ABSTRAK

Saat ini banyak sekali tempat yang menyediakan kursus untuk melatih kemampuan berbahasa inggris di Jogja. Salah satunya Lembaga Bahasa Universitas Sanata Dharma Yogyakarta. Lembaga Bahasa USD memiliki banyak program kursus bahasa inggris, salah satunya yaitu *Center of English for International Communication* (*CEIC*). Peserta yang akan mengikuti tes ini akan ditempatkan di level yang sesuai dengan hasil tes. Level-level yang ada yaitu *Real Beginner*, *Mid Beginner*, *Upper Beginner* dan *Pre Intermediate*. Pihak Lembaga Bahasa harus melakukan penempatan level yang selama ini dilakukan dengan cara yang manual. Pada penelitian ini data program *CEIC* tahun 2019 diolah menggunakan salah satu teknik *Data Mining* dengan menggunakan *Naive Bayes*. Data yang digunakan sebanyak 240 data, terdiri dari 6 atribut (*Question* 1-10, *Question* 11-20, *Question* 21-30, *Question* 31-40, *Reading* dan *Listening*) dan 4 label (Level 2, Level 3, Level 4 dan Level 5).

Pengujian dilakukan dengan dua skenario yaitu menggunakan berbagai jumlah *fold*, dengan atau tanpa *outlier*. Secara keseluruhan pada setiap skenario dilakukan dengan menguji berbagai jumlah atribut dan menggunakan semua label yang ada. Pada skenario pertama menggunakan 240 data, dan dilakukan dengan 3, 4 dan 5 *fold*. Dari skenario pertama menghasilkan akurasi tertinggi pada pengujian dengan menggunakan 3 atribut dan menggunakan 4-*fold* dan 5-*fold*, yaitu 65%. Sedangkan pada skenario kedua menggunakan 3-*fold* dan *outlier* dengan 226 data, diperoleh akurasi tertinggi pada uji coba menggunakan 3 atribut, yaitu 67.5556%.

Kata kunci : Level Bahasa Inggris, *Naive Bayes*, Klasifikasi, *Cross Validation*.

ABSTRACT

Currently, many places are providing courses to practice English skills in Jogja. One of them is the Yogyakarta Sanata Dharma Language Institute. The Sanata Dharma Language Institute has many English courses, one of which is the Center of English for International Communication (CEIC). Participants who sign up for the course, need to take the English test first and then will be placed at the level that matches the test results. Participants who will take this test will be placed at the level that matches the test results. The levels are Basic, Real Beginner, Mid Beginner, Upper Beginner and Pre Intermediate. The Language Institution must place the level in a manual way. From the 2019 CEIC program data will be processed using one of the Data Mining techniques using Naive Bayes. The data used are 240 data, consisting of 6 attributesn (Question 1-10, Question 11-20, Question 21-30, Question 31-40, Reading dan Listening) and 4 labels (Level 2, Level 3, Level 4 dan Level 5).

There are two scenarios used, namely using various folds, with or without outliers. Overall, each scenario is done by testing various numbers of attributes and using all existing labels. In the first scenario using 240 data, and performed with 3, 4 and 5-fold, from the first scenario the highest accuracy in testing using 3 attributes using 5-fold and 4-fold is 65%. While in the second scenario using 3-fold with outlier. The data used were 226 data. As in the first scenario, the highest accuracy is in the trial using 3 attributes, namely 67.5556%.

Keywords : Level, *Language Institute*, *Naive Bayes*, Classification, Cross Validation.