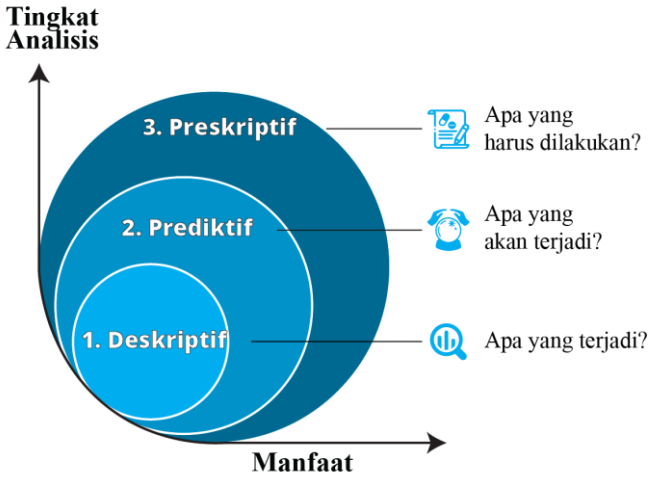
BAB 1
PENDAHULUAN

1.1. Latar Belakang Masalah

Pada sebuah perusahaan, karyawan merupakan aset yang berharga dan dapat menunjang kesuksesan perusahaan tersebut. Sehingga, apabila perusahaan tersebut kehilangan sejumlah karyawannya, maka perusahaan tersebut juga akan kehilangan sejumlah asetnya. Kondisi ini juga disebut dengan pergantian karyawan, yakni kondisi dimana karyawan keluar dari perusahaan dan perlu digantikan dengan karyawan baru [1], [2]. Hal ini seringkali sulit untuk diprediksi dan seringkali menghambat kinerja perusahaan. Dengan demikian, sangat penting bagi perusahaan untuk berusaha mengurangi atau bahkan mengatasi *Employee Churn* dengan solusi seperti perekrutan karyawan yang baik [3], strategi retensi karyawan yang tepat [4], ataupun perencanaan sumber daya manusia yang baik [5]. Penelitian tentang *Employee Churn* menggunakan *Machine Learning* telah banyak dilakukan pada penelitian sebelumnya menggunakan model *Churn Prediction* (CP) dalam segi analitik prediktif seperti yang diilustrasikan pada Gambar 1.1 [6]–[12].



Gambar 1.1 Tiga tingkat analitik

Selain *Employee Churn*, model CP juga telah diteliti untuk memprediksi *Customer Churn* pada sektor perbankan [13], telekomunikasi [14], dan *e-commerce* [15]. Selain itu, model CP juga telah diteliti menggunakan berbagai algoritma *Machine Learning* seperti algoritma *Decision Tree* yang menggunakan metode pohon keputusan

[16], [17], algoritma *Boosting Tree* yang memperkuat kelemahan pohon keputusan [18], [19], serta algoritma *Extreme Gradient Boosting* (XGB) yang merupakan gabungan antara *Decision Tree* dan *Boosting Tree* [1], [20]. Namun, penelitan-penelitian sebelumnya hanya meneliti dalam segi prediktif dan tidak meneliti dalam segi preskriptif seperti yang diilustrasikan pada gambar 1.1.

Oleh karena itu, penelitian ini tidak hanya menganalisa secara prediktif, namun juga secara preskriptif. Penelitian yang dilakukan oleh F. Devriendt *et al* [21] menyimpulkan bahwa model *Uplift* menghasilkan performa preskriptif yang lebih baik daripada model CP. Model *Uplift* sendiri memiliki beberapa metode seperti *Two-model*, *Pessimistic* model, dan *Lai's Generalized Weighted Uplift Model* (LGWUM). Model *Uplift* juga telah diteliti untuk memecahkan berbagai masalah, seperti penargetan kampanye politik Barack Obama [22], penargetan retensi pelanggan [21], [23], dan sistem rekomendasi perawatan medis [24], [25]. Terdapat beberapa algoritma *Machine Learning* yang telah diteliti dengan model *Uplift* seperti algoritma *Support Vector Machine* yang merupakan algoritma klasifikasi linear biner yang bersifat non-probabilistik [25], [26], algoritma *Logistic Regression* [25], [27] yang menggunakan garis regresi dalam memprediksi suatu nilai biner, dan algoritma *Random Forest* yang merupakan nilai acak saat melakukan seleksi input pohon keputusan [21], [28].

Pada penelitian yang dilakukan oleh Y. Zhao et al, algoritma XGB telah dibandingkan dengan delapan algoritma lainnya, antara lain *Naïve Bayes*, *Neural Net*work, *Support Vector Machine*, *Logistic Regression*, *Random Forest*, *Decision Tree*, *Gradient Boosting Tree*, dan *Linear Discriminant Analysis.* Penelitian tersebut menghasilkan bukti empiris bahwa algoritma XGB lah yang paling dapat diandalkan dalam memprediksi *Employee Churn* [1]. Oleh karena itu, pada penelitian ini penulis menerapkan algoritma XGB pada model *Uplift Modeling* dan *Churn Prediction* untuk mengatasi masalah *Employee Churn*. Untuk membandingkan kedua model tersebut, maka metrik akurasi digunakan untuk mengukur performa prediksi, sedangkan kurva *Qini* digunakan untuk mengukur perfoma preskripsi.

Terdapat tiga *Dataset* yang digunakan dalam penelitian ini. *Dataset* pertama adalah *Dataset* sintetik bernama *HR Analytics* yang berisi tentang data pekerjaan karyawan. *Dataset* kedua adalah *Dataset* sintetik bernama *IBM HR Analytics* yang memiliki isi berupa data pribadi dan pekerjaan karyawan. Kemudian yang terakhir

Dataset ketiga adalah *Dataset* SDM asli bernama *Employee Turnover* yang dibagikan oleh E. Babushkin, *Dataset* ini berisi data pribadi dan psikologis karyawan. Diharapkan hasil penelitian ini, dapat menambah wawasan tentang perbandingan antara model Churn Prediction dan *Uplift* dalam memprediksi dan mempreskripsi Employee Churn.

1.2. Rumusan Masalah

Rumusan masalah dalam penelitian ini, antara lain:

a. Analisis model *Uplift Modeling* dan *Churn Prediction* dalam mengatasi masalah *Employee Churn*
b. Menerapkan model *Extreme Gradient Boosting* untuk memprediksi dan mempreskripsi data karyawan

1.3. Tujuan

Penelitian ini bertujuan untuk melakukan analisis prediksi dan preskriptif pada model *Churn Prediction* dan *Uplift* untuk mengatasi masalah *Employee Churn* dengan menggunakan algoritma *Extreme Gradient Boosting*.

1.4. Manfaat

Manfaat yang hendak dicapai pada penelitian ini adalah menambah wawasan tentang perbandingan model *Churn Prediction* dan model *Uplift* dalam mengatasi masalah *Employee Churn.*

1.5. Batasan Masalah

Ruang lingkup penelitian ini, antara lain:

1. *Dataset* karyawan yang dibagikan oleh G. Pujar [29], Pavansubhash [30], dan E. Babushkin [31].
2. Melakukan analisis eksperimental untuk memprediksi dan mempreskripsi *Employee Churn* menggunakan algoritma *Extreme Gradient Boosting*
3. Menggunakan *Python* 3.0 pada *Jupyter Notebook*

1.6. Keterbaruan

Algoritma *Machine Learning* telah berkembang dengan berbagai metode yang meningkatkan akurasi dan mengurangi overfitting. Algoritma *Machine Learning* terbaru yang digunakan model *Churn Prediction* maupun model *Uplift* untuk memprediksi *Employee Churn* antara lain algoritma *Logistic Regression* yang menggunakan regresi dalam memprediksi [1], [21], algoritma *Random Forest* yang

menggunakan nilai acak dalam seleksi input pohon keputusan [21], [32], serta algoritma *Extreme Gradient Boosting* (XGB) yang menggabungkan *Decision Tree* dengan *boosting tree* [1], [20]. Kemudian dalam segi preskriptif, metode preskripsi yang digunakan model *Churn Prediction* dilakukan dengan mengurutkan resiko *churn* tertinggi [21], [33].

Di lain sisi, terdapat beberapa metode preskripsi terbaru yang digunakan pada model *Uplift* antara lain metode *Treatment-Covariates Interactions Approach* (TCIA) yang menargetkan efek dari variabel *Treatment* dengan mengurangi nilai probabilitas pada *Treatment* and *control group* [34], [35], metode *Pessimistic Uplift Modeling* yang merupakan gabungan antara metode *Lai's Weighted Uplift Model* (LWUM) dan *Reflective* [25], [35], serta metode *Lai's Generalized Weighted Uplift Method* (LGWUM) yang merupakan hasil pengembangan metode LWUM [25], [36].

BAB 2
METODOLOGI PENELITIAN

2.1. Jenis Penelitian

Jenis penelitian yang dipilih oleh penulis terbagi menjadi dua, yakni studi pustaka terhadap teori-teori dan penelitian yang berkaitan, serta analisis eksperimental dengan membandingkan model *Churn Prediction* dengan model *Uplift* dalam memprediksi dan mempreskripsi tiga *Dataset* karyawan. Tiga *Dataset* tersebut digunakan untuk memperkuat bukti empiris yang dihasilkan dari analisis eksperimental yang dilakukan pada penelitian ini.

2.2. Studi Pustaka

2.2.1. Algoritma Extreme gradient boosting

Fungsi dari *Extreme Gradient Boosting* (XGBoost) adalah sebagai *algoritma Machine Learning* yang sangat efisien dan fleksibel untuk memprediksi data tanpa menyebabkan masalah *overfitting*. Secara singkat, algoritma ini merupakan gabungan antara *Decision Tree* dan *boosting tree.* Algoritma ini diperkenalkan oleh T. Chen dan C. Guestrin [37] dan juga telah terbukti dapat diandalkan untuk memprediksi *Employee Churn* [38].

2.2.2. Model Churn prediction

Fungsi Model *Churn Prediction* (CP) adalah memprediksi *churn* [21]. Sehingga dalam memprediksi *Employee Churn*, model ini menggunakan algoritma *Machine Learning* untuk memprediksi apakah seorang karyawan akan keluar dari perusahaan (*churn*) atau tidak (*not churn*). Dalam segi preskriptif, model ini mengurutkan karyawan dengan resiko *churn* paling tinggi untuk kemudian ditargetkan dengan *Employee Retention Strategy* (ERS).

2.2.3. *Uplift Modeling*

Fungsi dari *Uplift Modeling* adalah memprediksi empat *Target Class* yang merupakan gabungan antara variabel dependen dan variabel *Treatment* [21]. Variabel dependen dalam penelitian ini adalah *Employee Churn*, sedangkan variabel *Treatment* [39] dalam penelitian ini adalah *Employee Retention Strategy* (ERS). Kemudian, empat *Target Class* yang akan diprediksi antara lain *Control Non-responder* (CN), *Control Responder* (CR), *Treatment Non-responder* (TN), dan *Treatment Responder* (TR) seperti yang digambarkan pada tabel sebelah kiri dalam Gambar 2.1. Dengan empat