PENGEMBANGAN APLIKASI PREDIKSI RISIKO DIABETES MENGGUNAKAN ALGORITMA RANDOM FOREST

Tiear Rafa komara[1], Widha Dwi Yanti[2] dan Wildan Aufa Rafid[3]
[1] S1 Data Sains, Universitas Telkom
email: tiearrafa@student.telkomuniversity.ac.id
[2] S1 Data Sains, Universitas Telkom
email: widhadwiyanti@student.telkomuniversity.ac.id
[3] S1 Data Sains, Universitas Telkom
email: waarrr@student.telkomuniversity.ac.id

Abstrak. Penelitian Diabetes melitus adalah penyakit metabolik yang terjadi akibat peningkatan kadar glukosa dalam darah, yang disebabkan oleh kelainan metabolik yang terjadi sebagai akibat dari gangguan hormonal. Diabetes melitus menjadi ancaman serius bagi kesehatan global, salah satu faktor utama yang berkontribusi pada peningkatan kasus diabetes adalah faktor genetik. Faktor gaya hidup juga memiliki peran yang signifikan dalam munculnya diabetes, terutama dengan adanya perubahan kebiasaan makan yang tidak sehat dan kurangnya aktivitas fisik. Para peneliti di bidang bioinformatika telah berusaha untuk mengatasi penyakit ini dan membuat sistem untuk membantu dalam prediksi diabetes. Pada penelitian ini kami menggunakan metode random forest dan k-means. Random forest adalah salah satu jenis algoritma *mechine learning,* pengembangan dari metode *decision tree* yang dapat meningkatkan hasil akurasi. K-Means Clustering adalah teknik pengelompokan data non-hirarki yang memisahkan data ke dalam cluster dan mengelompokkan data dengan karakteristik yang berbeda ke dalam kelompok yang berbeda. Peneliatian ini bertujuan untuk memprediksi apakah penderita/pasien dapat terkena penyakit diabetes atau tidak dengan menggunakan algoritma random forestc dan k-means.

Kata kunci: *Diabetes Melitus, Prediksi Diabetes, Bioinformatika, Random forest, Algortima Machine Learning*

I. PENDAHULUAN

Pada era modern saat ini, pola makan dan gaya hidup manusia cenderung beralih ke arah yang kurang sehat, ditandai dengan konsumsi makanan cepat saji yang tinggi lemak, gula, dan garam. Salah satu dampak negatif yang serius dari tren ini adalah peningkatan prevalensi penyakit diabetes di seluruh dunia. Menurut World Health Organization (WHO), pada tahun 2019 saja, lebih dari 1,5 juta kematian dilaporkan akibat diabetes. Diabetes mellitus menjadi ancaman serius bagi kesehatan global, salah satu faktor utama yang berkontribusi pada peningkatan kasus diabetes adalah faktor genetik. Individu yang memiliki riwayat keluarga dengan diabetes memiliki risiko yang lebih tinggi untuk mengembangkan penyakit tersebut. Namun, penting untuk dicatat bahwa faktor gaya hidup juga memiliki peran yang signifikan dalam munculnya diabetes, terutama dengan adanya perubahan kebiasaan makan yang tidak sehat dan kurangnya aktivitas fisik. Dalam kondisi diabetes melitus, tubuh tidak dapat menghasilkan atau menggunakan insulin secara efektif, yang menyebabkan penumpukan glukosa dalam darah dan berbagai komplikasi jangka panjang.

Penelitian ini menggunakan pendekatan bioinformatika dengan menerapkan metode Random Forest dan K-Means Clustering untuk memprediksi kemungkinan terjadinya diabetes dan mengelompokkan data medis pasien. Dataset yang digunakan adalah dataset diabetes dari Kaggle. Hasil penelitian menunjukkan bahwa model prediksi menggunakan Random Forest dapat membantu dalam deteksi dini diabetes, sementara K-Means Clustering dapat mengidentifikasi kelompok pasien berdasarkan atribut yang ada dalam dataset.

II. TINJAUAN PUSTAKA

2.1. Diabetes Melitus

Diabetes melitus adalah suatu kondisi

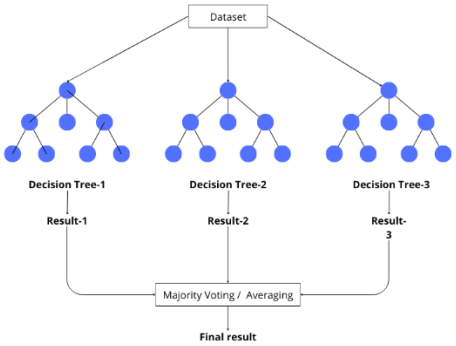
kronis yang ditandai dengan tingginya kadar gula dalam darah. Diabetes disebabkan ketika tubuh tidak dapat menghasilkan insulin atau tidak dapat menggunakan insulin secara efektif. Diabetes dibedakan menjadi dua jenis utama, yaitu diabetes tipe 1 dan tipe 2. Diabetes tipe 1 terjadi ketika sistem kekebalan tubuh menyerang dan menghancurkan sel-sel pankreas yang memproduksi insulin. Sementara itu, diabetes tipe 2 terjadi ketika sel-sel tubuh menjadi kurang sensitif terhadap insulin sehingga insulin yang dihasilkan tidak bisa digunakan dengan baik. Insulin sendiri adalah hormon yang diproduksi oleh pankreas dan berfungsi untuk mengendalikan kadar gula (glukosa) dalam darah. Tanpa insulin, sel-sel tubuh tidak dapat menyerap dan mengolah glukosa menjadi energi. Akibatnya, glukosa yang tidak diserap sel tubuh dengan baik akan menumpuk dalam darah.

2.2. Machine Learning

Machine Learning adalah suatu metode analisis data yang memungkinkan sistem untuk mempelajari data secara mandiri. Teknologi ini memungkinkan mesin untuk belajar dari data dan pengalaman, tanpa di-program secara eksplisit. Machine Learning sendiri dikembangkan berdasarkan disiplin ilmu lainnya seperti statistika, matematika dan data mining. Dengan Machine Learning, mesin dapat belajar dengan menganalisa data tanpa perlu diperintah. Mesin ini memiliki kemampuan untuk memperoleh data yang ada dengan perintah ia sendiri. Tugas yang dapat dilakukan oleh ML sangat beragam, Misalnya, dapat digunakan untuk analisis pola data, prediksi hasil analisis, dan banyak lagi.

2.3. Random Forest

Random Forest adalah algoritma pada machine learning itu sendiri, yang menggabungkan keluaran dari beberapa decision tree untuk mencapai satu hasil. Sesuai namanya, Forest atau 'hutan' dibentuk dari banyak tree (pohon) yang diperoleh melalui proses bagging atau bootstrap aggregating. Setiap tree pada Random Forest akan mengeluarkan prediksi kelas. Prediksi kelas dengan vote terbanyak menjadi kandidat prediksi pada model. Semakin banyak jumlah tree maka akan menghasilkan akurasi yang lebih tinggi dan mencegah masalah overfitting. Algoritma Random Forest. didasarkan pada konsep ensemble learning, yakni proses menggabungkan beberapa pengklasifikasi untuk memecahkan masalah yang kompleks dan untuk meningkatkan kinerja model.



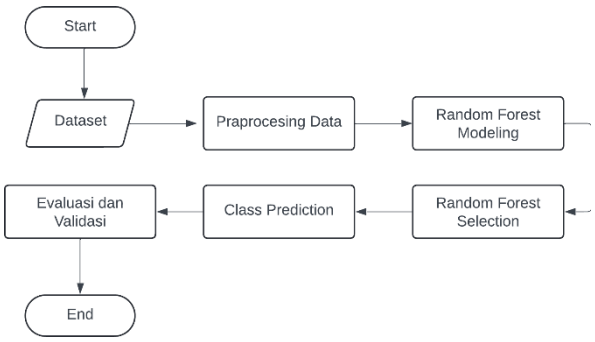
Gambar 1. Struktur Random Forest

Random Forest bekerja dalam dua fase. Fase pertama yaitu menggabungkan sejumlah N decision tree untuk membuat Random Forest. Kemudian fase kedua adalah membuat prediksi untuk setiap tree yang dibuat pada fase pertama.

2.4. K-Means

K-Means Clustering adalah teknik pengelompokan data non-hirarki yang memisahkan data ke dalam cluster dan mengelompokkan data dengan karakteristik yang berbeda ke dalam kelompok yang berbeda. Kelebihan dari penerapan K-Means yaitu mampu mengelompokkan objek besar serta dapat meningkatkan kecepatan proses pengelompokan, Selain itu K-Means memiliki beberapa kekurangan yang salah satunya adalah kegagalan dalam converge atau pergerakan data ke pusat cluster.

III. METODE PENELITIAN



Gambar 1. Perancangan Sistem

Metode penelitian ini dimulai dengan pengumpulan data, di mana dataset diabetes diambil dari UCI Machine Learning Repository. Dataset ini berisi informasi pasien dengan berbagai fitur seperti Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, dan Age. Data kemudian di-load menggunakan Pandas dan dibagi menjadi dua bagian: fitur (x) dan label (y). Selanjutnya, data dibagi menjadi data training (80%) dan data testing (20%) menggunakan metode ”train_test_split”.

Pengembangan model dilakukan menggunakan algoritma Random Forest. Model dilatih dengan data training, dan akurasinya diuji menggunakan data testing dengan metrik ”accuracy_score”. Aplikasi prediksi diabetes dikembangkan menggunakan Streamlit, yang memungkinkan pengguna untuk memasukkan data melalui antarmuka yang mudah digunakan. Setelah data dimasukkan, model prediksi yang telah dilatih digunakan untuk memberikan prediksi apakah pengguna berisiko menderita diabetes atau tidak.

Evaluasi dan validasi model dilakukan dengan menghitung akurasi model dan menampilkannya dalam aplikasi. Hasil prediksi diberikan kepada pengguna dengan pesan yang sesuai, menunjukkan apakah mereka sehat atau berisiko diabetes. Penelitian ini berhasil mengembangkan aplikasi yang mampu memprediksi risiko diabetes dengan akurasi yang memadai, membantu pengguna dalam memahami risiko kesehatan mereka.

IV. HASIL PENELITIAN DAN PEMBAHASAN

4.1. Prediksi

Prediksi Diabetes dengan Random Forest model prediksi diabetes menggunakan Random Forest mampu memberikan hasil akurasi yang baik dalam mendeteksi risiko diabetes pada individu dengan riwayat keluarga rentan. Namun, terdapat kekurangan seperti interpretasi yang sulit.

4.2. Visualisasi

Visualisasi Atribut pada Dataset Diabetes **v**isualisasi atribut pada Dataset diabetes memberikan pemahaman yang lebih baik tentang hubungan antara variabel seperti Glukosa, Tekanan Darah, dan BMI dengan risiko diabetes.

4.3. K-means

Pengelompokan Data Medis dengan K-Means Clustering Analisis kelompok data medis menggunakan K-Means Clustering berhasil mengidentifikasi kelompok pasien berdasarkan atribut yang ada pada Dataset diabetes. Namun, terdapat kekurangan dalam konvergensi data ke pusat cluster

V. KESIMPULAN DAN SARAN

Laporan ini mengungkap dampak negatif dari pola makan dan gaya hidup modern terhadap peningkatan prevalensi diabetes di seluruh dunia. Faktor genetik dan gaya hidup memiliki peran yang signifikan dalam munculnya penyakit ini. Diabetes melitus, sebuah gangguan metabolik, terjadi karena peningkatan kadar glukosa dalam darah dan kesulitan tubuh dalam menggunakan insulin secara efektif. Dalam penelitian ini, algoritma machine learning seperti Random Forest dan K-Means Clustering digunakan untuk memprediksi dan mengelompokkan data medis pasien diabetes. Meskipun Random Forest memiliki keunggulan dalam akurasi prediksi, ia juga memiliki beberapa kelemahan, seperti interpretasi yang sulit. Demikian pula, K-Means Clustering mampu mengelompokkan data dengan cepat, tetapi dapat mengalami masalah konvergensi. Penelitian ini bertujuan untuk memprediksi risiko diabetes, terutama pada individu dengan riwayat keluarga, menggunakan model Random Forest dan K-Means Clustering. Dataset yang digunakan diperoleh dari Kaggle dan terdiri dari berbagai atribut yang berkaitan dengan diabetes. Metode penelitian meliputi tahapan pengumpulan data dan preprocessing. Dengan demikian, laporan ini memberikan wawasan penting dalam upaya deteksi dan manajemen penyakit diabetes melalui pendekatan analitis dan teknologi informas.

DAFTAR PUSTAKA

[1] J. Biologi et al., "Diabetes Melitus: Review Etiologi." [Online]. Available: http://journal.uin-alauddin.ac.id/index.php/psb

[2] G. Nursa, Y. Fauzi, J. Habibi, P. Studi, K. Masyarakat, and I. Kesehatan, "Faktor-Faktor Yang Mempengaruhi Kejadian Diabetes Melitus Di Puskesmas Bintuhan Kabupaten Kaur Tahun 2022 Factors Affecting The Event Diabetes Mellitus In Bintuhan Puskesmas Kaur District Year 2022," 2022.

[3] R. Supriyadi, W. Gata, N. Maulidah, A. Fauzi, I. Komputer, and S. Nusa Mandiri Jalan Margonda Raya No, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah," vol. 13, no. 2, pp. 67–75, 2020, [Online]. Available: http://journal.stekom.ac.id/index.php/E-Bisnis■page67

[4] E. Renata and M. Ayub, "Penerapan Metode Random forest untuk Analisis Risiko pada dataset Peer to peer lending," Jurnal Teknik Informatika dan Sistem Informasi, vol. 6, no. 3, Dec. 2020, doi: 10.28932/jutisi.v6i3.2890.

[5] T. Amalina, D. Bima, A. Pramana, and B. N. Sari, "Metode K-Means Clustering Dalam Pengelompokan Penjualan Produk Frozen Food," Jurnal Ilmiah Wahana Pendidikan, vol. 8, no. 15, pp. 574–583, 2022, doi: 10.5281/zenodo.7052276.

[6] F. Pramataning Dewi, P. Siwi Aryni, and Y. Umaidah, "Implementasi Algoritma K-Means Clustering Seleksi Siswa Berprestasi Berdasarkan Keaktifan dalam Proses Pembelajaran," MEI, 2011.

[7] R. Syarif, M. T. Furqon, and S. Adinugroho, "Perbandingan Algoritme K-Means Dengan Algoritme Fuzzy C Means (FCM) Dalam Clustering Moda Transportasi Berbasis GPS," 2018. [Online]. Available: http://j-ptiik.ub.ac.id