	5.46
61.	4.59	4.63
62.	5.23	4.86
63.	4.39	4.38
64.	5.2	5
65.	6.11	5.86
66.	7.44	7.23
67.	5.47	5.26
68.	4.29	4.58
69.	3.93	4.2
70.	6.96	6.38
71.	4.84	4.82
72.	5.18	5.45
73.	5.52	5.15
74.	4.38	4.45
75.	5.28	5.19
76.	4.82	4.91
77.	4.86	4.86
78.	4.11	5.05
79.	4.11	4.36
80.	5.02	4.82
81.	9.17	9.87

82.	5.11	4.8
83.	5.07	5.41
84.	4.84	5.03
85.	4.1	4.21
86.	4.79	4.94
87.	7.79	7.14
88.	5.15	5.3
89.	10.15	10.68
90.	4.17	4.31
91.	4.05	4.23
92.	5.44	5.5
93.	4.31	4.5
94.	9.77	9.71
95.	3.89	4.32
96.	5.17	4.98
97.	5.71	6.72
98.	4.22	4.58
99.	5.17	4.84
100.	3.76	4.26
101.	4.31	4.31
102.	4.61	4.83
103.	5.35	6.86
104.	4.36	4.54
105.	4.41	4.45
106.	8.45	7.22
107.	3.98	4.25
108.	9.76	8.12
109.	4.83	4.7
110.	4.01	4.21
111.	3.84	4.24
112.	4.95	4.8

113.	6.39	6.75
114.	7.53	7.84
115.	4.62	4.71
116.	4.79	4.61
117.	5.09	4.97
118.	6.51	8.37
119.	4.86	5.83
120.	4.41	4.48
121.	6.13	5.92
122.	6.49	6.18
123.	7.51	7.98
124.	6.97	6.49
125.	4.9	4.88
126.	4.81	4.74
127.	4.58	4.85
128.	9.18	7.58
129.	5.46	5.11
130.	4.04	4.53
131.	5.23	5.29
132.	6.34	8.1
133.	5.37	5.12
134.	4.51	4.41
135.	9.58	9.28
136.	5.55	5.21
137.	4.87	5.29
138.	5.6	5.2
139.	12.97	11.95
140.	5.63	5.61
141.	4.5	4.48
142.	5.56	5.33
143.	4.03	4.34

144.	4.38	4.88
145.	4.57	5.76
146.	4.66	4.71
147.	4.71	4.68
148.	4.4	4.31
149.	9.25	8.18
150.	4.74	4.69
151.	4.4	4.5
152.	5.49	5.16
153.	3.44	3.83
154.	9.82	7.72
155.	4.96	4.99
156.	11.59	10.64
157.	4.41	4.54
158.	4.44	5.4
159.	7.14	8.45
160.	8.81	7.39
161.	5.35	5.91
162.	4.69	5.27
163.	7.4	6.6
164.	4.73	4.8
165.	4.52	4.51
166.	4.95	5.5
167.	4.21	4.99
168.	4.56	5.2
169.	5.35	5.05
170.	4.04	4.36
171.	5.49	5.45
172.	4.64	4.56
173.	4.4	4.66
174.	5.24	5.67

175.	4.03	4.42
176.	4.81	7.57
177.	5.23	5.31
178.	5.22	5.35
179.	8.25	8.09
180.	4.95	4.89
181.	4.01	4.41
182.	4.67	4.66
183.	6.17	6.97
184.	4.76	5.01
185.	4.66	4.79
186.	4.82	5.36
187.	4.97	4.73
188.	5.5	5.56
189.	3.55	4.11
190.	10.09	10.19
191.	4.01	4.11
192.	5.1	4.86
193.	12.67	11.4
194.	12.07	11.23
195.	4.17	4.65
196.	3.55	4.15
197.	4.44	4.95
198.	4.92	5.31
199.	4.59	5.11
200.	4.73	4.64
201.	5.14	5.59
202.	5.13	5.16
203.	5.41	5.09
204.	5.13	5.16
205.	4.64	4.51

206.	6.96	6.87
207.	5.36	5.12
208.	5.53	5.19
209.	4.26	4.91
210.	4.34	5.19
211.	8.57	8.67
212.	5.02	4.92
213.	4.36	4.44
214.	3.33	3.82
215.	4.18	4.72
216.	4.78	5.46
217.	4.74	4.73
218.	3.97	4.5
219.	6.42	5.58
220.	6.48	5.98
221.	5.6	5.54
222.	4.85	4.7
223.	4.61	4.49
224.	4.1	4.12
225.	7.51	7.08
226.	4.24	4.46
227.	4.76	4.62
228.	3.75	4.56
229.	5.04	4.9
230.	5.34	5.18
231.	4.5	4.4
232.	4.37	4.36
233.	7.48	7.49
234.	4.36	4.61
235.	4.74	5.74
236.	5.26	5.07

237.	10.47	9.93
238.	5.17	5.36
239.	5.7	5.43
240.	3.59	4.36
241.	6.42	6.18
242.	5.03	5.22
243.	4.41	4.46
244.	5.68	7.03
245.	4.39	4.69
246.	4.07	4.38
247.	4.62	4.67
248.	3.96	4.09
249.	4.75	4.61
250.	4.31	4.14
251.	5.35	5.11
252.	4.8	4.7
253.	4.81	4.99
254.	4.88	4.69
255.	5.05	5.21
256.	4.87	5.2
257.	4.43	4.76
258.	4.27	4.71
259.	5.03	5.51
260.	4.4	4.82
261.	4.67	4.55
262.	4.21	4.24
263.	5.24	5.45
264.	4.61	4.95
265.	4.4	4.53
266.	5.01	4.97
267.	6.06	6.07

268.	10.75	11.1
269.	4.69	4.66
270.	5.63	5.46
271.	3.69	3.87
272.	3.85	4.26
273.	4.09	5.27
274.	4.1	4.32
275.	9.28	9.62
276.	3.88	4.11
277.	5.7	5.24
278.	2.85	3.54
279.	13.06	10.72
280.	3.97	4.24
281.	3.03	3.76
282.	5.51	5.59
283.	5.68	5.44
284.	4.2	4.42
285.	6.44	6.02
286.	9.62	7.38
287.	4.54	4.74
288.	7.51	6.82
289.	5.63	5.22
290.	5.77	5.58
291.	4.4	4.56
292.	4.7	4.96
293.	2.85	3.64
294.	5.34	5.03
295.	4.95	4.71
296.	3.59	4.48
297.	8.23	7.63
298.	5.06	4.82

299.	8.06	7.08
300.	4.9	4.84
301.	4.71	5.07
302.	3.8	4.14
303.	10.97	7.29
304.	4.43	4.44
305.	5.91	6.45
306.	5.12	5.15
307.	3.78	4.08
308.	5.12	5.11
309.	3.62	4.04
310.	4.66	4.71
311.	5.01	4.86
312.	3.75	4.23
313.	4.55	4.59
314.	4.92	4.87
315.	4.09	4.36
316.	5.58	5.59
317.	7.22	6.89
318.	3.97	4.12
319.	4.65	4.69
320.	4.83	4.97
321.	4.88	4.81
322.	4.93	4.89
323.	5.16	5
324.	6.78	6.33
325.	4.16	4.25
326.	12.16	11.98
327.	4.28	4.69
328.	4.16	4.65
329.	5.22	4.9

330.	14.94	10.18
331.	10.16	9.48
332.	10.07	10.35
333.	3.66	4.03
334.	6.48	6.03
335.	4.9	5.23
336.	11.18	9.72
337.	4.33	4.39
338.	3.75	4.09
339.	4.97	5.82
340.	4.04	4.45
341.	4.68	4.78
342.	4.17	4.34
343.	4.14	4.36
344.	4.98	5.33
345.	4.66	4.77
346.	4.29	4.36
347.	5.56	5.51
348.	11.41	11.16
349.	4.07	4.35
350.	5.01	4.93
351.	4.48	4.57
352.	4.65	4.7
353.	4.05	4.27
354.	8.11	9.24
355.	4.13	4.3
356.	3.58	4.17
357.	9.37	7.73
358.	5.53	5.73
359.	4.96	5.05
360.	4.82	4.77

361.	4.99	4.79
362.	4.28	4.91
363.	16.11	13.73
364.	4.39	4.54
365.	13.63	11.78
366.	4.49	4.54

Lampiran 4. Output MSE OOB masing – masing pohon

```
Call:
 randomForest(formula = Glikogen.Hemoglobin ~ ., data = data,      mtry = 5, ntree = 200, importance = TRUE, na.action $
               Type of random forest: regression
                     Number of trees: 200
No. of variables tried at each split: 5

          Mean of squared residuals: 2.317982
                    % Var explained: 53.31
```

```
Call:
 randomForest(formula = Glikogen.Hemoglobin ~ ., data = data,      mtry = 5, ntree = 300, importance = TRUE,
                Type of random forest: regression
                      Number of trees: 300
No. of variables tried at each split: 5

           Mean of squared residuals: 2.302475
                     % Var explained: 53.62
```

```
Call:
 randomForest(formula = Glikogen.Hemoglobin ~ ., data = data,      mtry = 5, ntree = 400, importance = TRUE,
                Type of random forest: regression
                      Number of trees: 400
No. of variables tried at each split: 5

           Mean of squared residuals: 2.302052
                     % Var explained: 53.63
```

```
Call:
 randomForest(formula = Glikogen.Hemoglobin ~ ., data = data,      mtry = 5, ntree = 500, importance = TRUE,
                Type of random forest: regression
                      Number of trees: 500
No. of variables tried at each split: 5

           Mean of squared residuals: 2.30546
                     % Var explained: 53.56
```

Lampiran 5. Tampilan CART

```
> getTree(rf1,500,labelVar=TRUE)
    left daughter right daughter          split var split point status
1               2              3       Gula.Darah        181.0     -3
2               4              5       Gula.Darah        118.5     -3
3               6              7       Lemak.Baik         26.0     -3
4               8              9              Umur         55.5     -3
5              10             11       Lemak.Baik         63.5     -3
6               0              0             <NA>          0.0     -1
7              12             13       Gula.Darah        304.5     -3
8              14             15  Lingkar.Pinggul         58.5     -3
9              16             17 Lingkar.Pinggang         33.5     -3
10             18             19       Gula.Darah        119.5     -3
11              0              0             <NA>          0.0     -1
12             20             21       Gula.Darah        201.5     -3
13             22             23       Kolesterol        259.5     -3
14             24             25  Lingkar.Pinggul         40.5     -3
15              0              0             <NA>          0.0     -1
16             26             27       Gula.Darah         73.5     -3
17             28             29     Tinggi.Badan        124.0     -3
18              0              0             <NA>          0.0     -1
19             30             31       Gula.Darah        131.5     -3
20             32             33       Lemak.Baik         43.0     -3
21             34             35  Lingkar.Pinggul         47.5     -3
22              0              0             <NA>          0.0     -1
23              0              0             <NA>          0.0     -1
```

Output di atas merupakan salah satu dari 500 pohon CART yang telah dibangun.Pohon ini merupakan pohon ke 500. Maksud dari output ini adalah terlihat di output paling atas, split var nya adalah variabel gula darah dan split point nya adalah 181.0. Left daughter menunjukkan ke node 2, maksudnya adalah jika sampel di node 1 memiliki nilai pada variabel angka melek huruf ≤ 181.0, maka turun ke percabangan sebelah kiri (node 2). Right daughter menunjukkan ke node 3, maksudnya adalah jika sampel di node 1 memiliki nilai pada variabel Gula Darah > 181.0, maka turun ke percabangan sebelah kanan (node 3). Proses dilakukan terus menerus hingga tidak ada lagi split var dan split point nya, atau pohon mencapai terminal node. Nilai prediksi ada pada terminal node, pada output nilai prediksi tertulis di kolom prediction. Berikut ini adalah ilustrasi pohonnya jika digambarkan