BAB II

TINJAUAN PUSTAKA

2.1 Pembangunan Manusia

Pembangunan manusia adalah suatu proses untuk memperbesar pilihan-pilihan bagi manusia. Definisi pembangunan manusia tersebut pada dasarnya mancakup dimensi pembangunan yang sangat luas. Dalam konsep pembangunan manusia, pembangunan seharusnya dianalisis serta dipahami dari sudut manusianya, bukan hanya dari pertumbuhan ekonominya. Untuk menjamin tercapainya tujuan pembangunan manusia, ada empat hal pokok yang perlu diperhatikan yaitu (UNDP, 1995) :

a. Produktifitas

Penduduk harus meningkatkan produktifitas dan partisipasi penuh dalam proses penciptaan pendapatan dan nafkah. Sehingga pembangunan ekonomi merupakan bagian dari model pembangunan manusia.

b. Pemerataan

Penduduk memiliki kesempatan yang sama untuk mendapatkan akses terhadap sumber daya ekonomi dan sosial. Semua hambatan yang memperkecil kesempatan untuk memperoleh akses tersebut harus dihapus, sehingga mereka dapat mengambil manfaat dari kesempatan yang ada dan berpartisipasi dalam kegiatan prosuktif yang dapat meningkatkan kualitas

hidup.

c. Kesinambungan

Akses terhadap sumber daya ekonomi dan sosial harus dipastikan tidak hanya untuk generasi-generasi yang akan datang. Semua sumber daya fisik, manusia, dan lingkungan selalu diperbarui.

d. Pemberdayaan

Penduduk harus berpartisipasi penuh dalam keputusan dan proses yang akan menentukan (bentuk/arah) kehidupan mereka serta untuk berpartisipasi dan mengambil keputusan dalam proses pembangunan.

2.2 Indeks Pembangunan Manusia (IPM)

Indeks Pembangunan Manusia merupakan pengukuran capaian pembangunan manusia berbasis sejumlah komponen dasar kualitas hidup. Sebagai ukuran kualitas hidup, IPM dibangun melalui pendekatan tiga dimensi dasar. Dimensi tersebut mencakup umur panjang dan sehat; pengetahuan dan kehidupan yang layak (Badan Pusat Statistik, 2008).

Untuk mengukur dimensi kesehatan, digunakan angka harapan hidup waktu lahir. Selanjutnya untuk mengukur dimensi pengetahuan digunakan gabungan indikator harapan lama sekolah dan rata-rata lama sekolah. Adapun untuk mengukur dimensi hidup layak digunakan indikator pengeluaran perkapita disesuaikan (Badan Pusat Statistik, 2014).

Sebelum menghitung IPM, setiap komponen IPM harus dihitung indeksnya. Formula yang digunakan dalam penghitungan indeks komponen IPM adalah

sebagai berikut :

$$I_{AHH} = \frac{AHH - AHH_{min}}{AHH_{max} - AHH_{min}} \quad (2.1)$$

$$I_{HLS} = \frac{HLS - HLS_{min}}{HLS_{max} - HLS_{min}} \quad (2.2)$$

$$I_{Rls} = \frac{RLS - RLS_{min}}{RLS - RLS_{min}} \quad (2.3)$$

$$I_{pengeluaran} = \frac{\ln(pengeluaran) - \ln(pengeluaran)_{min}}{\ln(pengeluaran_{max}) - \ln(pengeluaran_{min})} \quad (2.4)$$

Untuk menghitung indeks masing-masing komponen IPM digunakan batas maksimum dan minimum berikut ini :

Tabel 2.1 Batas Minimum dan Maksimum IPM (BPS, 2014)

Komponen IPM	Minimum	Maksimum	Satuan
Angka Harapan Hidup Saat Lahir (AHH)	20	85	Tahun
Harapan Lama Sekolah (HLS)	0	18	Tahun
Rata-rata Lama Sekolah (RLS)	0	15	Tahun
Pengeluaran per Kapita	1.007.436	26.572.352	Rupiah

Keterangan :

* Daya beli minuman merupakan garis kemiskinan terendah kabupaten tahun 2010 (data empiris) yaitu di Tolikara Papua

** Daya beli maksimum merupakan nilai tertinggi kabupaten yang diproyeksikan hingga 2025 (akhir RPJPN) yaitu perkiraan pengeluaran per kapita Jakarta Selatan tahun 2025

Selanjutnya nilai IPM dapat dihitung sebagai berikut :

$$IPM = \sqrt[3]{I_{Kesehatan} \times I_{Pendidikan} \times I_{Pengeluaran}} \quad (2.5)$$

Pengelompokkan Indeks Pembangunan Manusia di suatu wilayah yaitu (BPS, 2014) :

IPM ≥ 80 : IPM sangat tinggi

70 ≤ IPM < 80 : IPM tinggi

60 ≤ IPM < 70 : IPM sedang

IPM < 60 : IPM rendah

Variabel atribut yang digunakan adalah sebagai berikut :

2.2.1 Angka Harapan Hidup

Angka Harapan Hidup (AHH) merupakan rata-rata perkiraan banyak tahun yang dapat sitempuh oleh seseorang selama hidup. Angka Harapan Hidup digunakan untuk mengevaluasi kinerja pemerintah dalam meningkatkan kesejahteraan penduduk pada umumnya, dan meningkatkan derajat kesehatan pada khususnya. Indeks harapan hidup dihitung dengan menghitung nilai maksimum dan minimum harapan hidup sesuai standar UNDP, yaitu angka tertinggi sebagai batas atas untuk prnghitungan indeks dipakai 85 tahun dan terendah adalah 20 tahun (BPS, 2014).

2.2.2 Tingkat Pendidikan

Salah satu komponen pembentuk IPM adalah dari dimensi pengetahuan yang diukur melalui tingkat pendidikan. Dalam hal ini, indikator yang digunakan adalah rata-rata lama sekolah dan harapan lama sekolah. Rata-rata lama sekolah menggambarkan jumlah tahun yang digunakan oleh penduduk usia 25 tahun ke atas dalam menjalani pedidikan formal. Penghitungan rata-rata lama sekolah menggunakan dua batasan yang dipakai sesuai kesepakatan UNDP. Rata-rata lama sekolah memiliki batas maksimumnya 15 tahun dan batas minimum sebesar 0 tahun. Harapan lama sekolah didefinisikan sebagai lamanya sekolah (dalam tahun) yang diharapkan akan dirasakan oleh anak pada umur tertentu di masa mendatang. Batas maksimum untuk harapan lama sekolah adalah 18 tahun, sedangkan batas minimumnya 0 tahun (BPS, 2014).

2.2.3 Standar Hidup Layak

Dimensi lain dari ukuran kualitas hidup manusia adalah standar hidup layak. Dengan kata lain, standar hidup layak menggambarkan tingkat kesejahteraan yang dinikmati oleh penduduk sebagai dampak semakin membaiknya ekonomi. UNDP mengukur standar hidup layak menggunakan Produk Nasional Bruto (PNB) per kapita yang disesuaikan, sedangkan BPS dalam menghitung standar hidup layak menggunakan rata-rata pengeluaran per kapita riil yang disesuaikan dengan paritas daya beli berbasis formula Rao.

Rumus penghitungan Paritas Daya Beli (PPP) :

$$PPP_j = \prod_{i=1}^{m} \left(\frac{p_{ij}}{p_{ik}}\right)^{1/m} \quad (2.6)$$

Keterangan :

$\boldsymbol{PPP_j}$: Paritas daya beli di wilayah *j*

$\boldsymbol{p_{ij}}$: Harga komoditas *i* di kabupaten/kota *j*

$\boldsymbol{p_{ik}}$: Harga komoditas *i* di Jakarta Selatan

m : Jumlah Komoditas

2.3 *Data Mining*

Data Mining adalah proses mengolah atau merangkum data yang berjumlah besar melalui proses analisis agar dapat mengambil kesimpulan data yang berharga. Selain itu, dapat diartikan dengan gabungan antara metode statistik dan *artifical intelligence*/kecerdasan buatan yang terus berkembang (Gorunescu, 2011).

Machine Learning, kecerdasan buatan/*artifical intelligence*, statistika dan matematika merupakan teknik yang digunakan dalam *data mining* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* yang besar (Wyatt & Taylor, 2008).

Tujuan *data mining* adalah untuk meningkatkan pemasaran, penjualan, dan dukungan pelanggan melalui teknik *data mining*. *Data mining* dibagi menjadi beberapa kelompok, antara lain (Buttle, 2008) :

a. Klasifikasi

   Salah satu proses *data mining* yang paling umum, untuk memahami data kita harus mengekategorikan untuk membuat model dugaan

klasifikasi dari beberapa jenis faktor yang dapat diterapkan pada data yang belum terklasifikasikan.

b. Estimasi

Digunakan untuk melakukan klasifikasi nilai variabel untuk membuat model yang menghasilkan nilai dari variabel terget sebagai nilai prediksi.

c. Clustering

Segmentasi data heterogen menjadi beberapa kelompok yang homogen, yang membedakan pengelompokan dari klasifikasi yaitu bergantung pada kelas yang telah ditetapkan.

2.4 *Random Forest*

Metode *Random Forest* adalah pengembangan dari metode CART, yaitu dengan menerapkan metode *Bootstrap Aggregating* dan *Random Feature Selection* (Wulansari, 2018). Oleh karena itu, sebelum membahas algoritma *Random Forest*, akan dijelaskan terkait CART (*Classification And Regression Trees*) serta pembentukan pohon klasifikasi.

2.4.1 CART (*Classification And Regression Trees*)

CART merupakan salah satu metode atau algoritma dari salah satu teknik eksplorasi data yaitu teknik pohon keputusan. CART terbilang sederhana namun merupakan metode yang kuat. CART bertujuan untuk mendapatkan suatu kelompok data yang akurat sebagai penciri dari sautu pengklasifikasian, selain itu

CART digunakan untuk menggambarkan hubungan antara variabel respon (variabel dependen atau tak bebas) dengan satu atau lebih variabel prediktor (variabel independen atau bebas). Model pohon yang dihasilkan bergantung pada skala variabel respon, jika variabel data berbentuk kontinu maka model pohon yang dihasilkan adalah regression trees (pohon regresi) sedangkan bila variabel respon mempunyai skala kategorik maka pohon yang dihasilkan adalah classification trees (pohon klasifikasi) (Pratiwi & Zain, 2014).

2.4.2 Pembentukan Pohon Klasifikasi

Proses pembentukan pohon klasifikasi terdiri atas 3 tahapan, yaitu :

a. Pemilihan (*Classifier*)

Data yang digunakan pada tahap ini adalah sampel data *training/learning* (*L*) yang kemudain dipilah berdasarkan aturan pemilihan dari kriteria *goodness of split*. Himpunan yang dihasilkan dari proses pemilihan harus lebih homogen dibandingkan simpul induknya. Hal ini dapat dilakukan dengan mendefinisikan fungsi keheterogenan simpul(*impurity* atau *i(t)*). Fungsi heterogenitas yang umum digunakan adalah Indeks Gini. Metode ini memiliki kelebihan yaitu proses perhitungan yang sederhana dan relatif cepat, serta mudah dan sesuai untuk diterapkan dalam berbagai kasus (Breiman *et al*., 1993). Fungsi Indeks Gini adalah berikut ini :

$$i(t) = \sum_{i,j=1} p(j|t)p(i|t), i \neq j \qquad (2.7)$$

Dengan $p(j|t)$ adalah proporsi kelas j pada simpul t dan $p(j|t)$ adalah proporsi kelas i pada simpul t. Setelah dilakukan pemilihan dari semua

kemungkinan pemilah, maka tahapan berikutnya adalah menentukan kriteria *goodness of split* ($\emptyset(s,t)$) untuk mengevaluasi pemilah dari pemilah *s* pada simpul *t*. *Goodness of split* ($\emptyset(s,t)$)didefinisikan sebagai penurunan heterogenitas, berikut adalah rumusnya :

$$\emptyset(s,t) = \Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \quad (2.8)$$

dengan

$i(t)$: Fungsi heterogenitas pada simpul *t*

p_L : proporsi pengamatan simpul kiri

p_R : proporsi pengamatan menuju simpul kanan

$i(t_L)$: Fungsi heterogenitas pada simpul anak kiri

$i(t_R)$: Fungsi heterogenitas pada simpul anak kanan

Pemilah yang menghasilkan nilai $\Delta i(s,t)$lebih tinggi merupakan pemilah yang lebih baik karena hal ini memungkinkan untuk mereduksi keheterogenan secara lebih signifikan (Breiman, 1984).

b. Penentuan Simpul Terminal

Suatu simpul *t* akan menjadi terminal atau tidak, akan dipilah kembali bila pada simpul *t* tidak terdapat penurunan keheterogenan secara berarti atau adanya batasan minimum *n* seperti hanya terdapat satu pengamatan pada tiap simpul anak. Jumlah kasus minimum dalam suatu terminal akhir umumnya adalah 5, dan apabila hal itu terpenuhi maka pengembangan pohon dihentikan (Pratiwi et al., 2014).

c. Penandaan Label Kelas

Penandaan label kelas pada simpul terminal dilakukan berdasarkan aturan jumlah terbanyak. Label kelas simpul terminal *t* adalah $j_0$yang memberi nilai dugaan kesalahan pengklasifikasian simpul *t* terbesar. Proses pembentukan pohon klasifikasi berhenti saat terdapat hanya satu pengamatan dalam tiap-tiap simpul anak atau adanya batasan minimum *n*, semua pengamatan dalam tiap simpul anak identik, dan adanya batasan jumlah level/kedalaman pohon maksimal.

$$p(j_0|t) = max_{jp}(j|t) = max_j \frac{N_j(t)}{N(t)} \quad (2.9)$$

dengan

$p(j_0|t)$: proporsi kelas j pada simpul

$N_j(t)$: jumlah pengamatan kelas j pada simpul t

$N(t)$: jumlah pengamatan pada simpul t

2.4.3 Penentuan Jumlah Data *Training* dan Data *Testing*

Data *training* digunakan oleh algoritma untuk membentuk sebuah model klasifikasi. Model ini merupakan alat yang akan digunakan untuk mengukur sejauh mana klasifikasi berhasil melakukan prediksi dengan benar. Proporsi yang digunakan untuk data *training* dan data *testing* tidak mengikat tetapi agar variasi dalam model tidak terlalu besar maka disarankan data *training* lebih besar dibandingkan data testing (Yahya, 2018).

2.5 Pengertian *Random Forest*

Random Forest adalah suatu metode klasifikasi yang terdiri dari gabungan

pohon klasifikasi (CART) yang saling independen yang berasal dari distribusi yang sama melalui proses *voting* (jumlah terbanyak) untuk memperoleh prediksi klasifikasi. *Random Forest* merupakan salah satu metode *ensemble* (gabungan)yang berguna untuk meningkatkan akurasi klasifikasi dari sebuah pemilah tunggal yang tidak stabil dengan cara mengkobinasikan banyak pemilah dari suatu metode yang sama melalui proses *voting* (aturan pemilihan jumlah terbanyak) untuk memperoleh prediksi klasifikasi akhir (Jatmiko *et al*., 2019). *Random Forest* dikembangkan oleh Leo Breiman dari proses bagging. Bila dalam proses *bagging* digunakan *resampling bootstrap* untuk membangkitkan pohon klasifikasi dengan banyak versi yang kemudian mengkombinasikannya untuk memperoleh prediksi akhir, maka dalam *Random Forest* proses pengacakan untuk membentuk pohon klasifikasi tidak hanya dilakukan untuk data sampel saja melainkan juga pada pengambilan variabel prediktor. Sehingga, proses ini akan menghasilkan kumpulan pohon klasifikasi dengan ukuran dan bentuk yang berbeda-beda (Seftiana, 2014).

2.6 Karakteristik *Random Forest*

Salah satu kekuatan yang dimiliki oleh *Random Forest* adalah meminimumkan korelasi yang dapat menurunkan hasil kesalahan prediksi *Random Forest*. Hasil *Random Forest* memberikan keakuratan sebaik Adabost. Karakteristik *Random Forest* adalah sebagai berikut :

a. Keakuratan akurasi sebaik Adaboost dan kadang-kadang lebih baik dari Adaboost.

b. *Random Forest* relatif kuat untuk mengatasi *outlier* dan pengganggu yang lain.
c. *Random Forest* prosesnya lebih cepat daripada *bagging* atau *boosting*.
d. *Random Forest* berguna dalam hal mengestimasi *error*, kekuatan, korelasi dan variabel yang penting.
e. *Random Forest* sederhana dan mudah.

Adaboost (*Adaptive Boosting*) adalah salah satu metode *ensemble* seperti *bagging* dan *Random Forest*. *Bagging* dan *Random Forest* mendapatkan banyak pohon dari anak gugus data yang berbeda-beda dari proses *booststrap*. Namun jika Adaboost, setiap kali pembuatan pohon, data yang digunakan tetap seperti semua tetapi memiliki sebaran bobot yang berbeda dalam setiap iterasi. Penggunaan bobot juga dilakukan pada saat proses penggabungan dugaan akhir dari banyak pohon yang dihasilkan (Sartono & Syafitri, 2010).

2.7 Algoritma *Random Forest*

Secara umum, pengembangan *Random Forest* yang dilakukan dari proses *bagging* yaitu terletak pada proses pemilihan pemilah. Pada *Random Forest* pemilihan pemilah hanya melibatkan beberapa variabel prediktor yang terambil secara acak. Menurut (Rakhmawati, 2015) algoritma *Random Forest* dijelaskan sebagai berikut :

a. Mengambil *n* data sampel dari *dataset* awal dengan menggunakan teknik *resampling bootstrap* dengan pengembalian.
b. Menyusun pohon klasifikasi dari setiap *dataset* hasil *resampling*

bootstrap, dengan penentuan pemilah terbaik didasarkan pada variabel prediktor yang diambil secara acak. Jumlah variabel yang diambil secara acak dapat ditentukan melaui perhitungan $\frac{1}{2}\sqrt{m}$ atau $\sqrt{m}$ atau $2\sqrt{m}$ (Seftiana, 2014) dimana *M* adalah banyak variabel prediktor.

c. Melakukan prediksi klasifikasi data sampel berdasarkan pohon klasifikasi yang terbentuk.

d. Mengulangi langkah a-c hingga diperoleh sejumlah pohon klasifikasi yang diinginkan. Perulangan dilakukan sebanyak *K* kali.

e. Melakukan prediksi klasifikasi data sampel akhir dengan mengkombinasikan hasil prediksi pohon klasifikasi yang diperoleh.

Perhatikan bahwa setiap kali pembentukan pohon, kandidat peubah penjelas yang digunakan untuk melakuan pemisahan bukanlah seluruh peubah yang terlibat namun hanya sebagian saja hasil pemilihan secara acak. dapat dibayangkan bahwa proses ini menghasilkan kumpulan pohon tunggal dengan ukuran dan bentuk yang berbeda-beda. Hasil yang diharapkan adalah kumpulan pohon tunggal memiliki korelasi yang kecil antar pohonnya. Korelasi kecil tersebut mengakibatkan ragam dugaan *Random Forest* menjadi kecil dan lebih kecil dibandingkan ragam dugaan hasil *bagging* (Yahya, 2018).

Jika dilihat algoritma *Random Forest*, salah satu yang dapat diubah adalah nilai *m*, yaitu banyaknya peubah penjelas yang digunakan sebagai kandidat pemisah dalam pembentukan pohon. Nilai *m* yang semakin besar akan menyebabkan korelasi semakin besar.

2.8 Ketepatan Klasifikasi

Pengukuran ketepatan klasifikasi diukur melalui *total accuracy rate* (1-APER) yang dihitung berdasarkan Tabel Klasifikasi. 1-APER menunjukkan akurasi keseluruhan suatu klasifikasi (Jatmiko *et a*., 2019). Bentuk umum Tabel Klasifikasi adalah berikut ini :

Tabel 2.2 Ketepatan Klasifikasi

Aktual	Prediksi		Total
	1	2	
1	n_{11}	n_{12}	$n_1.$
2	n_{21}	n_{22}	$n_2.$
Total	$n_1.$	$n_2.$	N

Dengan,

$\boldsymbol{n_{11}}$: jumlah observasi dari kelas 1 yang tepat diprediksi sebagai kelas 1

$\boldsymbol{n_{22}}$: jumlah observasi dari kelas 2 yang tepat diprediksi sebagai kelas 2

$\boldsymbol{n_{12}}$: jumlah observasi dari kelas 1 yang salah diprediksi sebagai kelas 2

$\boldsymbol{n_{21}}$: jumlah observasi dari kelas 2 yang salah diprediksi sebagai kelas 1

$\boldsymbol{n_1.}$: jumlah observasi dari kelas 1

$\boldsymbol{n_2.}$: jumlah observasi dari kelas 2

$\boldsymbol{N}$: jumlah observasi

Perhitungan ketepatan klasifikasi adalah sebagai berikut :

Total accuracy rate (1 – APER) (dalam %) = $\frac{\boldsymbol{n_{11}+n_{22}}}{\boldsymbol{N}} \times \mathbf{100}\%$ (2.10)