## **2.3 *Explore Data***

*Ekploratory Data Analysis* (EDA) diperlukan sebagai sebuah pendekatan dalam menganalisis dataset untuk meringkas karakteristik utama *dataset*. Biasanya dilakukan dengan menggunakan metode visual. EDA digunakan untuk memahami data, mendapatkan konteks data, memahami variabel dan hubungan di antara variabel, dan merumuskan hipotesis yang berguna dalam membangun model prediksi. Atribut atau fitur pada dataset tidak semua diperlukan dalam mengananalisis. Fitur atau atribut yang digunakan merupakan atribut yang relevan dan sesuai dengan tujuan proyek yaitu fitur yang berpengaruh pada kinerja karyawan serta faktor-faktor lainnya yang dapat meningkatkan kinerja karyawan. Variabel yang relevan terkait kinerja karyawan pada proyek ini adalah variabel *performance\_rating* dan hubungannya dengan variabel lain yang relevan dalam dataset akan dipelajari secara mendalam. Terdapat beberapa variabel terkait yang diprediksi berpengaruh terhadap kinerja dari karyawan. Hipotesis-hipotesis tersebut dirumuskan sebagai berikut:

1. *Education* berpengaruh terhadap kinerja karyawan. Karyawan dengan *education* ‘Master’s & above’ akan memiliki kinerja yang lebih baik.
2. *No\_of\_trainings* berpengaruh terhadap kinerja karyawan. Karyawan dengan jumlah training (*no\_of\_trainings*) yang lebih banyak akan memiliki kinerja yang lebih baik.
3. KPIs\_met>80% berpengaruh terhadap kinerja karyawan. Karyawan dengan KPIs lebih dari 80% akan memiliki kinerja yang lebih baik.
4. *Awards\_won* berpengaruh terhadap kinerja karyawan, karyawan yang memenangkan hadiah akan memiliki kinerja yang lebih baik.
5. Rata-rata nilai karyawan selama pelatihan berpengaruh terhadap kinerja karyawan. Karyawan dengan *avg\_training\_score* yang tinggi akan memiliki kinerja yang lebih baik juga.

Dari beberapa hipotesis yang dirumuskan, terdapat beberapa atribut yang relevan yang selanjutnya akan digunakan untuk menentukan kinerja dari karyawannya, yaitu variabel *education, no\_of\_trainings, KPIs\_met >80%, awards\_won, avg\_training\_score* berpengaruh terhadap kinerja dari karyawan tersebut.

Setelah dilakukan eksplorasi pada data tersebut, ditemukan karakteristik-karakteristik baru yang berpengaruh terhadap hipotesis yang sebelumnya telah dirumuskan. Hal ini membuat perlu mengidentifikasi kembali subset data yang relevan untuk digunakan pada tahapan selanjutnya yang sesuai dengan tujuan data mining task pada proyek ini, yaitu untuk melakukan prediksi terhadap kinerja karyawan dalam perusahaan.