Proposal Project

MACHINE LEARNING-
Kelas B

Anggota Kelompok:
1. 123200125 - M. Handi Rachmawan
2. 123200167 - Iman Abdurrahman

DAFTAR ISI

1. Definisi Permasalahan

1.1. Latar Belakang dan Identifikasi Permasalahan Bisnis

Obesitas merupakan sebuah penyakit dimana tubuh mengalami kelebihan berat badan dan memiliki lemak tubuh berlebih. Obesitas dapat memicu timbulnya masalah kesehatan dan penyakit lainya, seperti penyakit jantung, diabetes, tekanan darah tinggi, kanker tertentu, dan untuk kasus terburuknya, obesitas dapat menyebabkan kematian. Menurut WHO, obesitas merupakan sebuah pandemi, dengan lebih dari 650 juta (> 13%) penduduk dunia mengalami obesitas. Umumnya obesitas terjadi di negara-negara dengan pendapatan perkapita tinggi maupun negara-negara berkembang. Kasus obesitas banyak didominasi oleh orang dewasa. Obesitas terjadi karena faktor genetik, faktor lingkungan dan fisiologis, pola makan, aktivitas fisik, dan pilihan olahraga.

Obesitas erat kait kaitannya dengan *body mass index* (BMI). BMI merupakan pengukuran lemak tubuh berdasarkan tinggi dan berat badan. Rumus dari BMI adalah berat badan dalam kilogram dibagi dengan dibagi dengan kuadrat dari tinggi badan dalam meter. BMI sering digunakan untuk menentukan seseorang apakah tergolong obesitas atau tidak. Akan tetapi karena, obesitas tidak dipengaruhi oleh tinggi dan berat badan saja, maka penentuan obesitas dengan BMI tidaklah terlalu akurat.

Ditinjau dari segi ekonomi, obesitas memberikan dampak yang cukup signifikan bagi Indonesia. Setiap tahunnya negara mengeluarkan lebih dari 900 miliar rupiah untuk kasus kelebihan berat badan (termasuk obesitas). Rincian pengeluaran negara tersebut adalah Rp842 miliar, Rp55.546 miliar per tahun tahun untuk biaya rawat inap, dan Rp78.478 miliar jika kerugian ekonomi dijumlah secara nasional.

Berdasarkan pemaparan diatas, akan dibuat Sistem Klasifikasi Obesitas, adalah sistem yang berfungsi untuk membantu klasifikasi obesitas berdasarkan, gaya hidup, aktivitas harian, genetika, tinggi badan, dan berat badan..

1.2. Tujuan Proyek

Tujuan dari adanya Sistem Klasifikasi Obesitas ini adalah memberikan prediksi yang tepat mengenai kelas obesitas seseorang dan mempercepat proses

pendataan warga mengenai kasus obesitas. Selain itu, dengan adanya sistem klasifikasi ini, diharapkan penyintas obesitas dapat melakukan diagnosa mandiri dan melakukan upaya mandiri untuk menyembuhkan obesitas sehingga mengurangi pengeluaran negara yang besar untuk pasien rawat inap maupun rawat jalan akibat obesitas.

1.3. Asumsi, Kebutuhan Data, dan Limitasi

Dari proyek ini, kami berasumsi bahwa tinggi badan dan berat badan memiliki faktor yang besar dalam menentukan kelas-kelas obesitas.. Kemudian, level obesitas memiliki korelasi positif dengan jenis kelamin pria dan juga seseorang yang sering mengonsumsi makanan berkalori tinggi.

Untuk keberlangsungan proyek, data sangat amat dibutuhkan. Data yang akan digunakan merupakan data yang diperoleh dari laman UCI Machine Learning Repository, untuk tepatnya ada di tautan berikut, https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+. Data yang diperoleh terdiri dari 2111 baris dan 17 kolom. Kolom-kolom yang ada di dalam data tersebut merepresentasikan suatu atribut. Karakteristik dari atribut-atribut tersebut adalah seperti pada tabel berikut (untuk detail lebih mendalam dapat diketahui melalui laman berikut, https://linkinghub.elsevier.com/retrieve/pii/S2352340919306985):

No.	Atribut	Nilai	Keterangan
1	Gender	Male, Female	Jenis kelamin
2	Age	Numerik	Umur
3	Height	Numerik dalam meter	Tinggi badan
4	Weight	Numerik dalam kilogram	Berat badan
5	family_history_with_overweight	yes, no	Obesitas dari keturunan keluarga

6	FAVC	yes, no	Seringnya makan makan yang berkalori tinggi
7	FCVC	[1,3]	Kebiasaan dalam makan sayur
8	NCP	[1,4]	Makanan utama yang dimakan setiap harinya
9	CAEC	'Sometimes', 'Frequently', 'Always', 'no'	Makan makanan lain dalam waktu antara makan utama sekarang ke makanan utama berikutnya
10	SMOKE	yes, no	Perokok atau bukan perokok
11	CH2O	yes, no	Seringnya minum air putih
12	SCC	yes, no	Memonitor kalori yang masuk dan keluar
13	FAF	[0,3]	Frekuensi dalam aktivitas fisik
14	TUE	[0,2]	Penggunaan gadget
15	CALC	'no', 'Sometimes', 'Frequently', 'Always'	Frekuensi minum minuman alkohol
16	MTRANS	'Public_Transportation', 'Walking', 'Automobile', 'Motorbike', 'Bike'	Transportasi yang biasanya dipakai
17	NObeyesdad	'Normal_Weight', 'Overweight_Level_I', 'Overweight_Level_II', 'Obesity_Type_I',	Kategori obesitas

		'Insufficient_Weight', 'Obesity_Type_II', 'Obesity_Type_III'	

Limitasi atau batasan masalah pada proyek ini adalah sistem dapat proyek tidak membahas lebih lanjut mengenai sejarah dan latar belakang obesitas. Selain itu, data yang digunakan dalam proyek ini merupakan data tahun 2019, sehingga hasil akhir dalam proyek memiliki kemungkinan besar untuk mengabaikan fakta-fakta terbaru mengenai obesitas yang dapat menyebabkan penyimpangan hasil prediksi. Kemudian, proyek ini hanya akan berfokus pada tugas klasifikasi. Akan tetapi, jika ingin menggunakan regresi maka beberapa fitur perlu diproses terlebih dahulu untuk bisa dilakukan pemodelan.

2. Solusi dan Perspektif Bisnis

2.1. Analytical Approach: Data Cleansing, Data Preparation

Dalam menggunakan data, data tidak selalu pada kondisi yang siap digunakan perlu dilakukan data cleansing. Data cleansing yang dilakukan diantaranya adalah mengatasi missing values dan menghapus atribut yang tidak dipakai dalam pemodelan. Mengatasi missing values pada suatu baris ketika dalam satu baris terlalu banyak akan dihapus. Disamping itu, ketika baris *missing values* sedikit diberikan nilai null dan atau diberi label nol. atribut akan dihapus apabila dalam satu kolom terlalu banyak yang tidak ada nilainya dan atau atribut tersebut tidak dibutuhkan dalam pengerjaan model. Semua atribut yang ada pada dataset akan diaplikasi dengan *SimpelImputer* untuk mengatasi *missing values*.

Data preparasi pada dataset yang digunakan ada 2 jenis numerik dan kategorik. Data kategorikal pun nilainya ada yang melebihi 2 kategori. Atribut numerik atau non-kategorikal, seperti age, height, dan weight akan distandarisasi menggunakan *StandardScaler*. *StandardScaler* ini dapat menghindari bobot yang tidak proporsional pada nilai yang ditetapkan. Pada atribut, seperti CAEC, CALC akan diproses dengan *OrdinalEncoder*, sama hal nya dengan *LabelEncoder*, namun kategori yang nilainya lebih dari dua kategori. Pada *atribut* yang bersifat non-ordinal diproses dengan *OneHotEncoder*.

2.2. Analytical Approach: Model Selection and Metric Analysis

Pada proyek ini model yang akan digunakan adalah model *decision tree*. *Decision tree* atau yang sering disebut juga pohon keputusan merupakan model pembelajaran mesin untuk jenis *supervised learning* (klasifikasi dan regresi). *Decision tree* menggunakan struktur data pohon, yang terdiri dari *node* dan *edge*. *Decision tree* mengklasifikasikan suatu data dengan mengurutkan data dari akar sampai ke daun, dengan setiap *node* nya memberikan hasil klasifikasi. *Decision tree* dipilih dalam proyek ini karena algoritma ini tergolong algoritma yang mudah dibandingkan dengan algoritma *logistic regression* maupun *support vector machines*. Kemudian, *decision tree* tidak memerlukan proses normalisasi data dan juga data kosong sampai tingkat tertentu tidak mempengaruhi proses untuk membangun *decision tree*. Terlepas dari beberapa kelebihan *decision tree* yang telah disebutkan sebelumnya, tentunya *decision tree* juga memiliki kekurangan, kekurangan tersebut adalah untuk beberapa kasus *decision tree* dapat menjadi sangat kompleks dan juga perubahan kecil dalam data dapat mempengaruhi struktur *decision tree* yang mana nantinya dapat menyebabkan ketidakstabilan.

Untuk memastikan bahwa model yang telah dilatih bekerja dengan baik, diperlukan sebuah evaluasi model. Jenis evaluasi model yang dipakai adalah *confusion matrix*. *Confusion matrix* merupakan model evaluasi yang digunakan untuk mengukur kinerja model untuk tugas-tugas klasifikasi. Jenis evaluasi ini dipilih karena selain dapat menentukan *accuracy*, dengan *confusion matrix* dapat diketahui juga *precision* dan *recall*-nya.

3. Instrumen Pengukuran Keberhasilan

Proyek ini dikatakan berhasil jika memenuhi asumsi-asumsi yang telah diberikan. Untuk mengetahui asumsi-asumsi tersebut benar, proyek ini akan menghitung kepentingan setiap fitur dalam klasifikasi (*feature importance*). Selain itu, tingkat *accuracy* yang minimal yang harus dicapai adalah 80%, dengan nilai *precision* lebih dari 70% dan nilai *recall* lebih dari 80% untuk setiap kelasnya.

DAFTAR PUSTAKA

Palechor, Fabio M., and Alexis de la Hoz Manotas. *Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico*, 2019, https://www.sciencedirect.com/science/article/pii/S2352340919306985.