DNA ploidy status directly reflects the cellular neoplasm activity and the abnormal cell division can be detected when the aneusomy or aneuploidy is observed. Several methods have been developed to directly measure the DNA content and further convert to the ratio of G0/G1. This enables us to determine the ploidy status, which can serve as a reliable marker of cell proliferation, even before the clear histopathological sign is observed. Exfoliative cytology, which offers a simple and non-invasive procedure, has emerged as a prominent technology in early oral cancer diagnosis, though many technical hurdles largely limited this method from becoming an automated and robust clinical standard protocol. In this research, we explored a novel data analysis procedure which allows automatically processing the data and re-constructing informative new variables by integrating the expert guided parameters. We then leveraged the modern machine learning technique to efficiently utilize the newly constructed data to build statistical prediction models. Using resampling methods for pruning the model core parameters, we tested a series of models and a successful Support Vector Machine (SVM) model was finally determined. Our method showed high sensitivity (median > 0.98) and specificity (median > 0.99) obtained both during the training process and in predicting on a hold-off test data. Finally, we proposed a risk index metrics for the oral leukoplakis (OLK) diagnosis, clinically defined lesions likely leading to oral squamous cell carcinoma (OSCC). Such an index reflects the probability leading to OSCC predicted from our statistical model, and it will provide a valuable guide for the clinical professionals to develop a meaningful patient’s follow up plan.